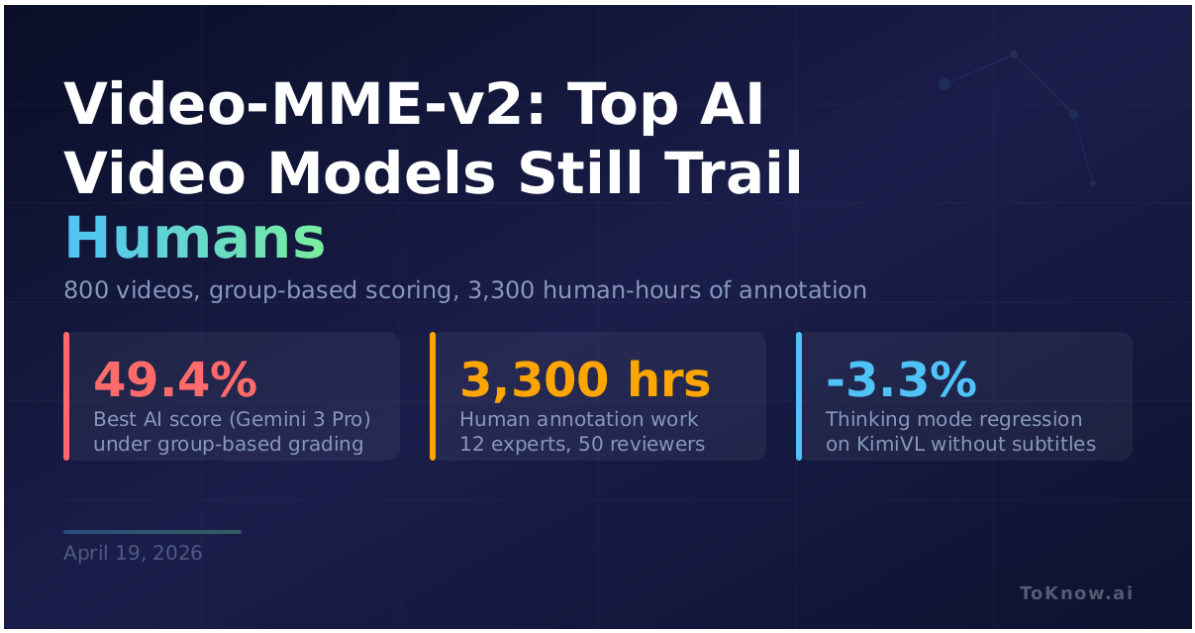
Video-MME-v2: Top AI Video Models Still Trail Humans by a Wide Margin

Kabui, Charles

2026-04-19



A team led by Chaoyou Fu released Video-MME-v2, a video understanding benchmark built to expose the gap between leaderboard scores and real model capability. It contains 800 videos averaging 10.4 minutes, with 4 linked questions per video and 8 answer options each, built over 3,300 human-hours by 12 annotators and 50 independent reviewers. Two design choices matter. A tri-level hierarchy progresses from finding scattered visual cues, to tracking actions

and time, to multi-step reasoning over plot, physics, and social behavior. And questions are scored in groups of 4 with a non-linear formula $(N/4)^2$, so getting one right by luck while missing related questions earns almost nothing. Gemini 3 Pro, the strongest model tested, scores 66.1% on plain per-question accuracy but only 49.4% under group scoring.

For anyone picking a model for sports analysis, surveillance review, or instructional video QA, this is the first benchmark that punishes lucky single-question hits and rewards stable understanding. On Action & Motion and Physical World Reasoning, even Gemini 3 Pro scores below 30%. Open-source models lag further: Qwen3.5-397B-Think reaches 39.1% with 512 frames but drops to 30.6% at 64 frames. The most interesting finding is about thinking mode. Extended reasoning helps when subtitles are present but often hurts on purely visual tasks: KimiVL-16B loses 3.3% overall with thinking on, and 4% on the hardest Level 3 questions. Current "thinking" in video models leans on text cues, not pixels.

Read More: WildWorld takes the opposite angle, building structured video data so models can learn what pixels actually mean.

Sources:

- Video-MME-v2 project page and leaderboard
- arXiv: Video-MME-v2 paper
- GitHub: MME-Benchmarks/Video-MME-v2
- Hugging Face dataset
- Original Video-MME (2024)
