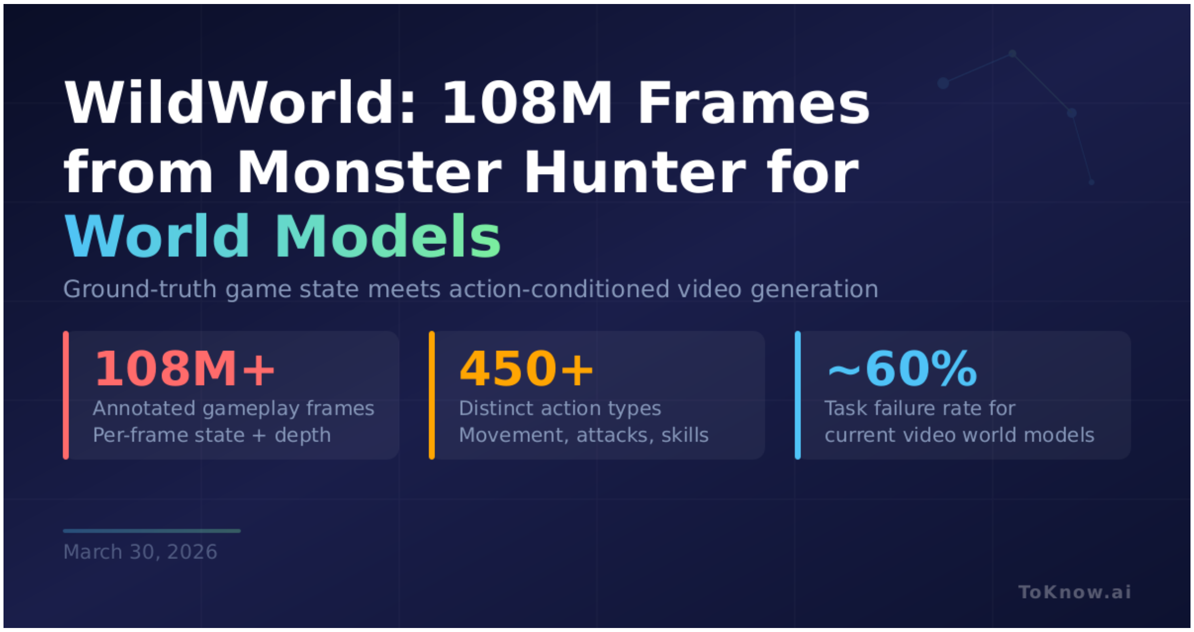
WildWorld: 108 Million Frames from Monster Hunter for Training World Models

Kabui, Charles

2026-03-30

Read at ToKnow.ai


WildWorld: 108M Frames from Monster Hunter for World Models
Ground-truth game state meets action-conditioned video generation
108M+
Annotated gameplay frames
Per-frame state + depth
450+
Distinct action types
Movement, attacks, skills
~60%
Task failure rate for current video world models
March 30, 2026
ToKnow.ai

Shanda AI Research Tokyo released WildWorld, the largest open dataset for training action-conditioned world models: 108 million frames of gameplay from Monster Hunter: Wilds, a photorealistic AAA action RPG. What makes it different from existing video datasets is the inclusion of ground-truth game state. Every frame comes with synchronized annotations: character skeletons, world states (health, ammo, animation phase), camera poses, and depth maps, all extracted automatically from the game engine. The dataset spans 450+ distinct actions across 29 monster species, 4 player characters, and 4 weapon types, with clips running up to 30 minutes of continuous play. Alongside the data, the team built WildBench, a benchmark that tests models on whether they faithfully simulate actions ("Action Following") and whether internal state stays consistent over time ("State Alignment"). Current models fail roughly 60% of tasks.

That 60% failure rate is the real takeaway. Existing video generation models can produce convincing pixels, but they fall apart when actions need to cause specific, state-dependent consequences. A "shoot" action should reduce ammo count, which later determines whether the character can fire again. Without explicit state, models can't learn these dependencies. WildWorld gives researchers the data to fix that, at zero annotation cost since everything is engine-extracted.

This is the clearest sign yet that world models need structured state, not just pixels. The field has been training on flat video for years. WildWorld suggests the next breakthroughs will come from data that separates what happened (state) from what it looked like (pixels). For related work on world model evaluation, see Trinity of Consistency.

Sources:

- WildWorld Paper (arXiv)
- WildWorld GitHub
- WildWorld Project Page
- HuggingFace Daily Papers (March 25)

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more: /terms-of-service***