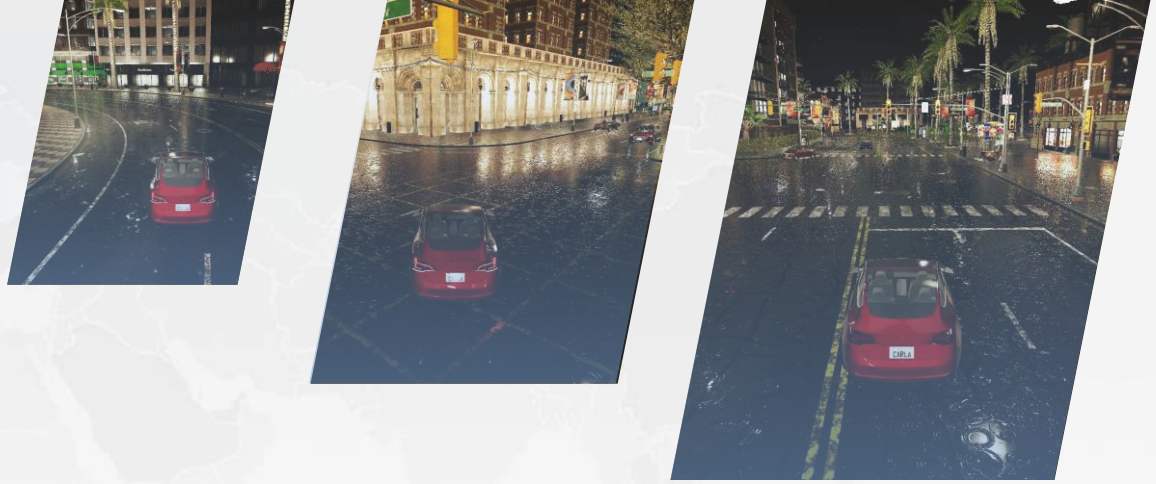
CVPR JUNE 3-7, 2026 DENVER COLORADO

SHENZHEN UNIVERSITY 深圳大学



Reliable Policy Transfer for Safety-Aware End-to-End Driving with Deep Reinforcement Learning

Uddin, Md. Borhan[1]; Raza, Arif[1]; Lin, Zhiliang[1]; Wang, Lu[1]; Li, Jianqiang[1,2,3]; Chen, Jie[1,2*]

[1]College of Computer Science and Software Engineering, [2]School of Artificial Intelligence,
[3]National Engineering Laboratory for Big Data System Computing Technology
Shenzhen University, Shenzhen, Guangdong Province, China-518060

[*]Corresponding Author

Roadmap

Motivation, implementation, results, and ablations.

1

Why safe transfer?

Closed-loop E2E RL fails under weather, traffic, and town shift.

2

What is proposed?

A unified DRL framework with one shared reliability interface.

3

How does it work?

Ego-relational state, dense reward, uncertainty-gated exploration, causal transfer.

4

Does it work?

CARLA 0.9.15 validation, trajectory evidence, graphs, transfer, ablation.

Motivation

Safe transfer is the bottleneck.

Closed-loop failure modes

- Late braking under fog or occlusion
- Lateral drift near intersections
- Overreaction to noisy detections
- Collision risk at crossings
- Overfitting to source geometry

The issue is not only perception accuracy; it is unsafe control under uncertainty.

Why existing E2E RL is brittle

State	Global tensors hide causal influence
Reward	Sparse events give weak gradients
Exploration	Fixed entropy ignores confidence
Transfer	Features adapt but control semantics drift

Research question

How can a safety-aware E2E driving policy be reliably transferred under distribution shift by aligning causality and uncertainty at the control layer?

Answer:
Build a single uncertainty interface used by state attention, reward, entropy, and transfer.

Research Gaps

Autonomous-driving E2E Research

TransFuser / sensor fusion	Strong multimodal fusion, but control state is not explicitly causal or confidence-weighted.
ST-P3 / planning heads	Adds interpretable scene representations and cost optimization, but safety & transfer limited.
UniAD / task coordination	Unifies perception-prediction-planning through queries, but the work focuses on RL transfer.
RaSc / risk-aware imitation	Improves risk-aware planning, but can still drift or brake late under ambiguity.

Key Contributions

Reliable signal and compatible with real-time inference.

1

Ego-centric relational state

Encodes vehicles, pedestrians, signals, lane geometry, and uncertainty as edges into an ego node.

2

Dense multi-objective reward

Smoothly optimizes safety, progress, comfort, and uncertainty so gradients exist before crashes.

3

Uncertainty-gated exploration

Combines aleatoric variance and critic-ensemble epistemic variance to gate policy entropy.

4

Causal-uncertainty transfer

Aligns policy KL, attention MMD, and uncertainty moments with MAML-style meta-initialization.

01 Method: Unified Reliability Interface

Representation, Reward, Exploration, and Transfer.

Closed-loop Driving as an MDP

The controller learns continuous throttle, brake, and steering actions from an ego-centric decision state.

MDP and Action Space

Closed-loop driving is modeled as an MDP, M.

- State describes the ego-centric scene.
- Action contains continuous control commands for throttle, brake, and steering.
- Transition model captures vehicle dynamics and interactions.

$$a_t = [a_t^{thr}, a_t^{brk}, a_t^{str}] \in \mathcal{A}$$

$$\max_{\theta} J(\theta) = \mathbb{E}_{\pi_\theta, P}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1})\right]$$

Why this matters for transfer

- The policy should not only maximize source-domain return.
- Preserve safety when P changes due to new town, lighting, weather, traffic, and noise.
- Training objective needs state features, reward terms, and exploration behavior.

