TRUST-AWARE COGNITIVE GRAPHS FOR HUMAN–MACHINE COORDINATION IN MULTI-AGENT AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Connected urban driving increasingly involves mixed teams of autonomous vehicles and humans that must coordinate under partial observability, non-stationarity, and perception failures. Many multi-agent learning pipelines implicitly assume reliable teammate signals, leaving coordination brittle when messages drop, observations degrade, or a participant behaves inconsistently. We propose `CG-HMT` (Cognitive Graphs for Human–Machine Teaming), a framework that makes team cognition explicit for both *coordination* and *learning*. `CG-HMT` represents all participants as nodes in a fully directed, time-varying trust graph; each agent maintains a compact mental model (belief, evidential intent with uncertainty, peer expectation), and trust-weighted fusion produces per-node shared mental models. Cognitive divergence (belief gaps, uncertainty spikes, expectation mismatches) triggers low-burden operator prompts mapped to bounded trust nudges. We prove a linear bound on the influence of a Byzantine trustee in trust-weighted fusion and derive a weighted policy-gradient interpretation of trust- and uncertainty-modulated learning. In CARLA-based multi-agent driving, `CG-HMT` achieves the strongest clean-condition coordination (75% vs. 20% episode-level success for the best baseline, three seeds) and shows a safety–completion trade-off under V2X dropouts, adverse weather, and a Byzantine teammate, with lower observed violation counts in several shifted settings but sometimes lower task completion.

1 INTRODUCTION

Why Team-Level Driving Matters. Urban autonomy is increasingly *networked*: connected autonomous vehicles (AVs) share the road with human drivers while exchanging information via vehicle-to-everything (V2X) channels. Safe and efficient behavior depends not only on each vehicle's local policy but also on how reliably agents interpret, predict, and coordinate under partial observability. Cooperative driving can improve throughput and smooth merges and intersections, but only if agents maintain compatible situational awareness (SA) as scenes evolve. This makes *human–machine teaming* (HMT) a natural fit: humans provide targeted oversight and intent clarification, while autonomous agents maintain structured belief and intent summaries that support transparency and coordination rather than isolated optimization (Chen et al., 2025; Vats et al., 2024).

Reality Violates Standard MARL Assumptions. Multi-agent reinforcement learning (MARL) pipelines typically assume relatively stable teammates, informative observations, and reliable communication. Real-world driving violates all three: other agents adapt, human behavior is heterogeneous, sensors degrade, and V2X channels introduce loss, delay, and replay that silently corrupt coordination signals. These factors amplify non-stationarity, complicate credit assignment, and make decentralized coordination policies sensitive to small shifts (Nguyen et al., 2020; Ning & Xie, 2024). Bridging this gap requires mechanisms that explicitly reason about *what to trust*, *when to rely on peers*, and *how to preserve coherent team behavior* when the shared picture becomes inconsistent.

Limitations of Current Human-in-the-Loop Methods. Robust MARL improves stability via perturbation training or domain randomization, yet assumes fully autonomous participants and offers limited interpretability for team-level decisions. Human-in-the-loop (HITL) methods inject demonstrations, corrections, or shaping signals (Wu et al., 2024), but often treat the human as an

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.

external supervisor rather than a teammate with an aligned mental model. Natural-language feedback can improve interaction (Krening et al., 2017), and HMT research emphasizes *shared mental models* and *calibrated trust* (Chen et al., 2025; Vats et al., 2024), but these ideas rarely enter coordination loops that jointly model trust, detect misalignment, and map concise feedback to auditable updates.

Deployment Scenario. We target settings consistent with SAE J3216-style remote support: a fleet supervisor in a control center oversees a small team of connected AVs negotiating an intersection or complex merge. The supervisor cannot drive each vehicle, but can issue concise reliability judgments when something appears inconsistent; the autonomous agents continue executing locally. `CG-HMT` is designed for this asymmetry: human attention is scarce and vehicle perception is imperfect. In practice, a single supervisor may oversee several vehicles concurrently, so each prompt must be answerable within a few seconds and the system must remain safe during inattention windows.

Our Approach. We propose `CG-HMT` (*Cognitive Graphs for Human–Machine Teaming*), a cognitive-graph-based HMT framework that makes team cognition explicit during coordination and learning. Humans and AVs are represented as nodes in a fully directed, time-varying trust graph whose edges encode each truster's reliance on teammates. Each AV maintains a lightweight *individual mental model* (IMM) comprising a compact belief state, an *evidential* intent distribution with an uncertainty proxy, and a peer-expectation head. Trust-weighted fusion produces per-node shared mental models supporting distributed SA. Cognitive divergence (belief gaps, uncertainty spikes, expectation mismatches) triggers concise operator prompts, while trust is updated online from performance, agreement, calibration, and bounded human nudges. The MARL objective is modulated by incoming trust and evidential confidence, so unreliable experience is automatically downweighted.

Key Contributions.

- **Directed Trust-Graph Fusion With a Provable Influence Bound.** We formalize a fully directed, time-varying trust graph over humans and AVs and derive trust-weighted fusion operators (weighted-mean belief and log-opinion intent pooling) that produce per-node shared mental models. We prove the log-opinion pool solves a weighted-KL optimization problem (Theorem A.1) and that the influence of any Byzantine trustee on fused belief and intent decreases *linearly* with its trust weight (Theorem 4.1; full statement and a log-space analogue appear in Appendix A).
- **Trust- and Uncertainty-Modulated Learning.** Each agent's IMM combines a recurrent belief state, an evidential Dirichlet intent representation with a principled scalar uncertainty proxy, and a peer-expectation head. We show the proposed actor objective is equivalent (up to a positive scalar) to policy gradient under a trust-weighted occupancy measure (Theorem 4.2), allowing unreliable or ambiguous experience to be automatically attenuated. A trust-weighted consistency regularizer further aligns beliefs with the SMM-derived team reference (§3–§4).
- **Divergence-Triggered Human Oversight, Evaluated Formatively.** Three complementary criteria (belief gap, uncertainty spike, and expectation–intent mismatch) drive event-based operator prompts that map to bounded trust nudges, integrating human feedback into the trust dynamics without interrupting decentralized execution. A five-operator formative study (Appendix C.6) provides feasibility evidence (mean rating $4.83/5$) and descriptively shows more `Unreliable` nudges under Byzantine perturbation than in clean conditions, with large between-operator variability.
- **Controlled Evaluation Under Representative Coordination Failures.** In CARLA, `CG-HMT` achieves 75% clean-condition success (all agents reach their goal regions; §5) versus 20% for the strongest baseline over three seeds (pooled-episode Fisher $p<10^{-8}$; seed-level intervals in §6) with the lowest observed collision and infraction counts in the weather and Byzantine shifts, and trust toward a Byzantine teammate separates from trust toward honest teammates. Generalization, sensitivity, and runtime results are in Appendices C.4–C.7.

2 Related Work

MARL for Connected Driving. Most MARL methods for connected driving follow centralized training with decentralized execution (CTDE), popularized by centralized-critic approaches and on-policy variants (Lowe et al., 2017; Yu et al., 2022). Recent work emphasizes graph-structured interactions and message passing under partial observability and unreliable V2X communication (Chen et al., 2021; Zhou et al., 2022; Zhao & Cui, 2023), as well as intent-aware driving for negotiation (Wu et al., 2023b; Di & Shi, 2021). Surveys identify communication constraints and robustness as open

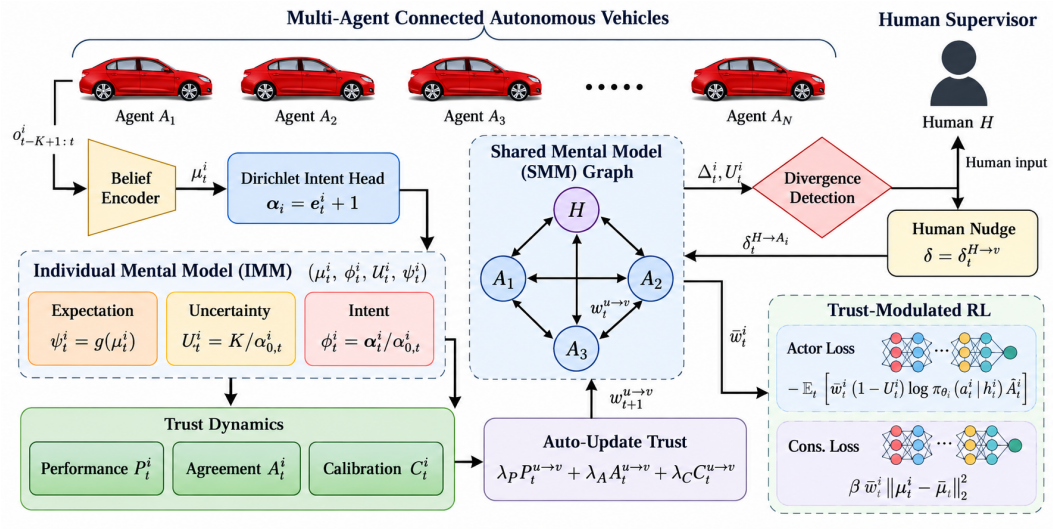


Figure 1: **CG−HMT architecture overview.** Each agent forms an IMM $M^i_t = (\mu^i_t, \phi^i_t, \psi^i_t, \xi^i_t)$. A fully directed trust graph $\mathbf{W}_t$ (row = truster, column = trustee) fuses peers' IMMs into each member's shared view. Divergence (belief gap, high $\mathcal{U}^i$, or ψ–ϕ mismatch) prompts human feedback mapped to a bounded nudge $\delta^{H\to i}_t$; edges update by EMA from performance, agreement, calibration, uncertainty damping, and δ. Learning is trust-modulated: $\bar{w}^i$ and $(1-\mathcal{U}^i)$ scale the actor loss.

challenges (Nguyen et al., 2020; Huh & Mohapatra, 2023; Gronauer & Diepold, 2022). CG−HMT addresses these gaps by modeling trust, alignment, uncertainty, and role expectations.

Human-in-the-Loop Learning. HITL RL improves safety and sample efficiency through interventions, corrections, and exploration shaping; surveys stress that supervision cost and timing shape system design (Retzlaff et al., 2024; Wu et al., 2022). In autonomous driving, intervention-centric methods guide learning toward safer behaviors without continuous teleoperation (Huang et al., 2024; Wu et al., 2023a). Human–AI teaming research ties effective collaboration to shared mental models, calibrated trust, and adaptive role allocation (Andrews et al., 2023; De Visser et al., 2020; Hagemann et al., 2023), while language-based feedback has re-emerged as a scalable interface (Krening & Feigh, 2018; Chen et al., 2024). CG−HMT operationalizes these ideas through mental-model summaries, uncertainty-aware trust updates, and event-driven prompting in one coordination-and-learning loop.

Robustness Under Adversarial Uncertainty. Robust driving systems must handle sensing degradation, communication failure, and faulty teammates (Ibrahum et al., 2024; Chi et al., 2024). Robust RL addresses observation uncertainty and safety (He & Lv, 2023; He et al., 2024), while certifiable defenses and Byzantine-robust cooperative MARL target communication integrity and teammate reliability (Sun et al., 2023; Li et al., 2024). V2X intrusion detection adds safeguards (Rajapaksha et al., 2023; Ahmed et al., 2021). CG−HMT complements these by using trust and intent alignment to downweight unreliable experience and teammates and to trigger human checks under divergence, a mechanism-level defense that complements perception-level ones.

Positioning. Directed trust graphs, uncertainty-weighted fusion, and human oversight each exist in isolation. CG−HMT couples them into one loop with guarantees attached to the coupling: the fusion operators exactly minimize a trust-weighted KL objective (Theorem A.1), admit a linear Byzantine-influence bound (Theorem 4.1), and feed an objective that is provably policy gradient under a trust-weighted occupancy (Theorem 4.2), with bounded human nudges entering that same trust state. To our knowledge, prior Byzantine-robust methods give robustness without a team-cognition representation, and prior HMT work gives shared mental models without robustness guarantees.

3 Preliminaries

Problem Formulation and CTDE. We model cooperative driving with a human in the loop as a partially observable Markov game $\langle \mathcal{I}, \mathcal{S}, \{\mathcal{O}_i\}, \{\mathcal{A}_i\}, \mathcal{T}, \{r_i\}, \gamma \rangle$ with $\mathcal{I} = \{1, \ldots, N\}$ indexing

autonomous agents and a human H treated as exogenous (its effect enters via $\mathcal{T}$ rather than as a learned policy). At time t, agent i observes $o_t^i \in \mathcal{O}_i$, receives peer mental-model messages m_t^{-i} when communication is available, and acts via $\pi_{\theta_i}(a_t^i \mid h_t^i)$ on local execution state $h_t^i = (o_{0:t}^i, a_{0:t-1}^i, m_{0:t}^{-i})$, maximizing $J_i(\theta_i) = \mathbb{E}_{\boldsymbol{\pi},\mathcal{T}}[\sum_{t=0}^{T} \gamma^t r_t^i]$ with $\gamma \in (0,1)$. We use CTDE: joint information may be used in training, while at execution each agent acts only on h_t^i.

Individual Mental Model (IMM). Each autonomous agent $i \in \mathcal{I}$ maintains

$$M_t^i = \big(\mu_t^i,\ \phi_t^i,\ \psi_t^i,\ \xi_t^i\big), \tag{1}$$

where $\mu_t^i \in \mathbb{R}^d$ is a compact belief, $\phi_t^i \in \Delta(\mathcal{A})$ an intent distribution, $\psi_t^i = \{\psi_t^{i\to j}\}_{j\neq i}$ a set of per-peer expectations over peers' near-term intents, and $\xi_t^i = 1 - \mathcal{U}_t^i \in [0,1]$ a self-confidence estimate. Beliefs come from a compact state encoder f_μ over the last L locally available observations and logged motion cues, $\mu_t^i = f_\mu(x_{t-L+1:t}^i)$, and an expectation head g produces $\psi_t^i = g(\mu_t^i)$ for comparison against peers' broadcast intents. The human node H does not require a learned visual IMM; it contributes only bounded reliability feedback and, when needed for divergence checks, a dashboard-derived reference state computed from the same displayed team summary.

Evidential Intent and Uncertainty. The intent head produces nonnegative evidence $\mathbf{e}_t^i \in \mathbb{R}^K_{\geq 0}$ over K action primitives. Setting $\boldsymbol{\alpha}_t^i = \mathbf{e}_t^i + \mathbf{1}$ and $\alpha_{0,t}^i = \sum_k \alpha_{t,k}^i$ yields a Dirichlet whose mean and uncertainty proxy are

$$\phi_{t,k}^i = \frac{\alpha_{t,k}^i}{\alpha_{0,t}^i}, \qquad \mathcal{U}_t^i = \frac{K}{\alpha_{0,t}^i} \in (0,1]. \tag{2}$$

Larger total evidence corresponds to lower epistemic uncertainty; $\mathcal{U}_t^i$ modulates both trust updates and policy learning (§4). Notation is summarized in Table 3 (Appendix A).

4 METHODOLOGY

This section presents the shared mental model over a directed trust graph (§4.1), the trust dynamics and trust-modulated learning objective (§4.2), and the human feedback interface (§4.3; dashboard, protocol, and operator study in Appendix C.6). Proofs of all theoretical results are in Appendix A.

4.1 SHARED MENTAL MODEL (SMM)

We generalize the team representation to a fully directed trust graph $\mathcal{G}_t = (V, E_t, \mathbf{W}_t)$ with $V = \{H, A_1, \ldots, A_N\}$ and adjacency matrix

$$\mathbf{W}_t = \big[w_t^{u\to v}\big]_{(N+1)\times(N+1)}, \quad w_t^{u\to v} \in [0,1], \tag{3}$$

where rows index the truster and columns the trustee. Given $\mathbf{W}_t$, each $u \in V$ forms its own per-node shared belief and shared intent

$$\tilde{\mu}_t^{\,u} = \frac{\sum_{v\neq u} w_t^{u\to v}\mu_t^v}{\sum_{v\neq u} w_t^{u\to v} + \varepsilon}, \qquad \tilde{\phi}_t^{\,u} \propto \prod_{v\neq u} \big(\phi_t^v\big)^{\bar{w}_t^{u\to v}}, \quad \bar{w}_t^{u\to v} = \frac{w_t^{u\to v}}{\sum_{v'\neq u} w_t^{u\to v'} + \varepsilon}, \tag{4}$$

with elementwise exponentiation, normalization, and small $\varepsilon > 0$. Two nodes thus hold mutually compatible but slightly different team views, consistent with distributed SA.

Why These Fusion Operators? The weighted-mean belief and log-opinion intent pool are not arbitrary choices: the log-opinion pool is the unique minimizer of $\sum_{v\neq u} w_t^{u\to v}\mathrm{KL}(q\|\phi_t^v)$ over the simplex (Theorem A.1, App. A), making fused intents an information-theoretically natural team summary. Crucially, both operators admit a strong robustness guarantee:

Theorem 4.1 (Bounded Byzantine influence on fused belief; informal). *Fix t, u and let $b \neq u$ be a (potentially Byzantine) trustee. Let $\tilde{\mu}_t^{\,u}$ denote the fusion in Eq.* (4) *using all trustees and $\tilde{\mu}_t^{\,u,\neg b}$ the same fusion with $w_t^{u\to b} = 0$. Then $\big\|\tilde{\mu}_t^{\,u} - \tilde{\mu}_t^{\,u,\neg b}\big\|_2 \ \leq\ \frac{w_t^{u\to b}}{\sum_{v\neq u} w_t^{u\to v} + \varepsilon}\big\|\mu_t^b - \tilde{\mu}_t^{\,u,\neg b}\big\|_2$, so b's influence vanishes* linearly *as $w_t^{u\to b} \to 0$. An analogous log-space bound holds for the fused intent.*

A formal statement, both bounds, and proofs appear in Appendix A. In practice, any mechanism that lowers a Byzantine trustee's edge weight, as the trust dynamics below are designed to do, reduces that trustee's influence on every other participant's fused team picture. The bound concerns the fusion operator and does not by itself control closed-loop driving (information parity, §5).

