

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,†}
¹KAUST ²VILA Lab, MBZUAI ³DocPanel Technologies

Project page: <https://timeblindness.github.io/>

Abstract

*Recent advances in vision-language models (VLMs) have made impressive strides in understanding spatio-temporal relationships in videos. However, when spatial information is obscured, these models struggle to capture purely temporal patterns. We introduce **SpookyBench**, a benchmark where information is encoded solely in temporal sequences of noise-like frames, mirroring natural phenomena from biological signaling to covert communication. Interestingly, while humans can recognize shapes, text, and patterns in these sequences with over 98% accuracy, state-of-the-art VLMs achieve 0% accuracy. This highlights a critical limitation: an over-reliance on frame-level spatial features and an inability to extract meaning from temporal cues. Overcoming this limitation will require novel architectures or training paradigms that decouple spatial dependencies from temporal processing. Our systematic analysis shows that this issue persists across model scales and architectures. We release SpookyBench to catalyze research in temporal pattern recognition and bridge the gap between human and machine video understanding. Dataset and code are available on the project website.*

1. Introduction

Large multimodal models have revolutionized visual understanding in both images [5, 20, 26, 28, 51, 80] and videos [3, 54, 84, 85, 99]. Recent Video-Vision Language Models (Video-VLMs) demonstrate impressive capabilities across action recognition [38, 89, 100], visual question answering [4, 58, 61, 96, 102], dense captioning [13, 17, 18, 39, 63, 91, 93], and temporal grounding [15, 79, 92]. Yet even in tasks labeled as temporal, strong per-frame spatial cues often allow models to succeed without genuinely reasoning over time [10, 25, 46]. This paper introduces **SpookyBench**, a benchmark that closes this loophole. Every stimulus is constructed so that individual frames appear as noise; the signal exists only in the

temporal relationship between frames. Existing temporal benchmarks [10, 44, 47, 95] still entangle spatial and temporal cues, making it difficult to diagnose where a model’s understanding breaks down. **SpookyBench** eliminates this confound entirely. In spirit it parallels ARC-AGI [23], which uses synthetic puzzles to isolate core reasoning; here, controlled stimuli isolate a single fundamental capability, extracting meaning from change over time.

Current approaches to video understanding [59, 75] typically follow a hierarchical paradigm: extract frame-level features using ViTs [6, 29, 65], integrate these features temporally, and fuse them with language for downstream tasks [45, 81, 88, 98]. This paradigm has yielded significant advances in general video understanding [31, 43, 59, 75]. However, our findings reveal a critical blind spot: when information exists purely in the temporal domain without reliable frame-level features, state-of-the-art models fail catastrophically (Figure 1). The inability to decode temporal patterns has significant implications for real-world applications. In nature, organisms such as fireflies communicate through precise temporal sequences of bioluminescence [11, 60, 66], encoding information exclusively through timing rather than spatial arrangements. Similarly, in medical domains, preictal EEG patterns for epileptic seizure prediction emerge only through temporal analysis [73]. These natural examples demonstrate how temporal patterns can carry rich information even when individual observations contain minimal static content. Similarly, various human technologies from Morse code to digital communication protocols rely on temporal encoding, yet current Video-VLMs lack the fundamental mechanisms to process such information.

The human visual system has evolved mechanisms for processing temporal information without relying solely on spatial cues [1, 37, 76]. Neuroscience research has revealed that temporal processing is distributed across neural structures rather than centralized in a single area [56], and the brain uses intrinsic network dynamics to perform temporal computations [62]. Areas such as the parietal cortex integrate temporal information along with spatial and numeric magnitudes [8]. Our experiments confirm humans’ remarkable temporal perception: participants achieve over 98% accuracy

*Equal contribution; Author ordering determined by a coin toss.

†Corresponding authors

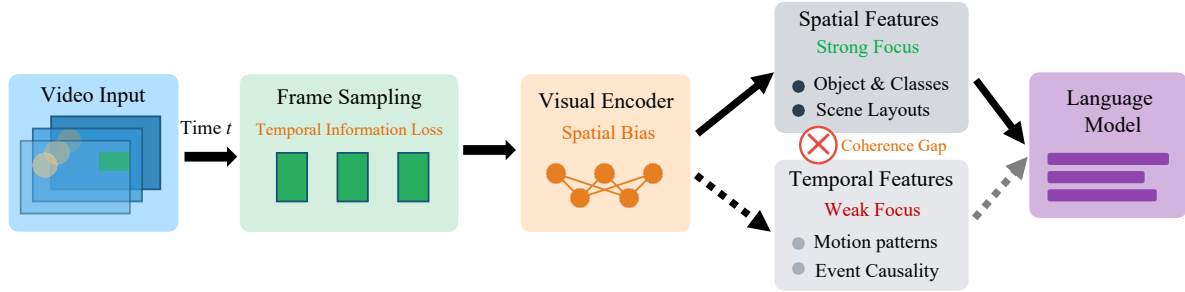


Figure 1. Illustration of the current video-language models’ limitations: over reliance on spatial visual features within individual frame. Frame sampling results in temporal information loss, while the visual encoder exhibits a strong spatial bias. This creates a coherence gap (×) between well-represented spatial features (objects, scene layouts) and under-represented temporal features (motion patterns, causality), limiting video understanding capabilities.

on **SpookyBench** tasks without training. In stark contrast, our evaluation of 15 state-of-the-art Video-VLMs, including closed-source commercial systems such as GPT-4o [36], and Gemini 2.0 Flash [27], reveals near-zero accuracy on these same tasks.

