From Recorded Traffic to Steering and Acceleration: A Gated Hybrid of Convolutional and Vision-Language AI

Abolfazl Afshari, Joyoung Lee

John A. Reif, Jr. Dept. of Civil & Environmental Engineering, New Jersey Institute of Technology

aa3377@njit.edu | jo.y.lee@njit.edu

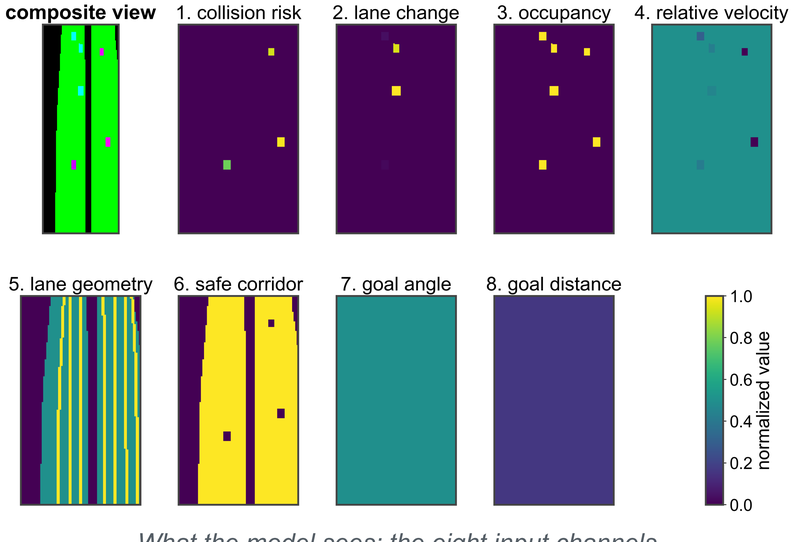
NJIT New Jersey Institute of Technology | ITSRC Intelligent Transportation Systems Resource Center

1. WHY THIS STUDY?

- Recorded human driving holds millions of examples of how people actually behave in traffic.
- Most trajectory models predict which maneuver comes next, not the control commands a vehicle executes.
- Instead of:
  - Keep lane | Change lane | Brake
- We predict directly:
  - **Acceleration (m/s²) + Steering angle (°)**
- **Goal: learn continuous vehicle control directly from real human driving.**

2. DATA FROM HUMAN DRIVERS

- 2.66 million driving samples
- 93 highway recordings of human drivers, filmed from a drone (exiD)
- 8-channel ego-centric bird's-eye view, 2-second prediction horizon
- Training and test sets split by recording, so no scene appears in both
- Each scene becomes a compact bird's-eye image of vehicles, road geometry, risk, relative motion, and route context.



What the model sees: the eight input channels

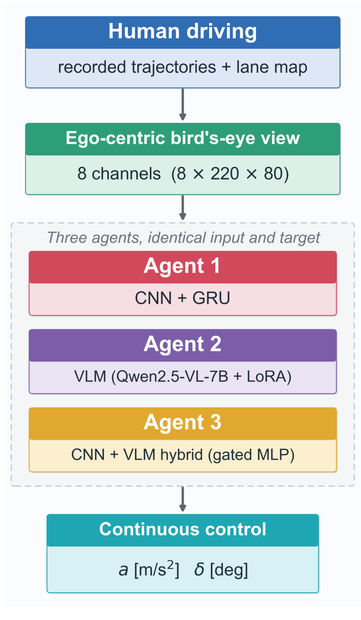
3. THE MAIN METHOD

- **Same scene. Same target. Three different models.**
  - **Agent 1 - CNN + GRU**
    - Five scenes in, acceleration and steering out.
  - **Agent 2 - Vision-language model**
    - Fine-tuned Qwen2.5-VL-7B writes control commands.
  - **Agent 3 - CNN + VLM hybrid**
    - Corrects the CNN only where the VLM helps.

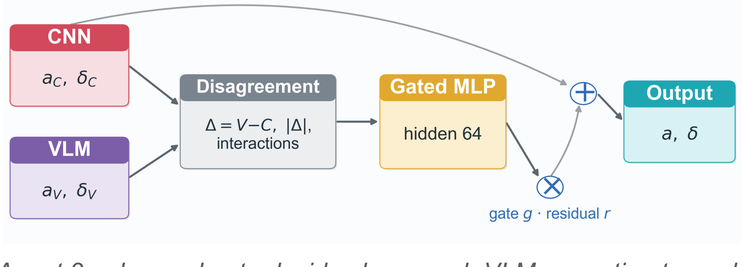
THE BOTTOM LINE

A compact CNN matched human control on 95.6% of unseen samples, and 76.7% on a highway it never saw.

The 7B model was weaker alone, but a learned gate turned it into a 14.4% error cut on the hardest cases.



Study design: one input, one target, three agents



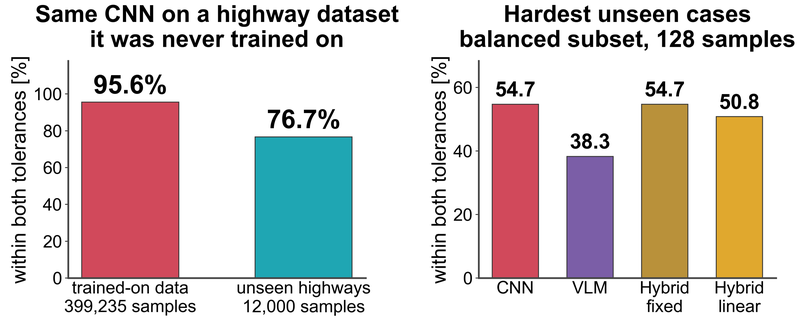
Agent 3: a learned gate decides how much VLM correction to apply

4. WHAT HAPPENED?

- **Result 1 - The compact CNN is the accurate one**
  - 0.161 m/s² and 0.18° average error on 399,235 unseen samples
  - 95.6% inside both tolerances at once
- **Result 2 - Bigger is not automatically better**
  - Alone, the 7B model was the least precise
  - Reasoning ability ≠ precise low-level control
- **Result 3 - But the VLM still helps**
  - Hardest cases: 14.4% less acceleration error
  - +5.1 points more inside both tolerances



Harder balanced subset (3,994 samples): the gated hybrid leads



average error on the unseen data: acceleration 0.16 → 0.35 m/s², steering 0.18 → 0.21 deg

Transfer to an unseen highway dataset (highD)

5. WHAT THIS MEANS

Trajectory data agencies already collect can train human-like controllers for traffic simulation, driver-behavior models, and AV evaluation.

Model size is not the deciding factor: the small model was both more accurate and far cheaper to run.

Limits and next steps: highway scenes, evaluated offline. Urban intersections and closed-loop testing come next.