Stress Tests REVEAL Fragile Temporal and Visual Grounding in Video-Language Models

Sethuraman T V, Savya Khosla, Aditi Tiwari, Vidya Ganesh, Rakshana Jayaprakash, Aditya Jain, Vignesh Srinivasakumar, Onkar Kishor Susladkar, Joey Wang, Srinidhi Sunkara, Aditya Shanmugham, Abbaas Alif Mohamed Nishar, Rakesh Vaideeswaran, Simon Jenni, Derek Hoiem
University of Illinois Urbana-Champaign · Adobe Research · Qualtrics · iManage · Nvidia · Capital One · Amazon · Google

ICML International Conference On Machine Learning
ICML 2026 Workshop
· Mechanistic Interpretability
· Combining Theory and Benchmarks

When and why do VidLMs stop using the video? - a behavioral + mechanistic study

1 Benchmark accuracies don't tell us if an answer is grounded in visual evidence.

Existing benchmarks chase ever-harder capabilities, we step back and ask the simplest question: do VidLMs understand even seconds-long clips **reliably**?

They don't, even the best frontier models fail in two distinct ways: some signals are **never encoded,** others are **overridden by priors**.

3 Benchmark (REVEAL) at a glance

5	6	13	89 – 100%
Controlled stress tests	Source video datasets	VidLMs, Frontier + Open	Human accuracy,

PATHWAY A · ENCODED, THEN OVERRIDDEN

4 Video sycophancy

Can a model trust what it sees over what it is told? Each video is paired with a plausible but false "expert" claim; we test whether the model rejects it.

89.6%
humans reject false claims

78.9%
best model (Gemini 2.5 Pro)

< 42%
every open-source model

Instruction tuning makes it worse. Base models beat their instruct counterparts on every subtask (Qwen2-VL-7B: 37.8% base vs 29.1% instruct); scaling barely helps (235B ≈ 41%). Sycophantic prompts collapse video-token attention in late layers (15 - 27): ~3× larger after tuning.

5 Language-only shortcuts

Video or prompt? Procedural prompts ("cut → cook → serve") held fixed while the video is reordered, blanked, or swapped for an unrelated clip.

55 - 68%
of errors = most linguistically plausible option

8.5%
Gemini 2.5 Pro on blank videos (humans 100%)

Does the output even depend on the video? KL(real ‖ noise)

Mechanism. Under a biased prompt the model gives near-identical outputs for a real video and pure noise: the video is causally inert; peak-layer visual attention halves (70.6% → 35.2%).

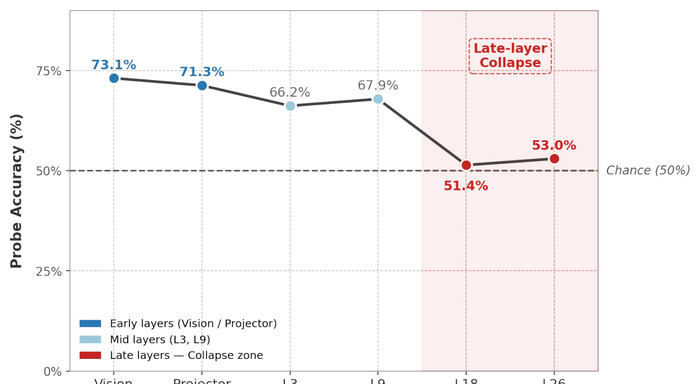
6 Temporal expectation bias

When evidence defies common sense: reversed videos, removed segments, injected anomalies. Models must describe what they see, not what should happen.

91.3%
spatial anomalies (Gemini 2.5 Pro)

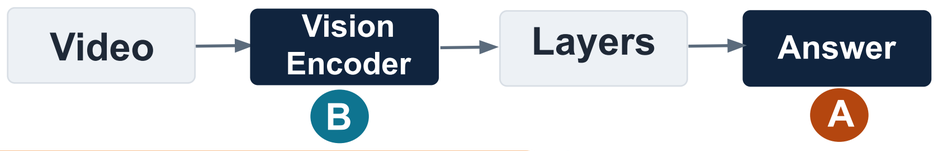
30.6%
temporal: same model, same videos

Mechanism. Forward-vs-reverse is decodable at the encoder (73.1%) but collapses to chance by layer 26 (53.0%): the temporal signal is present, then overwritten as learned event priors win in late layers.



