SAGIVTECH IMVC 2026

15th Israel Machine Vision Conference (IMVC) 2026

Monday | April 27, 2026 | Pavilion 10, EXPO Tel Aviv

HERBench: A Benchmark for Multi-Evidence Integration in Video Question Answering

Dan Ben Ami, Ben Gurion University

Video Large Language Models (Video-LLMs) are rapidly improving, yet current Video Question Answering (VideoQA) benchmarks often allow questions to be answered from a single salient cue, under-testing reasoning that must aggregate multiple, temporally separated visual evidence. In this direction, we present HERBench, a VideoQA benchmark purpose-built to assess multi-evidence integration across time. Each question is constructed to require aggregating at least three non-overlapping evidential cues across distinct video segments (so neither language priors nor a single snapshot can suffice). HERBench comprises 26K five-way multiple-choice questions organized into twelve compositional tasks that probe identity binding, cross-entity relations, temporal ordering, co-occurrence verification, and counting. To make evidential demand measurable, we introduce the Minimum Required Frame-Set (MRFS)—the smallest number of frames a model must fuse to answer correctly—and show that HERBench imposes substantially higher demand than prior datasets (mean MRFS 5.5 vs. 2.6-4.2). Evaluating 13 state-of-the-art Video-LLMs on HERBench reveals pervasive failures: accuracies of 31-42% are only slightly above the 20% random-guess baseline. We disentangle this failure into two critical bottlenecks: (1) a retrieval deficit, where frame selectors overlook key evidence, and (2) a fusion deficit, where models fail to integrate information even when all necessary evidence is provided. By making cross-time evidence both unavoidable and quantifiable, HERBench establishes a principled target for advancing robust, compositional video understanding.