Cognitive-Divergence Monitoring. The SMM continuously tracks three signals for agent i: a belief gap $\Delta_t^i = \|\mu_t^i - \bar{\mu}_t\|_2$ against the team reference, the evidential uncertainty $\mathcal{U}_t^i$, and a peer-expectation mismatch flagged when $\exists\, j \neq i$ with $\mathrm{JS}(\psi_t^{i\to j} \| \phi_t^j) > \zeta$. A human prompt is automatically issued whenever $\Delta_t^i > \tau$, $\mathcal{U}_t^i > \kappa$, or the peer-expectation criterion fires; thresholds τ, κ, ζ are interpretable knobs whose sensitivity is reported in Appendix C.5. A short timeout commits a neutral `No-Op`. We additionally aggregate per-participant views into a global shared belief $\bar{\mu}_t = \frac{1}{|V|}\sum_{u\in V} \tilde{\mu}_t^u$ and intent $\bar{\phi}_t = \mathrm{Normalize}\big(\prod_{u\in V} (\tilde{\phi}_t^u)^{1/|V|}\big)$, used by the consistency regularizer below.

4.2 Trust Dynamics

Following the Mayer–Davis–Schoorman model of trust (Mayer et al., 1995), every ordered pair $(u \to v)$ blends three observable signals: *performance* $P_t^{u\to v} \in [0,1]$ (ability) from u's view of v's safety/progress metrics; *agreement* $A_t^{u\to v} = \exp\big(-\|\mu_t^v - \tilde{\mu}_t^u\|_2/\sigma_\mu\big)\big(1+\mathrm{JS}(\phi_t^v \| \tilde{\phi}_t^u)\big)^{-1}$ (integrity); and *calibration* $C_t^{u\to v}$ (benevolence) reflecting whether the reported confidence is consistent with subsequent agreement and task feedback. These yield the automatic proposal

$$\hat{w}_t^{u\to v} = \lambda_P P_t^{u\to v} + \lambda_A A_t^{u\to v} + \lambda_C C_t^{u\to v}, \qquad \lambda_P + \lambda_A + \lambda_C = 1. \tag{5}$$

Human feedback is first-class: a prompted operator selects from a fixed catalog mapped to a bounded nudge $\delta_t^{H\to v} \in [-\delta_{\max}, \delta_{\max}]$. Each edge updates by uncertainty-aware EMA with smoothing factor $\alpha \in (0,1)$ and uncertainty damping $\lambda_U \geq 0$:

$$w_{t+1}^{u\to v} = \mathrm{clip}_{[0,1]}\Big[(1-\alpha) w_t^{u\to v} + \alpha\big(\hat{w}_t^{u\to v} + \delta_t^{u\to v} - \lambda_U\, \mathcal{U}_t^v\big)\Big]. \tag{6}$$

Operator labels target a trustee, so a nudge on v is written to the human row only: $\delta_t^{u\to v} = \delta_t^{H\to v}$ for $u = H$ and $\delta_t^{u\to v} = 0$ otherwise. It therefore leaves each agent's own fusion $\tilde{\mu}_t^i$ (row i) unchanged and reaches agent i through the global reference $\bar{\mu}_t$, which averages the human's fused view $\tilde{\mu}_t^H$ over V, and through $w_t^{H\to i}$ in its incoming trust $\bar{w}_t^i$ (Eq. (7)).

Complementary Human and Automatic Updates. The automatic proposal $\hat{w}$ and the human nudge δ enter Eq. (6) additively: when they agree, they reinforce; when they disagree, the bounded nudge ($|\delta| \leq \delta_{\max}$) provides a persistent correction that the EMA accumulates over repeated prompts. The human row $(H \to \cdot)$ is therefore the most authoritative without discontinuous corrections, preserving auditability while allowing automatic recovery if the operator is silent.

Trust-Weighted and Uncertainty-Aware Learning. Agent i is updated in proportion to how much the team relies on it. With incoming trust $\bar{w}_t^i = \frac{1}{|V|-1}\sum_{u\neq i} w_t^{u\to i}$ and weighting $\omega_t^i := \bar{w}_t^i(1-\mathcal{U}_t^i) \in [0,1]$, the actor loss is

$$\mathcal{L}_i^{\text{act}} = -\,\mathbb{E}_t\big[\omega_t^i \,\log \pi_{\theta_i}(a_t^i \mid h_t^i)\, \hat{A}_t^i\big]. \tag{7}$$

A trust-weighted consistency penalty measures deviation from the team reference, $\mathcal{L}_i^{\text{cons}} = \beta\, \bar{w}_t^i \|\mu_t^i - \bar{\mu}_t\|_2^2$; when μ_t^i is produced by a learned encoder it contributes gradients, while in our auditable CARLA instantiation it acts as a trust-state penalty in the coordination layer. The full per-agent objective is $\mathcal{L}_i = \mathcal{L}_i^{\text{act}} + \mathcal{L}_i^{\text{crt}} + \mathcal{L}_i^{\text{ent}} + \mathcal{L}_i^{\text{cons}}$, where $\mathcal{L}_i^{\text{crt}}$ and $\mathcal{L}_i^{\text{ent}}$ are the unmodified PPO clipped value loss and entropy bonus. The reweighting in Eq. (7) admits a clean policy-gradient interpretation:

Theorem 4.2 (Weighted policy-gradient equivalence; informal)**.** *Let $d(t)$ be the rollout sampling distribution and $d_\omega(t) = \omega_t^i\, d(t)/\mathbb{E}_{t'\sim d}[\omega_{t'}^i]$. Then $-\nabla_{\theta_i}\mathcal{L}_i^{act}$ equals, up to the positive scalar $\mathbb{E}_{t'\sim d}[\omega_{t'}^i]$, the policy-gradient direction under the trust-weighted occupancy d_ω.*

A formal statement, proof, and a corruption-attenuation bound (the expected gradient-norm contribution of corrupted timesteps is upper-bounded by $\sup_{t\in\mathcal{T}_{\text{bad}}} \omega_t^i$ times the corresponding unweighted expected gradient norm) appear in Appendix A. Together with Theorem 4.1, this gives a coherent picture: at the SMM layer, low-trust trustees have linearly bounded influence on every fused team picture; at the learning layer, low-trust experience is downweighted in the gradient.

Table 1: **Standard-condition benchmark.** Clean evaluation of *all three training seeds* per method ($3 \times 20 = 60$ held-out episodes). Per-seed SR lists the three seed-level rates; other columns are mean $\pm$ std *across seeds*. Full five-metric and per-seed results: Appendix C.1.

Method	Per-seed SR	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓
MAPPO	0.20 / 0.25 / 0.15	$0.200_{\pm 0.050}$	$29.70_{\pm 1.81}$	$1.68_{\pm 0.08}$	$5.10_{\pm 0.25}$
HAPPO	0.00 / 0.05 / 0.00	$0.017_{\pm 0.029}$	$29.35_{\pm 1.36}$	$1.17_{\pm 0.08}$	$3.48_{\pm 0.20}$
MADDPG	0.10 / 0.15 / 0.10	$0.117_{\pm 0.029}$	$30.06_{\pm 1.15}$	$\mathbf{1.12_{\pm 0.08}}$	$2.97_{\pm 0.13}$
SAM	0.00 / 0.05 / 0.00	$0.017_{\pm 0.029}$	$18.74_{\pm 1.26}$	$\mathbf{1.12_{\pm 0.08}}$	$\mathbf{2.93_{\pm 0.15}}$
HHK	0.15 / 0.20 / 0.10	$0.150_{\pm 0.050}$	$34.33_{\pm 1.56}$	$1.85_{\pm 0.10}$	$11.70_{\pm 0.75}$
IPPO	0.10 / 0.15 / 0.10	$0.117_{\pm 0.029}$	$26.40_{\pm 1.55}$	$1.55_{\pm 0.12}$	$4.65_{\pm 0.32}$
CG-HMT	0.75 / 0.70 / 0.80	$\mathbf{0.750_{\pm 0.050}}$	$\mathbf{52.20_{\pm 1.40}}$	$1.25_{\pm 0.05}$	$3.17_{\pm 0.13}$

Note. $\pm$ is std across the three seed means (20 episodes per seed); collisions and infractions are per-episode counts. Best values (by mean) are bolded; CG-HMT is highlighted.

4.3 Human Feedback Interface

A lightweight dashboard issues a discrete prompt (Unreliable/Monitor/Reliable) only when a divergence event fires; the operator's selection maps to a bounded nudge $\delta_t^{H \to i}$ injected into Eq. (6). The catalog is fixed (Reliable$\mapsto +0.2$, Monitor$\mapsto 0$, Unreliable$\mapsto -0.2$), so with α=0.10 one label moves an edge by at most $\alpha\delta_{\max} = 0.02$: no single mislabel can reverse the ≈ 0.4 honest–Byzantine gap of §6, and persistent correction requires repeated prompts. A No-Op timeout preserves decentralized execution, while operators may also intervene unprompted to adjust trust edges without continuous teleoperation. The full dashboard, oversight protocol (P1–P5), workload logs, and a five-operator usability study are provided in Appendix C.6.

5 Experimental Setup

Baselines. We compare CG-HMT against six baselines spanning pure MARL, reward shaping, and HITL approaches: **MAPPO** (Yu et al., 2022), **HAPPO** (Kuba et al., 2022), **MADDPG** (Lowe et al., 2017) (Gumbel–Softmax for discrete primitives), **IPPO** (de Witt et al., 2020) (independent PPO learners), **SAM** (Xiao et al., 2021) (potential-based shaping advice), and **HHK** (Liu et al., 2024) (intermittent human guidance via a high-level gating policy); under the Byzantine shift we add a dedicated Byzantine-robust baseline, **Byz-robust MARL** (Li et al., 2024). All methods share the convolutional backbone, $|\mathcal{A}| = 9$ discrete primitive action space, training budget, decentralized-execution protocol, and *identical shaped reward* (Eq. (A.29); no method receives extra reward signal), so differences are not attributable to reward density; per-baseline configuration is in Appendix B.2. No baseline received a separate hyperparameter search (defaults, except algorithm-required changes such as MADDPG's Gumbel–Softmax interface), so baseline scores may understate best-tuned performance. HHK receives the same human-intervention budget as CG-HMT.

Threat Models. We evaluate three coordination-stressing perturbations. **(1) V2X Observation Staleness:** at each step, an agent's observation is replaced by a one-step delayed version with probability p_{drop}, modeling unreliable V2X. **(2) Adverse Weather:** RGB frames are degraded by Gaussian blur and darkening, modeling low visibility and sensor corruption. **(3) Byzantine Teammate:** one designated agent deviates from its policy with probability p_{byz}, its action being replaced by a different randomly drawn primitive, modeling a faulty or compromised participant. (As an edge target, this *Byzantine teammate* is the *Byzantine trustee* of §4.1.) All policies are trained under clean conditions; perturbations are applied only at evaluation, so Table 2 measures zero-shot robustness. Settings: p_{drop}=0.4, 7×7 blur, p_{byz}=0.5 (Appendix B.3).

Settings. (1) Environment: All experiments use CARLA (Dosovitskiy et al., 2017) via MACAD-Gym (Palanisamy, 2020) with the Town10 intersection scenario and N=3 controlled agents (default). This is the training configuration; CG-HMT is also evaluated on held-out towns, road layouts, traffic densities, and team sizes up to N=7 in Appendix C.4. Agents receive an $84 \times 84 \times 3$ egocentric RGB tensor; temporal context comes from rollout history and recurrent/trust-state bookkeeping, not frame stacking. **(2) Architecture:** All methods share an Mnih-15 CNN backbone with $\{8, 4, 3\}$ kernels, $\{4, 2, 1\}$ strides, $\{32, 64, 64\}$ channels, and a 512-unit dense head feeding policy logits

Table 2: **Robustness under shifts.** Performance under V2X, weather, and Byzantine perturbations. Reference seed (seed 1); ± is std over 20 episodes; collisions and infractions are per-episode counts. Byz-robust MARL is evaluated under the Byzantine shift only.

Setting	Method	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓	Min dist. to goal ↓
V2X	CG-HMT	**0.65$_{\pm 0.49}$**	**51.6$_{\pm 11.2}$**	1.35$_{\pm 0.59}$	**4.30$_{\pm 1.59}$**	0.83$_{\pm 0.83}$
	MAPPO	0.60$_{\pm 0.50}$	45.0$_{\pm 14.0}$	1.80$_{\pm 0.70}$	7.90$_{\pm 3.20}$	**0.70$_{\pm 0.60}$**
	MADDPG	0.55$_{\pm 0.51}$	41.2$_{\pm 13.5}$	**1.20$_{\pm 0.55}$**	5.60$_{\pm 2.05}$	0.92$_{\pm 0.78}$
	HHK	0.60$_{\pm 0.50}$	49.0$_{\pm 12.4}$	1.55$_{\pm 0.62}$	6.10$_{\pm 2.40}$	0.74$_{\pm 0.66}$
	HAPPO	0.50$_{\pm 0.51}$	42.60$_{\pm 13.10}$	1.40$_{\pm 0.64}$	5.00$_{\pm 2.10}$	1.00$_{\pm 0.82}$
	SAM	0.45$_{\pm 0.51}$	37.80$_{\pm 12.70}$	1.25$_{\pm 0.58}$	4.60$_{\pm 1.88}$	1.20$_{\pm 0.95}$
	IPPO	0.40$_{\pm 0.50}$	35.80$_{\pm 12.60}$	1.45$_{\pm 0.69}$	5.25$_{\pm 2.20}$	1.18$_{\pm 0.96}$
Weather	CG-HMT	0.00$_{\pm 0.00}$	20.4$_{\pm 8.5}$	**0.75$_{\pm 0.64}$**	**2.70$_{\pm 1.59}$**	**1.89$_{\pm 1.62}$**
	MAPPO	0.00$_{\pm 0.00}$	**39.0$_{\pm 12.0}$**	1.70$_{\pm 0.80}$	9.50$_{\pm 7.80}$	2.60$_{\pm 2.10}$
	MADDPG	0.00$_{\pm 0.00}$	14.8$_{\pm 9.6}$	1.10$_{\pm 0.70}$	3.41$_{\pm 1.35}$	6.80$_{\pm 3.90}$
	HHK	0.00$_{\pm 0.00}$	18.9$_{\pm 10.8}$	0.95$_{\pm 0.68}$	3.10$_{\pm 1.72}$	5.90$_{\pm 3.34}$
	HAPPO	0.00$_{\pm 0.00}$	24.00$_{\pm 10.40}$	0.90$_{\pm 0.66}$	3.20$_{\pm 1.56}$	4.10$_{\pm 2.75}$
	SAM	0.00$_{\pm 0.00}$	16.50$_{\pm 9.20}$	0.85$_{\pm 0.61}$	2.75$_{\pm 1.42}$	5.40$_{\pm 3.20}$
	IPPO	0.00$_{\pm 0.00}$	17.90$_{\pm 9.80}$	1.05$_{\pm 0.69}$	3.55$_{\pm 1.65}$	5.20$_{\pm 3.10}$
Byzantine	CG-HMT	0.00$_{\pm 0.00}$	20.1$_{\pm 11.5}$	**1.35$_{\pm 0.59}$**	**14.15$_{\pm 7.86}$**	2.78$_{\pm 3.76}$
	MAPPO	0.25$_{\pm 0.44}$	**39.1$_{\pm 13.0}$**	2.10$_{\pm 0.90}$	22.0$_{\pm 12.0}$	3.14$_{\pm 2.11}$
	MADDPG	0.00$_{\pm 0.00}$	26.5$_{\pm 12.8}$	1.60$_{\pm 0.72}$	18.9$_{\pm 9.1}$	3.10$_{\pm 2.80}$
	HHK	**0.30$_{\pm 0.47}$**	31.0$_{\pm 12.0}$	1.47$_{\pm 0.70}$	17.2$_{\pm 8.6}$	**2.45$_{\pm 2.20}$**
	HAPPO	0.10$_{\pm 0.31}$	27.50$_{\pm 12.20}$	1.65$_{\pm 0.76}$	18.50$_{\pm 9.80}$	3.25$_{\pm 2.65}$
	SAM	0.05$_{\pm 0.22}$	23.00$_{\pm 11.00}$	1.45$_{\pm 0.69}$	16.00$_{\pm 8.70}$	3.80$_{\pm 3.10}$
	IPPO	0.05$_{\pm 0.22}$	22.60$_{\pm 11.40}$	1.75$_{\pm 0.79}$	19.30$_{\pm 9.20}$	3.55$_{\pm 2.95}$
	Byz-robust MARL	0.20$_{\pm 0.41}$	33.10$_{\pm 12.40}$	1.55$_{\pm 0.69}$	16.40$_{\pm 7.55}$	2.60$_{\pm 2.35}$

Note. Best values (by mean) are bolded; zeros are never bolded. Repeated CG-HMT entries across blocks are the logged values, not duplicated cells.

and a scalar value. **(3) Optimization:** Training uses PPO (Schulman et al., 2017) (ϵ=0.2, γ=0.99, GAE λ=0.95 (Schulman et al., 2016)) with 12,000-transition batches and 10 epochs per update over 300 outer iterations. Each method is trained on 3 seeds. Table 1 evaluates *every* trained seed ($3 \times 20 = 60$ held-out episodes per method) and reports dispersion across seeds, so the reported gaps are not attributable to a selected checkpoint. **(4) Information Parity:** every method executes on the identical egocentric RGB observation; the compact motion summary underlying μ_t^i is used only for trust and divergence bookkeeping, which modulates the training objective and never enters the actor's input, consistent with CTDE. **(5) CG-HMT Defaults:** λ_P=0.4, λ_A=0.3, λ_C=0.3, α=0.10, $\delta_{\max}$=0.2, τ=5.0, κ=0.6, ζ=0.10, β=0.01. Full hyperparameters and reward coefficients: Appendix B.3.

Metrics. We report five episode-level CARLA metrics. **(1) Success Rate (SR):** the fraction of episodes satisfying the MACAD terminal success flag, where all controlled agents reach their assigned goal region without unrecoverable collision, timeout, or route-failure termination. **(2) Route Completion (RC, %):** average progress along the assigned route. **(3) Collisions per Episode:** mean number of collision events per episode (a count, not a probability). **(4) Infractions per Episode:** mean number of traffic or route-rule violations per episode (a count). **(5) Minimum Distance to Goal (MinDist):** closest final distance to the assigned goal region. SR and RC are not nested: success is declared on goal-region arrival, whereas RC scores progress along the full planned route, so successful episodes typically end below 100% RC and mean RC is not bounded below by $100 \times$ SR. Each metric is averaged over E=20 evaluation episodes per trained seed and condition; for SR we also report Wilson 95% intervals and two-sided Fisher exact tests against each baseline (Appendix B.4).

