Time Blindness: Why Video-Language Models Can't See What Humans Can?

Ujjwal Upadhyay*, Mukul Ranjan*, Zhiqiang Shen, Mohamed Elhoseiny

We expose a major flaw in top Video-Language Models like GPT-4o & Gemini:

They're blind to completely temporal patterns.

Humans score 98%.
These models? 0%.

timeblindness.github.io

Why does this happen?

Today's Video-Language Models (VLMs) aren't really watching videos.

1. They just look at frames.
2. Extract spatial features.
3. Then try to guess what's happening between them.

They don't truly see through time. They see time through spatial lens.

Model	Direct Prompt	CoT	Params
Human Performance	**98.0% ± 0.6**	N/A	N/A
Open-Source Models			
VideoLLaMA3-7B [97]	0% ± 0.0	0% ± 0.0	7B
VideoLLaMA3-2B [97]	0% ± 0.0	0% ± 0.0	2B
TimeChat-7B [68]	0% ± 0.0	0% ± 0.0	7B
MiniGPT4-Video [3]	0% ± 0.0	0% ± 0.0	7B
MovieChat [74]	0% ± 0.0	0% ± 0.0	7B
Video-ChatGPT-7B [54]	0% ± 0.0	0% ± 0.0	7B
VideoGPT-plus-Phi3-mini-4k [55]	0% ± 0.0	0% ± 0.0	7B
VILA1.5-13b[50]	0% ± 0.0	0% ± 0.0	13B
ShareGPT4Video-8B [13]	0% ± 0.0	0% ± 0.0	8B
VideoLLaMA2-7B [22]	0% ± 0.0	0% ± 0.0	7B
Video-LLaVA [99]	0% ± 0.0	0% ± 0.0	7B
LLaVA-NeXT-Video [45]	0% ± 0.0	0% ± 0.0	8B
InternVL2-40B [20]	0% ± 0.0	0% ± 0.0	40B
InternVL2-8B [20]	0% ± 0.0	0% ± 0.0	8B
InternVL2.5-78B [19]	0% ± 0.0	0% ± 0.0	78B
InternVL2.5-8B [19]	0% ± 0.0	0% ± 0.0	8B
InternVideo2.5-Chat-8B [84]	0% ± 0.0	0% ± 0.0	8B
InternVideo2-Chat-8B [82]	0% ± 0.0	0% ± 0.0	8B
Qwen2-VL-2B-Instruct [80]	0% ± 0.0	0% ± 0.0	2B
Qwen2-VL-7B-Instruct [80]	0% ± 0.0	0% ± 0.0	7B
Qwen2-VL-72B-Instruct [80]	0% ± 0.0	0% ± 0.0	72B
Qwen2.5-VL-3B-Instruct [5]	0% ± 0.0	0% ± 0.0	3B
Qwen2.5-VL-7B-Instruct [5]	0% ± 0.0	0% ± 0.0	7B
Qwen2.5-VL-72B-Instruct [5]	0% ± 0.0	0% ± 0.0	72B
Qwen3-VL-8B-Instruct [78]	0% ± 0.0	0% ± 0.0	8B
Closed-Source Models			
Gemini 2.5 Pro [24]	0% ± 0.0	0% ± 0.0	N/A
Gemini 1.5 Pro [77]	0% ± 0.0	0% ± 0.0	N/A
Gemini 2.0 Flash[27]	0% ± 0.0	0% ± 0.0	N/A
GPT-4o [36]	0% ± 0.0	0% ± 0.0	N/A

“SpookyBench”

We designed a benchmark: videos made entirely of noise.

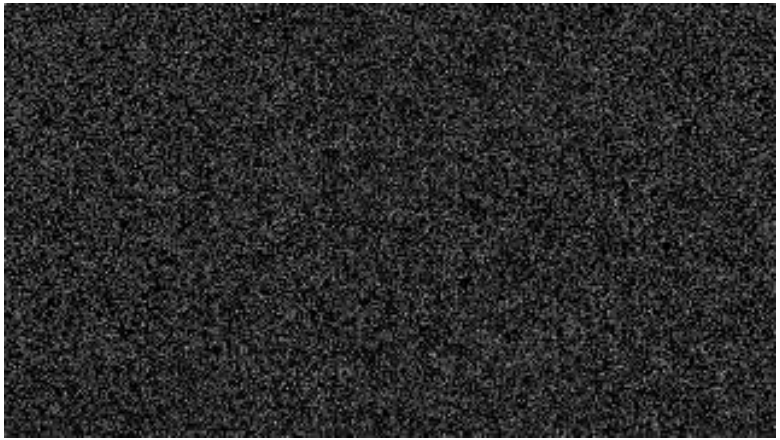
But hidden in that noise?
Text. Shapes. Objects. Depth Map.

