



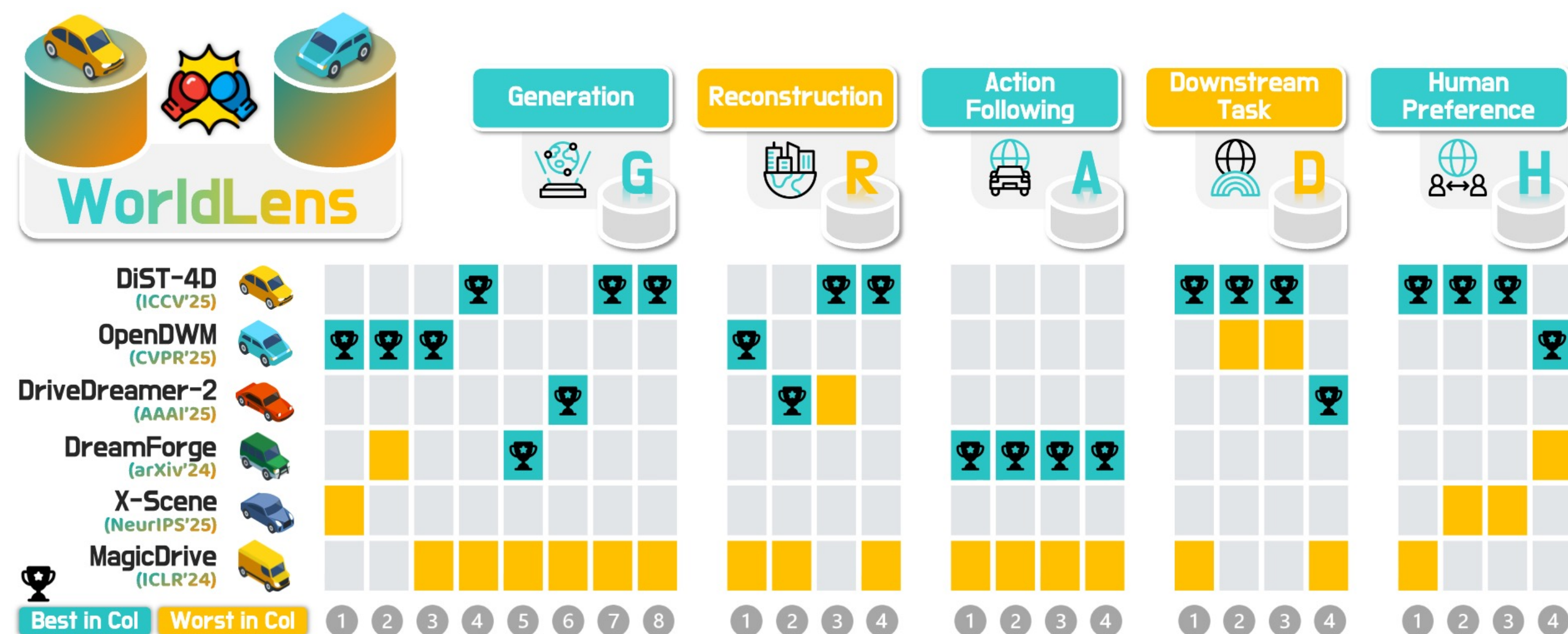
Full-Spectrum Evaluations of Driving World Models in Real World

WorldBench Team Ao Liang*, Lingdong Kong*†, Tianyi Yan*, Hongsi Liu*, Yu Yang*, Ziqi Huang, Wei Yin, Jialong Zuo, Yixuan Hu, Dekai Zhu, Dongyue Lu, Youquan Liu, Guangfeng Jiang, Linfeng Li, Xiangtai Li, Long Zhuo, Lai Xing Ng, Benoit R. Cottureau, Changxin Gao, Liang Pan, Wei Tsang Ooi, Ziwei Liu



Motivation & Contribution

- ❖ **Observation:** Generative world models are capable of generating realistic visual environments but often fail in terms of geometry, physics, and behavior. Looking **real** does not mean behaving **real**.
- ❖ Existing metrics focus on 2D image quality, cross-frame fluency, and text-to-video alignments, while overlooking the importance of **3D Consistency**, **Physical Plausibility**, **Controllability**.