Closed-loop Constraints

Lane, safety, and comfort.

Lane Adherence

The ego vehicle is constrained to stay near the active route segment.

A context-dependent corridor tightens when the scene is fragile.

$$d_L(p_t; r_j, r_{j+1}) = \frac{\|(r_{j+1} - r_j) \times (p_t - r_j)\|_2}{\|r_{j+1} - r_j\|_2}$$

$$d_L \leq \epsilon(\mu_A) = \epsilon_{min} + (\epsilon_{max} - \epsilon_{min})\mu_A$$

Safety Separation

Safety is expressed using distance to the actor and a threshold.

This converts near-miss behavior into a learnable signal.

$$\mathbf{dist}(p_t, p_t^{(i^\star)}) \geq d_{min}$$

$$\tau_t = \frac{\|\Delta p_{i^\star}\|_2}{\max(\epsilon_v, (-\Delta p_{i^\star} \cdot \Delta v_{i^\star})_+)} \geq \tau_{min}$$

Comfort And Lagrangian

The policy should avoid oscillatory steering and harsh acceleration.

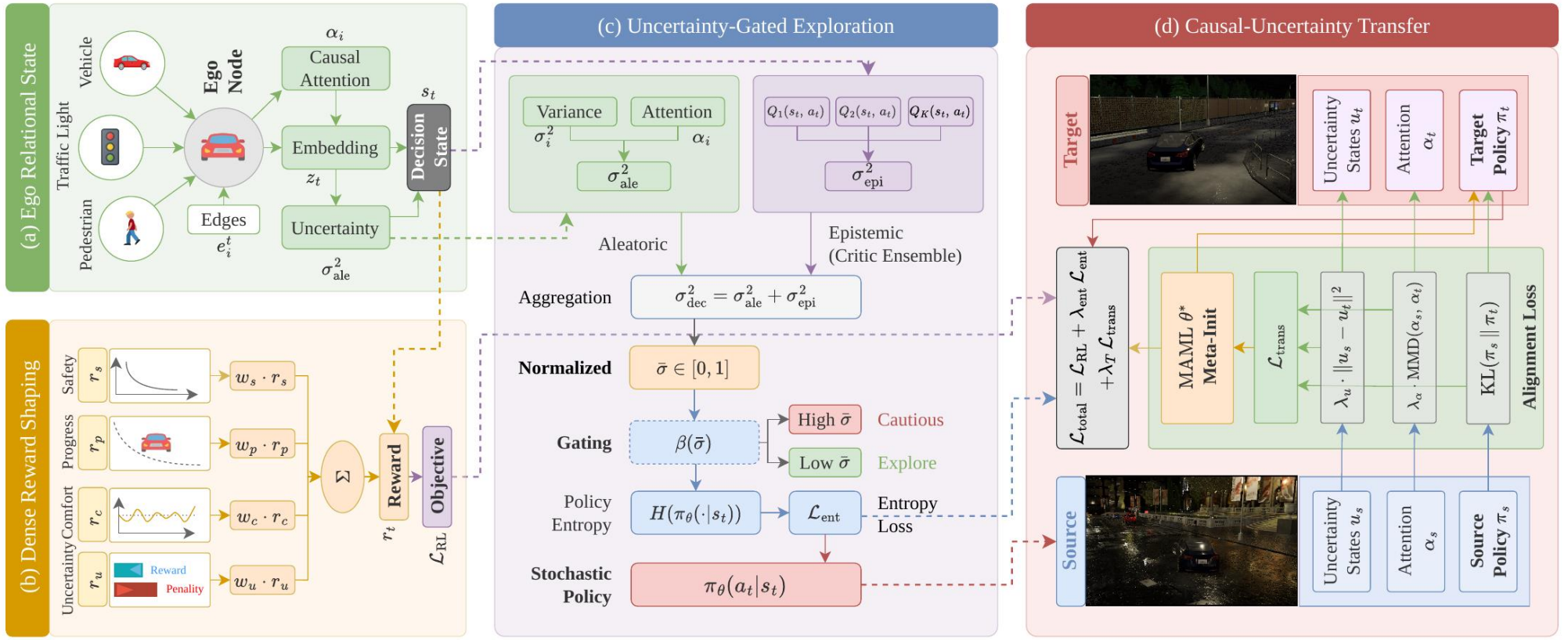
Jerk and steering-rate penalties enter the Lagrangian objective.

$$j_t = \|a_t^{veh} - a_{t-1}^{veh}\|_2/\Delta t, \quad \dot{\delta}_t = (a_t^{str} - a_{t-1}^{str})/\Delta t$$

$$\max_\theta \mathbb{E}\left[\sum_{t\geq 0} \gamma^t (r_d - \boldsymbol{\lambda}^\top \boldsymbol{\phi}_t)\right]$$