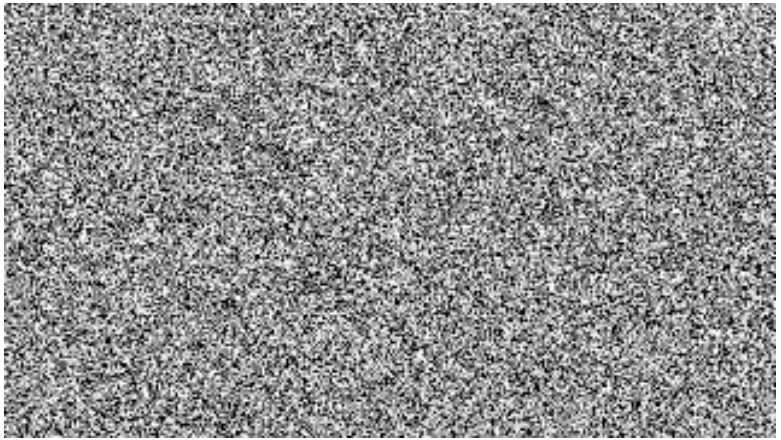
The catch? You can only see them through motion.

Each frame looks meaningless, but together, they reveal something spooky.

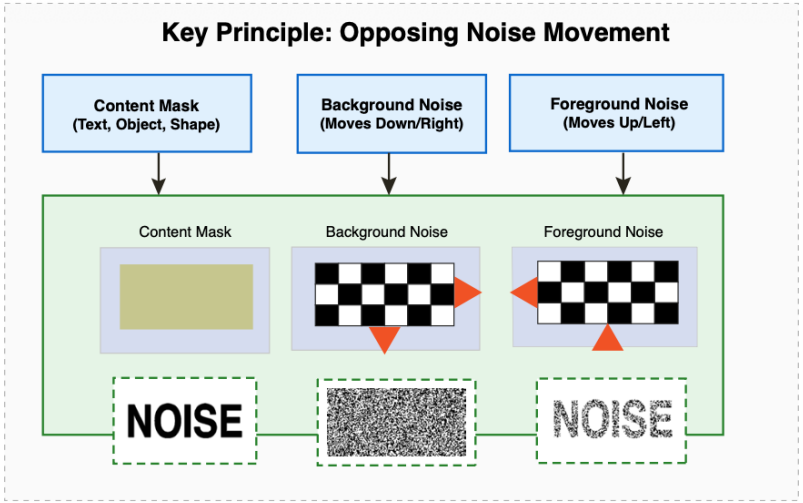
Examples

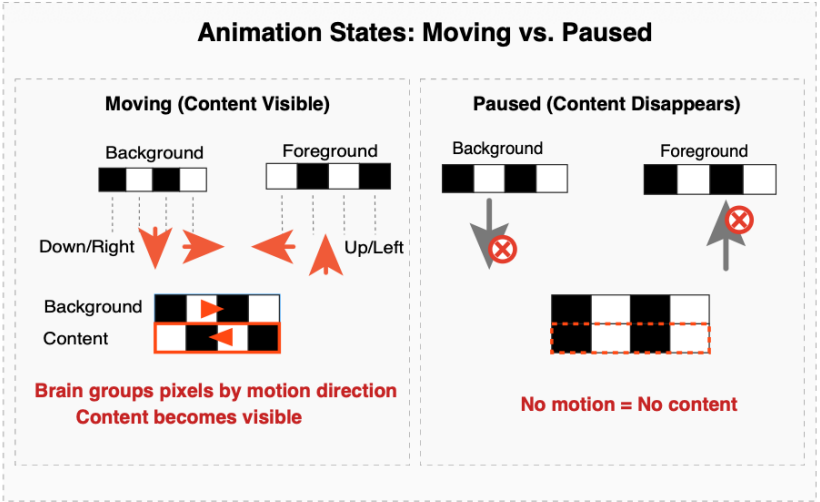
A paused video will feel like a completely random noise.





How we made it?





Brain Perceives Content Through Motion

Brain groups motion → Content becomes visible

Machine, Human Test

The test is simple.

- Show humans a video made of moving noise.
- Ask: What do you see?

Humans: “It says CAT”
GPT-4o: “I don’t understand” (Mostly some random thing based on prompt)

Humans crushed it: 98% accuracy.
All 17 state-of-the-art models? 0%.

Even when prompted how to look.

