جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Mohamed bin Zayed University of Artificial Intelligence

docpanel

CVPR JUNE 3-7, 2026 DENVER COLORADO

VISION-LANGUAGE MODELS · TEMPORAL REASONING · BENCHMARK

Time Blindness.

Why Video-Language Models can't see what humans can.

Ujjwal Upadhyay[1,3,*] · Mukul Ranjan[2,*] · Zhiqiang Shen[2,†] · Mohamed Elhoseiny[1,†]

[1]KAUST [2]VILA Lab, MBZUAI [3]DocPanel Technologies · *Equal contribution · †Corresponding authors

98%
HUMANS
6 annotators · perceptibility 4.7 / 5

0%
EVERY VIDEO-VLM
25+ models · 2 B – 78 B · open + closed

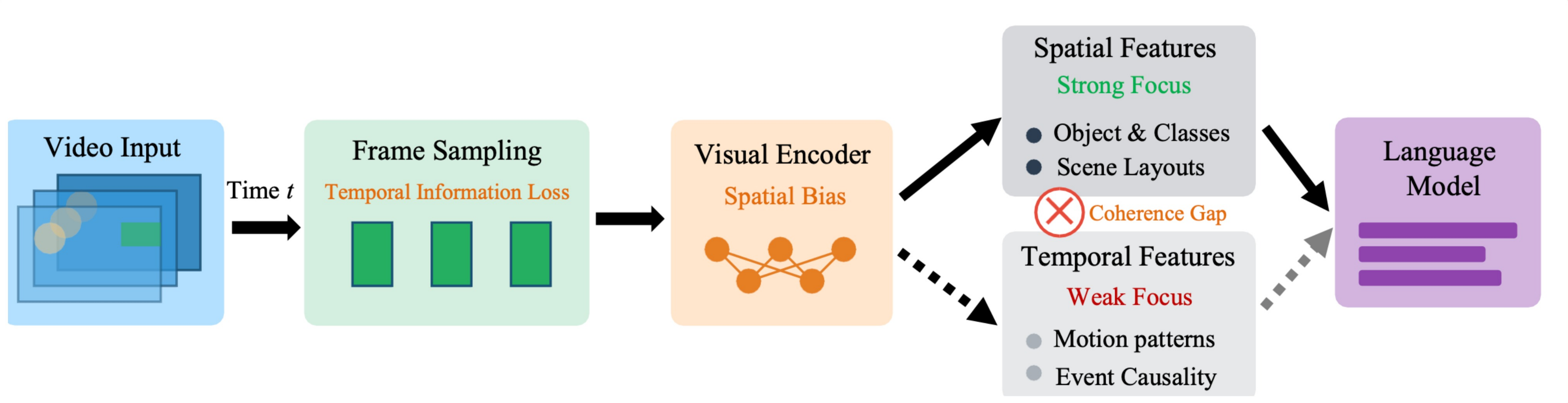
timeblindness.github.io

01 THE QUESTION

*If a video carries information only in **how** frames change over time — and each frame is pure noise — can today's Video-VLMs see it?*

Nature already does this. Fireflies communicate through flash sequences, EEG pathologies emerge from temporal patterns, Morse code carries language without spatial form. Yet modern video understanding pipelines extract frame features first and treat time as an afterthought.

02 THE SPATIAL-RELIANCE GAP



Frame sampling discards time. Visual encoders bias toward spatial features (objects, layouts); temporal features (motion, causality) stay under-represented — opening the coherence gap (×) we exploit.

03 SPOOKYBENCH — TIME-ONLY STIMULI

A synthetic benchmark where **each frame is pure noise** and the content emerges only from *how* the noise moves.

- **451 videos** · 960×540 · avg 7.1 s · ~333 frames
- **3 categories**: Text (46.6%), Object Images (40.8%), Dynamic Scenes (12.6%)
- **Opposing-motion** noise for text/images; **threshold-gated** motion for depth videos
- Generator releases an *indefinitely* extendable dataset

04 FRAMES CARRY NOTHING

CATEGORY	BASIC SNR	PERCEPTUAL	TEMP. COH.	MOTION C.
Images	−46.95	−47.28	8.00	7.17
Dynamic Scenes	−48.95	−63.43	**21.91**	−3.18
Text	−39.27	−49.18	7.84	8.26

Read: Dynamic Scenes have the highest temporal coherence (21.9) yet near-zero motion contrast — the regime current VLMs cannot exploit.