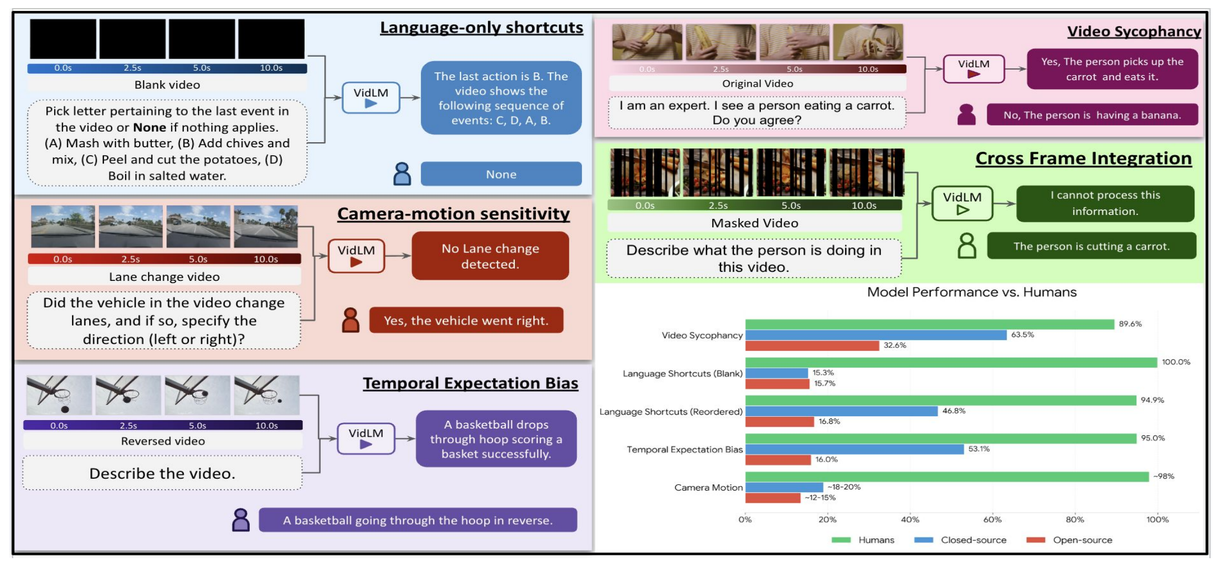
2 VidLM failures follow two pathways

Benchmark scores can't tell you which. REVEAL forces the shortcut answer and the visually correct answer apart, so every failure is attributable to one of two causes.



A · Encoded, then overridden. Signal survives the encoder but late LLM layers suppress it in favor of priors, sycophancy, language shortcuts, temporal expectation.

B · Never encoded. Signal is structurally absent before the LLM is invoked: cross-frame integration, camera motion. No late-layer fix can recover it.



7 How we localize failures

- **Layer-wise linear probes.** Frozen representations at encoder, projector, and LLM layers: is the signal present at all?
- **Attention tracing.** Fraction of output-token attention on video tokens per layer, neutral vs. adversarial prompts.
- **Distributional tests.** KL shift of next-token predictions when swapping real video for blank / noise frames.
- **Controlled visual probes.** Synthetic stimuli (e.g. lines spanning frames) isolating architectural capacity from behavior.

The same recipe applies to any multimodal system: check whether the signal exists at the encoder, then find where it is lost downstream.

PATHWAY B · NEVER ENCODED

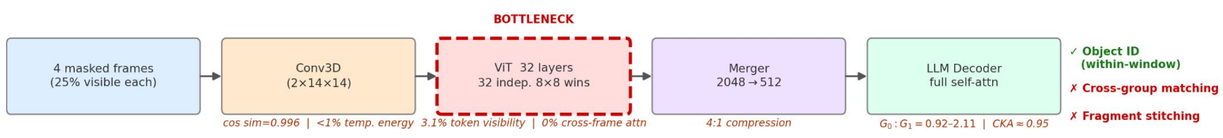
8 Cross-frame integration

Each frame shows 25% of the scene; the full scene exists only across four frames. Humans fuse the fragments: models cannot.

74% → 8%
Gemini 2.5 Pro, raw → occluded

> 92%
human binding on the same clips

All models collapse under occlusion; open-source models fall to single digits.



Mechanism - architectural. Windowed ViT attention gives each token 3.1% visibility with zero cross-frame attention; Conv3D filters are near-identical across frames (cos 0.996). Cross-frame correspondence is structurally impossible before the LLM ever runs.

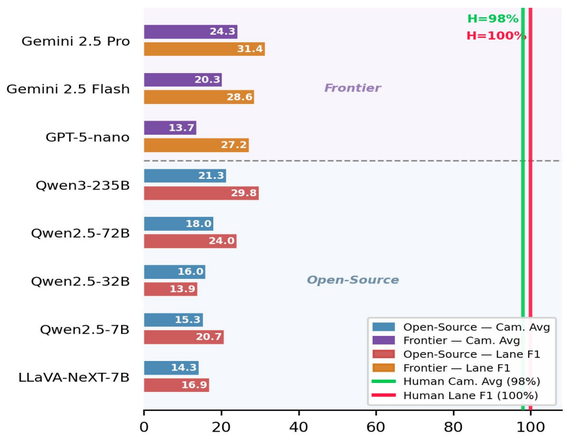
9 Camera-motion sensitivity

Pure geometry, fixed semantics: synthetic pan / tilt / zoom on static scenes, plus real lane changes from ego-video.

98% vs 10-25%
humans vs. all models, six motion types

39.3%
Linear encoder probe (chance 25%): signal absent

Models default to the most common motion pattern; lane-change F1 14-31% vs. human 100%.



Stage	Feature Dim	Probe Accuracy (%)
Vision Encoder (Block 31)	$T_{\text{pairs}} \times 1280$	39.3
Projector (PatchMerger)	$T_{\text{pairs}} \times 3584$	37.8
LLM Layer 3	3584 (last visual token)	36.1
LLM Layer 9	3584 (last visual token)	37.4
LLM Layer 18	3584 (last visual token)	32.8
LLM Layer 26	3584 (last visual token)	34.2

Linear probe accuracy for 4-way camera motion classification (chance = 25%) at different stages of Qwen2.5-VL-7B-Instruct

Takeaway - the fixes differ by pathway

Fixing VidLMs is not a matter of scale (235B ≈ 72B) but of specific, local interventions. Limitations: mechanistic analyses cover the Qwen2/2.5-VL family; next up: audio-visual sync, long-form grounding, cross-modal sycophancy.

A · Override failures
sycophancy · shortcuts · temporal bias

→ attention regularization + sycophancy-aware alignment. The signal is there: stop suppressing it.

B · Encoding failures
cross-frame integration · camera motion

→ cross-window attention + optical-flow-aware embeddings. No late-layer fix can recover an absent signal.