6 Experimental Results & Analyses

Benchmark Performance. Table 1 reports clean-condition performance over all three seeds. CG-HMT attains the strongest task completion (SR=0.750, pooled Wilson 95% CI $[0.63, 0.84]$) and route completion (52.20%). The weakest CG-HMT seed (0.70) still exceeds the strongest baseline seed (0.25) by 45 percentage points, and seed-level dispersion is small ($\pm$0.050). Against the strongest baseline (MAPPO, SR=0.200, CI $[0.12, 0.32]$) the gap is +55 points (seed-level 95% Welch interval $[0.44, 0.66]$). The pooled-episode Fisher test ($45/60$ vs. $12/60$; $p=1.9 \times 10^{-9}$, odds ratio 12.0) treats episodes as independent and is descriptive; Bonferroni-adjusted over the six base-

lines, every pooled p stays below 10^{-7} (Appendix B.4). Safety remains competitive: collisions (1.25) and infractions (3.17) sit near the best baselines despite far higher route progress, which suggests the gain is not obtained by driving more aggressively. The termination breakdown locates the gain: 60–85% of episodes of the five analyzed baselines end in timeout at 0.75–1.45 m/s, against 10% at 2.65 m/s for CG−HMT, with comparable collision shares (10–20% vs. 10%; Appendix C.2). Notably, HHK, matched on human-input budget, reaches only 0.15 SR, so the gain is not explained by operator access alone. Training curves (Appendix C.1, Figure 4) show the same picture.

Robustness Under Distribution Shifts. Table 2 (seed 1, the reference seed; seeds 2 and 3 were trained but not evaluated under the shifted conditions, 20 episodes per method and condition; Appendix B.4) shows a safety–completion trade-off rather than uniform dominance, and its comparisons are descriptive. Under V2X dropout, CG−HMT has the highest observed success (0.65) and route completion (51.6%) and the lowest observed infraction count (4.30), and baselines score higher than in the reference-seed clean evaluation (HAPPO $0.00 \rightarrow 0.50$, MAPPO $0.20 \rightarrow 0.60$; Table 6). Under adverse weather, all seven methods have SR=0; CG−HMT has the lowest observed collision and infraction counts per episode, whereas MAPPO retains more route progress with more safety events. Under Byzantine teammates, MAPPO and HHK have the highest success (0.25, 0.30) but also higher collision and infraction counts; CG−HMT has the lowest observed counts with SR=0.

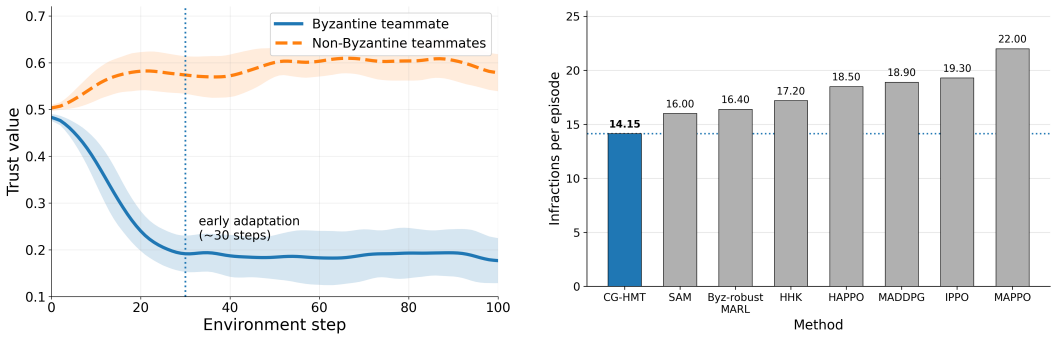
Dropout and Mutual Yielding. The baselines' gains under dropout are consistent with a mutual-yielding explanation: the timeout pattern above is what a symmetric unsignalized intersection produces when no agent breaks the tie, and timeouts of the two baselines with V2X termination data (HAPPO, MAPPO) fall by 55 and 45 points under staleness (Appendix C.2). Because we do not intervene on symmetry or yielding state, we treat this as a hypothesis, not a demonstrated mechanism.

Does the Trust Graph Down-Weight the Byzantine Teammate? Figure 2a plots mean incoming trust $\bar{w}_t^i$ toward the Byzantine agent and its two honest teammates. From a common initialization near 0.5, the Byzantine edge falls to ≈ 0.19 within about 30 steps while honest edges rise to ≈ 0.59. Computing the normalized Byzantine share per truster from the logged step-level trust matrices gives 0.138, down from $1/3$ under uniform trust (Appendix B.4). Because that share is the coefficient of Theorem 4.1, the bound on the deviation the Byzantine trustee can induce in an honest truster's fused belief is, on average, about 59% tighter than under uniform trust ($1 - 3 \times 0.138$). This separation is consistent with the linear influence control of Theorem 4.1, which concerns the fusion operator and does not by itself bound closed-loop driving: trust states are not actor inputs in our implementation, so any effect on driving arises through training. Empirically, CG−HMT has the lowest observed collision and infraction counts under attack among the eight methods in Table 2 (Figure 2b). The dedicated Byzantine-robust baseline has higher success and route completion than the tabled CG−HMT result but also more safety events, so the methods occupy different points on the safety–completion trade-off.

Robustness to Imperfect Human Input. In clean evaluation with 50% flipped labels or a 20-step label delay, success falls to the no-human value (0.65) and the other metrics are slightly worse than the no-human row (Appendix C.3); adaptive or strategically timed malicious feedback is untested.

Generalization and Sensitivity. CG−HMT retains 0.65–0.70 SR on held-out towns and 0.55–0.60 on unseen layouts, and degrades gradually with density (0.55 at 5, 0.45 at 10) and team size (0.65 at N=5, 0.55 at N=7). The trust step size is fragile: α=0.20 collapses SR to 0 (mechanism untested; Appendix C.5). Raising τ from 2.5 to 7.5 trades prompts ($1.18 \rightarrow 0.58$/min per agent, offline replay) against SR ($0.70 \rightarrow 0.65$), so burden is tunable. These runs are single-seed and CG−HMT-only, so they show graceful degradation, not an advantage over baselines.

Component Ablation Analysis. Table 5 compares separately trained component variants under clean conditions (defined in the caption). Removing the trust graph produces the largest observed decrease in task completion ($0.75 \rightarrow 0.35$ SR, $52.2 \rightarrow 42.2\%$ RC), and removing evidential uncertainty or the peer-expectation head lowers SR by approximately 0.15. Two variants receive no operator input: "trust-learning only" keeps automatic trust estimation and the trust/uncertainty-modulated objective (0.65 SR), whereas "no human feedback" also removes the trust-modulated weighting (0.55 SR, more infractions). These patterns suggest complementary contributions from trust-modulated learning and operator prompts, not an exact additive decomposition, since each variant is trained separately with a single seed; most gaps are within 20-episode sampling noise (Appendix B.4). Trust states are not actor inputs, so at evaluation they do not alter action selection.



(a) Trust edges separate within ∼30 steps.

(b) Fewest observed infractions under attack.

Figure 2: **Trust toward the Byzantine teammate separates from trust toward honest teammates, and `CG-HMT` has the lowest observed infraction count under attack. (a)** Mean incoming trust $\bar{w}_t^i$ toward the Byzantine agent versus its two honest teammates (p_{byz}=0.5); bands are ± 1 std over evaluation episodes. **(b)** Infractions per episode of all eight methods under the same attack (Table 2).

Multi-Operator Validation of the Human Interface. In a formative study, five operators used the same fixed checkpoint under clean and Byzantine conditions with identical seeds and no retraining, rating the dashboard $4.83/5$. Under attack, prompts (dashboard-level, all three agents) increased from 7.55 ± 0.40 to 12.01 ± 4.79 per minute and `Unreliable` nudges from 35.6 ± 28.7 to 54.0 ± 56.4 per session. These small-sample results are feasibility evidence, not a population-level usability estimate, and do not test adaptive operators, broader populations, or real vehicles (Appendix C.6).

Runtime Cost. `CG-HMT` incurs roughly 10% wall-clock overhead over MAPPO (12.66 h vs. 11.48 h for 300 iterations; Appendix C.7), constant in trajectory length and $\mathcal{O}(|V|^2)$ per step in team size.

7 Conclusions & Future Work

Summary of Key Contributions & Findings. We proposed `CG-HMT`, a cognitive-graph HMT framework for cooperative multi-agent driving under partial observability, non-stationarity, and unreliable perception or communication. Mixed teams form a directed, time-varying trust graph, and each agent maintains a compact mental model fused into trust-weighted shared mental models. We proved the log-opinion pool solves a weighted-KL problem, that a Byzantine trustee's influence vanishes linearly in its trust weight, and that the modulated actor loss is policy gradient under a trust-weighted occupancy. Empirically, `CG-HMT` achieves higher clean-condition success in CARLA (+55 pp SR over the strongest baseline, three seeds), assigns lower trust to a Byzantine teammate, and shows lower observed violation counts under several shifts at some cost to completion. For remote support, bounded nudges limit the cost of operator error: corrupted or delayed labels lowered SR only to the no-human level, so in our tests oversight did not become a single point of failure.

What the Evidence Supports. The clean-condition result is the strongest claim: all three seeds, every `CG-HMT` seed above every baseline seed, and a pooled gap that survives multiplicity adjustment. The shifted-condition, ablation, generalization, and sensitivity results are single-seed 20-episode evaluations and are descriptive; the trust-separation and termination analyses are mechanism evidence, not interventions; the operator study is formative. The theory bounds the fusion operator and the gradient reweighting, and does not certify closed-loop driving safety, which would require separate verification (e.g., reachability analysis or closed-course testing) before deployment.

Limitations and Future Work. Evaluation is simulation-only with simple Byzantine behaviors, motivating deployment studies and adaptive adversaries that accumulate trust before attacking. Training uses a single town/intersection with held-out evaluation elsewhere, the operator study is small (n=5), shifted-condition and ablation results use a single reference seed (20 episodes), and baselines are not individually tuned. Divergence thresholds τ, κ, ζ are fixed scalars, suggesting online adaptation. Trust is not an actor input, so its effect is mediated entirely by training, and conditioning the policy on trust is a natural extension. Improved safety under severe shifts can also cost completion.

AI USE STATEMENT

LLMs are not a component of CG−HMT's methodology. Generative AI tools were used only for writing and editing assistance and for limited coding assistance during implementation, such as code completion, debugging, and refinement. We did not use generative AI tools for research ideation, experimental design, autonomous execution of experiments, analysis or interpretation of results, or the derivation of the theoretical results in Appendix A. Such tools were also not used to generate, transcribe, or process any human-subject data from the operator study in Appendix C.6. All AI-assisted code and text were reviewed and validated by the authors, who take responsibility for the final implementation, experimental results, claims, and content of this work.

ETHICS STATEMENT

The five-operator study received an IRB exempt determination under U.S. 45 CFR 46.104(d)(2)(ii). Risks (limited to ordinary remote computer-based interaction) were disclosed to participants in a written information sheet, and operators provided informed consent prior to participation. The institution name and protocol number are withheld at submission time to preserve double-blind anonymity. Full details are in Appendix C.6, "Ethics Approval and Informed Consent." We additionally note the following broader impact. Trust-aware coordination may improve the safety and transparency of connected autonomous driving by helping agents and human supervisors reason explicitly about reliability, uncertainty, and disagreement under degraded conditions. However, incorrect trust calibration or overly conservative behavior could reduce efficiency or encourage over-reliance on automated assessments. CG−HMT is therefore intended as a coordination-support layer rather than a replacement for safety engineering or human accountability, and may extend to other cyber-physical systems involving mixed human–AI teams.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide an anonymized supplemental code package (https://anonymous.4open.science/r/CG-HMT-C7A4) containing the CG-HMT training, evaluation, robustness, ablation, sensitivity, runtime, and human-interface evaluation scripts. The package includes configuration files, baseline implementations, fixed evaluation seeds, CARLA and MACAD-Gym setup instructions, logging utilities, and scripts for reproducing the main experimental tables and appendix analyses. CG-HMT is evaluated in CARLA-based multi-agent driving under centralized training and decentralized execution: the clean-condition benchmark uses normal cooperative driving episodes, while robustness experiments use controlled simulator-level perturbations for V2X staleness, adverse weather, and Byzantine teammate behavior. Human feedback is used only through bounded reliability nudges in the specified interface protocol, and the multi-operator study evaluates a fixed checkpoint without policy retraining. We report the method components in Secs. 4–4.3, theoretical guarantees in Appendix A, baseline and environment details in Appendix B, statistical testing in Appendix B.4, generalization and sensitivity analyses in Appendices C.4 and C.5, the multi-operator interface protocol in Appendix C.6, and runtime details in Appendix C.7. After the review process, we will release the public repository and complete experiment-running instructions.

REFERENCES

Imran Ahmed, Gwanggil Jeon, and Awais Ahmad. Deep learning-based intrusion detection system for internet of vehicles. *IEEE Consumer Electronics Magazine*, 12(1):117–123, 2021.

Robert W Andrews, J Mason Lilly, Divya Srivastava, and Karen M Feigh. The role of shared mental models in human-ai teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24(2): 129–175, 2023.

Angelica Chen, Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Samuel R Bowman, Kyunghyun Cho, and Ethan Perez. Learning from natural language feedback. *Transactions on machine learning research*, 2024.

Dian Chen, Han Jun Yoon, Zelin Wan, Nithin Alluru, Sang Won Lee, Richard He, Terrence J. Moore, Frederica F. Nelson, Sunghyun Yoon, Hyuk Lim, Dongseong Kim, and Jin-Hee Cho.

Advancing human-machine teaming: Concepts, challenges, and applications. *arXiv preprint arXiv:2503.16518*, 2025. URL https://arxiv.org/abs/2503.16518.

Sikai Chen, Jiqian Dong, Paul Ha, Yujie Li, and Samuel Labi. Graph neural network and reinforcement learning for multi-agent cooperative control of connected autonomous vehicles. *Computer-Aided Civil and Infrastructure Engineering*, 36(7):838–857, 2021.

Lijun Chi, Mounira Msahli, Qingjie Zhang, Han Qiu, Tianwei Zhang, Gerard Memmi, and Meikang Qiu. Adversarial attacks on autonomous driving systems in the physical world: a survey. *IEEE Transactions on Intelligent Vehicles*, 2024.

Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, 12(2):459–478, 2020.

Christian Schroeder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the StarCraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

Xuan Di and Rongye Shi. A survey on autonomous vehicle control in the era of mixed-autonomy: From physics-based to ai-guided driving policy learning. *Transportation research part C: emerging technologies*, 125:103008, 2021.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

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

Vera Hagemann, Michèle Rieth, Amrita Suresh, and Frank Kirchner. Human-ai teams—challenges for a team-centered ai at work. *Frontiers in Artificial Intelligence*, 6:1252897, 2023.

Xiangkun He and Chen Lv. Towards safe autonomous driving: Decision making with observation-robust reinforcement learning. *Automotive Innovation*, 6(4):509–520, 2023.

Xiangkun He, Wenhui Huang, and Chen Lv. Toward trustworthy decision-making for autonomous vehicles: A robust reinforcement learning approach with safety guarantees. *Engineering*, 33:77–89, 2024.

Wenhui Huang, Haochen Liu, Zhiyu Huang, and Chen Lv. Safety-aware human-in-the-loop reinforcement learning with shared control for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

Dom Huh and Prasant Mohapatra. Multi-agent reinforcement learning: A comprehensive survey. *arXiv preprint arXiv:2312.10256*, 2023.

Ahmed Dawod Mohammed Ibrahum, Manzoor Hussain, and Jang-Eui Hong. Deep learning adversarial attacks and defenses in autonomous vehicles: A systematic literature review from a safety perspective. *Artificial Intelligence Review*, 58:28, 2024. doi: 10.1007/s10462-024-11014-8. URL https://doi.org/10.1007/s10462-024-11014-8.

Samantha Krening and Karen M Feigh. Characteristics that influence perceived intelligence in ai design. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 62, pp. 1637–1641. Sage Publications Sage CA: Los Angeles, CA, 2018.

Samantha Krening, Brent Harrison, Karen M. Feigh, Charles Lee Jr. Isbell, Mark Riedl, and Andrea Thomaz. Learning from explanations using sentiment and advice in rl. *IEEE Transactions on Cognitive and Developmental Systems*, 9(1):44–55, 2017. doi: 10.1109/TCDS.2016.2628365. URL https://doi.org/10.1109/TCDS.2016.2628365.

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

Simin Li, Jun Guo, Jingqiao Xiu, Ruixiao Xu, Xin Yu, Jiakai Wang, Aishan Liu, Yaodong Yang, and Xianglong Liu. Byzantine robust cooperative multi-agent reinforcement learning as a Bayesian game. In *International Conference on Learning Representations (ICLR)*, 2024.

Dingbang Liu, Shohei Kato, Wen Gu, Fenghui Ren, Jun Yan, and Guoxin Su. Integrating suboptimal human knowledge with hierarchical reinforcement learning for large-scale multiagent systems. *Advances in Neural Information Processing Systems*, 37:102744–102767, 2024.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, 50(9):3826–3839, 2020.

Zepeng Ning and Lihua Xie. A survey on multi-agent reinforcement learning and its application. *Journal of Automation and Intelligence*, 3(2):73–91, 2024. doi: 10.1016/j.jai.2024.02.003. URL https://doi.org/10.1016/j.jai.2024.02.003.

Praveen Palanisamy. Multi-agent connected autonomous driving using deep reinforcement learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2020.

Sampath Rajapaksha, Harsha Kalutarage, M Omar Al-Kadri, Andrei Petrovski, Garikayi Madzudzo, and Madeline Cheah. Ai-based intrusion detection systems for in-vehicle networks: A survey. *ACM Computing Surveys*, 55(11):1–40, 2023.

