Technical Summary: ActionAtlas: A VideoQA Benchmark for Domain-specialized Action Recognition

Part I: Problem Formulation, Methods, and Evidence

Kai-Yu Lu

2025/10/10

1 Research Problem and Motivation

Many video understanding benchmarks allow high accuracy using a single frame, which reveals a static appearance bias and weakens assessment of temporal motion understanding in Vision–Language Models. The paper introduces ActionAtlas v1.0 to evaluate whether modern models can recognize fine-grained, domain-specialized actions that require identifying a specific actor and discriminating subtle motion patterns within short sports videos.

2 Related Work

Classical action datasets such as UCF101, HMDB51, ActivityNet, AVA, Kinetics, Moments-in-Time and several fine-grained sports datasets primarily emphasize coarse actions or settings where static frames are informative. Video–Language benchmarks and multiple-choice conversions also tend to emphasize appearance. ActionAtlas differs by coupling fine-grained domain actions with natural-language disambiguation of actor identity and a narrow temporal window, aligning evaluation with common Vision–Language usage.

3 Dataset Construction

Scope and Scale

ActionAtlas v1.0 comprises 934 short videos that cover 580 unique actions across 56 sports. The answer-choice pool spans 1,896 actions. Average clip length is 6.07 seconds with mean frame rate 32.18 FPS.

Collection Pipeline

The authors compile action lists per sport using GPT-4 to broaden coverage, then query YouTube and download metadata at scale. Exact and soft lexical filters and CLIP similarity screening retain likely candidates. For long videos, Whisper transcripts and GPT-4 text prompts localize candidate 30-second segments without using frames. Crowd workers then validate action presence, propose tighter temporal spans, and provide actor attributes. The team generates multiple-choice questions and hard negatives with GPT-4, then conducts extensive quality checks, rewriting any items that could be answered from text alone and blurring in-frame text that might leak answers.

Videos	934
Unique actions	580
Sports	56
Total actions in choices	1,896
Average duration (seconds)	6.07
Average frame rate (FPS)	32.18

Table 1: ActionAtlas v1.0 dataset statistics.

Dataset Summary

4 Query Protocol and Task Definitions

Each item is a multiple-choice Video Question Answering sample. The question uses natural language to pinpoint a target person and a narrow time span, then asks which option *best* describes the performed action. Answers require tracking the correct actor and discriminating motions that can look similar in single frames.

5 Modeling Approach

Evaluation Setup

The benchmark evaluates proprietary and open-weight Vision–Language Models. For models that do not accept native video, uniformly sampled frames are fed as images together with the question and options. Gemini video mode samples frames internally at one frame per second; an additional evaluation converts videos to 1 FPS to expose all frames under that mode. The study reports top-1 accuracy with bootstrap 95% confidence intervals and audits efficiency factors, including number of frames, tokens, and approximate inference FLOPs.

Baselines

A non-expert human baseline is measured by providing brief action descriptions for each choice. CLIP is evaluated using class prompts constructed from the question and options.

6 Empirical Results

Overall Difficulty

Random chance is approximately 20.91%. Non-expert humans reach 61.64% with access to short action descriptions. The best proprietary model, GPT-4o, attains up to 45.52%. Open-weight models perform close to chance, with Qwen2-VL-7B peaking at 30.24%.

Effect of Frame Sampling

For GPT-4o, increasing frames from 1 to 16 improves accuracy from 33.08% to 42.95%. Gemini 1.5 Pro video mode operates at 1 FPS by default and shows limited improvement when converted videos expose all frames under that mode.

Ablations

Providing concise textual descriptions of action choices does not significantly improve model accuracy, although such descriptions help humans. Chain-of-thought prompting without descriptions reduces accuracy; combining chain-of-thought with descriptions yields a modest increase for GPT-4o but not a significant improvement over the no-chain-of-thought setting.

Headline Numbers

Model and setting	Frames	Accuracy (%)
Random chance	–	20.91
Non-expert humans (with descriptions)	–	61.64
CLIP ViT-L-14-336	16	23.71
Qwen2-VL-7B (best open)	16	30.24
GPT-4o	1 / 8 / 16	33.08 / 41.55 / **42.95**
Gemini 1.5 Pro (video, 1 FPS)	–	32.37
Gemini 1.5 Pro (video, all frames via 1 FPS conversion)	–	35.59

Table 2: Selected results on ActionAtlas. Confidence intervals are reported in the paper.

7 Summary

Methodological Strengths

The benchmark enforces actor grounding and temporal localization through natural-language questions, stresses fine-grained motion discrimination across many sports, and uses a scalable pipeline that combines LLM-driven action discovery, metadata and CLIP filtering, ASR-guided segment search, crowd validation, and rigorous leakage control. The evaluation isolates visual signals and reports uncertainty and efficiency.

Key Limitations

The current release focuses on sports and lacks a formal action taxonomy. The scale is moderate relative to some multimodal benchmarks. Future iterations plan to expand domains and incorporate expert-informed taxonomies.

Main Findings

Open-weight models remain near chance on fine-grained action recognition. GPT-4o benefits from denser frame sampling. Gemini's default sampling under-utilizes motion cues. Textual access to action definitions does not close the gap, indicating that visual recognition, not label knowledge, is the primary bottleneck.

Terminology and Definitions

Video Question Answering is a task where a model answers natural-language questions about a video. In this benchmark it is multiple-choice with actor and time-span grounding.
Vision–Language Model denotes a model jointly processing visual and textual inputs.

Static appearance bias refers to high performance achieved with single frames, implying weak temporal modeling.