Framework

Unified E2E policy transfer



1. Relational state
Builds a control-ready ego graph; actors dominate attention.

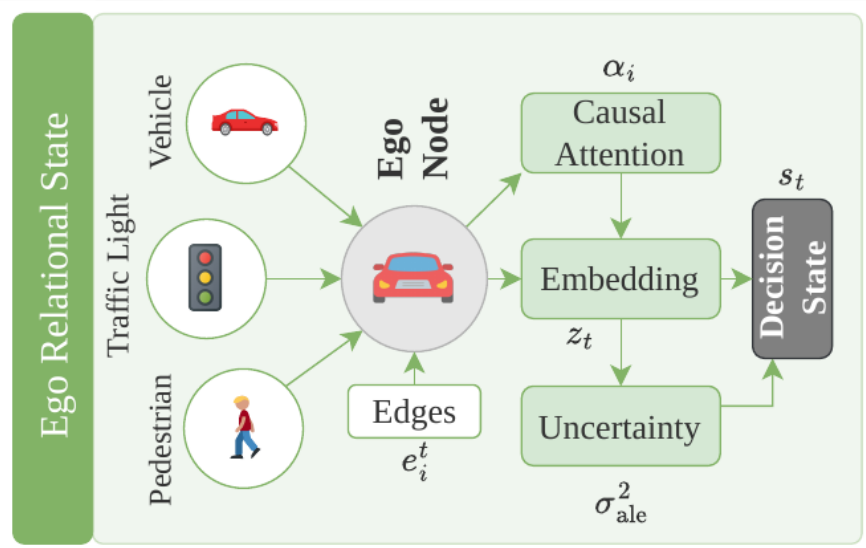
2. Dense reward
Turns safety, progress, comfort, and uncertainty into continuous.

3. Entropy gate
High $\bar{\sigma}$ reduces stochasticity in fog, occlusion, and critic disagreement.

4. Transfer
Aligns action, attention, and confidence for target adaptation.

Module 1

Ego-relational state



What the representation changes

1. Actors become directed edges into the ego node: vehicles, pedestrians, traffic lights, and local lane cues.
2. Attention becomes uncertainty-weighted: nearer and more reliable entities receive higher influence.
3. The decision state is compact: attended interaction embedding + ego speed + previous action + goal progress + lane geometry + aleatoric uncertainty.

Goal: Expose safety-critical influence and confidence.

Ego-relational State

Equations

Edge features

$$\mathbf{e}_i^t = [\Delta p_i,\ \Delta v_i,\ c_i,\ \kappa_i,\ \sigma_i^2]$$

Relative position, relative velocity, semantic class, lane geometry, aleatoric variance

Causal attention

$$\alpha_i = \mathbf{softmax}_i\left(-\frac{\|\Delta p_i\|_2^2}{\sigma_i^2 + \varepsilon}\right), \qquad z_t = \sum_{i=1}^{M} \alpha_i W_e \mathbf{e}_i^t$$

Near + confident actors dominate; noisy or distant actors are down-weighted

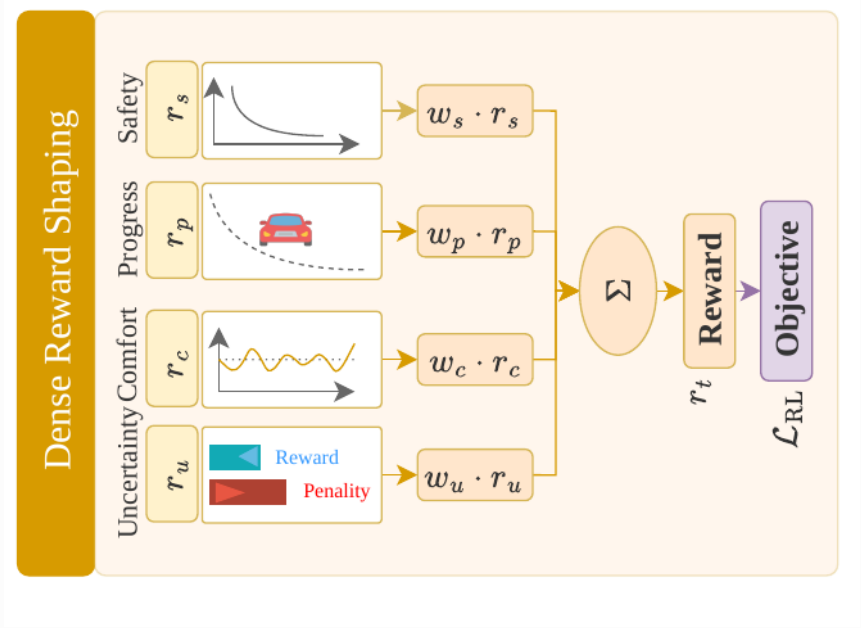
Decision state

$$\boldsymbol{s}_t = [z_t;\ v_{ego};\ a_{ego}^{t-1};\ d_{goal};\ \phi_{lane};\ \sigma_{ale}^2]$$

This is the state used by the stochastic policy and critics.

Module 2

Dense multi-objective reward shaping



Why dense reward is needed

Event-only rewards are sparse: the policy receives a large penalty after a collision or lane departure but little guidance in near-miss states.

This causes unstable gradients and slow transfer. The proposed reward provides continuous feedback.

Reward Components

- Safety: lane adherence, proximity, red-light/stop-line compliance
- Progress: forward route advancement
- Comfort: jerk and steering-rate smoothness
- Uncertainty: preference for well-observed states

Dense reward links optimization stability to closed-loop safety.