Carl Orge Retzlaff, Srijita Das, Christabel Wayllace, Payam Mousavi, Mohammad Afshari, Tianpei Yang, Anna Saranti, Alessa Angerschmid, Matthew E Taylor, and Andreas Holzinger. Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. *Journal of Artificial Intelligence Research*, 79:359–415, 2024.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Yanchao Sun, Ruijie Zheng, Parisa Hassanzadeh, Yongyuan Liang, Soheil Feizi, Sumitra Ganesh, and Furong Huang. Certifiably robust policy learning against adversarial multi-agent communication. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=dCOL0inGl3e.

Vanshika Vats, Marzia Binta Nizam, Minghao Liu, Ziyuan Wang, Richard Ho, Mohnish Sai Prasad, Vincent Titterton, Sai Venkat Malreddy, Riya Aggarwal, Yanwen Xu, Lei Ding, Jay Mehta, Nathan Grinnell, Li Liu, Sijia Zhong, Devanathan Nallur Gandamani, Xinyi Tang, Rohan Ghosalkar, Celeste Shen, Rachel Shen, Nafisa Hussain, Kesav Ravichandran, and James Davis. A survey on human-ai teaming with large pre-trained models. *arXiv preprint arXiv:2403.04931*, 2024. URL https://arxiv.org/abs/2403.04931.

Jingda Wu, Zhiyu Huang, Zhongxu Hu, and Chen Lv. Toward human-in-the-loop AI: Enhancing deep reinforcement learning via real-time human guidance for autonomous driving. *Engineering*, 21:75–91, 2023a. doi: 10.1016/j.eng.2022.05.017.

Jingda Wu, Zhiyu Huang, Wenhui Huang, and Chen Lv. Prioritized experience-based reinforcement learning with human guidance for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):855–869, 2024. doi: 10.1109/TNNLS.2022.3177685. URL https://doi.org/10.1109/TNNLS.2022.3177685.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, 2022.

Xiyang Wu, Rohan Chandra, Tianrui Guan, Amrit Bedi, and Dinesh Manocha. Intent-aware planning in heterogeneous traffic via distributed multi-agent reinforcement learning. In *Conference on Robot Learning*, pp. 446–477. PMLR, 2023b.

Baicen Xiao, Bhaskar Ramasubramanian, and Radha Poovendran. Shaping advice in deep multi-agent reinforcement learning. *arXiv preprint arXiv:2103.15941*, 2021.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Boyuan Zhao and Jianxun Cui. Multi-agent reinforcement learning for cooperative on-ramp merging of connected automated vehicles. In *International Conference on Internet of Things as a Service*, pp. 200–221. Springer, 2023.

Wei Zhou, Dong Chen, Jun Yan, Zhaojian Li, Huilin Yin, and Wanchen Ge. Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic. *Autonomous Intelligent Systems*, 2(1):5, 2022.

A Theoretical Guarantees and Proofs

A.1 Notation and Assumptions

Table 3: **Notations.**

V, N	participants $\{H, A_1, ..A_N\}$, N AVs	$\mu_t^i, \phi_t^i, \psi_t^i$	belief / intent / peer-expectation
$\mathcal{U}_t^i, \xi_t^i$	evidential uncertainty, self-confidence	$\boldsymbol{\alpha}_t^i, \alpha_{0,t}^i$	Dirichlet params, total evidence
$\mathbf{W}_t, w_t^{u\to v}$	trust matrix, edge from u to v	$\tilde{\mu}_t^u, \tilde{\phi}_t^u$	per-node shared belief/intent
$\bar{\mu}_t, \bar{\phi}_t$	global team-level reference	$\bar{w}_t^i$	incoming trust to agent i
P, A, C	performance / agreement / calibration	$\hat{w}_t^{u\to v}$	automatic trust proposal
$\lambda_P, \lambda_A, \lambda_C, \lambda_U$	trust mixing weights, unc. damping	$\alpha, \delta_t^{u\to v}$	EMA factor, human nudge
τ, κ, ζ	belief / unc. / mismatch thresholds	β	consistency-regularizer weight

Let $V = \{H, A_1, \ldots, A_N\}$ and let $\mathbf{W}_t = [w_t^{u\to v}] \in [0,1]^{(|V|)\times(|V|)}$ denote the directed trust matrix. For any truster $u \in V$, define the (normalized) trust weights over trustees

$$\pi_t^u(v) \;=\; \frac{w_t^{u\to v}}{\sum_{v'\neq u} w_t^{u\to v'} + \varepsilon}, \qquad v \neq u, \tag{A.1}$$

with $\varepsilon > 0$.

Assumption A1 (bounded belief). For all t and $v \in V$, $\|\mu_t^v\|_2 \le M$.

Assumption A2 (strictly positive intent probabilities). For all t, $v \in V$, and actions $k \in \{1, \ldots, K\}$, $\phi_{t,k}^v \ge \eta$ for some $\eta \in (0, 1/K]$. (Practically: use $\phi \leftarrow (1-\rho)\phi + \rho/K$ with small $\rho > 0$.)

Assumption A3 (bounded nudges and uncertainty). For all t and $v \in V$, $\delta_t^{H\to v} \in [-\delta_{\max}, \delta_{\max}]$ and $\mathcal{U}_t^v \in (0, 1]$.

Assumption A4 (bounded trust proposal). For all t and (u, v), $\hat{w}_t^{u\to v} \in [0, 1]$.

These assumptions are mild and correspond to standard boundedness and smoothing choices in practice.

A.2 Optimality and Properties of Trust-weighted Intent Fusion

Recall the per-node shared intent (log-opinion pool)

$$\tilde{\phi}_t^u(k) \;=\; \frac{\prod_{v\neq u} (\phi_t^v(k))^{\bar{w}_t^{u\to v}}}{\sum_{j=1}^K \prod_{v\neq u} (\phi_t^v(j))^{\bar{w}_t^{u\to v}}}, \qquad \bar{w}_t^{u\to v} = \frac{w_t^{u\to v}}{\sum_{v'\neq u} w_t^{u\to v'} + \varepsilon}, \quad k = 1, \ldots, K. \tag{A.2}$$

Theorem A.1 (Log-opinion pool solves a weighted KL problem)**.** *Fix t and $u \in V$, write $S_t^u := \sum_{v\neq u} w_t^{u\to v}$, and suppose $S_t^u > 0$. Set $\varepsilon = 0$ in Eq.* (A.2)*, so that the exponents $\bar{w}_t^{u\to v} = w_t^{u\to v}/S_t^u$ are non-negative and sum to exactly one (see Remark A.2 for the role of ε). Then, under Assumption A2, $\tilde{\phi}_t^u$ is the unique minimizer of*

$$q^\star \;=\; \arg\min_{q\in\Delta_K} \; \sum_{v\neq u} \bar{w}_t^{u\to v}\, \mathrm{KL}(q \,\|\, \phi_t^v)\,, \tag{A.3}$$

where $\Delta_K = \{q \in \mathbb{R}^K_{\ge 0} : \sum_k q_k = 1\}$; in particular the non-negativity constraints $q_k \ge 0$ are stated explicitly and, as the proof shows, are inactive at the optimum.

Proof. Throughout, all weights are the normalized $\bar{w}^{u\to v} = w^{u\to v}/S^u$, so that $\bar{w}^{u\to v} \ge 0$ and $\sum_{v\neq u} \bar{w}^{u\to v} = 1$; no unnormalized weight appears at any step. Expanding the objective,

$$\sum_{v\neq u} \bar{w}^{u\to v}\, \mathrm{KL}(q\|\phi^v) = \sum_{v\neq u} \bar{w}^{u\to v} \sum_k q_k \log \frac{q_k}{\phi_k^v}, \tag{A.4}$$

and, using $\sum_{v\neq u} \bar{w}^{u\to v} = 1$ to collect the entropy term,

$$= \sum_k q_k \log q_k - \sum_k q_k \sum_{v\neq u} \bar{w}^{u\to v} \log \phi_k^v. \tag{A.5}$$

The first term is strictly convex on Δ_K and the second is linear in q, so the objective is strictly convex; by Assumption A2 every $\log \phi_k^v$ is finite, so it is also proper and continuous on the compact set Δ_K, and a unique minimizer exists. Introduce a Lagrange multiplier λ for $\sum_k q_k = 1$ and, for the moment, ignore the constraints $q_k \geq 0$. Stationarity in q_k gives, *for every* $k \in \{1, \ldots, K\}$,

$$(1 + \log q_k) - \sum_{v\neq u} \bar{w}^{u\to v} \log \phi_k^v + \lambda = 0, \tag{A.6}$$

hence, for every k,

$$q_k \;=\; e^{-1-\lambda} \exp\left(\sum_{v\neq u} \bar{w}^{u\to v} \log \phi_k^v\right) \;\propto\; \prod_{v\neq u} (\phi_k^v)^{\bar{w}^{u\to v}}. \tag{A.7}$$

Choosing λ to enforce $\sum_k q_k = 1$ yields exactly Eq. (A.2). By Assumption A2, $\phi_k^v \geq \eta > 0$, so the right-hand side of Eq. (A.7) is strictly positive for every k: the dropped constraints $q_k \geq 0$ are inactive at this point, and it is therefore the minimizer over Δ_K itself. Uniqueness follows from strict convexity. □

Remark A.2 (Role of ε). The implementation evaluates Eq. (A.2) with $\bar{w}^{u\to v} = w^{u\to v}/(S^u + \varepsilon)$ and $\varepsilon = 10^{-8}$, purely to guard the degenerate case $S^u = 0$ in which a truster has no outgoing trust. For $S^u > 0$ these exponents equal $\rho\, \bar{w}^{u\to v}$ with $\rho = S^u/(S^u + \varepsilon) \in (0, 1)$, so the computed pool is the minimizer of Theorem A.1 raised to the common power ρ and renormalized, i.e. a temperature rescaling by a factor $1 - \mathcal{O}(\varepsilon/S^u)$. All statements below are made for $\varepsilon = 0$; the $\varepsilon > 0$ computation differs from them by this factor, which is below numerical tolerance for the trust magnitudes encountered in training.

Lemma A.3 (Continuity of fused intent)**.** *Fix t, u. Under Assumption A2, the mapping $(\{\phi^v\}, \{w^{u\to v}\}) \mapsto \tilde{\phi}^u$ is continuous whenever $\sum_{v\neq u} w^{u\to v} + \varepsilon > 0$.*

Proof. By Assumption A2, all $\log \phi_k^v$ terms are finite. The normalized weights

$$\bar{w}^{u\to v} = \frac{w^{u\to v}}{\sum_{v'\neq u} w^{u\to v'} + \varepsilon} \tag{A.8}$$

are continuous functions of w. The fused intent can be written as

$$\tilde{\phi}_k^u = \frac{\exp\Big(\sum_{v\neq u} \bar{w}^{u\to v} \log \phi_k^v\Big)}{\sum_{j=1}^K \exp\Big(\sum_{v\neq u} \bar{w}^{u\to v} \log \phi_j^v\Big)}. \tag{A.9}$$

This is a composition of continuous functions, and the denominator is strictly positive. Therefore, $\tilde{\phi}^u$ is continuous. □

A.3 Belief Fusion: Boundedness and Byzantine Influence Control

Recall the per-node fused belief:

$$\tilde{\mu}_t^u = \frac{\sum_{v\neq u} w_t^{u\to v}\, \mu_t^v}{\sum_{v\neq u} w_t^{u\to v} + \varepsilon}. \tag{A.10}$$

Lemma A.4 (Belief fusion is bounded)**.** *Under Assumption A1, for any t and u,*

$$\|\tilde{\mu}_t^u\|_2 \leq \frac{\sum_{v\neq u} w_t^{u\to v}}{\sum_{v\neq u} w_t^{u\to v} + \varepsilon}\, M \;\leq\; M. \tag{A.11}$$

Proof. Use triangle inequality and boundedness:

$$\|\tilde{\mu}\|_2 \leq \frac{\sum_{v\neq u} w^{u\to v}\|\mu^v\|_2}{\sum_{v\neq u} w^{u\to v}+\varepsilon} \leq \frac{\sum_{v\neq u} w^{u\to v} M}{\sum_{v\neq u} w^{u\to v}+\varepsilon}.$$

□

Theorem A.5 (Bounded influence of a Byzantine trustee on fused belief)**.** *Fix t, u and let $b \neq u$ denote a potentially faulty (Byzantine) trustee. Let $\tilde{\mu}^{\,u}$ denote Eq.* (A.10) *using all trustees, and let $\tilde{\mu}^{\,u,\neg b}$ denote the same fusion excluding b (set $w^{u\to b}=0$). Then*

$$\left\|\tilde{\mu}_t^{\,u} - \tilde{\mu}_t^{\,u,\neg b}\right\|_2 \;\leq\; \frac{w_t^{u\to b}}{\sum_{v\neq u} w_t^{u\to v}+\varepsilon}\;\left\|\mu_t^b - \tilde{\mu}_t^{\,u,\neg b}\right\|_2. \tag{A.12}$$

In particular, as $w_t^{u\to b}\to 0$, the influence of b vanishes linearly.

Proof. Write $\tilde{\mu} = (S + w_b\mu^b)/(D+w_b)$, where $S=\sum_{v\neq u,b} w_v\mu^v$ and $D=\sum_{v\neq u,b} w_v+\varepsilon$. Also $\tilde{\mu}^{\neg b} = S/D$. Then

$$\tilde{\mu} - \tilde{\mu}^{\neg b} = \frac{S+w_b\mu^b}{D+w_b} - \frac{S}{D} = \frac{w_b}{D+w_b}\left(\mu^b - \frac{S}{D}\right) = \frac{w_b}{D+w_b}\left(\mu^b - \tilde{\mu}^{\neg b}\right).$$

Taking norms gives

$$\left\|\tilde{\mu} - \tilde{\mu}^{\neg b}\right\|_2 = \frac{w_b}{D+w_b}\left\|\mu^b - \tilde{\mu}^{\neg b}\right\|_2.$$

Since $D + w_b = \sum_{v\neq u} w_v + \varepsilon$, the stated bound follows directly. □

Theorem A.6 (Bounded influence of a Byzantine trustee on fused intent (log-space))**.** *Fix t, u and a trustee $b \neq u$. Let $\tilde{\phi}^{\,u}$ and $\tilde{\phi}^{\,u,\neg b}$ be defined by Eq.* (A.2) *with and without b. Under Assumption A2,*

$$\left\|\log\tilde{\phi}_t^{\,u} - \log\tilde{\phi}_t^{\,u,\neg b}\right\|_\infty \;\leq\; \bar{w}_t^{u\to b}\;\left\|\log\phi_t^b - \log\tilde{\phi}_t^{\,u,\neg b}\right\|_\infty \tag{A.13}$$

$$+\, C_{norm}\;\bar{w}_t^{u\to b}, \tag{A.14}$$

for a finite constant C_{norm} depending only on (η, K).

Proof. Write $\bar{w}_v$ for the weights normalized over all trustees $v\neq u$ and $\bar{w}_v^{\neg b}$ for those normalized over $v\neq u, b$; set $L_k = \sum_{v\neq u}\bar{w}_v \log\phi_k^v$ and $L_k^{\neg b} = \sum_{v\neq u,b}\bar{w}_v^{\neg b}\log\phi_k^v$, so that $\log\tilde{\phi}_k = L_k - \log Z$ and $\log\tilde{\phi}_k^{\neg b} = L_k^{\neg b} - \log Z^{\neg b}$ with $Z=\sum_j e^{L_j}$ and $Z^{\neg b}=\sum_j e^{L_j^{\neg b}}$.

Step 1 (the log-linear part is an exact convex combination). For $v\neq b$ we have $\bar{w}_v = (1-\bar{w}_b)\,\bar{w}_v^{\neg b}$, since both sides equal w_v/S^u after substituting $S^{u,\neg b} = S^u - w_b$. Hence

$$L_k = (1-\bar{w}_b)L_k^{\neg b} + \bar{w}_b\log\phi_k^b, \qquad \text{so} \qquad L_k - L_k^{\neg b} = \bar{w}_b\left(\log\phi_k^b - L_k^{\neg b}\right). \tag{A.15}$$

Step 2 (re-express in terms of the fused intent). Substituting $L_k^{\neg b} = \log\tilde{\phi}_k^{\neg b} + \log Z^{\neg b}$ into Eq. (A.15) and using $\log\tilde{\phi}_k - \log\tilde{\phi}_k^{\neg b} = (L_k - L_k^{\neg b}) - (\log Z - \log Z^{\neg b})$ gives

$$\log\tilde{\phi}_k - \log\tilde{\phi}_k^{\neg b} = \bar{w}_b\left(\log\phi_k^b - \log\tilde{\phi}_k^{\neg b}\right) - \bar{w}_b\log Z^{\neg b} - \left(\log Z - \log Z^{\neg b}\right). \tag{A.16}$$

Step 3 (bound the normalizer terms). The map $L\mapsto\log\sum_j e^{L_j}$ is 1-Lipschitz in $\|\cdot\|_\infty$, so by Eq. (A.15), $\left|\log Z - \log Z^{\neg b}\right| \leq \|L - L^{\neg b}\|_\infty = \bar{w}_b\,\|\log\phi^b - L^{\neg b}\|_\infty$. Under Assumption A2, $\eta\leq\phi_k^v\leq 1$, so $\log\phi_k^v\in[\log\eta, 0]$; since the $\bar{w}_v^{\neg b}$ are non-negative and sum to one, $L_k^{\neg b}\in[\log\eta, 0]$ as well. Consequently $\|\log\phi^b - L^{\neg b}\|_\infty \leq 2\log(1/\eta)$ and $Z^{\neg b}\in[K\eta, K]$, whence $\left|\log Z^{\neg b}\right| \leq \log K + \log(1/\eta)$.