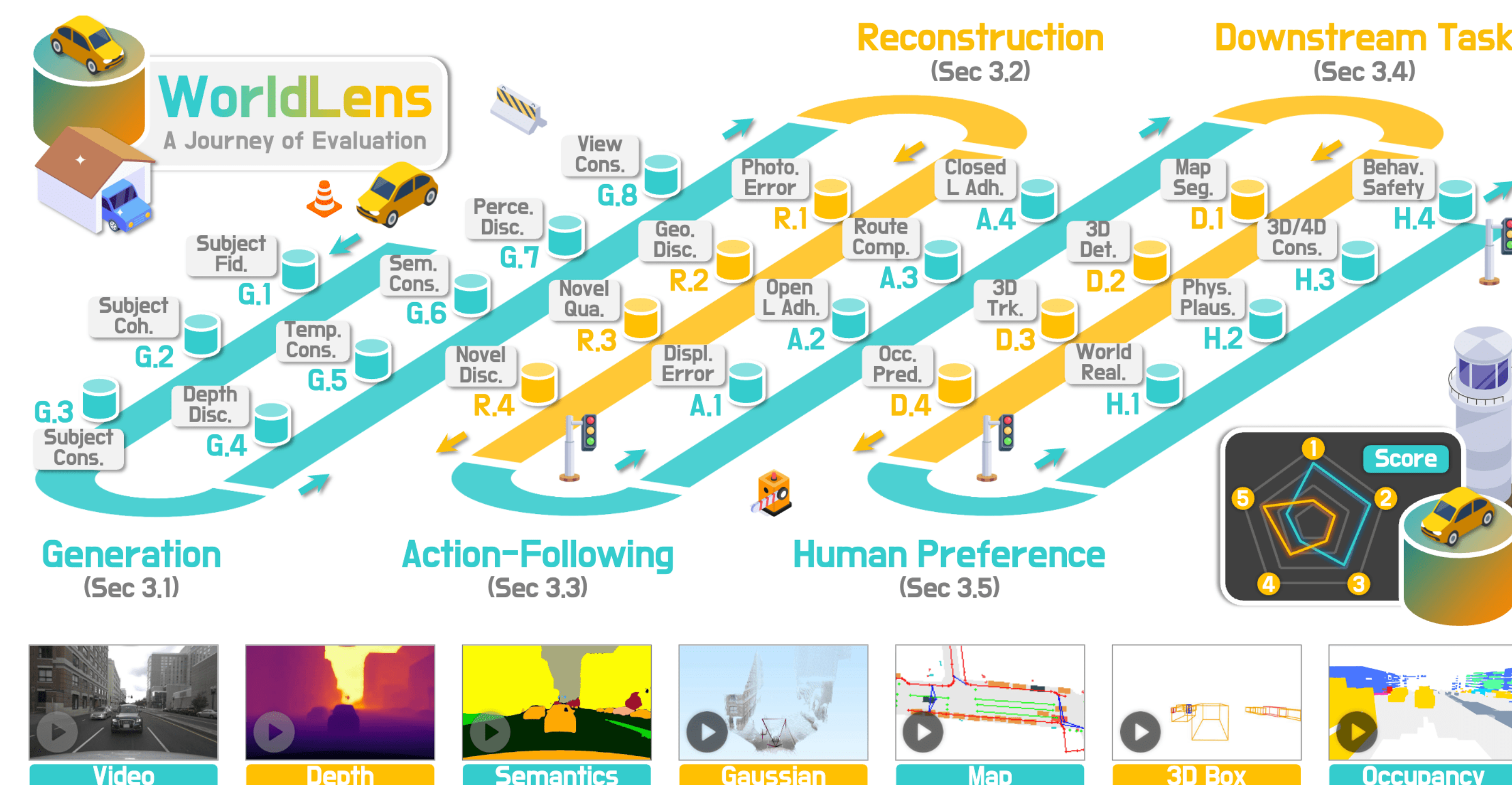


- ❖ **WorldLens** is the **first** benchmark that evaluates how well a model can **build**, **understand**, and **behave** within the generated environments, spanning 5 aspects and a total of 25 dimensions for generation quality, reconstruction fidelity, action-following capability, downstream utility, and human preference alignment.
- ❖ **Key Findings:** Evaluations reveal that **no existing world model excels in all dimensions**. Models with strong perceptual realism often suffer from geometric instability, unsafe behaviors, or poor downstream performance, while geometry-aware approaches improve reconstruction fidelity but may sacrifice visual quality.



Benchmark Construction

- ❖ **WorldLens** establishes a unified evaluation framework through unique & complementary aspects: **Generation**, **Reconstruction**, **Action-Following**, **Downstream**, and **Human Preference**.
- ❖ Across aspects, we define a total of 24 fine-grained dimensions covering visual realism, temporal consistency, geometric fidelity, physical plausibility, behavioral safety, and task utility.