Reward Shaping

Equations

$$r_t = w_s r_s + w_p r_p + w_c r_c + w_u r_u, \qquad \sum_m w_m = 1$$

The uncertainty term uses the same $\bar{\sigma}$ that also appears in entropy gating and transfer alignment.

$$r_s = 1 - \kappa_L \psi_L(d_L, \mu_A) - \kappa_P \psi_P - \kappa_R \rho_t$$

Safety: Penalizes lateral deviation, proximity, and rule violation.

$$r_p = \tanh(\Delta s_t / \tau_s) \quad \text{or} \quad r_p = \tanh(\hat{v}_t / \tau_v)$$

Progress: Encourages route advancement.

$$r_c = -\kappa_j j_t^2 - \kappa_\delta \dot{\delta}_t^2$$

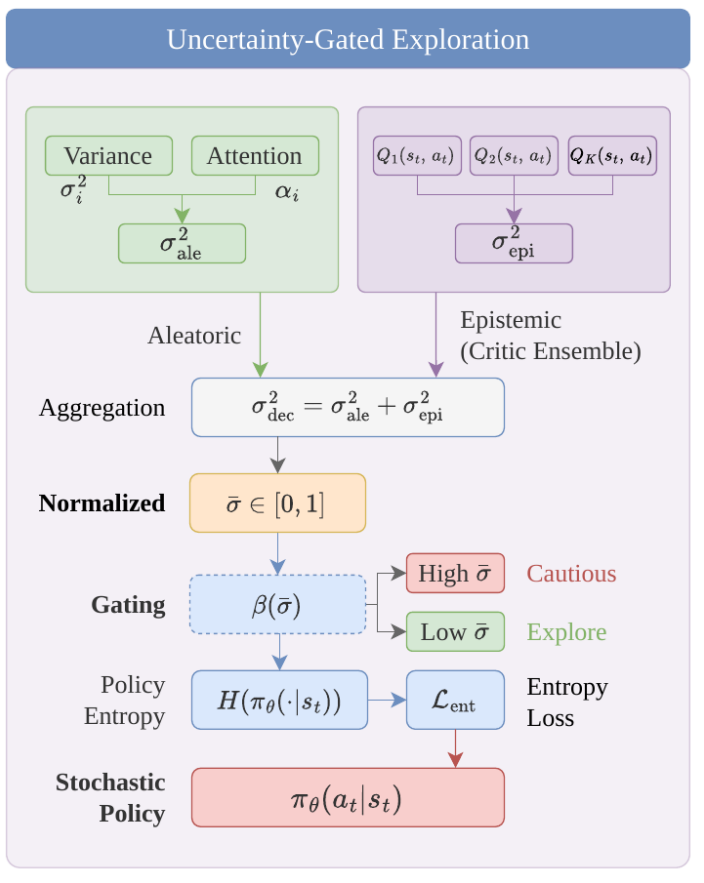
Comfort: Reduces oscillatory acceleration and steering.

$$r_u = 1 - \bar{\sigma}$$

Uncertainty: Rewards confident operation.

Module 3:

Uncertainty-gated exploration



Decision-time Uncertainty

Aleatoric uncertainty captures noisy or ambiguous observations and enters per-edge attention.

Epistemic uncertainty captures critic disagreement and detects unfamiliar state-action regions.

Behavioral Effect:

- High $\bar{\sigma}$: reduce entropy → cautious driving
- Low $\bar{\sigma}$: restore exploration → efficient learning
- Training only: critic ensemble estimates epistemic uncertainty
- Inference: single actor pass; no ensemble overhead

Uncertainty changes action sampling, not just logging.

Uncertainty-gated Exploration

Equations

$$\sigma_{dec}^2 = \sigma_{ale}^2 + \sigma_{epi}^2$$

Decision-time variance combines observation ambiguity and model uncertainty.

$$\sigma_{epi}^2(s_t, a_t) = \mathrm{Var}_k[Q_{\phi_k}(s_t, a_t)]$$

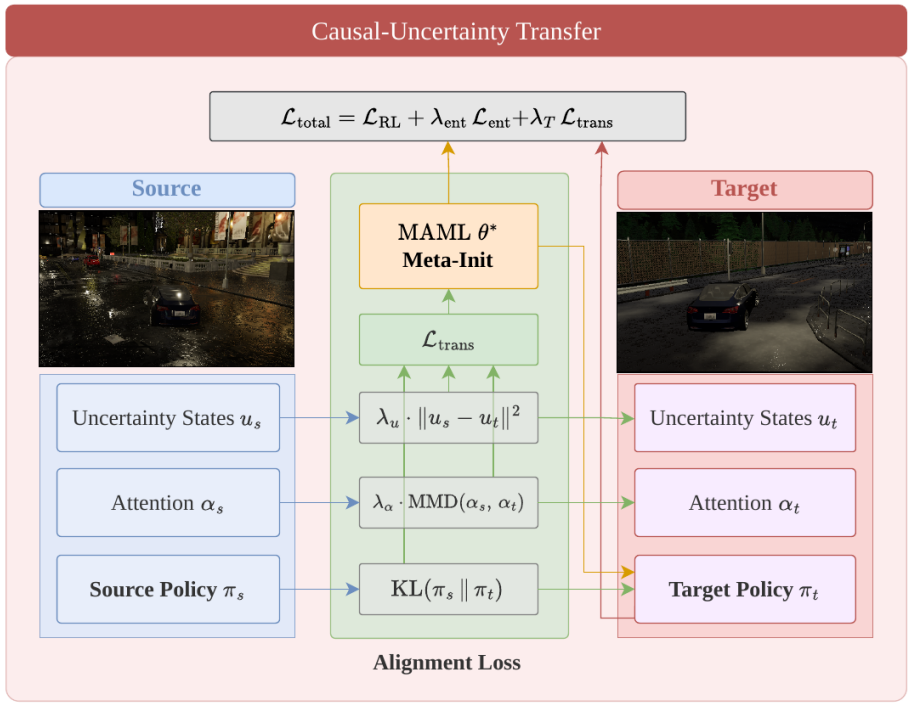
Critic ensemble disagreement estimates epistemic uncertainty during training.

$$\mathcal{L}_{ent} = -\beta(\bar{\sigma}) H(\pi_\theta(\cdot|s_t)), \qquad \beta(\bar{\sigma}) = \beta_0(1 - \bar{\sigma})$$

When $\bar{\sigma}$ increases, β decreases, reducing stochastic exploration under uncertain scenes.