BASIC SNR Motion energy ÷ frame variance.

PERCEPTUAL SNR Weighted to human contrast sensitivity.

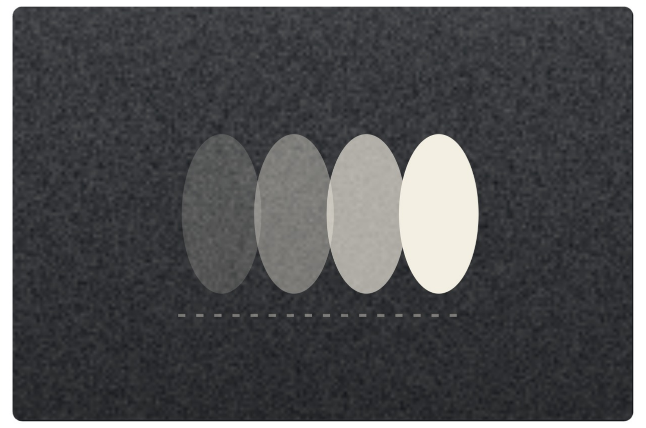
TEMPORAL COHERENCE Motion vectors aligned over time.

MOTION CONTRAST Foreground vs background motion.

05 THREE STIMULUS FAMILIES

TEXT · 46.6%
Words encoded as opposing-motion noise. **210 videos.**

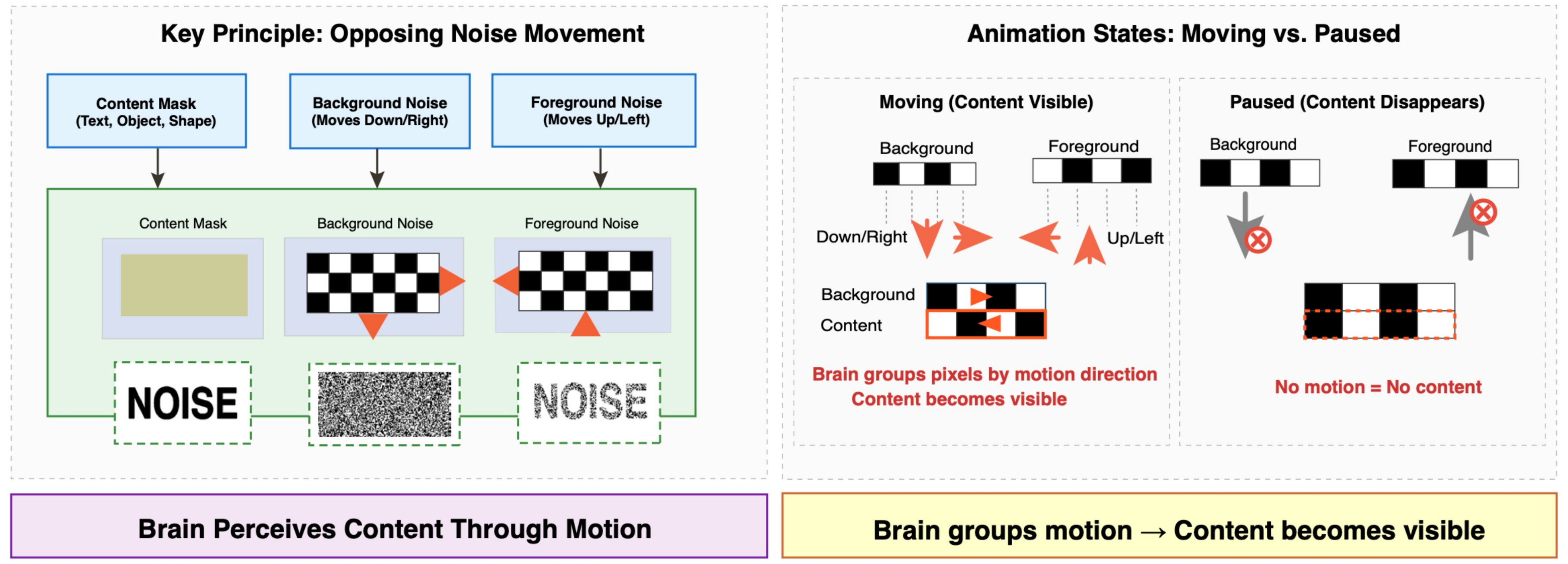
OBJECT IMAGES · 40.8%
SAM2 silhouettes of generated objects. **184 videos.**



