香港科技大学（广州）
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

USTBench

Benchmarking and dissecting spatiotemporal reasoning capabilities of LLMs as urban agents

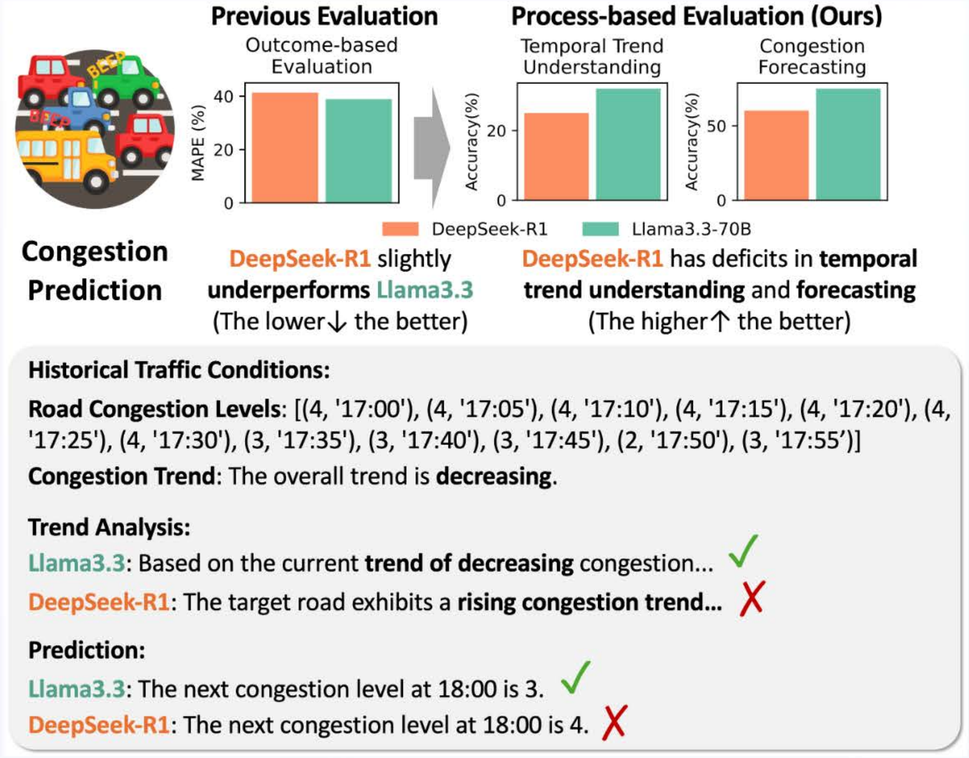
ICLR 2026

Siqi Lai · Yansong Ning · Zirui Yuan · Zhixi Chen · Hao Liu

— HKUST *Guangzhou*

Motivation

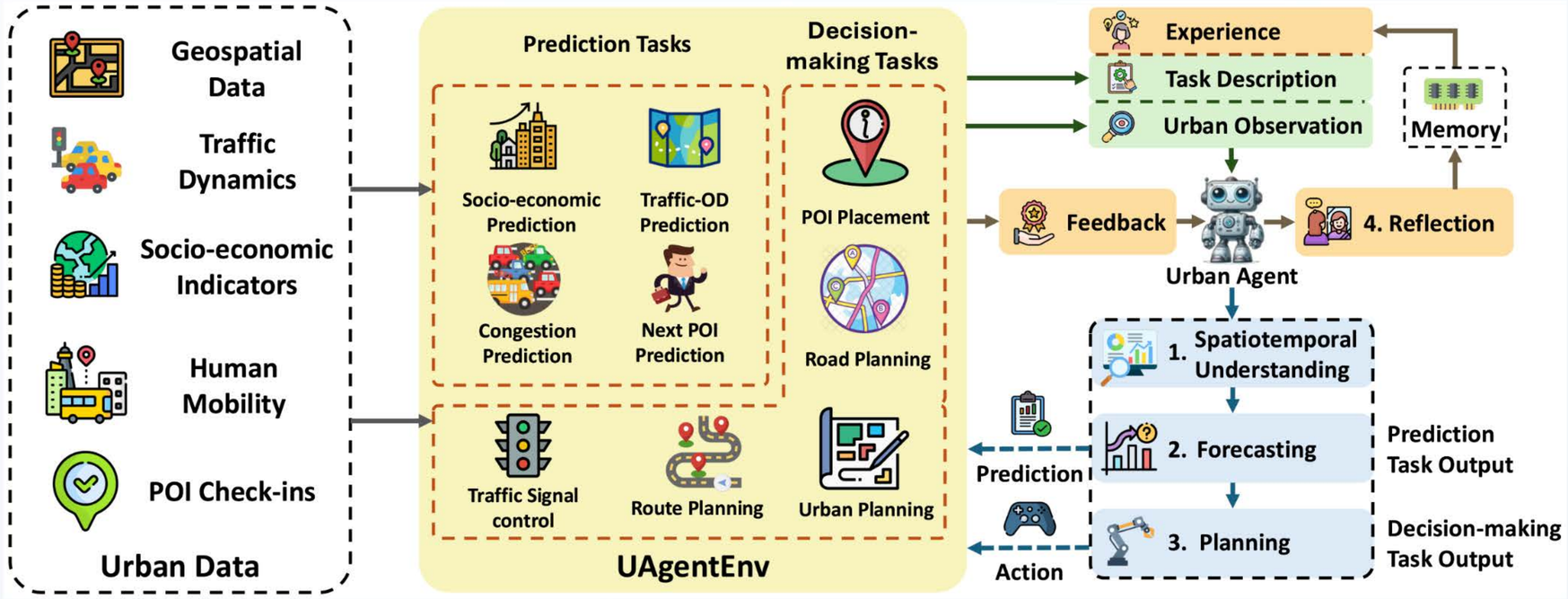
- Outcome-only scores miss flawed intermediate reasoning.
- Prior urban benchmarks rarely test process or reflection.
- We jointly evaluate reasoning QA and full tasks.



Outcome metrics can disagree with process-level diagnosis.

Contributions

- **USTBench** — four reasoning dimensions with 62k+ structured QAs.
- **UAgentEnv** — unified interactive environment for nine real urban tasks $OSM, traffic, mobility, socio-economics, POI$.