Module 4

Causal-uncertainty transfer



What is transferred?

1. Policy behavior: Match action distributions using KL divergence.
2. Causal attention: Match which actors and cues the policy prioritizes using MMD.
3. Uncertainty statistics: Match confidence moments so entropy gating remains calibrated after domain shift.
4. Meta-initialization: MAML-style initialization supports few-shot adaptation.

Transfer operates where decisions are made: policy, attention, and uncertainty.

Policy Transfer

Equations

$$\mathcal{L}_{trans} = \mathcal{L}_{KL} + \lambda_{\alpha}\,\mathrm{MMD}(\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_t) + \lambda_u \|u_s - u_t\|_2^2$$

Aligns action distributions, causal attention vectors, and uncertainty moments across source and target domains.

$$\theta^{\star} = \arg\min_{\theta} \sum_{d \in \mathcal{D}} \mathcal{L}_{RL}^{(d)}\left(\theta - \alpha \nabla_{\theta} \mathcal{L}_{RL}^{(d)}(\theta)\right)$$

MAML-style initialization accelerates few-shot adaptation while preserving source-domain safety structure.

$$\min_{\theta,\phi} \; \mathcal{L}_{RL}(\theta, \phi; r_d) + \lambda_{ent}\mathcal{L}_{ent} + \lambda_T \mathcal{L}_{trans}$$

A single actor pass with top-K relational aggregation; critic ensemble and MAML are training-only.

Training

Inference pipeline

1

Collect transition

CARLA synchronous stepping at 20 Hz; state, action, reward, next state.

2

Build relational state

Construct ego-node graph with top-K safety-relevant entities.

3

Update actor-critic

Use shaped reward, Lagrangian penalties, and entropy gate.

4

Align domains

Apply KL + MMD + uncertainty matching with meta-initialization.

5

Deploy actor

Single policy pass; no critic ensemble or meta-gradient at inference.

02 Closed-loop Validation

Results show safety, stability, and transfer improvements under conditions.

Simulation

Protocol

Environment

- Carla 0.9.15
- Synchronous stepping at 20 hz
- Tesla model 3 ego vehicle
- Town10hd source domain
- Town02 cross-town transfer
- Town05 zero-shot transfer

Adverse Source Regime

- Heavy rain at night
- Dense fog
- Cloudiness 90%
- Precipitation 90%
- Fog density 40%
- Sun altitude -25°
- Moderate-to-high NPC traffic

SAC Training Details

- Replay buffer 2×10^5
- Batch size 512
- $\Gamma = 0.99$
- $\mathrm{T} = 5\times10^{-3}$
- Adam $l_r = 3\times10^{-4}$
- $B_0 \in \{0.5, 1.0\}$
- 5×10^5 steps
- Evaluation every 10k steps

Baselines: ST-P3 · ThinkTwice · TransFuser · RaSc

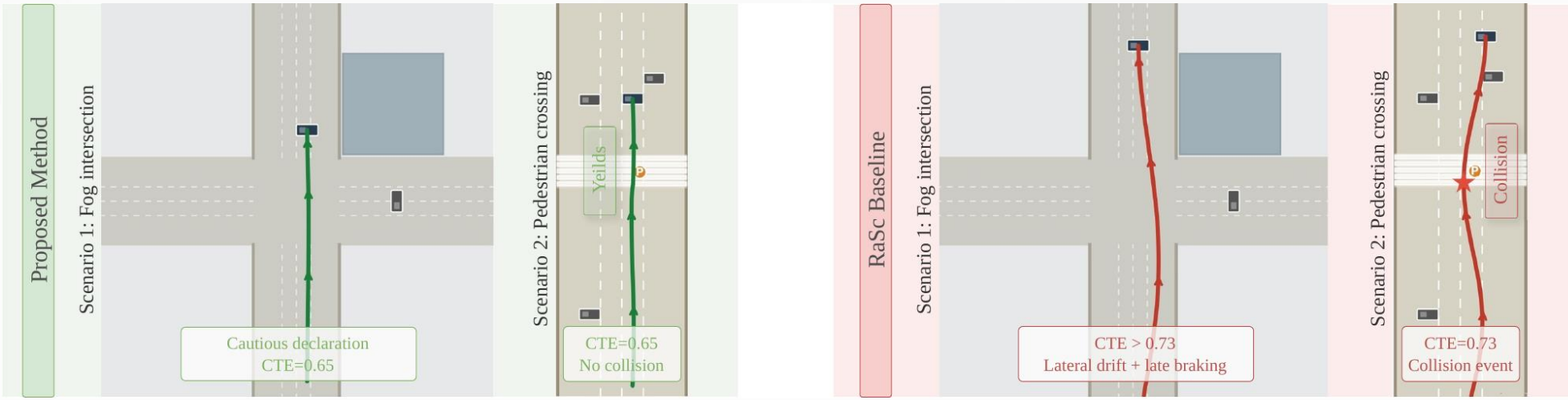
Evaluation

Metrics

Metric	Definition	What it shows
SR (%)	N_{succ} / N_{ep}	End-to-end task success
RC (%)	100 × traversed arc-length / route arc-length	Route completion
Q/km	N_q / D_{km} for collision, off-road, timeout	Infraction frequency
DS	Overall driving score with infraction penalties	Safety-aware driving quality
IS	$\prod_Q (1 - \min(1, q/km / \kappa_q))$, κ_q =0.02	Independent infraction score
CTE / Heading	Lateral deviation and heading error	Lane stability and control smoothness