Human accuracy for various category of data

Annotator	Text		Images		Dynamic Scenes	
	Acc(%)	Perc(1-5)	Acc(%)	Perc(1-5)	Acc(%)	Perc(1-5)
Annotator 1	99.5	4.7	99.5	4.7	96.5	4.3
Annotator 2	98.6	4.8	98.4	4.9	91.2	4.0
Annotator 3	99.5	4.9	97.2	4.5	94.7	4.4
Annotator 4	97.6	4.6	96.7	4.5	91.2	4.0
Annotator 5	100.0	4.8	99.5	4.7	99.0	4.7
Annotator 6	98.0	4.7	97.8	4.5	93.0	4.2
Mean	**98.9±0.7**	**4.8±0.0**	**98.2±1.1**	**4.7±0.1**	**94.3±3.1**	**4.3±0.1**

Human accuracy at different frames per second

Category	1 FPS	5 FPS	10 FPS	20 FPS	30 FPS
Images	0.0	12.5	80.0	95.8	97.5
Words	0.0	10.8	35.8	95.8	95.8
Videos	0.0	15.0	62.5	93.3	93.3
Average	0.0	12.8	59.4	95.0	95.6

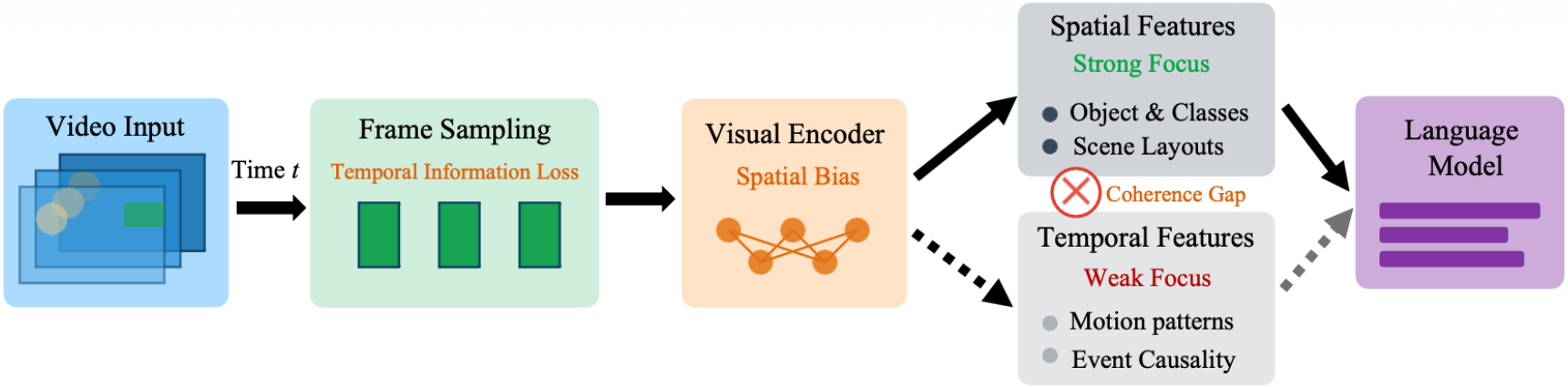
Can it be a out of domain problem?

1. We fine-tuned InternVL2.5-8B, Qwen2-VL-7B, and Qwen2.5-VL-7B models.
2. We ran all the models on 60fps video to check if frame drop was causing the issue.

All still yielded 0% accuracy.

This indicates fundamental issue with current VLMs.

