Learning to Drive via Asymmetric Self-Play

Chris Zhang, Sourav Biswas, Kelvin Wong, Kion Fallah, Lunjun Zhang, Dian Chen, Sergio Casas, Raquel Urtasun

https://waabi.ai/selfplay

Goal: Learn realistic driving policies that handle complex, safety-critical scenarios.

How can we scale training data beyond real-world collection?

Challenges

The problem with relying solely on real data:

- Most nominal driving is boring, with little learning signal.
- Collecting real safety-critical scenarios is dangerous.
- Upsampling existing scenarios lacks diversity.

Existing synthetic data approaches:

- MARL often converges to cooperative, nominal driving.
- Manually designed scenarios are difficult to scale.
- Adversarial optimization might not always discover useful training scenarios; challenging to control difficulty.

Our approach

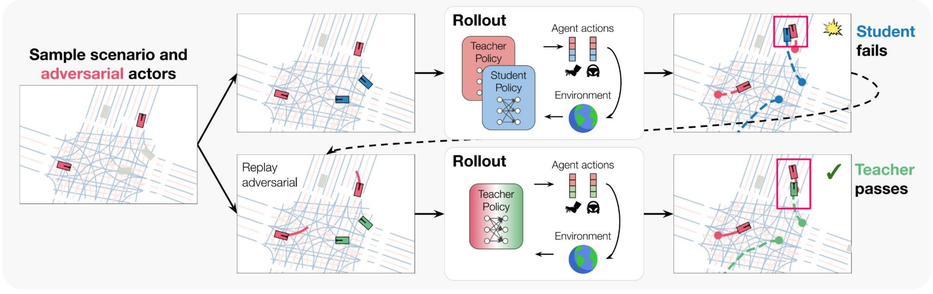
An **asymmetric self-play** mechanism in which

1) *challenging*
2) *solvable*
3) *realistic*

scenarios naturally emerge from interactions between teacher and student policies with differing objectives.

Asymmetric Self-Play

Main Idea: The teacher (**red**, **green**) learns to generate realistic scenarios where the student (**blue**) makes a mistake (top) while demonstrating a solution itself (bottom).



Learning: Optimize challenging and solvability terms under realism regularization:

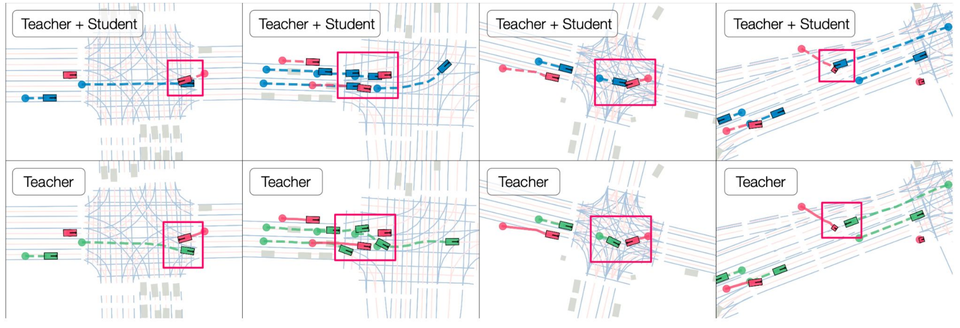
$$R_T(\boldsymbol{s}_1, \boldsymbol{m}) = \underbrace{C(\pi_{TS}, \mathcal{S})}_{\text{Challenging}} - \underbrace{C(\pi_T, N)}_{\text{Solvable}} + \beta \underbrace{\left(I_{\text{data}}(\pi_T) + I_{\text{data}}(\pi_{TS})\right)}_{\text{Realistic}}$$

$$R_S(\boldsymbol{s}_1, \boldsymbol{m}) = -C(\pi_{TS}, \mathcal{S}) + \beta I_{\text{data}}(\pi_{TS})$$

$$C(\pi, \mathcal{A}) = \mathbb{E}_{\pi|\boldsymbol{s}_1,\boldsymbol{m}}\left[\sum_{i \in \mathcal{A}} c_i(\boldsymbol{s}_{\leq T})\right] \quad \textit{(collision)}$$

$$I_{\text{data}}(\pi) = \mathbb{E}_{\pi|\boldsymbol{s}_1,\boldsymbol{m}}\left[-\log p_{\text{data}}(\boldsymbol{s}_{\leq T}|\boldsymbol{m})\right] \quad \textit{(likelihood)}$$

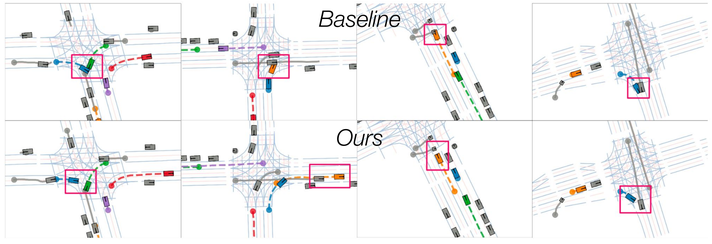
Example scenarios discovered over the course of training:



Traffic Modeling

The **student** policy achieves SOTA on Argoverse2 motion dataset and synthetic safety-critical scenarios.

Model	Safety Col.	Argoverse2 FDE	Col.	Offroad	JSD
Closed-loop *(IL)* [68]	40.41	**4.95**	1.02	**3.14**	0.436
TrafficSim *(IL+Prior)* [68]	26.69	5.13	0.33	3.36	0.436
SMARTS *(MARL)* [90]	13.65	16.3	8.12	17.2	0.528
Emb. Syn. *(Curation)* [11]	27.75	6.89	2.02	4.30	0.449
KING *(Adversarial)* [28]	12.65	6.33	1.16	3.29	0.465
Ours	**8.16**	5.04	**0.24**	3.39	**0.433**



End-to-end Autonomy

Teacher policy can be zero-shot deployed to interact with autonomy in simulation.

These training scenarios result in more robust autonomy policy.

Train Data	Priv	Safety GSR (↑)	Col (↓)	mTTC (Δ)	Prog (Δ)	P2E (Δ)	Accel (Δ)
Expert	✓	90.6	0.0	5.82	232	0.17	0.85
Safety	✓	80.1	0.0	5.83	236	0.35	0.91
Highway		40.2	58.3	3.33	280	1.01	1.41
IL [68]		45.6	59.7	3.61	277	0.90	1.39
Adv. [28]		83.1	6.2	5.54	253	0.45	0.99
Ours		**92.6**	**0.0**	**5.77**	**247**	**0.36**	**0.88**