Qualitative Safety

Earlier caution and fewer collisions



Proposed method

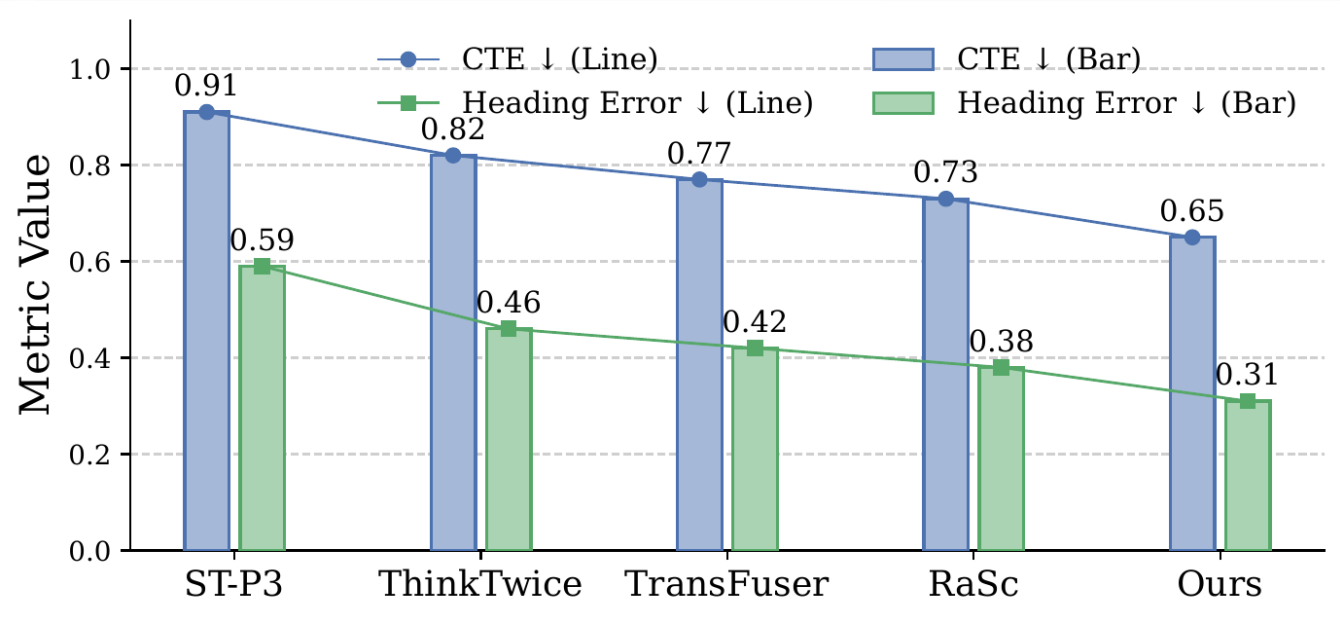
Maintains lane, decelerates cautiously, yields at pedestrian crossing, and avoids collision under ambiguity.

RaSc baseline

Shows lateral drift and delayed braking; pedestrian-crossing scenario ends in collision.

Result 1

Ego-relational state improves lane stability



0.65

CTE ↓

0.31

heading error ↓

Interpretation

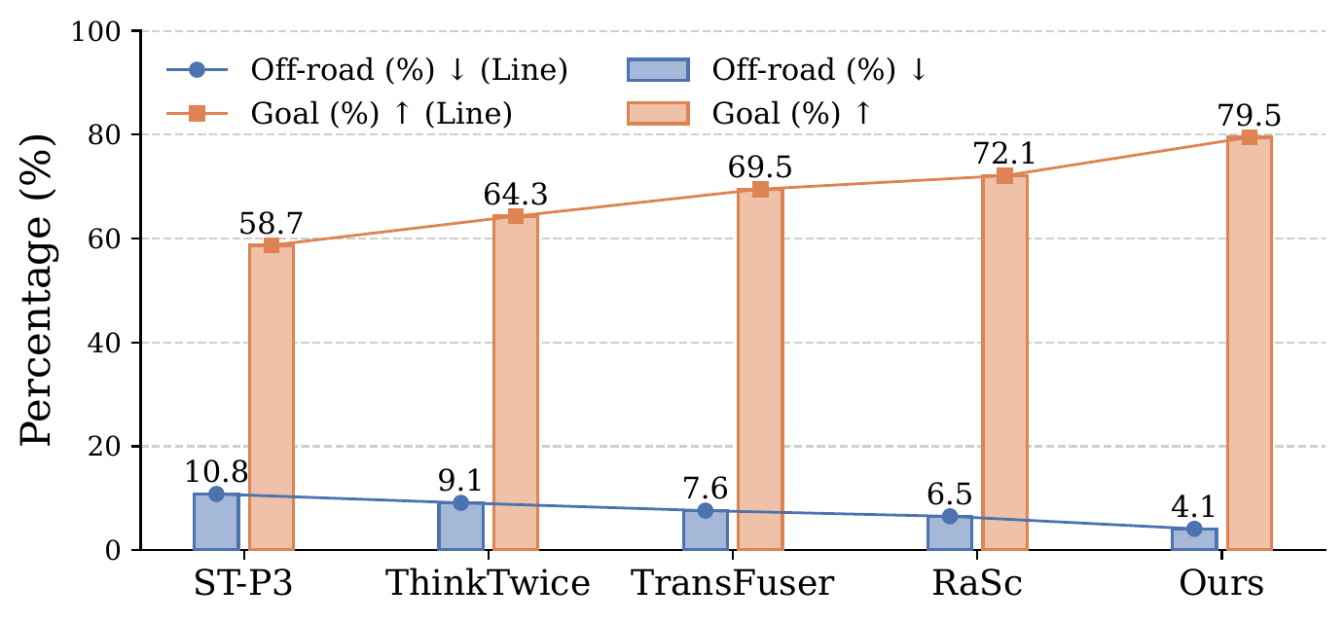
The causal state drops CTE to 0.65 and heading error to 0.31.

