WFM-Eval: An Evaluation Framework for *Video World Models* in Robotic Manipulation

Sahil Khose[1] · Mengqi Zhang[1] · Prithvijit Chattopadhyay[2] · Judy Hoffman[1]

[1] UC Irvine · [2] NVIDIA · skhose@uci.edu

VWM @ CVPR 2026 | FMEA @ CVPR 2026

WEBSITE | CODE

1 The Gap

World models can imagine robot rollouts from one image plus an instruction. But do the generated videos teach the ***right* behavior?**

FVD/CLIP-Score measure realism, not faithfulness. Video can show the wrong action, hallucinate objects, or teleport them.

VLM judges aren't enough

- No single VLM reliably predicts task completion.
- Judges show **opposing biases** that don't cancel out.
- Best judge (Kimi-K2.5): only **69.5% F1**.

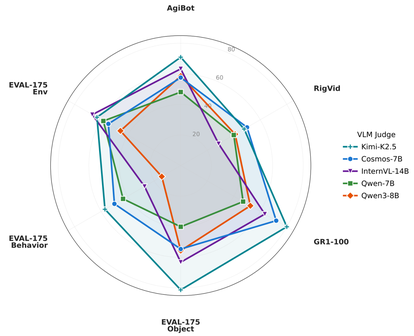

Fig. 1 · VLM judge accuracy (F1) across 6 datasets for task completion

2 WFM-Eval Framework

A multi-dimensional taxonomy across **three axes** and **four event types**, scored by an VLM-grounding pipeline.

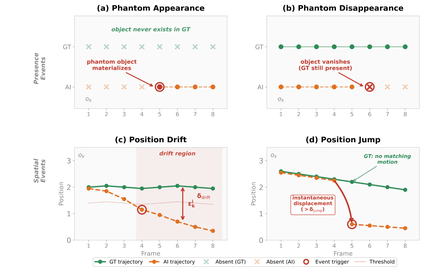

Fig. 2a · Error taxonomy.

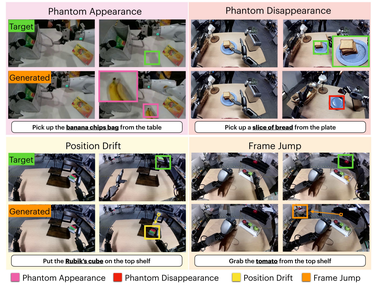

Fig. 2b · Failure modes. Real target vs. generated rollouts

Metrics · OHR hallucination · PAR position-anomaly · TCS temporal consistency · **HSS** harm-weighted severity

3 Rankings Reverse

Cross-family hallucination on **GR1** vs. **AgiBot**. The same model is best-in-class on one dataset, worst on the other.

Table 1 · HSS / OHR / PAR per dataset (lower HSS = better)

	GR1			AgiBot		
Model	HSS	OHR	PAR	HSS	OHR	PAR
Predict2	**.396**	.108	.067	**.699**	.341	.128
Veo 3.1	.436	.171	.050	.630	.274	.131
Hunyuan	.482	.200	.048	**.596**	.256	.128
Predict2.5	.501	.210	.049	.682	.322	.128
Wan2.2	**.518**	.207	.058	.661	.301	.145

- **Predict2:** rank 1 on GR1 → rank 5 on AgiBot. **Hunyuan** does the opposite (3 → 1).
- Domain training trades generalization for peak: Predict2's HSS degrades **81%** across datasets vs. **19%** for Hunyuan.
- Size doesn't explain it: 2B Predict2 > 14B Wan2.2 on GR1.

Single-dataset benchmarks mislead. Veo 3.1 is the only family that holds rank across both.

4 It Transfers Downstream

Predict2.5 regresses on GR1 (+0.095 HSS): a **52% rise in phantom appearances** mistrains the policy.

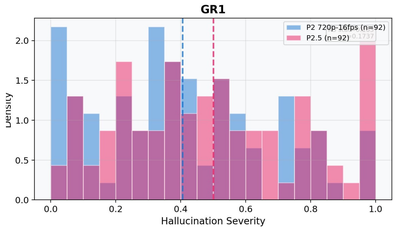

Fig. 3 · Severity, GR1. P2.5 (pink) shifts right with a spike at 1.0; no gap on AgiBot (KS p=0.82).

+8.75 LIBERO success-rate points, Predict2 over Predict2.5

95.45% vs. 86.70% avg; largest gap on LIBERO-Long (89.6 vs 69.4).

Takeaways

1. **Object hallucination** drives model differences.
2. **Rankings reverse** across datasets, so use more than one.
3. Diagnostics **predict downstream** policy success.