- **Empirical study** — 14 LLMs; planning and reflection lag; general "reasoning" models not uniformly best.



Unified pipeline for perception, reasoning, and reflection.

USTBench — what we measure

Process-level QA *diagnostics and* separate **end-to-end** evaluation on the same nine UAgentEnv tasks.

62,466 structured QA pairs across nine urban tasks.

Four abilities: **understand→forecast→plan→reflect** $full interaction loop$.

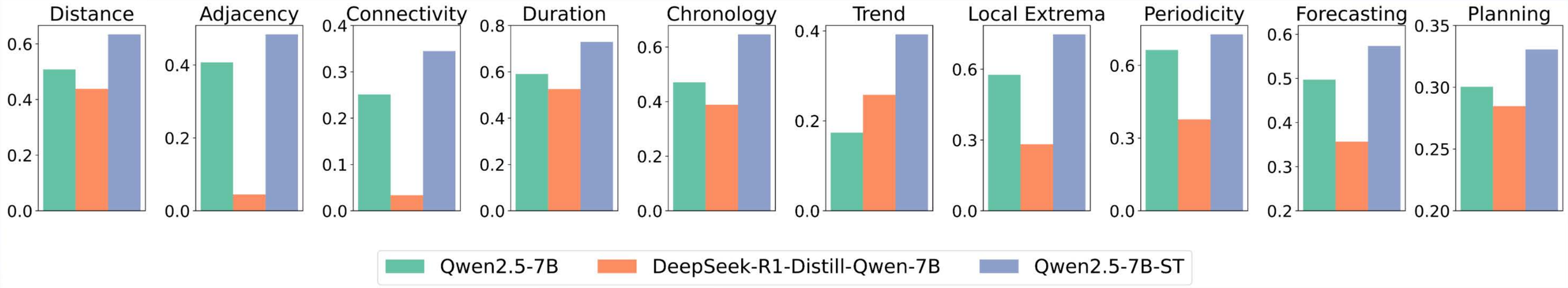
~40% spatiotemporal understanding · **~60%** higher-level reasoning.

Process QA scored by **accuracy**; end-to-end uses task-specific metrics $MAPE, accuracy, service access, ecology, cost, \ldots$.

