Trinity of Consistency: A Framework for Measuring What Makes a General World Model

Kabui, Charles

2026-03-12

Read at ToKnow.ai

Researchers from Shanghai AI Lab, Westlake University, NUS, and several other institutions published "Towards General World Models", a 119-page survey with 50 figures that proposes a formal framework for evaluating world models. The core argument: a general world model must

satisfy three consistency axes simultaneously. Modal consistency ensures that different input types (text, image, video, audio) produce coherent shared representations. Spatial consistency preserves 3D geometry, so objects maintain their shape, size, and relative position across viewpoints. Temporal consistency enforces cause-and-effect, meaning actions have plausible consequences that persist over time. The paper traces the field's evolution from isolated specialist modules toward unified architectures and introduces CoW-Bench (Consistency of World Benchmark), a new evaluation suite with 1,485 samples across 18 sub-tasks spanning all three axes and their interactions.

CoW-Bench reveals a pattern the authors call "constraint backoff." Current video generation models and Unified Multimodal Models score well on local motion plausibility but fail when a task requires maintaining a persistent, goal-directed world state. A model can make a ball roll convincingly in isolation yet lose track of it the moment two physical constraints interact. The benchmark gives the community a shared protocol for measuring these failures rather than relying on subjective visual quality.

Building models that look realistic is increasingly a solved problem. The open challenge is building models that reason about the physical world consistently across modalities, space, and time, which is what separates a video generator from an actual world simulator.

Sources:

- Towards General World Models: A Survey (arXiv)
- Hugging Face Daily Papers Discussion
- CoW-Bench Project Page
- Awesome World Model Evolution (GitHub)