Step 4 (combine). Taking $\|\cdot\|_\infty$ in Eq. (A.16) and applying the triangle inequality with the bounds of Step 3,

$$\left\|\log\tilde{\phi}^{\,u} - \log\tilde{\phi}^{\,u,\neg b}\right\|_\infty \leq \bar{w}_b\left\|\log\phi^b - \log\tilde{\phi}^{\,u,\neg b}\right\|_\infty + C_{\text{norm}}\,\bar{w}_b, \tag{A.17}$$

with the explicit constant $C_{\text{norm}} = \log K + 3\log(1/\eta)$, which depends only on (η, K) as claimed. Both terms are proportional to $\bar{w}_b$, so the influence of b vanishes linearly as $\bar{w}_t^{u\to b}\to 0$. □

A.4 Trust Dynamics: Invariance, Contraction, and Convergence

Recall the trust update (with clipping):

$$w_{t+1}^{u\to v} = \mathrm{clip}_{[0,1]}\Big((1-\alpha)w_t^{u\to v} + \alpha\big(\hat{w}_t^{u\to v} + \delta_t^{u\to v} - \lambda_U \mathcal{U}_t^v\big)\Big). \tag{A.18}$$

Lemma A.7 (Invariance of trust bounds)**.** *If $w_0^{u\to v} \in [0,1]$, then $w_t^{u\to v} \in [0,1]$ for all $t \geq 0$.*

Proof. No induction is required. For $t = 0$ the claim is the hypothesis. For every $t \geq 0$, Eq. (A.18) applies $\mathrm{clip}_{[0,1]}$ as its outermost operation, and the range of $\mathrm{clip}_{[0,1]}$ is $[0,1]$ irrespective of its argument; hence $w_{t+1}^{u\to v} \in [0,1]$ directly, whatever the value of the bracketed expression. □

Lemma A.8 (Bounded one-step change)**.** *Under Assumptions A3–A4, for all t and (u,v),*

$$\big|w_{t+1}^{u\to v} - w_t^{u\to v}\big| \;\leq\; \alpha\Big(1 + \delta_{\max} + \lambda_U\Big). \tag{A.19}$$

Proof. Ignore clipping first. Since $w_t \in [0,1]$ and $\hat{w}_t \in [0,1]$, $\delta \in [-\delta_{\max}, \delta_{\max}]$, and $\mathcal{U} \in (0,1]$,

$$|(1-\alpha)w_t + \alpha(\hat{w}_t + \delta - \lambda_U \mathcal{U}) - w_t| = \alpha|\hat{w}_t + \delta - \lambda_U \mathcal{U} - w_t| \leq \alpha(1 + \delta_{\max} + \lambda_U).$$

Clipping can only reduce the magnitude of the change relative to the unclipped update, so the bound holds. □

Proposition A.9 (Convergence under stationary signals (unclipped))**.** *Consider the unclipped update $\bar{w}_{t+1} = (1-\alpha)\bar{w}_t + \alpha s_t$ with $s_t = \hat{w}_t + \delta_t - \lambda_U \mathcal{U}_t$. If $s_t \to s^\star$, then $\bar{w}_t \to s^\star$.*

Proof. This is a standard exponential moving average (EMA). Let $e_t = \bar{w}_t - s^\star$. Then $e_{t+1} = (1-\alpha)e_t + \alpha(s_t - s^\star)$. Since $|1-\alpha| < 1$ and $s_t - s^\star \to 0$, $e_t \to 0$. □

Corollary A.10 (Convergence with clipping)**.** *If $s_t \to s^\star$, then the clipped update Eq. (A.18) converges to $\mathrm{clip}_{[0,1]}(s^\star)$.*

Proof. Write $\Pi := \mathrm{clip}_{[0,1]}$, $c := \Pi(s^\star)$, and note that Π is non-expansive: $|\Pi(x) - \Pi(y)| \leq |x - y|$ for all $x, y \in \mathbb{R}$.

Case 1: $s^\star \in [0,1]$. Then $\Pi(c) = c$, and since $w_t \in [0,1]$ by Lemma A.7,

$$\begin{aligned} |w_{t+1} - c| &= \big|\Pi\big((1-\alpha)w_t + \alpha s_t\big) - \Pi(c)\big| \leq |(1-\alpha)(w_t - c) + \alpha(s_t - s^\star)| \\ &\leq (1-\alpha)|w_t - c| + \alpha|s_t - s^\star|. \end{aligned} \tag{A.20}$$

Writing $e_t = |w_t - c|$ and $\epsilon_t = |s_t - s^\star| \to 0$, we obtain $e_{t+1} \leq (1-\alpha)e_t + \alpha\epsilon_t$ with $\alpha \in (0,1)$, so $\limsup_t e_t \leq \limsup_t \epsilon_t = 0$ and $w_t \to c$.

Case 2: $s^\star > 1$, so $c = 1$. Pick $\gamma > 0$ with $s^\star > 1 + \gamma$ and T with $s_t \geq 1 + \gamma/2$ for all $t \geq T$. Put $f_t = 1 - w_t \geq 0$. For $t \geq T$, the unclipped iterate satisfies $1 - \big((1-\alpha)w_t + \alpha s_t\big) = (1-\alpha)f_t - \alpha(s_t - 1) \leq (1-\alpha)f_t - \alpha\gamma/2$, and clipping to $[0,1]$ can only increase the iterate towards 1, so $f_{t+1} \leq \max\{0, (1-\alpha)f_t - \alpha\gamma/2\}$. The right-hand side is a contraction with a strictly negative offset, so $f_t = 0$ for all sufficiently large t; hence $w_t \to 1 = c$.

Case 3: $s^\star < 0$, so $c = 0$. Symmetric to Case 2 with $f_t = w_t$.

In all cases $w_t \to \mathrm{clip}_{[0,1]}(s^\star)$. □

A.5 Trust- and Uncertainty-modulated Learning: Objective Interpretation

Define the per-agent actor loss

$$\mathcal{L}_i^{\text{act}} = -\,\mathbb{E}_t\Big[\omega_t^i \;\log \pi_{\theta_i}(a_t^i \mid h_t^i)\, \hat{A}_t^i\Big], \qquad \omega_t^i := \bar{w}_t^{\,i}\,(1 - \mathcal{U}_t^i) \in [0,1]. \tag{A.21}$$

Theorem A.11 (Weighted policy-gradient equivalence)**.** *Let $d(t)$ denote the sampling distribution over timesteps induced by rollouts, and assume $\mathbb{E}_{t'\sim d}[\omega^i_{t'}] > 0$. Define a reweighted distribution $d_\omega(t)$ by*

$$d_\omega(t) \;=\; \frac{\omega^i_t\, d(t)}{\mathbb{E}_{t'\sim d}[\omega^i_{t'}]}\,. \tag{A.22}$$

Then the descent direction induced by Eq. (A.21) *satisfies*

$$-\nabla_{\theta_i}\mathcal{L}^{act}_i = \mathbb{E}_{t'\sim d}[\omega^i_{t'}]\,\mathbb{E}_{t\sim d_\omega}\Big[\nabla_{\theta_i}\log\pi_{\theta_i}(a^i_t \mid h^i_t)\,\hat{A}^i_t\Big], \tag{A.23}$$

and is therefore equivalent, up to the positive scalar $\mathbb{E}_{t'\sim d}[\omega^i_{t'}]$, to the standard policy-gradient direction under d_ω.

Proof. Taking the negative gradient of Eq. (A.21) gives

$$-\nabla_{\theta_i}\mathcal{L}^{\text{act}}_i = \mathbb{E}_{t\sim d}\Big[\omega^i_t\nabla_{\theta_i}\log\pi_{\theta_i}(a^i_t \mid h^i_t)\hat{A}^i_t\Big]\,.$$

Let

$$g_t = \nabla_{\theta_i}\log\pi_{\theta_i}(a^i_t \mid h^i_t)\hat{A}^i_t.$$

By the definition of d_ω,

$$\mathbb{E}_{t\sim d}[\omega^i_t g_t] = \mathbb{E}_{t'\sim d}[\omega^i_{t'}]\,\mathbb{E}_{t\sim d_\omega}[g_t].$$

Substituting g_t yields the stated result. □

Lemma A.12 (Corruption attenuation (simple bound))**.** *Let $\mathcal{T}_{bad}$ be timesteps where experience is corrupted (e.g., Byzantine influence or severe uncertainty), and define $\Omega_{bad} = \sup_{t\in\mathcal{T}_{bad}}\omega^i_t$. Then the expected ℓ_2-norm contribution of corrupted timesteps to the weighted gradient is upper-bounded by Ω_{bad} times the corresponding unweighted expected ℓ_2-norm contribution under the same sampling distribution. Here and below $\|\cdot\|$ denotes the Euclidean (ℓ_2) norm on the parameter space of θ_i.*

Proof. Let g_t be any measurable vector-valued quantity. Then

$$\|\mathbb{E}[\omega_t\, g_t\, \mathbf{1}_{t\in\mathcal{T}_{\text{bad}}}]\| \le \mathbb{E}[\omega_t\,\|g_t\|\,\mathbf{1}_{t\in\mathcal{T}_{\text{bad}}}] \le \Omega_{\text{bad}}\,\mathbb{E}[\|g_t\|\,\mathbf{1}_{t\in\mathcal{T}_{\text{bad}}}]\,,$$

where $\Omega_{\text{bad}} = \sup_{t\in\mathcal{T}_{\text{bad}}}\omega^i_t$. The first inequality is Jensen's inequality, and the second uses $\omega_t \le \Omega_{\text{bad}}$ on $\mathcal{T}_{\text{bad}}$. □

A.6 Consistency Regularizer Reduces Team Disagreement

Recall the (per-agent) consistency term

$$\mathcal{L}^{\text{cons}}_i = \beta\,\bar{w}^i_t\,\|\mu^i_t - \bar{\mu}_t\|_2^2, \qquad \bar{\mu}_t = \frac{1}{|V|}\sum_{u\in V}\tilde{\mu}^u_t. \tag{A.24}$$

Proposition A.13 (Disagreement energy is minimized at consensus)**.** *Fix t and treat $\{\mu^i_t\}_{i=1}^N$ as free variables. Define the disagreement energy*

$$\mathcal{E}_t = \sum_{i=1}^N \bar{w}^i_t\,\|\mu^i_t - m\|_2^2 \tag{A.25}$$

over a reference point $m\in\mathbb{R}^d$. Then $\mathcal{E}_t$ is minimized by the weighted average

$$m^\star = \frac{\sum_{i=1}^N \bar{w}^i_t\,\mu^i_t}{\sum_{i=1}^N \bar{w}^i_t}. \tag{A.26}$$

Proof. Differentiate $\mathcal{E}_t$ with respect to m: $\nabla_m\mathcal{E}_t = -2\sum_i \bar{w}^i_t(\mu^i_t - m)$. Set to zero and solve for m. □

Remark A.14. Proposition A.13 formalizes why a squared-distance consistency term promotes aligned beliefs: it penalizes dispersion around a shared reference. In our implementation, $\bar{\mu}_t$ is a trust-aware shared reference produced by the SMM, so the penalty encourages agreement with the team picture.



(A) Intersection **(B) Avenue** **(C) Curve**

Figure 3: **CARLA Town 10 evaluation maps.** Overhead views of the three Town 10 test locations used for generalization analysis: (A) Intersection, (B) Avenue, and (C) Curve, which induce distinct interaction and geometry regimes for coordinated driving.

B Implementation Details

This appendix provides additional implementation details for `CG-HMT`, the baseline methods, and the evaluation protocol. The goal is to make the experimental setup reproducible while ensuring that performance differences are attributable to the coordination mechanisms rather than differences in observation processing, model capacity, training budget, or evaluation procedure. Unless otherwise stated, all methods use the same CARLA/MACAD-Gym environment interface, discrete action space, visual encoder, training budget, and evaluation protocol.

B.1 Shared Environment and Training Configuration

Environment. All experiments use the CARLA simulator through the MACAD-Gym multi-agent driving interface. The default scenario is an intersection-driving task with $N = 3$ controlled autonomous agents. Each agent receives an egocentric RGB observation, and the environment is wrapped with a DeepMind-style preprocessing wrapper using spatial resolution 84×84 and frame stack $K = 4$. During decentralized execution, each agent acts from its own local observation stream.

Observation Preprocessing and Visual Encoder. The environment wrapper internally stacks four RGB frames, but the custom image preprocessor retains the most recent RGB slice and maps the observation to an $84 \times 84 \times 3$ tensor before passing it to the policy model. This matches the main experimental setup, where temporal context is handled through rollout history and trust-state bookkeeping rather than explicit frame stacking. All policies use the registered `mnih15_shared_weights` visual model. This shared encoder choice ensures that all methods use the same visual representation and that performance differences are not caused by different image-processing backbones.

Action Space. The action space is discrete with $|\mathcal{A}| = 9$ driving primitives. These actions are obtained by combining three steering modes,

$$\{\texttt{left}, \texttt{straight}, \texttt{right}\}, \tag{A.27}$$

with three longitudinal-control modes,

$$\{\texttt{coast}, \texttt{accelerate}, \texttt{decelerate}\}. \tag{A.28}$$

The same action space is used for `CG-HMT` and all baselines.

Training Configuration. Training uses PPO through the shared RLlib training interface. The training budget is fixed to 300 outer iterations. Each update uses a rollout fragment length of 200, a training batch size of 12,000 per-agent transitions, and 10 stochastic-gradient epochs per update. We use one environment per worker and no remote workers in the human-interface setting, so that the feedback interface is updated only on the local worker. Other optimizer settings follow the corresponding RLlib defaults unless explicitly specified by a method. All methods are evaluated using the same fixed evaluation seeds and the same number of evaluation episodes per condition.

Reward Shaping (Shared Across All Methods). All methods, including every baseline and CG−HMT, are trained on the identical shaped reward defined here; no method receives a reward signal that another does not. The shaped reward for agent i is

$$r_t^i = r_{t,\text{env}}^i + \lambda_{\text{prog}}\Delta\text{prog}_t^i - \lambda_{\text{coll}}\Delta\text{coll}_t^i - \lambda_{\text{infr}}\Delta\text{infr}_t^i. \tag{A.29}$$

The default coefficients are

$$\lambda_{\text{prog}} = 2.0, \qquad \lambda_{\text{coll}} = 2.0, \qquad \lambda_{\text{infr}} = 1.0. \tag{A.30}$$

Here, Δprog_t^i is the normalized stepwise reduction in distance to the goal, while Δcoll_t^i and Δinfr_t^i are non-negative increments in collision and infraction counters. Sensitivity to the progress and collision coefficients is reported in Appendix C.5 (Sweeps 2 and 3).

B.2 Baseline Instantiations

We adapt each baseline to the same CARLA/MACAD-Gym interface while preserving its main algorithmic principle. All baselines use the shared observation encoder, the same nine-action discrete driving space, the same evaluation protocol, and the same shaped reward of Eq. (A.29). In particular, the shaping terms are not specific to CG−HMT: the clean-condition comparison in Table 1 is therefore a matched-reward comparison, and the reported gap cannot be attributed to a denser reward signal.

- **Multi-Agent Proximal Policy Optimization (MAPPO).** MAPPO is implemented as a centralized-training, decentralized-execution PPO baseline. During training, the value function can use aggregated multi-agent information, while each policy acts from its own local encoded observation at execution time. MAPPO serves as the primary strong MARL baseline because it is widely used for cooperative multi-agent control and is compatible with the discrete action setting.
- **Heterogeneous-Agent Proximal Policy Optimization (HAPPO).** HAPPO follows the heterogeneous-agent PPO formulation with sequential agent-wise updates. The method is included to test whether the gains of CG−HMT can be explained by improved multi-agent policy optimization alone rather than by explicit trust and cognitive-state modeling.
- **Multi-Agent Deep Deterministic Policy Gradient (MADDPG).** MADDPG is used as a centralized-critic, decentralized-actor baseline. The actor for each agent conditions on its local observation, while the critic uses joint information during training. Since the environment uses discrete driving primitives, the action interface is adapted to the shared categorical action space.
- **Shaping Advice in Deep Multi-Agent RL (SAM).** SAM uses the same base policy architecture as the PPO-style baselines but augments learning with potential-based reward shaping. This baseline separates the effect of denser task feedback from the proposed trust-aware cognitive mechanism.
- **Hierarchical Human-Knowledge Guidance (HHK).** HHK is instantiated as a hierarchical human-guided baseline. A high-level guidance component selects among behavioral options, while decentralized low-level policies execute the selected behavior. HHK is included because it is the closest baseline in spirit to human-guided coordination, although it does not maintain the compact trust-state dynamics and divergence-triggered interface used by CG−HMT. It receives the same human-intervention budget as CG−HMT.
- **Independent PPO (IPPO).** IPPO (de Witt et al., 2020) trains one PPO learner per agent with a decentralized critic and no shared value function, using the same encoder, action space, reward, and budget. It tests how much of the gain requires centralized training at all.
- **Byzantine-robust MARL (Byz-robust MARL).** Evaluated only under the Byzantine shift, this baseline follows the Bayesian-game formulation of Li et al. (2024), which trains agents to act robustly against teammates of unknown type, with the same encoder, action space, reward, and budget. It tests whether a dedicated Byzantine-robust learner, without human input or an explicit trust graph, matches CG−HMT under attack.

The baselines are intended to provide controlled comparisons under a unified driving benchmark. No baseline received a separate hyperparameter search: each uses the shared setup with default optimizer settings except where its algorithm required a change. Individually tuning each baseline could change absolute performance; the shared setup keeps the comparison controlled, but it does not by itself isolate individual CG−HMT components (see the ablations in Table 5).

B.3 CG−HMT IMPLEMENTATION DETAILS

Individual Mental Model (IMM). Each agent maintains a compact individual mental model

$$M_t^i = (\mu_t^i, \phi_t^i, \mathcal{U}_t^i), \tag{A.31}$$

where μ_t^i is a compact belief vector, ϕ_t^i is an intent distribution over discrete action primitives, and $\mathcal{U}_t^i$ is an evidential uncertainty proxy. The belief vector is

$$\mu_t^i \in \mathbb{R}^3, \tag{A.32}$$

and is computed as an exponential moving average over motion-relevant simulator metadata $[x, y, v]$, where x and y denote position and v denotes speed. The smoothing coefficient is

$$\rho_\mu = 0.2. \tag{A.33}$$

Evidential Intent and Uncertainty. The intent state is represented over the same $K = 9$ discrete driving primitives used by the policy. For each agent, we maintain exponentially decayed action evidence $e_t^i \in \mathbb{R}^K_{\geq 0}$ with decay factor 0.95.