The improvement indicates that the policy receives better context.

Representation quality becomes visible as closed-loop stability.

Result 2

Route completion improves while off-road decreases



4.1%

off-road ↓

79.5%

goal rate ↑

Interpretation

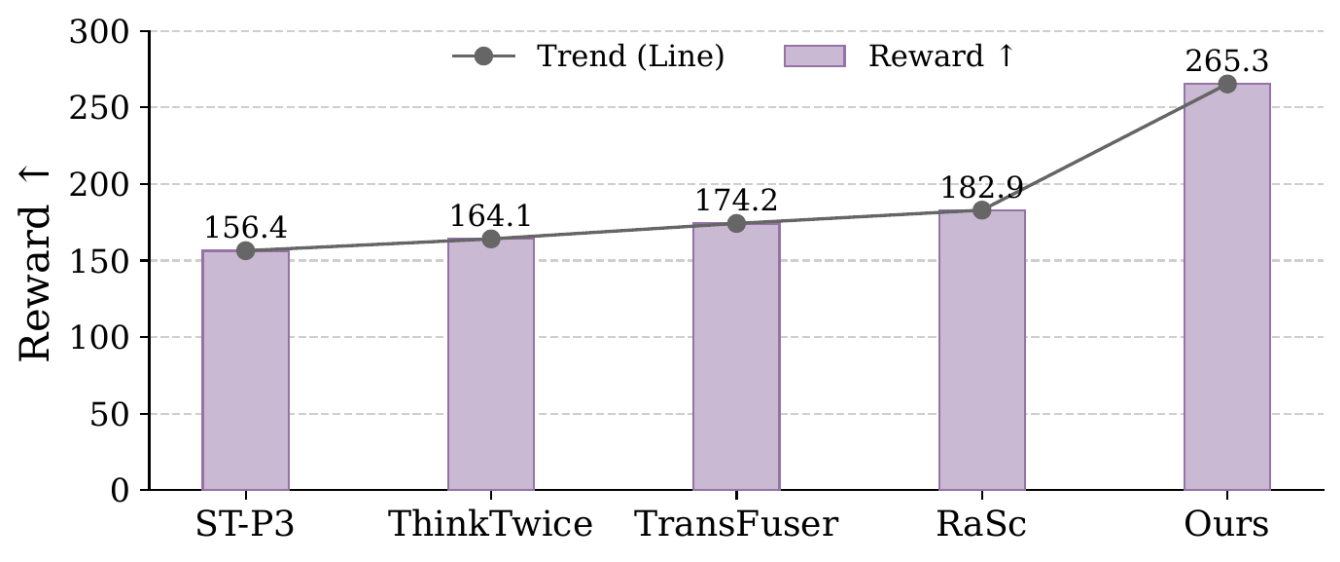
Off-road rate decreases to 4.1%. Goal completion rises to 79.5%.

This shows that the state improves route-level task completion.

Better local decisions accumulate into better route completion.

Result 3

dense reward shaping stabilizes learning



265.3

reward ↑

+45.1%

vs. RaSc

Interpretation

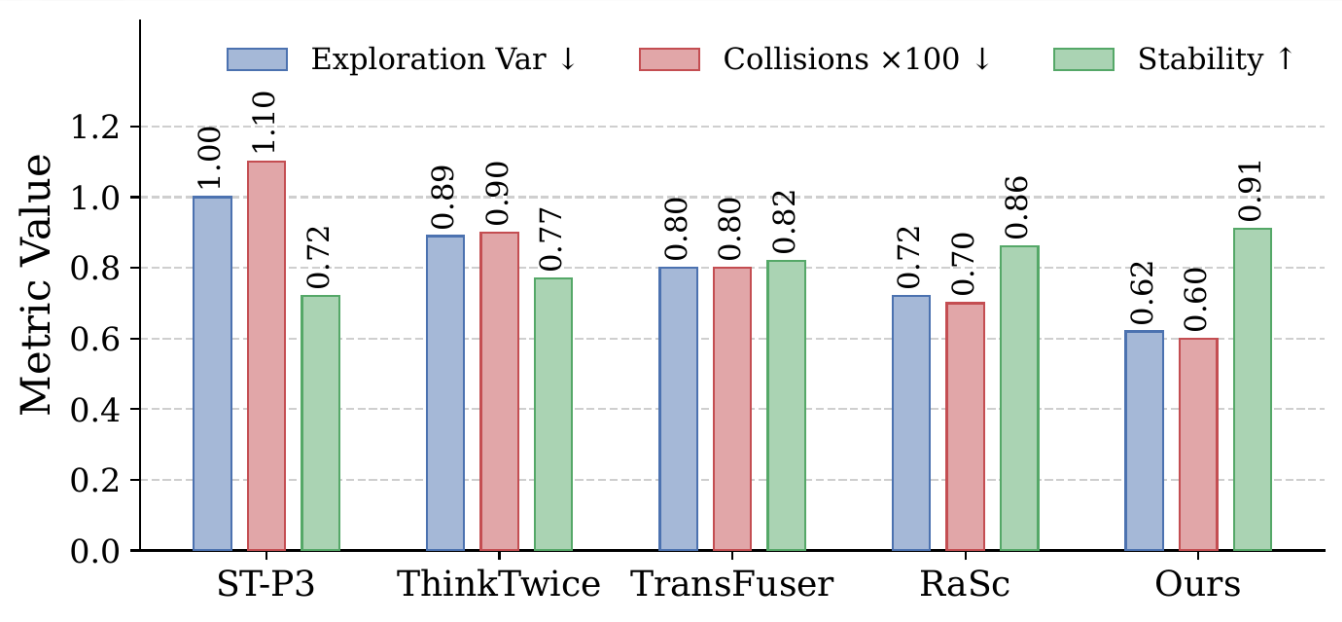
The framework reaches 265.3, showing that dense feedback helps the critic learn from near-miss states.