Unlike outcome-only benchmarks, USTBench diagnoses *where* reasoning fails in the loop.

Key findings *I*

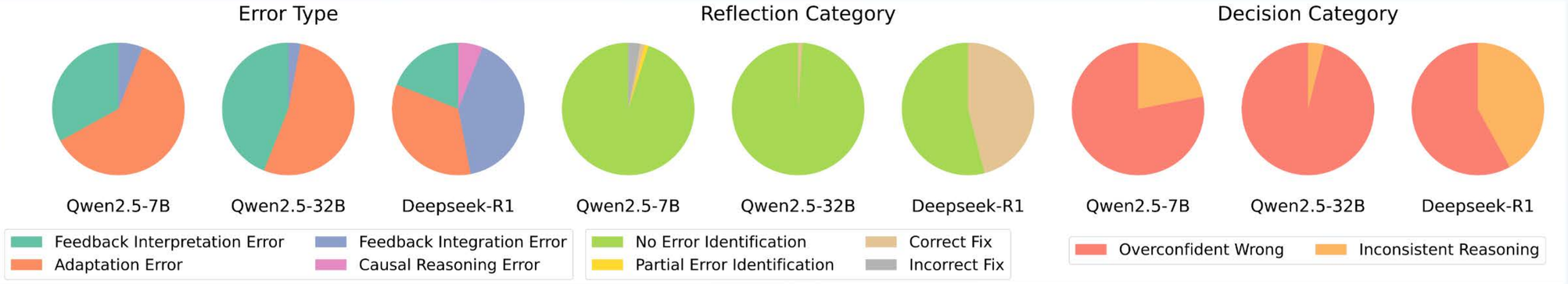
- General reasoning training does not guarantee strong urban ST reasoning.
- Forecasting often reaches high accuracy; planning remains much harder.
- Post-training on spatiotemporal understanding improves forecast and planning.



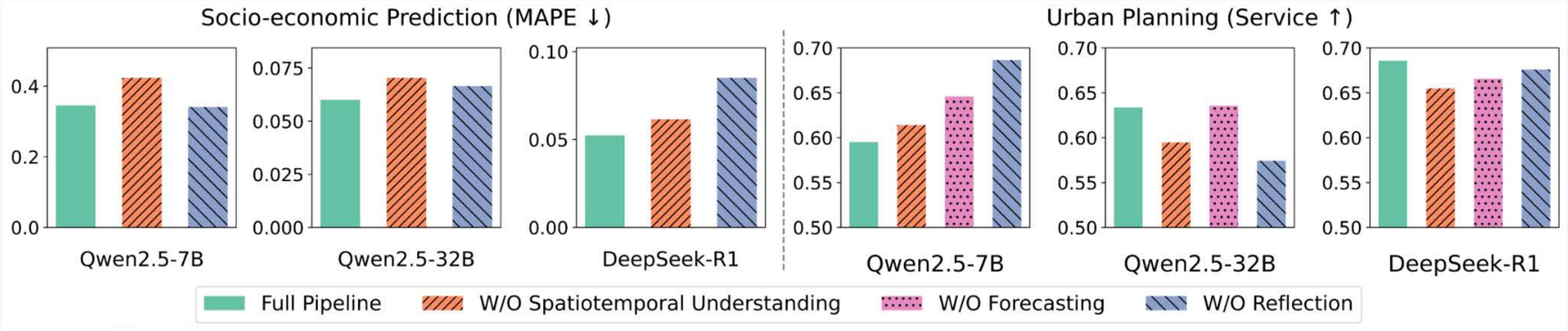
Targeted ST understanding training transfers to downstream reasoning.

Key findings *II*

- Reflection accuracy is often below 50%; integration of feedback remains brittle.
- Ablations: strong models rely on understanding, forecasting, and reflection; weak chains hurt weak models.



Reflection: error types and faithfulness.



Removing reasoning modules hurts capable agents most.

Conclusion

- USTBench couples process QA with end-to-end urban evaluation.
- Open challenges: long-horizon planning, reflection, domain-specialized adaptation.
- Limits: evaluation-centric scope; simulated decisions; broader agent skills future work.

Code & data: github.com/usail-hkust/USTBench

Thank you — questions?