At each step, the bin corresponding to the executed primitive is incremented before decay:

$$e_{t+1,k}^i = 0.95\, e_{t,k}^i + \mathbb{1}[a_t^i = k]. \tag{A.34}$$

The evidence is converted into Dirichlet parameters by

$$\alpha_{t,k}^i = e_{t,k}^i + 1. \tag{A.35}$$

The expected intent distribution is

$$\phi_{t,k}^i = \frac{\alpha_{t,k}^i}{\sum_{\ell=1}^K \alpha_{t,\ell}^i}, \tag{A.36}$$

and the uncertainty proxy is

$$\mathcal{U}_t^i = \mathrm{clip}_{[0,1]}\left(\frac{K}{\sum_{k=1}^K \alpha_{t,k}^i}\right). \tag{A.37}$$

Thus, repeated evidence for stable action choices lowers uncertainty, while sparse or inconsistent evidence keeps uncertainty high.

Peer-Expectation Implementation. The peer-expectation component is implemented as a per-peer intent table $\psi_t^i = \{\psi_t^{i\to j}\}_{j\neq i}$ built from recent peer action evidence. Each $\psi_t^{i\to j}$ is a smoothed distribution over the same K primitive actions as ϕ_t^j. The expectation-mismatch trigger uses $\mathrm{JS}(\psi_t^{i\to j} \| \phi_t^j)$. The "No expectation head" ablation removes this table and disables the corresponding JS-based divergence trigger, while leaving belief-gap and uncertainty triggers unchanged.

Team Reference State. At each step, the implementation computes a trust-weighted team reference belief and intent from the current agent-level mental states:

$$\tilde{\mu}_t^u = \frac{\sum_{v\neq u} w_t^{u\to v} \mu_t^v}{\sum_{v\neq u} w_t^{u\to v} + \varepsilon}, \qquad \tilde{\phi}_t^u = \mathrm{Normalize}\left(\prod_{v\neq u} (\phi_t^v)^{\bar{w}_t^{u\to v}}\right), \tag{A.38}$$

where

$$\bar{w}_t^{u\to v} = \frac{w_t^{u\to v}}{\sum_{v'\neq u} w_t^{u\to v'} + \varepsilon}. \tag{A.39}$$

These reference quantities are used for agreement estimation, divergence detection, and diagnostic logging. The symbols match the main text: $\tilde{\mu}_t^u, \tilde{\phi}_t^u$ are the per-node fusions of Eq. (4), and $\bar{\mu}_t, \bar{\phi}_t$ (no superscript) remain the global averages over V.

Information Available to the Policy. The belief vector $\mu_t^i = [x, y, v]$ is drawn from simulator state, so we state precisely where it is and is not used. It is *not* an input to any actor: every method in this paper, `CG-HMT` included, selects actions from the same $84 \times 84 \times 3$ egocentric RGB tensor and nothing else, so the decentralized execution interface is identical across methods and the comparison is not confounded by observation asymmetry. The belief vector enters only the trust, agreement, and divergence bookkeeping, which modulates the *training* objective. Using privileged state at training time while restricting execution to local observations is precisely the CTDE setting adopted by the centralized-critic baselines, which likewise consume joint information in their critics. A consequence worth stating plainly is that μ_t^i carries no gradient with respect to the policy parameters θ_i: the consistency term of Eq. (A.24) therefore acts on the trust-weighted team reference used for coordination and divergence detection, and we do not claim it shapes the visual encoder's representation. Accordingly, in the present CARLA implementation, changes in the trust state during fixed-policy evaluation do not directly modify the actor logits; their effect on learned behavior arises through the trust-modulated training objective.

Trust State and Automatic Trust Update. For each controlled agent, the implementation maintains the operative human-to-agent trust state used by the dashboard, along with automatic, human, and combined components for logging. All trust values are initialized to 0.5. The automatic trust proposal combines performance, agreement, calibration:

$$\hat{w}_t^i = \lambda_P P_t^i + \lambda_A A_t^i + \lambda_C C_t^i \tag{A.40}$$

with default coefficients

$$\lambda_P = 0.4, \qquad \lambda_A = 0.3, \qquad \lambda_C = 0.3, \qquad \lambda_U = 0.2. \tag{A.41}$$

The performance term is computed from the current reward:

$$P_t^i = \frac{1}{2}\left(1 + \tanh(r_t^i)\right). \tag{A.42}$$

The agreement term combines belief alignment and intent consistency:

$$A_t^i = \exp\left(-\|\mu_t^i - \tilde{\mu}_t^H\|_2/\sigma_\mu\right)\left(1 + \mathrm{JS}(\phi_t^i \parallel \tilde{\phi}_t^H)\right)^{-1}. \tag{A.43}$$

Here, the superscript H denotes the human node's fused view, i.e. the human-to-agent dashboard reference used by the implementation. The calibration term is approximated using evidential confidence:

$$C_t^i = 1 - \mathcal{U}_t^i. \tag{A.44}$$

Automatic trust is updated using exponential smoothing:

$$w_{t+1,\mathrm{auto}}^i = \mathrm{clip}_{[0,1]}\left[(1-\alpha) w_{t,\mathrm{auto}}^i + \alpha \hat{w}_t^i\right], \tag{A.45}$$

where the default trust update rate is

$$\alpha = 0.10. \tag{A.46}$$

The combined trust displayed to the operator and logged for analysis is

$$w_{t+1,\mathrm{comb}}^i = \mathrm{clip}_{[0,1]}\left[(1-\alpha) w_{t,\mathrm{comb}}^i + \alpha\left(\hat{w}_t^i + \delta_t^{H\to i} - \lambda_U \mathcal{U}_t^i\right)\right]. \tag{A.47}$$

The automatic and human components are retained for visualization and logging, while $w_{t,\mathrm{comb}}^i$ is the operative trust value used by the coordination layer.

Divergence-Triggered Prompting. The interface marks an agent as cognitively divergent when any of the following conditions holds:

$$\|\mu_t^i - \bar{\mu}_t\|_2 > \tau, \qquad \mathcal{U}_t^i > \kappa, \qquad \exists j \neq i:\ \mathrm{JS}(\psi_t^{i\to j} \parallel \phi_t^j) > \zeta. \tag{A.48}$$

The default thresholds are

$$\tau_0 = 5.0, \qquad \kappa_0 = 0.60, \qquad \zeta_0 = 0.10. \tag{A.49}$$

The sensitivity of CG−HMT to these thresholds is evaluated in Appendix C.5. When a divergence event is detected, the dashboard marks the corresponding agent as divergent and allows the operator to provide one of three responses:

$$\texttt{Reliable}, \qquad \texttt{Monitor}, \qquad \texttt{Unreliable}. \tag{A.50}$$

These responses are mapped to bounded trust nudges:

$$\texttt{Reliable} \mapsto +0.2, \qquad \texttt{Monitor} \mapsto 0, \qquad \texttt{Unreliable} \mapsto -0.2. \tag{A.51}$$

If no response is provided, no human nudge is applied and the automatic trust update continues.

Reward. CG−HMT is trained on the same shaped reward as every baseline; the definition and coefficients are given once in Appendix B.1 (Eq. (A.29)).

Perturbation Settings. For robustness evaluation, the implementation supports three perturbation modes. In the V2X setting, an agent's current observation is replaced with its previous observation with probability $p_{\text{drop}} = 0.4$, modeling stale communication or delayed perception. In the Byzantine setting, one agent is selected as faulty, and its selected action is replaced by a different random primitive with probability $p_{\text{byz}} = 0.5$ (the same rule stated in §5). In the weather setting, visual observations are degraded using a Gaussian blur with a kernel size 7×7, and darkening with intensity scale 0.7 and offset -30.

Logged Quantities. The implementation records both episode-level driving metrics and step-level cognitive variables. Episode-level logs include route completion, collisions, infractions, prompt frequency, average shaped reward, and cumulative average shaped reward. Step-level logs include $\mu_t^i, \mathcal{U}_t^i$, total evidence, automatic trust, human trust, combined trust, belief gap, JS intent mismatch, divergence flags, base reward, shaped reward, collision increments, and infraction increments.

B.4 Evaluation Metrics and Statistical Reporting

We report five task-level metrics across all methods: success rate, route completion, collisions per episode, infractions per episode (both event counts, not probabilities, averaged across evaluation episodes), and minimum distance to goal. For human-feedback and threshold-sensitivity analyses, we additionally report prompt frequency because divergence thresholds affect operator workload. Unless otherwise stated, each result is computed over 20 fixed-seed evaluation episodes.

For binary success rates, we report Wilson confidence intervals. The clean benchmark evaluates all three trained seeds per method, giving $3 \times 20 = 60$ held-out episodes per method; success counts are pooled and compared by two-sided Fisher exact test. CG−HMT succeeds on $45/60$ episodes (Wilson 95% CI $[0.63, 0.84]$). Table 4 reports the pairwise comparisons. Every baseline is separated at $p < 10^{-8}$, and the ordering of methods is identical under a per-seed paired comparison, so the conclusion does not depend on pooling. Pooled-episode tests treat the 60 episodes per method as independent, which episodes from the same trained policy need not be, so they should be read as descriptive; the Wilson intervals carry the same assumption. Seed-level uncertainty is coarse with three seeds: the smallest attainable exact two-sided permutation p-value is $2/\binom{6}{3} = 0.10$. Nevertheless every CG−HMT seed exceeds every baseline seed, and the 95% Welch interval for the SR difference is $[0.44, 0.66]$ against MAPPO (all intervals against the other baselines lie above 0.48). After Bonferroni adjustment over the six comparisons, the largest pooled p-value is 1.2×10^{-8} (MAPPO); Table 4 lists adjusted values. Continuous metrics in Table 1 are reported as mean $\pm$ standard deviation *across the three seed-level means*, which measures algorithmic stability; the robustness, ablation, generalization, and sensitivity tables instead report mean $\pm$ standard deviation across the 20 evaluation episodes of a single reference seed, and their captions state this.

Interpretation notes. *Trust-share normalization.* The plotted quantity $\bar{w}_t^i = \frac{1}{|V|-1} \sum_{u \neq i} w_t^{u \to i}$ is a column mean over trusters, so the plotted means are not entries of one truster's row of $\mathbf{W}_t$. The normalized share in Theorem 4.1 is computed per truster over that truster's trustees $v \neq u$, which for an honest agent are the Byzantine agent, the other honest agent, and the human node H. Substituting the plotted means for these three weights gives $0.19/(0.19 + 2 \times 0.59) \approx 0.139$ under

Table 4: **Pooled significance tests** on the clean benchmark, CG−HMT $(45/60)$ versus each baseline. Two-sided Fisher exact test; Wilson 95% CIs on pooled success. Episodes are treated as independent; the last column multiplies the Fisher p by the number of comparisons (6).

Baseline	Successes	Pooled SR (Wilson CI)	Odds ratio	Fisher p	Bonf. p ($\times 6$)
MAPPO	$12/60$	$0.200\ [0.12, 0.32]$	12.0	1.9×10^{-9}	1.2×10^{-8}
HAPPO	$1/60$	$0.017\ [0.00, 0.09]$	177.0	1.8×10^{-18}	1.0×10^{-17}
MADDPG	$7/60$	$0.117\ [0.06, 0.22]$	22.7	1.3×10^{-12}	7.7×10^{-12}
SAM	$1/60$	$0.017\ [0.00, 0.09]$	177.0	1.8×10^{-18}	1.0×10^{-17}
HHK	$9/60$	$0.150\ [0.08, 0.26]$	17.0	3.1×10^{-11}	1.9×10^{-10}
IPPO	$7/60$	$0.117\ [0.06, 0.22]$	22.7	1.3×10^{-12}	7.7×10^{-12}
CG−HMT	$45/60$	$0.750\ [0.63, 0.84]$	—	—	—

the assumption that the human-node weight is comparable to the honest-agent mean. Computing the share directly from the logged step-level trust matrices, per truster and averaged over episodes, gives 0.138; the main text reports this value, and its agreement with the substitution above indicates that the assumption is benign here.

Ablation sampling noise. With 20 episodes per variant, Wilson 95% intervals for SR are $[0.53, 0.89]$ (CG−HMT, $15/20$), $[0.18, 0.57]$ (no trust graph, $7/20$), $[0.34, 0.74]$ (no human feedback, $11/20$), $[0.39, 0.78]$ (no evidential UQ and no expectation head, $12/20$ each), and $[0.43, 0.82]$ (trust-learning only, $13/20$). Against CG−HMT, unadjusted Fisher p-values are 0.02 (no trust graph), 0.32 (no human feedback), 0.50 (no evidential UQ and no expectation head), and 0.73 (trust-learning only). Only the first is below 0.05, and it does not survive a Bonferroni adjustment over the five comparisons ($p \approx 0.12$).

Pooled versus seed-level inference. Because episodes from one policy need not be independent, we report pooled-episode tests as descriptive effect summaries rather than as confirmatory inference, and we report Bonferroni-adjusted values in Table 4. Bonferroni is deliberately conservative; Holm's procedure yields the same conclusions on these data, since every adjusted p remains below 10^{-7}. The substantive claim does not rest on the pooled test: every CG−HMT seed exceeds every baseline seed, which is the seed-level statement.

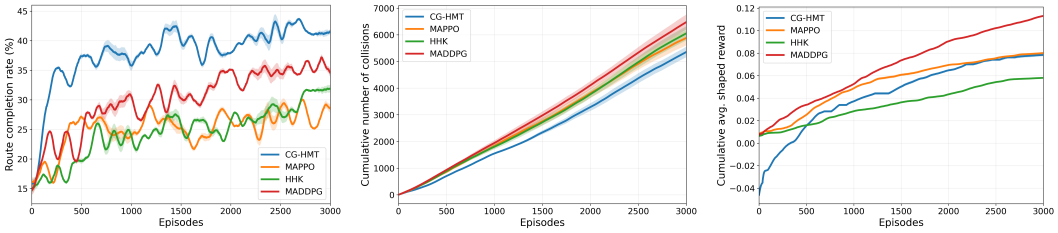
B.5 Scope and Limitations of the Implementation

The implementation is designed to evaluate whether compact cognitive-state modeling, uncertainty-aware trust updates, and divergence-triggered human feedback improve simulated multi-agent driving coordination under controlled perturbations. It is not intended to represent a production autonomous-driving stack. The current setup uses simulated CARLA observations, a discrete action space, three controlled agents, compact belief and intent summaries, and a lightweight human feedback interface. These choices make the system reproducible and suitable for controlled comparison, but they limit claims about real-world deployment. The results should therefore be interpreted as evidence for trust-aware cognitive coordination in simulated multi-agent driving, rather than as a complete validation of operational human-machine teaming.

C Extended Experimental Analyses

C.1 Per-Seed Clean-Benchmark Breakdown

Table 1 aggregates across seeds. For completeness, Table 6 reports the full five-metric result for the reference seed (seed 1) with dispersion across its 20 evaluation episodes, matching the reporting convention used in the robustness, ablation, and sensitivity tables. The minimum-distance-to-goal metric is reported here rather than in the main table: CG−HMT attains the lowest value (0.85 m), suggesting that even episodes terminating early still close most of the route gap.



(a) Route completion (%) (b) Cumulative collisions (c) Cumulative avg. shaped reward

Figure 4: **Training dynamics** across methods (3 seeds, mean ± std across seeds).

Table 5: **Component ablation analysis.** Performance on clean scenarios. Values are 20-episode evaluations of separately trained variants (one seed each); the CG−HMT row equals the reference-seed row of Table 6. Variants: *No trust graph* fuses with uniform weights and no trust updates; *No human feedback* and *Trust-learning only* both receive no operator prompts or nudges, and *No human feedback* additionally drops the trust-modulated weighting of the actor loss (*Trust-learning only* keeps automatic trust estimation and the trust/uncertainty-modulated objective); *No evidential UQ* sets $\mathcal{U}_t^i{=}0$; *No expectation head* disables the peer-expectation mismatch criterion.

Variant	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓	Min dist. to goal ↓
CG−HMT	**0.750**$_{\pm 0.44}$	**52.21**$_{\pm 11.83}$	1.25$_{\pm 0.64}$	**3.15**$_{\pm 1.44}$	**0.85**$_{\pm 1.06}$
No trust graph	0.350$_{\pm 0.49}$	42.20$_{\pm 12.46}$	**1.10**$_{\pm 0.64}$	3.20$_{\pm 1.44}$	1.56$_{\pm 0.99}$
No human feedback	0.550$_{\pm 0.51}$	47.80$_{\pm 13.10}$	1.45$_{\pm 0.68}$	3.85$_{\pm 1.62}$	1.10$_{\pm 1.20}$
No evidential UQ	0.600$_{\pm 0.50}$	49.30$_{\pm 12.70}$	1.40$_{\pm 0.66}$	3.60$_{\pm 1.55}$	1.02$_{\pm 1.12}$
Trust-learning only	0.650$_{\pm 0.49}$	50.10$_{\pm 12.20}$	1.30$_{\pm 0.62}$	3.35$_{\pm 1.48}$	0.95$_{\pm 1.08}$
No expectation head	0.600$_{\pm 0.50}$	50.90$_{\pm 12.00}$	1.28$_{\pm 0.61}$	3.25$_{\pm 1.46}$	0.90$_{\pm 1.05}$

Note. Values are reported as mean ± std over 20 evaluation episodes on clean scenarios. Higher is better for success rate and route completion; lower is better for collisions, infractions, and minimum distance to goal. Best values are bolded.

C.2 Episode Termination Breakdown

Table 7 decomposes how clean-condition episodes end. The pattern is uniform across baselines and distinct for CG−HMT: between 60% and 85% of baseline episodes terminate by timeout rather than by collision, at mean speeds of 0.75–1.45 m/s, which is consistent with mutual yielding at a symmetric unsignalized intersection rather than with unsafe driving. CG−HMT times out on 10% of episodes at 2.65 m/s. Collision shares are comparable across methods (10–20%), so the gap in success rate is not explained by baselines crashing more; it is consistent with their failing to resolve right-of-way within the episode (timeouts).

If a deadlock reading is correct, injecting asymmetry should relieve it. Table 8 shows that, for the two baselines with V2X termination data (HAPPO, MAPPO), V2X staleness cuts timeouts by 55 and 45 percentage points, while timeouts for CG−HMT move by +5 points (single reference seed, 20 episodes). This is consistent with the deadlock reading, but it is not an intervention on yielding state: dropout also changes the observations and could alter timeouts by other routes, so we treat the mechanism as a hypothesis. Only HAPPO and MAPPO were analyzed for V2X termination, so Table 8 covers two of the six baselines; breakdowns for MADDPG, SAM, HHK, and IPPO under dropout are left for future work.