Potential reason for failure of VLMs

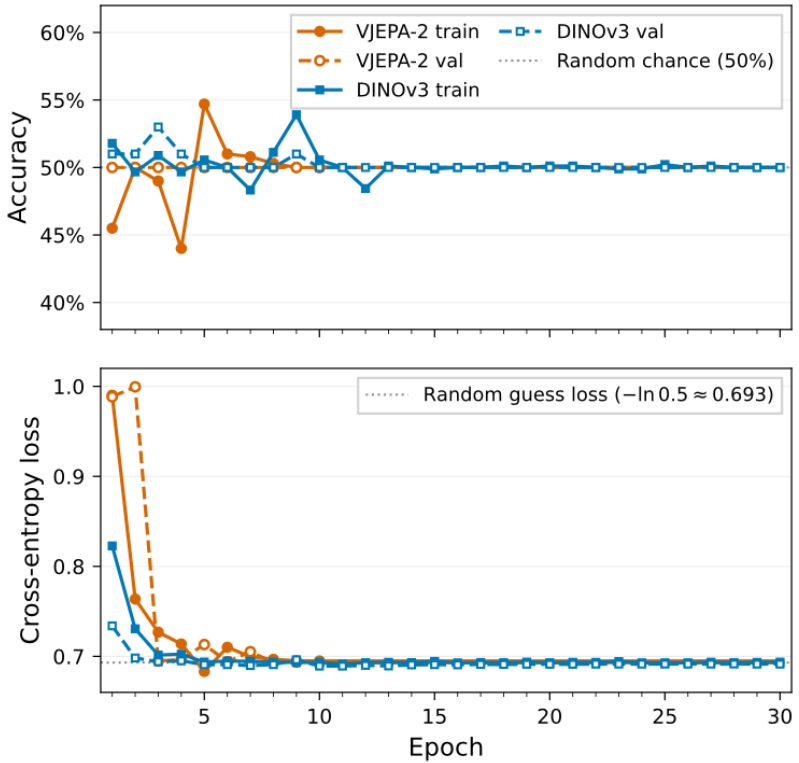


Can SOTA Visual Models Tell There’s an Object?

Total Failure, Models guess at random when forced to detect objects purely through motion.

Structural Flaw, Adding more training data doesn't fix this fundamental architectural limitation

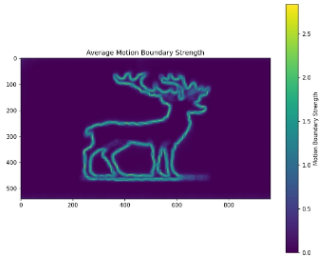
No Overfitting Observed, even when working with smaller dataset.

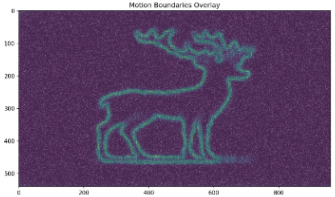


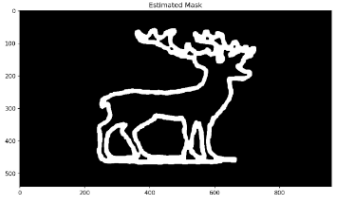
What unlocks VLMs performance?

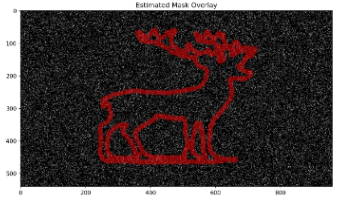
Pre-compute motion boundaries with classical optical flow, overlay them on the noisy frames, and the same models suddenly score in the 50s.

Model	Words	Images	Videos	Overall
Qwen2-VL-7B (Baseline)	0.0%	0.0%	0.0%	0.0%
Qwen2-VL-7B (Augmented)	50.95%	70.51%	1.75%	51.54%
GPT-4o (Baseline)	0.0%	0.0%	0.0%	0.0%
GPT-4o (Augmented)	56.19%	83.33%	3.51%	59.10%







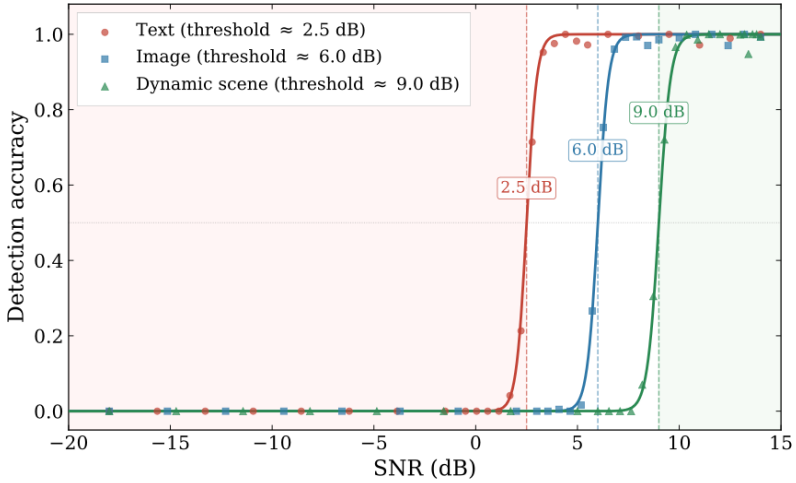


Why this matters?

Imagine:

- In medical imaging, key signals may only emerge over time, not in any single frame.
- Security systems miss suspicious behavior if they only see stills, not subtle patterns over long time.
- Low SNR input

Model performance vs SNR on SpookyBench



SpookyBench exposes below vulnerability among many:

1. Video-Language Models aren't truly temporal
2. They can't perceive motion-based meaning
3. They're vulnerable to temporal adversarial attacks

We'll be presenting our work physically in Denver on 7th June
Poster Session 5 & Exhibit Hall (ExHall F)

timeblindness.github.io