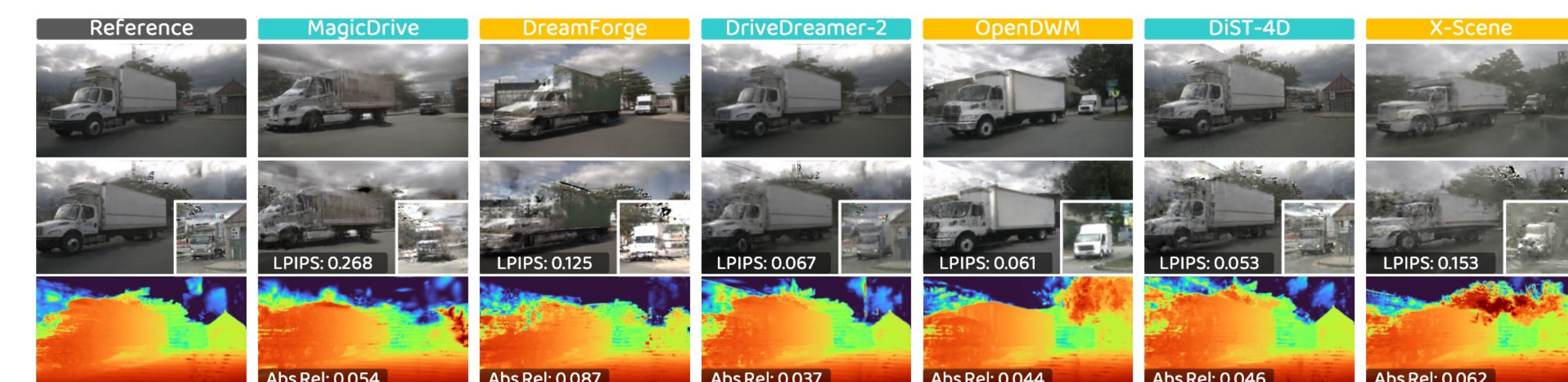


- ❖ We further conduct large-scale human evaluation study across physical plausibility, 4D consistency, and behavioral safety to measure how generated worlds **align** with human perception and expectations.
- ❖ We collect a large-scale dataset with human preference scores and textual rationales, aiming for scalable world-model evaluation.

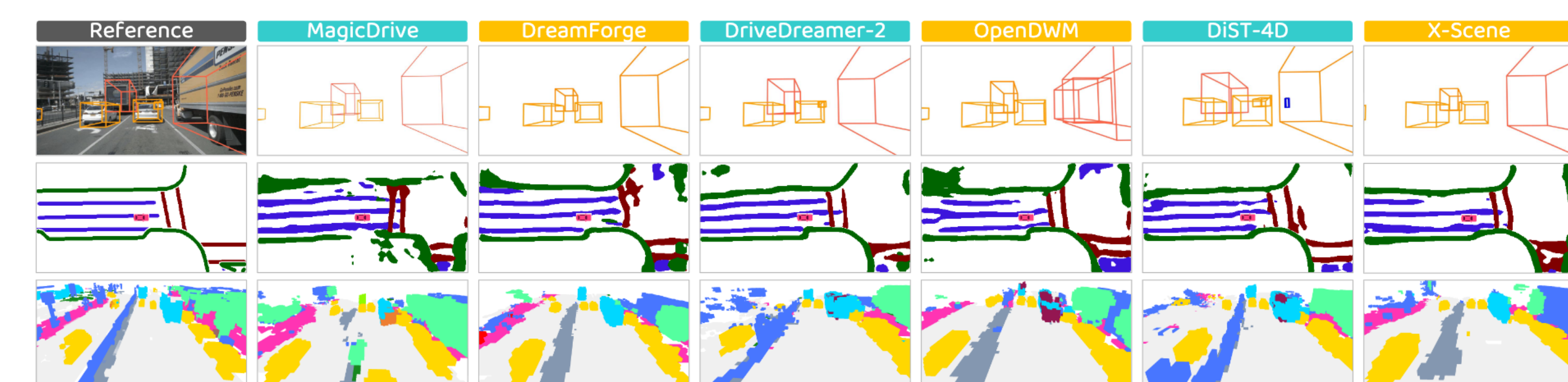


Benchmark Study & Experiments

- ❖ Experiments reveal that current models exhibit **highly uneven** capabilities across evaluation axes. World models that generate convincing videos still **collapse** under 4D reconstruction, **fail** to support safe planning, or provide **limited** downstream utility.



- ❖ We observe a clear tension between visual fidelity and world coherence. Those **appearance-driven** models generate sharp textures but unstable 3D layouts, whereas **geometry-aware** methods better preserve scene structure and motion continuity, though sometimes at the expense of fine visual details.



- ❖ Beyond visual, we observe large performance **gaps** in detection, tracking, depth, flow, occupancy prediction, and closed-loop simulation, indicating that synthetic **3D and 4D worlds** remain far from being dependable substitutes for real-world data.