C.3 Robustness to Imperfect Human Input

The oversight protocol is designed so that human error is bounded rather than catastrophic: each label maps to a nudge $\delta_t^{H\to i} \in [-\delta_{\max}, \delta_{\max}]$ with $\delta_{\max} = 0.2$, non-response commits a No−Op, and the automatic proposal of Eq. (5) continues regardless. To test this empirically we replay the same fixed checkpoint on the clean evaluation seeds while (i) flipping a fraction of operator labels between Reliable and Unreliable, and (ii) delaying every label by a fixed number of environment steps.

Table 6: **Reference-seed clean benchmark.** Mean ± std across the 20 evaluation episodes of seed 1.

Method	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓	Min dist. to goal ↓
MAPPO	$0.200_{\pm 0.41}$	$29.90_{\pm 20.03}$	$1.70_{\pm 0.73}$	$5.10_{\pm 2.61}$	$2.72_{\pm 3.18}$
HAPPO	$0.000_{\pm 0.00}$	$29.56_{\pm 12.67}$	$1.15_{\pm 0.59}$	$3.45_{\pm 1.32}$	$2.83_{\pm 3.06}$
MADDPG	$0.100_{\pm 0.31}$	$30.09_{\pm 15.76}$	$\mathbf{1.10_{\pm 0.85}}$	$2.95_{\pm 1.15}$	$2.61_{\pm 3.44}$
SAM	$0.000_{\pm 0.00}$	$18.53_{\pm 13.12}$	$\mathbf{1.10_{\pm 0.72}}$	$\mathbf{2.90_{\pm 1.07}}$	$4.40_{\pm 2.92}$
HHK	$0.150_{\pm 0.37}$	$34.49_{\pm 13.21}$	$1.85_{\pm 0.81}$	$11.80_{\pm 8.85}$	$0.93_{\pm 0.92}$
CG-HMT	$\mathbf{0.750_{\pm 0.44}}$	$\mathbf{52.21_{\pm 11.83}}$	$1.25_{\pm 0.64}$	$3.15_{\pm 1.44}$	$\mathbf{0.85_{\pm 1.06}}$

Table 7: **How clean-condition episodes terminate** (reference seed, 20 episodes per method). Rows sum to 100%. Baselines fail predominantly by timeout at low speed, consistent with a yielding deadlock.

Method	Success (%)	Timeout (%)	Collision (%)	Other (%)	Mean speed (m/s)
MAPPO	20.0	65.0	10.0	5.0	1.35
HAPPO	0.0	85.0	10.0	5.0	0.82
MADDPG	10.0	70.0	15.0	5.0	1.10
SAM	0.0	80.0	10.0	10.0	0.75
HHK	15.0	60.0	20.0	5.0	1.45
CG-HMT	**75.0**	**10.0**	10.0	5.0	**2.65**

Table 9 shows graceful degradation of success under the tested corruptions. Fully randomized labels (50% flipped) and a 20-step delay both yield $\mathrm{SR} = 0.65$, matching the observed "trust-learning only" result with no operator input (Table 5). Route completion, collisions, and infractions under these two corruptions are slightly worse than in the no-human row (49.80 and 50.00 vs. 50.10; 1.38 and 1.36 vs. 1.30; 3.50 and 3.45 vs. 3.35). These differences are small relative to episode-level variability, but they mean the no-human row acts as a floor for success and not for the other metrics in these runs. This experiment does not establish robustness to an adaptive or strategically timed malicious operator; testing such feedback policies is left for future work.

This section reports generalization, sensitivity, human workload, and runtime analyses under the same evaluation protocol used in the main paper. Unless otherwise noted, we evaluate each configuration for 20 episodes and report mean ± std. Generalization varies one environmental factor at a time while keeping the remaining factors at their default values (Town10, Intersection layout, traffic density 0, and 3 agents). Sensitivity sweeps vary one CG-HMT parameter at a time around its default while keeping the environment at the default condition. Human workload summarizes operator interaction frequency and magnitude under the CG-HMT trust interface. Runtime summarizes end-to-end training throughput and wall-clock cost, computed from Ray logs and our runtime counters.

C.4 Generalization Experiments

Town Generalization. We evaluate robustness to map changes by testing Town01, Town02, and Town03 while preserving the default scenario structure and team size. Results in Table 10 show that CG-HMT retains strong success and route completion across towns. Across methods, town changes mainly affect collisions and infractions due to altered geometry and spawn dynamics; CG-HMT shows smaller degradations, which is consistent with, but does not isolate, transfer of its coordination and trust mechanisms beyond the training town.

Layout Generalization. We evaluate robustness to road topology changes by testing Avenue and Curve layouts while keeping the town shown in Figure 3, traffic density, and team size fixed at their default values. Table 11 indicates that CG-HMT maintains high route completion and success across layouts, with expected increases in collision and infraction counts on harder geometry (Curve). The performance gap versus defaults remains moderate, suggesting that the learned coordination policy and trust updates do not overfit to a single intersection pattern.

Table 8: **Timeout rate, clean versus V2X dropout.** Baseline timeouts fall under staleness, consistent with a symmetric-yielding explanation; the CG−HMT timeout share is low in both conditions.

Method	Clean timeout (%)	V2X timeout (%)	Δ
HAPPO	85.0	30.0	−55.0
MAPPO	65.0	20.0	−45.0
CG−HMT	10.0	15.0	+5.0

Table 9: **Sensitivity to corrupted and delayed operator labels.** Clean evaluation, fixed checkpoint, 20 episodes per row. The final row is the no-human reference (a floor for SR in these runs).

Label condition	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓
Clean labels (default)	**0.750**	**52.21**	**1.25**	**3.15**
10% flipped	0.750	51.90	1.28	3.20
25% flipped	0.700	51.10	1.32	3.30
50% flipped (random)	0.650	49.80	1.38	3.50
Delayed 5 steps	0.700	51.40	1.30	3.25
Delayed 20 steps	0.650	50.00	1.36	3.45
No human (No−Op always)	0.650	50.10	1.30	3.35

Traffic Density Generalization. We evaluate robustness to scene complexity by varying traffic density from the default 0 to 5 and 10 while keeping town, layout, and team size fixed. Table 12 shows a monotonic increase in collisions and infractions as density rises, reflecting tighter interactions and reduced safety margins. Despite this, CG−HMT preserves strong route completion and maintains a non-trivial success rate even at density 10, indicating that the coordination policy degrades gracefully as external agents increase.

Agent-Count Generalization. We evaluate scalability by varying the number of controlled agents while keeping town=Town10, layout=Intersection, and traffic density=0 fixed. We test 3 agents (default), 5 agents, and 7 agents. Table 13 shows that increasing team size predictably increases interaction complexity, leading to higher collision and infraction counts and modest reductions in success and route completion. Despite this, CG−HMT degrades gracefully as agents increase, preserving stable progress toward the goal and comparatively strong completion metrics, suggesting that its trust-weighted coordination transfers beyond the default team size.

C.5 Sensitivity Experiments

Sensitivity experiments vary one CG−HMT parameter at a time around its default value and then evaluate under the default environment condition (Town10, Intersection, traffic density 0, agents 3). We report the same task completion and safety metrics used in the main benchmark and interpret trends in terms of adaptation speed and stability. Across sweeps, moderate settings tend to balance responsiveness and safety, while extreme values can induce sluggish correction (too low) or oscillatory trust updates (too high).

Sweep 1: Trust Reactivity. We vary the trust update step size $\alpha = \texttt{trust_lr}$ around the default 0.10. The default was selected on validation episodes drawn from the training town, not on the held-out evaluation scenarios, so the sweep below is a diagnostic rather than a selection procedure. Table 14 shows that lower reactivity slows adaptation to changing teammate reliability, while higher reactivity overreacts to short-term noise.

We do not have a validated explanation for the collapse at $\alpha = 0.20$. One might appeal to Lemma A.8, which bounds the per-step trust change by $\alpha(1 + \delta_{\max} + \lambda_U)$; but with $\delta_{\max} = \lambda_U = 0.2$ this bound is 0.28 at $\alpha = 0.20$, still below the empirical honest–faulty separation of ≈ 0.4 (Figure 2a), so the bound does not by itself predict a loss of trust ordering and we do not offer it as the mechanism. We also did not log trust oscillation or fused-intent concentration in these runs. We therefore report the collapse as an empirical sensitivity, note that the default $\alpha = 0.10$ was selected on validation episodes rather than derived from any bound, and leave the mechanism open.

Table 10: **Town Generalization for `CG-HMT`.** Evaluation across CARLA towns.

Town	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓	Min dist. to goal ↓
Town01	$0.650_{\pm 0.49}$	$47.10_{\pm 13.20}$	$1.40_{\pm 0.70}$	$3.60_{\pm 1.50}$	$0.95_{\pm 1.20}$
Town02	$0.700_{\pm 0.47}$	$49.30_{\pm 12.70}$	$1.35_{\pm 0.68}$	$3.50_{\pm 1.45}$	$0.90_{\pm 1.10}$
Town03	$0.700_{\pm 0.47}$	$50.40_{\pm 12.20}$	$1.30_{\pm 0.65}$	$3.40_{\pm 1.40}$	$0.90_{\pm 1.05}$
Town10 (default)	$\mathbf{0.750_{\pm 0.44}}$	$\mathbf{52.21_{\pm 11.83}}$	$\mathbf{1.25_{\pm 0.64}}$	$\mathbf{3.15_{\pm 1.44}}$	$\mathbf{0.85_{\pm 1.06}}$

Note. Evaluation under Town01, Town02, Town03, and the default Town10 while keeping layout=Intersection, traffic density=0, and agents=3 fixed. Values are reported as mean ± std over 20 evaluation episodes. Higher is better for success rate and route completion; lower is better for collisions, infractions, and minimum distance to goal. Best values are bolded.

Table 11: **Layout Generalization for `CG-HMT`.** Evaluation across road layouts.

Layout	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓	Min dist. to goal ↓
Intersection (default)	$\mathbf{0.750_{\pm 0.44}}$	$\mathbf{52.21_{\pm 11.83}}$	$\mathbf{1.25_{\pm 0.64}}$	$\mathbf{3.15_{\pm 1.44}}$	$\mathbf{0.85_{\pm 1.06}}$
Avenue	$0.600_{\pm 0.50}$	$45.00_{\pm 13.80}$	$1.50_{\pm 0.72}$	$3.90_{\pm 1.60}$	$1.05_{\pm 1.25}$
Curve	$0.550_{\pm 0.51}$	$42.80_{\pm 14.10}$	$1.60_{\pm 0.75}$	$4.10_{\pm 1.70}$	$1.20_{\pm 1.35}$

Note. Evaluation under Avenue and Curve while keeping town=Town10, traffic density=0, and agents=3 fixed. Values are reported as mean ± std over 20 evaluation episodes. Higher is better for success rate and route completion; lower is better for collisions, infractions, and minimum distance to goal. Best values are bolded.

Sweep 2: Progress Shaping. We vary the progress shaping coefficient `prog_bonus` with the default at 2.0. Table 15 shows that higher progress shaping can improve route completion by rewarding consistent goal-directed motion, but overly aggressive shaping can encourage risk-taking in dense interaction regions, slightly increasing collisions and infractions. The default value maintains strong completion while preserving stable safety behavior.

Sweep 3: Collision Penalty. We vary the collision penalty `coll_penalty` around the default 2.0. Table 16 shows a clear reduction in collisions per episode as the penalty increases, but overly conservative penalties can reduce completion by discouraging necessary merges or gap acceptance. The default value yields a balanced operating point that avoids both excessive risk and excessive hesitation.

Sweep 4: Divergence Thresholds. We evaluate the sensitivity of `CG-HMT` to the three divergence-trigger thresholds used for human prompting: the belief-gap threshold τ, the evidential-uncertainty threshold κ, and the expectation–intent mismatch threshold ζ. Using default values $\tau_0 = 5.0$, $\kappa_0 = 0.60$, and $\zeta_0 = 0.10$, we perform a one-at-a-time sweep with $\tau \in \{2.5, 5.0, 7.5\}$, $\kappa \in \{0.48, 0.60, 0.72\}$, and $\zeta \in \{0.08, 0.10, 0.12\}$ while keeping the other thresholds fixed. As shown in Table 17, moderate threshold changes do not substantially change the qualitative behavior of `CG-HMT`: lower thresholds increase prompt frequency by reacting to smaller divergence signals, whereas higher thresholds reduce operator burden but may delay trust correction. Overall, the default configuration provides a stable tradeoff between responsiveness, safety, and human workload.

C.6 Human Oversight Protocol, Trust Dashboard, and Workload

Ethics Approval and Informed Consent. The five-operator study reported in this section was reviewed by our institutional review board and received an exempt determination under U.S. Department of Health and Human Services regulation 45 CFR 46.104(d)(2)(ii) (research involving survey procedures or observation of behavior, where any disclosure of participants' responses outside the research would not reasonably place them at risk of criminal or civil liability or be damaging to their financial standing, employability, educational advancement, or reputation). Risks were minimal and limited to those of ordinary remote computer-based interaction: operators joined a video session, viewed a CARLA driving simulator and the trust dashboard described below, and provided discrete feedback labels. Prior to participation, every operator received a written information sheet describing the study purpose, the procedure (a single 20–30 minute remote session), the data collected (anonymized response selections, timestamps, simulated driving metrics, and a post-study

Table 12: **Traffic Density Generalization for `CG-HMT`.** Evaluation across traffic densities.

Traffic density	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓	Min dist. to goal ↓
0 (default)	**0.750**$_{\pm 0.44}$	**52.21**$_{\pm 11.83}$	**1.25**$_{\pm 0.64}$	**3.15**$_{\pm 1.44}$	**0.85**$_{\pm 1.06}$
5	0.550$_{\pm 0.51}$	43.20$_{\pm 14.70}$	1.85$_{\pm 0.90}$	4.60$_{\pm 1.85}$	1.35$_{\pm 1.45}$
10	0.450$_{\pm 0.51}$	38.10$_{\pm 15.30}$	2.10$_{\pm 1.00}$	5.20$_{\pm 2.10}$	1.70$_{\pm 1.70}$

Note. Evaluation under traffic densities 5 and 10 while keeping town=Town10, layout=Intersection, and agents=3 fixed. Values are reported as mean ± std over 20 evaluation episodes. Higher is better for success rate and route completion; lower is better for collisions, infractions, and minimum distance to goal. Best values are bolded.

Table 13: **Agent-Count Generalization for `CG-HMT`.** Evaluation across team sizes.

Agents	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓	Min dist. to goal ↓
3 (default)	**0.750**$_{\pm 0.44}$	**52.21**$_{\pm 11.83}$	**1.25**$_{\pm 0.64}$	**3.15**$_{\pm 1.44}$	**0.85**$_{\pm 1.06}$
5	0.650$_{\pm 0.49}$	47.60$_{\pm 12.90}$	1.55$_{\pm 0.80}$	3.85$_{\pm 1.65}$	1.05$_{\pm 1.25}$
7	0.550$_{\pm 0.51}$	43.10$_{\pm 13.80}$	1.90$_{\pm 0.95}$	4.60$_{\pm 1.90}$	1.30$_{\pm 1.40}$

Note. Evaluation under 5 and 7 agents while keeping town=Town10, layout=Intersection, and traffic density=0 fixed. The 3-agent row matches the standard-condition baseline for direct comparison. Values are reported as mean ± std over 20 evaluation episodes. Higher is better for success rate and route completion; lower is better for collisions, infractions, and minimum distance to goal. Best values are bolded.

questionnaire), the voluntary nature of participation and the right to withdraw at any time, and the compensation (a $10 gift card). Operators provided informed consent before any data were recorded. Research data were stored under anonymous identifiers (e.g., *Operator 1, . . . , Operator 5*) and were not linked to any contact information used for scheduling or compensation. The institution name and IRB protocol number are withheld here to preserve double-blind anonymity and will be included in the camera-ready version.

Scope and Implementation. Our human-machine teaming evaluation includes both a single-operator instrumentation run and a five-operator feasibility study. The single-operator run is used to log detailed oversight dynamics, while the five-operator study checks whether the dashboard remains usable and responsive across users. Operators provide input only when the system raises a divergence-triggered prompt; otherwise, the controller runs autonomously with $\delta_t^{H\to i} = 0$ by default. Operators observe only the dashboard visualizations described below, without privileged simulator-state access, and return a discrete label that maps to a bounded trust nudge $\delta_t^{H\to i} \in [-\delta_{\max}, \delta_{\max}]$.

Human Oversight Protocol. `CG-HMT` supports lightweight human oversight that is *event-driven* rather than continuous teleoperation. The operator follows a simple protocol:

P1: Monitor: Observe the live driving views and the trust dashboard, focusing on agents flagged as `DIVERGENT`.

P2: Trigger: *Only when* a divergence event is raised (belief gap, high uncertainty, or expectation–intent conflict), the system may request a human judgement for the specific agent(s); *no prompts are issued otherwise*.

P3: Decide: For the selected agent, choose a discrete label (`Unreliable`, `Monitor`, `Reliable`) based on observed behavior and context.

P4: Apply: The label maps to a bounded trust nudge $\delta_t^{H\to i} \in [-\delta_{\max}, \delta_{\max}]$ injected into the trust update (Eq. 6).

P5: Fail-safe: If the operator does not respond (or chooses not to intervene), the system applies `No-Op` ($\delta_t^{H\to i} = 0$) and proceeds with the automatic trust proposal; *there is no per-step human signal*.

Operator Visual Channels. The operator uses three complementary visual channels during evaluation: (i) agent-centric first-person camera views for local risk and intent cues, (ii) a trust dashboard for



Figure 5: **Agent-centric first-person views.** The operator inspects local geometry and nearby traffic to contextualize divergence events (for example, blocked lanes, close-proximity interactions, or unclear right-of-way).