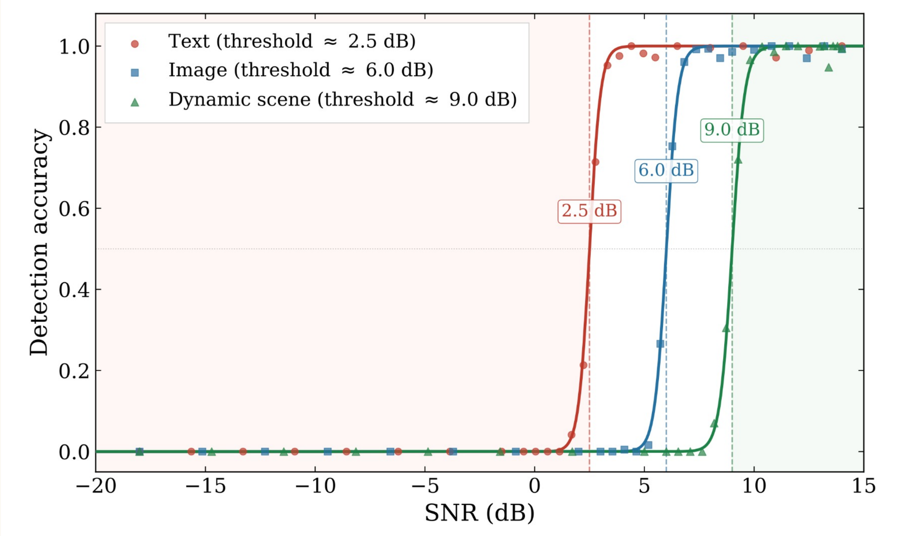
DYNAMIC SCENES · 12.6%
Depth maps from LaSOT / OTB2015. **57 videos.**

Each frame is noise. Content emerges only when played — content-mask animation for text/objects, threshold-gated motion for depth scenes.

06 HOW THE SIGNAL HIDES IN TIME



Opposing motion is the trick. A content mask drives foreground noise up and background noise down. The human visual system groups co-moving pixels (Gestalt *common fate*) and the content emerges — only while the video plays. Pause it, every frame looks like static.



SNR THRESHOLD · STEP FUNCTION

Detection is a step function. Below ~2.5 dB SNR, accuracy is ~0%; above, it jumps to ≥85%. Comparable thresholds: images at 6.0 dB, dynamic scenes at 9.0 dB.

07 THE 0% WALL

MODEL	PARAMS	DIRECT	COT
Human	—	98.0 ± 0.6	—
OPEN-SOURCE VIDEO-VLMS			
VideoLLaMA3 / TimeChat	7B	0.0	0.0
InternVL2.5	78B	0.0	0.0
InternVideo2.5	8B	0.0	0.0
Qwen2.5-VL	72B	0.0	0.0
Qwen3-VL	8B	0.0	0.0
CLOSED-SOURCE FRONTIER			
GPT-4o	—	0.0	0.0
Gemini 2.5 Pro / 2.0 Flash	—	0.0	0.0

Full table (25+ models) in paper. Failure is invariant to size, family, and prompting strategy.

08 IT'S ARCHITECTURAL — NOT DATA, NOT PROMPTING

- **Fine-tuning fails.** InternVL2.5-8B & Qwen2-VL-7B on 400 videos × 30 epochs → still 0%.
- **VJEPA-2 & DINOv3 can't even** *overfit* "is there content?" — loss saturates at chance.
- **Explicit motion cues unlock it.** Overlaying Farneback motion boundaries lifts Qwen2-VL-7B → **51.5%** and GPT-4o → **59.1%.**

Diagnosis: Today's VLMs lack a mechanism for *frame differencing / temporal integration*. The information is computationally extractable — it just never reaches the language model.

09 TAKE-AWAYS

- **"Time-blindness"** is a robust, architecture-level failure of every modern Video-VLM.
- Future models need **dedicated temporal pathways** — distributed timing, motion-energy estimation, motion-based figure-ground.
- SpookyBench + generator released to catalyze this direction.

10 WHAT UNLOCKS IT

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

Confirms the failure is *architectural*: temporal info is computationally extractable but VLMs can't reach it without spatial hand-holding. Dynamic Scenes still hard (≤3.5%).