Comfort penalties reduce oscillation.

Informative gradients before failure events improve safety.

Result 4

Uncertainty balances exploration and safety



0.62

exploration var ↓

0.91

stability ↑

Interpretation

Exploration variance drops to 0.62, and collision rate falls to 0.006.

The entropy gate makes cautious under fog, occlusion, and critic disagreement.

Lower variance + higher stability = risk-aware exploration.

Results

Cross-town and zero-shot transfer

Domain / Variant	SR (%)	RC (%)	DS	IS	Coll./km	OR/km	TO/km
Town10HD source	91.2	94.1	94.1	1.00	0.000	0.000	0.000
Town05 zero-shot	100.0	94.6	94.6	1.00	0.000	0.000	0.000
Town02 policy learning	72.1	75.2	188.6	0.88	0.007	0.005	0.003
Town02 source domain	80.3	82.6	205.7	0.92	0.006	0.004	0.002
Town02 target ours	85.0	84.1	214.3	0.94	0.005	0.003	0.001

85.0%
Town02 SR ↑

214.3
Town02 DS ↑

0.005
Coll./km ↓

100%
Town05 SR

Cross-town Town02
Full transfer improves SR to 85.0% and RC to 84.1%.

Zero-shot Town05
Agent reaches 100% success and 94.6% route completion.

Safety under shift
Collisions/km decrease to 0.005; OR/km, TO/km also decrease.

Interpretation
Action, attention, and uncertainty narrows the gap.

Ablation Study

Each module contributes

Variant	CTE↓	Coll./km↓	DS (Town02)↑	Stab.↑	Main lesson
w/o Unc. Attn.	0.76	0.008	203.5	0.84	Attention confidence is crucial
w/o Ensemble	0.68	0.009	208.1	0.87	Epistemic signal lowers collisions
w/o Ent. Gate	0.71	0.008	206.7	0.86	Fixed entropy weakens safety
Event Reward	0.74	0.009	195.9	0.83	Sparse reward destabilizes RL
w/o Transfer	0.65	0.006	194.1	0.91	Alignment drives cross-town DS
w/o MAML	0.65	0.006	200.8	0.91	Meta-init speeds adaptation
Full model	0.65	0.006	214.3	0.91	Best transfer and stability

State module
Removing uncertainty attention raises CTE to 0.76 and lowers stability to 0.84.

Uncertainty module
Removing critic ensemble or entropy gate increases collisions to 0.008 km^{-1}.

Reward module
Event-only reward gives the lowest stability (0.83) and low DS (195.9).

Transfer module
Without transfer/MAML, Town02 DS drops to 194.1/200.8.

Interpretation

Uncertainty-weighted attention improves stability; critic ensemble & entropy gating reduce collisions; reward improves optimization; transfer loss & MAML improve driving score.

Sensitivity

Real-time feasibility

Entropy gate β_0

$\beta_0 \in \{0.5, 1.0\}$

Collision rate changes by ≤ 0.001/km and CTE by 0.03.

This suggests the entropy gate is not overly sensitive within the validated range.

Reward weights

Default: $W_S = 0.4$, $W_p = 0.3$, $W_c = W_u = 0.15$

Shifting toward progress ($W_p = 0.5$) raises completion by 2.1% but increases collisions by 0.002/km.

Runtime

Inference runs at 20 Hz on RTX 3090. Top-K aggregation adds ≤ 3 ms over vanilla SAC.

Critic ensemble and MAML are training-only.

Results

Summary

Ego-relational state	CTE 0.65, heading error 0.31	Stable lane tracking
Dense reward shaping	Reward 265.3	Better learning signal
Uncertainty-gated entropy	Variance 0.62, stability 0.91	Cautious under ambiguity
Causal transfer	Town02 DS 214.3, SR 85.0%	Stronger cross-town adaptation

Conclusion

Reliable transfer comes from aligning causality and uncertainty where control decisions are made.

1

State

Ego-centric relational graph exposes safety-relevant actors, lane cues, and confidence.

2

Learning

Dense reward and uncertainty-gated entropy provide safer, closed-loop optimization.

3

Transfer

Policy KL, MMD, uncertainty matching, and MAML improve cross-town adaptation.

Future work: standardized benchmarks evaluation, stronger interventional causal modeling, reduced ensemble overhead, and Sim2Real transfer.

Thank you

Questions?

Reliable Policy Transfer
for Safety-Aware End-to-End Driving
with Deep Reinforcement Learning

深圳大學 SHENZHEN UNIVERSITY

CVPR JUNE 3-7, 2026 DENVER COLORADO