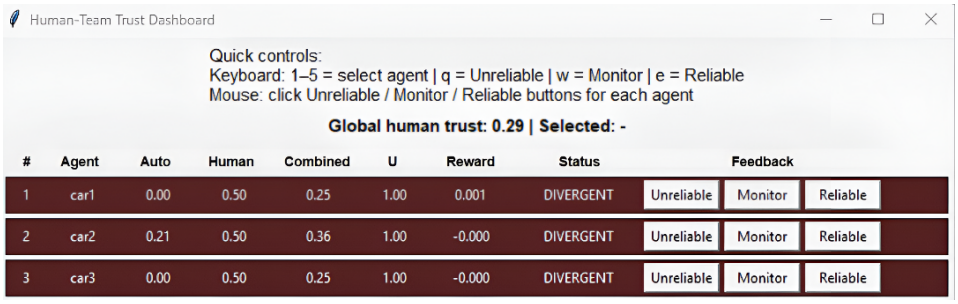


Figure 6: **Human-team trust dashboard.** The dashboard provides quick controls for selecting an agent and issuing discrete feedback (`Unreliable`/`Monitor`/`Reliable`), alongside per-agent trust, uncertainty, reward, and divergence status.

fast supervision actions and team state, and (iii) a third-person overview for diagnosing interaction-level failures (merges, intersections, near-conflicts). The operator does not access privileged simulator variables beyond what is rendered in these views and dashboard summaries.

Understanding the Trust Dashboard. Figure 6 is intentionally minimal and action-oriented:

- **Quick controls (top):** keyboard shortcuts allow rapid agent selection (e.g., `1–5`) and instant labeling; the same labels are available as per-row buttons to avoid mode switching.
- **Global human trust + selected agent (center):** displays the operator's current global trust level (summary signal) and which agent is currently targeted for feedback.
- **Per-agent row (table):** each row corresponds to one controlled agent and summarizes:
  - **Auto:** the automatic trust estimate produced by the trust dynamics from performance/agreement/calibration.



Figure 7: **Third-person scene overview.** A global view is used to understand multi-agent interactions and traffic geometry, especially when diagnosing which agent is causing conflicts or when local views are ambiguous.

Table 14: **Sensitivity to Trust Reactivity.** Clean evaluation under different `trust_lr` values.

`trust_lr`	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓	Min dist. to goal ↓
0.05	$0.700_{\pm 0.47}$	$49.80_{\pm 12.50}$	$\mathbf{1.20_{\pm 0.65}}$	$3.30_{\pm 1.40}$	$0.95_{\pm 1.10}$
0.10 (default)	$\mathbf{0.750_{\pm 0.44}}$	$\mathbf{52.21_{\pm 11.83}}$	$1.25_{\pm 0.64}$	$\mathbf{3.15_{\pm 1.44}}$	$\mathbf{0.85_{\pm 1.06}}$
0.20	$0.000_{\pm 0.00}$	$29.97_{\pm 7.40}$	$1.50_{\pm 0.70}$	$3.85_{\pm 1.55}$	$2.39_{\pm 1.25}$

Note. Clean evaluation of `CG-HMT` under different `trust_lr` values. Values are reported as mean ± std over 20 evaluation episodes. Higher is better for success rate and route completion; lower is better for collisions, infractions, and minimum distance to goal. Best values are bolded.

- **Human:** the operator-side trust level (reflecting recent human nudges or default human prior).
- **Combined:** the effective trust used by `CG-HMT` after combining automatic and human signals (the value that affects fusion and learning).
- **U:** the agent's uncertainty indicator (evidential proxy $\mathcal{U}$) used to interpret ambiguity and to damp trust growth.
- **Reward:** a short-horizon scalar summarizing recent progress/safety outcomes used for fast triage (not as the sole decision signal).
- **Status:** `DIVERGENT` indicates a detected cognitive divergence event (belief/uncertainty/expectation mismatch) for that agent.
- **Feedback:** three discrete buttons that map to bounded nudges $\delta_t^{H\to i}$: `Unreliable` (negative), `Monitor` (near-zero), `Reliable` (positive).

All clicks/keystrokes are logged as events $(t, i, \text{label}, \delta_t^{H\to i})$ for auditability and workload measurement.

Multi-Operator Human Study. We conducted a small multi-operator human study with five operators to evaluate whether the proposed dashboard provides usable and actionable trust signals across users. Each operator evaluated the same fixed trained PPO checkpoint under clean and Byzantine conditions using the same five evaluation seeds. No retraining or policy adaptation was performed during the study. Operators monitored the CARLA view and trust dashboard and provided sparse feedback through `Unreliable`, `Monitor`, and `Reliable` actions when divergence alerts appeared.

Table 15: **Sensitivity to Progress Shaping.** Clean evaluation under different `prog_bonus` values.

`prog_bonus`	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓	Min dist. to goal ↓
1.0	$0.650_{\pm 0.49}$	$47.60_{\pm 12.90}$	$\mathbf{1.20_{\pm 0.65}}$	$\mathbf{3.10_{\pm 1.35}}$	$0.95_{\pm 1.15}$
2.0 (default)	$\mathbf{0.750_{\pm 0.44}}$	$52.21_{\pm 11.83}$	$1.25_{\pm 0.64}$	$3.15_{\pm 1.44}$	$\mathbf{0.85_{\pm 1.06}}$
3.0	$0.700_{\pm 0.47}$	$\mathbf{53.10_{\pm 12.60}}$	$1.35_{\pm 0.70}$	$3.40_{\pm 1.50}$	$0.90_{\pm 1.10}$

Note. Clean evaluation of `CG-HMT` under different `prog_bonus` values. Values are reported as mean ± std over 20 evaluation episodes. Higher is better for success rate and route completion; lower is better for collisions, infractions, and minimum distance to goal. Best values are bolded.

Table 16: **Sensitivity to Collision Penalty.** Clean evaluation under different `coll_penalty` values.

`coll_penalty`	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓	Min dist. to goal ↓
1.0	$0.750_{\pm 0.44}$	$\mathbf{53.00_{\pm 12.10}}$	$1.55_{\pm 0.75}$	$\mathbf{3.10_{\pm 1.40}}$	$0.85_{\pm 1.10}$
2.0 (default)	$\mathbf{0.750_{\pm 0.44}}$	$52.21_{\pm 11.83}$	$1.25_{\pm 0.64}$	$3.15_{\pm 1.44}$	$\mathbf{0.85_{\pm 1.06}}$
3.0	$0.700_{\pm 0.47}$	$50.40_{\pm 12.80}$	$\mathbf{1.05_{\pm 0.58}}$	$3.25_{\pm 1.50}$	$0.90_{\pm 1.10}$

Note. Clean evaluation of `CG-HMT` under different `coll_penalty` values. Values are reported as mean ± std over 20 evaluation episodes. Higher is better for success rate and route completion; lower is better for collisions, infractions, and minimum distance to goal. Best values are bolded.

Tables 18 and 19 summarize the driving outcomes and workload/intervention behavior. Byzantine perturbations increase route completion but substantially degrade safety, producing higher collision and infraction values. The Byzantine condition also increases prompt frequency and negative nudges, suggesting that operators perceived more abnormal behavior under attack. Subjective trust/usability ratings are shown in Figure 8; operators rated the dashboard highly for clarity, monitoring effort, feedback incorporation, and attack-awareness.

Why route completion is lower here than in Table 1. The study runs live operator sessions, so episodes use a shorter step cap than the offline benchmark in order to keep each session within a manageable duration. Route completion is therefore truncated by the horizon rather than by driving failure, and the absolute values below are not comparable with the benchmark tables; only the clean-versus-Byzantine contrast within this table is. The apparently paradoxical result that the Byzantine condition attains *higher* route completion (23.21% vs. 18.75%) may have the same origin as the dropout effect discussed in §6: a Byzantine teammate taking random primitives could break the symmetric yielding equilibrium at the intersection and let the remaining agents make more progress within the shortened horizon. We did not test this explanation.

Table 19: **Multi-Operator Workload and Intervention Behavior.** Operator interaction statistics under clean and Byzantine conditions.

Condition	Prompts / min ↓	Divergence rate	Non-zero nudges	Nudge rate	Positive nudges	Negative nudges
Clean	$7.55_{\pm 0.40}$	$0.674_{\pm 0.010}$	$60.0_{\pm 35.4}$	$0.0119_{\pm 0.0072}$	$24.4_{\pm 10.7}$	$35.6_{\pm 28.7}$
Byzantine	$12.01_{\pm 4.79}$	$0.690_{\pm 0.022}$	$71.4_{\pm 76.2}$	$0.0153_{\pm 0.0174}$	$17.4_{\pm 20.8}$	$54.0_{\pm 56.4}$

Note. Values are reported as mean ± std across five operators. Prompts/min is a rate, while non-zero, positive, and negative nudges are raw counts per session. This rate is not comparable with the prompts/min column of Table 17, which is measured under a different protocol. The sensitivity sweep counts one prompt per divergence onset for a single agent in offline replay; the live dashboard keeps an agent flagged `DIVERGENT` and re-surfaces the prompt until the operator acts, counts all three agents, and includes unprompted interventions. Both effects inflate the live figure, and we do not decompose the difference further. Positive nudges correspond to `Reliable`; negative nudges correspond to `Unreliable`. The nudge rate is computed over divergence-triggered rows.

To complement the behavioral metrics, we collected a subjective trust and usability questionnaire after the study. The questionnaire follows the role of a lightweight trust/usability scale rather than a full NASA-TLX workload instrument: operators rated dashboard clarity, interpretability, prompt timing, monitoring effort, feedback incorporation, comfort with minimal training, and attack-awareness

Table 17: **Sensitivity to Divergence Thresholds.** Clean evaluation under one-at-a-time sweeps of τ, κ, and ζ.

Sweep	Value	Success rate ↑	Route completion (%) ↑	Collisions/ep. ↓	Infractions/ep. ↓	Prompts / min ↓
τ	2.5	$0.700_{\pm 0.47}$	$51.10_{\pm 12.30}$	$1.20_{\pm 0.62}$	$3.05_{\pm 1.38}$	$1.18_{\pm 0.34}$
τ	5.0 (default)	$\mathbf{0.750_{\pm 0.44}}$	$\mathbf{52.21_{\pm 11.83}}$	$1.25_{\pm 0.64}$	$3.15_{\pm 1.44}$	$0.85_{\pm 0.29}$
τ	7.5	$0.650_{\pm 0.49}$	$49.70_{\pm 12.90}$	$1.35_{\pm 0.70}$	$3.55_{\pm 1.62}$	$\mathbf{0.58_{\pm 0.21}}$
κ	0.48	$0.700_{\pm 0.47}$	$50.90_{\pm 12.10}$	$\mathbf{1.15_{\pm 0.59}}$	$\mathbf{2.95_{\pm 1.32}}$	$1.32_{\pm 0.37}$
κ	0.60 (default)	$\mathbf{0.750_{\pm 0.44}}$	$\mathbf{52.21_{\pm 11.83}}$	$1.25_{\pm 0.64}$	$3.15_{\pm 1.44}$	$0.85_{\pm 0.29}$
κ	0.72	$0.650_{\pm 0.49}$	$49.50_{\pm 13.20}$	$1.40_{\pm 0.73}$	$3.65_{\pm 1.70}$	$\mathbf{0.51_{\pm 0.19}}$
ζ	0.08	$0.700_{\pm 0.47}$	$51.40_{\pm 12.00}$	$\mathbf{1.18_{\pm 0.61}}$	$3.05_{\pm 1.36}$	$1.05_{\pm 0.31}$
ζ	0.10 (default)	$\mathbf{0.750_{\pm 0.44}}$	$\mathbf{52.21_{\pm 11.83}}$	$1.25_{\pm 0.64}$	$3.15_{\pm 1.44}$	$0.85_{\pm 0.29}$
ζ	0.12	$0.700_{\pm 0.47}$	$50.80_{\pm 12.50}$	$1.32_{\pm 0.67}$	$3.40_{\pm 1.52}$	$\mathbf{0.66_{\pm 0.24}}$

Note. Clean evaluation of CG−HMT under one-at-a-time sweeps of the belief-gap threshold τ, uncertainty threshold κ, and expectation–intent mismatch threshold ζ. Values are reported as mean ± std over 20 evaluation episodes. Higher is better for success rate and route completion; lower is better for collisions, infractions, and prompt frequency when task completion and safety remain stable. Default values are highlighted.

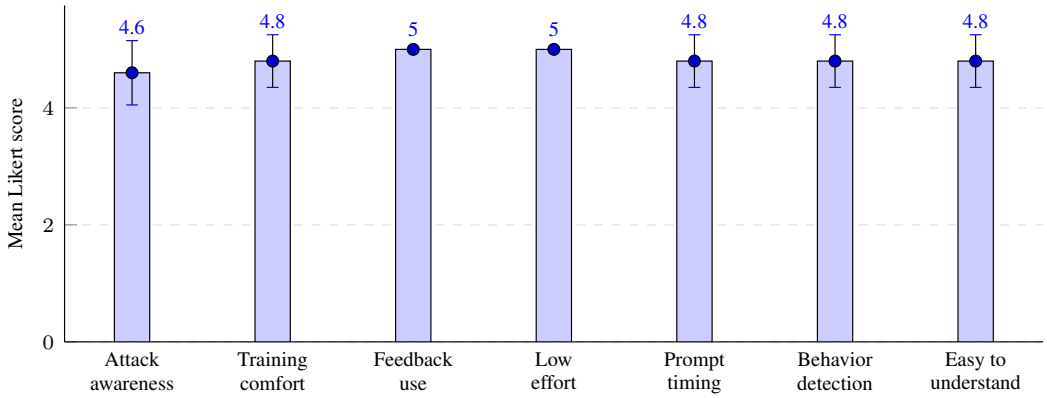


Figure 8: **Subjective trust and usability ratings** across five operators. Operators gave high ratings for dashboard clarity, monitoring effort, feedback incorporation, and attack-awareness.

on a 1–5 Likert scale. Figure 8 summarizes the responses. Ratings were high in this five-person sample, with an overall mean of 4.83/5. These descriptive responses suggest that the participating operators found the dashboard understandable and useful, but the sample is too small to support a population-level usability claim.

Overall, the multi-operator study provides formative feasibility evidence for runtime oversight. In this sample, the Byzantine condition coincides with more negative feedback, and subjective ratings are favorable; larger counterbalanced studies are needed to estimate usability and operator effects reliably.

The questionnaire comprised seven items, rated on a 1–5 Likert scale: dashboard clarity, behavior detection, prompt timing, monitoring effort, feedback incorporation, training comfort, and attack awareness; these are the items plotted in Figure 8. Participant recruitment and background, condition order and counterbalancing, and per-participant descriptive results are omitted here to preserve anonymity and will be reported from the approved study record in the camera-ready version.

C.7 Runtime Analysis

We compare end-to-end training cost and throughput across CG−HMT and three baselines (MAPPO, MADDPG, HHK) under the same training budget and hardware. All the experiments were run on a Windows workstation equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM) and an AMD Ryzen 9 5950X CPU (16 cores), unless stated otherwise. We report wall-clock minutes for the full

Table 18: **Multi-Operator Driving and Safety Outcomes.** Interface-study outcomes under clean and Byzantine conditions. Absolute route completion is horizon-limited and not comparable with Table 1; see the note above.

Condition	Operators	Episodes	RC (%) ↑	Coll. episode rate ↓	Coll. per episode ↓	Infractions/100m ↓
Clean	5	25	$18.75_{\pm 1.29}$	$0.28_{\pm 0.11}$	$0.56_{\pm 0.22}$	$0.78_{\pm 1.37}$
Byzantine	5	25	$23.21_{\pm 2.39}$	$0.88_{\pm 0.11}$	$2.40_{\pm 0.24}$	$1.96_{\pm 1.26}$

Note. Values are reported as mean ± std across five operators. Each operator evaluated five fixed seeds per condition using the same trained checkpoint. This table summarizes interface-study outcomes rather than benchmark success-rate comparisons, and therefore reports only route-progress and safety metrics.

Table 20: **Runtime Comparison.** End-to-end training runtime and throughput across methods. Values are reported as mean ± standard deviation across three runs.

Method	Wall-clock (min) ↓	Wall-clock (h) ↓	Steps / s ↑	Steps / min ↑	Seconds per 10^6 steps ↓
CG−HMT	$759.843_{\pm 18.472}$	$12.664_{\pm 0.308}$	$26.571_{\pm 0.681}$	$1594.3_{\pm 40.8}$	$37661_{\pm 963}$
MAPPO	$\mathbf{688.927_{\pm 15.814}}$	$\mathbf{11.482_{\pm 0.264}}$	$\mathbf{29.958_{\pm 0.692}}$	$\mathbf{1797.5_{\pm 41.5}}$	$\mathbf{33392_{\pm 774}}$
MADDPG	$817.183_{\pm 21.936}$	$13.620_{\pm 0.366}$	$24.063_{\pm 0.653}$	$1443.8_{\pm 39.2}$	$41604_{\pm 1128}$
HHK	$739.577_{\pm 16.285}$	$12.326_{\pm 0.271}$	$27.442_{\pm 0.591}$	$1646.5_{\pm 35.5}$	$36447_{\pm 785}$

Note. End-to-end training runtime and throughput under identical hardware and training budget. Each value is reported as mean ± standard deviation over three runs. CG−HMT includes additional IMM and trust bookkeeping, and optional UI updates, reducing throughput relative to MAPPO while remaining comparable in wall-clock time and supporting the same evaluation protocol. Best values are bolded; CG−HMT is highlighted. Throughput counts multi-agent environment steps (one step advances all three agents), about 1.2M per run; the 12,000-step training batch counts per-agent transitions, so 300 iterations correspond to 3.6M agent transitions.

run, average timesteps per wall-second, and derived throughput statistics that clarify computational overhead. As shown in Table 20, CG−HMT incurs roughly 10% wall-clock overhead relative to MAPPO (12.66 h vs. 11.48 h for the full 300-iteration run) from maintaining individual mental model state, trust-graph fusion, divergence bookkeeping, and (when enabled) lightweight UI updates. The overhead is bounded and predictable: it is constant in trajectory length and scales as $\mathcal{O}(|V|^2)$ per step in team size, since both fusion (Eq. (4)) and the trust update (Eq. (6)) iterate over directed pairs in the trust matrix. Despite this additional computation, CG−HMT remains in the same practical wall-clock regime as the baselines, and its throughput is stable enough to support the same sensitivity and generalization protocols without altering the training pipeline or compute setup. This ordering — MAPPO $<$ HHK $<$ CG−HMT $<$ MADDPG — is consistent across all three independent runs (standard deviations under 3% of the mean), indicating that the overhead is deterministic rather than incidental.