This performance gap persists across model architectures, parameter scales, and pre-training strategies, from compact systems like VideoLLaMA3-2B [97] to large-scale ones such as GPT-4o [36] and Qwen-VL [80], and extends to models designed specifically for video understanding including LongVLM [85], LLaVA-NeXT-Interleave [45], and InternVideo2.5 [84]. Efforts to enhance temporal reasoning through specialized temporal modeling [64, 68, 83] and fine-grained temporal localization [15, 53, 79] have not addressed the challenge of extracting meaning from purely temporal patterns without reliable spatial features.

Our findings suggest that achieving human-like video understanding requires rethinking how neural architectures process temporal information. Rather than treating temporal integration as secondary to spatial feature extraction, future models may need dedicated mechanisms for temporal pattern recognition, drawing inspiration from cognitive neuroscience research on distributed neural timing mechanisms [56, 62] and specialized brain regions for temporal processing [8, 57]. The gap between human and machine performance on **SpookyBench** indicates that current architectures remain “time-blind” despite strong performance on standard benchmarks. We release SpookyBench to catalyze research into temporal reasoning in Video-VLMs, with applications ranging from medical diagnostics to autonomous systems that must interpret temporal cues in complex environments.

2. Related Work

2.1. Temporal Reasoning in Video-VLMs

Video-VLMs have advanced through several architectural families, including LLaVA variants [43, 45, 51, 52, 54, 99], the Qwen series [5, 80], InternVL models [20, 21, 84], dual encoders [55], interleaved tokens [3, 103], compression tech-

niques [70], and multimodal fusion [81, 88, 98]. All of these exhibit limited temporal reasoning, with hallucinations [44], grounding failures [79], and reliance on linguistic shortcuts [40] persisting across action recognition [38, 89, 100], question answering [4, 58, 96], and captioning [13, 39, 93]. Targeted interventions such as timestamp-aware encoding [68], segment-level reasoning [64], direct token processing [53], temporal separation tokens [15], specialized temporal streams [79, 83], and synthetic training paradigms [59, 75, 96, 99] have not closed this gap, and dedicated video architectures like VideoGPT+ [55], TimeChat [68], LinVT [35], LongVLM [85], and Baichuan-Omni [49] extract spatial features first and treat temporal integration as an afterthought.

These limitations are further corroborated by temporal understanding benchmarks; TemporalBench [10] reveals a significant gap between model and human performance, TVBench [25], VITATECS [47], and Fateh et al. [34] confirm that many datasets reward spatial analysis over temporal reasoning, and targeted evaluations surface specific failures in temporal hallucinations [44], streaming video reasoning [95], and temporal location, object tracking, and anomaly detection [48]. Across these analyses, models such as LLaVA-Video [99], Video-ChatGPT [54], TemporalVLM [34], and VidChain [42] exploit spatial shortcuts instead of temporal reasoning [15, 40, 44, 79]. SpookyBench addresses this by obscuring spatial information and forcing models to extract meaning from temporal dynamics alone, exposing a “time-blindness” that conventional benchmarks leave undetected.

2.2. Neuroscience Insights on Temporal Processing

Human perception of temporally-defined objects relies on the Gestalt principle of common fate [86], whereby coherently moving elements are grouped together, enabling figure-ground segregation from motion alone [37]. Tangemann et al. [76] showed that a cortical motion energy model [71] matches human performance on random dot segmentation while 40 deep optical flow models fail, and relatedly, recent works [9, 12] encode motion into structured noise for video diffusion control. Beyond these perceptual grouping mecha-

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

Table 1. Benchmark results comparing model performance on **SpookyBench** across different prompting strategies along with model size. Human accuracy (98.0%) is the weighted average of accuracy across 3 different categories.

nisms, neuroscience research offers insights for addressing temporal limitations in Video-VLMs. Mauk and Buonomano [56] established that temporal processing is distributed across neural structures through intrinsic circuit properties, contrasting with current Video-VLMs’ sequential spatial processing. The human brain processes time at multiple granularities, with the cerebellum handling millisecond-to-second timing [57], the parietal cortex integrating temporal, spatial and numerical magnitudes [8], and neural patterns encoding time through “population clocks” [62]. Distributed temporal representations that evolve over time [62, 87] offer an alternative to treating temporal integration as secondary. The performance gap on temporal tasks [10, 25, 44] and our SpookyBench findings demonstrate that current architectures lack mechanisms for processing purely temporal patterns, a natural capability in humans through neural systems representing time as intrinsic dynamics.

3. SpookyBench

We introduce **SpookyBench**, a novel synthetic dataset specifically designed to isolate and evaluate pure temporal understanding in video language models. The key innovation of our benchmark lies in its unique design: All mean-

ingful information is encoded exclusively in the temporal domain through dynamic patterns of texts, images and video depth maps, while individual frames contain only structured noise. Our dataset is fundamentally different from the existing datasets used for training, fine-tuning, and evaluation of video-VLMs. Many state-of-the-art video language models employ advanced techniques, such as dynamic resolution strategies [5, 20, 80], specialized temporal encoding methods [5, 68, 80], hierarchical token merging [84, 85], and joint video-motion training frameworks [14] to capture temporal dynamics. However, these methods still rely on spatial representations extracted from individual frames, which currently remain the only viable mechanism for inferring temporal information. In contrast, **SpookyBench** forces models to depend only on temporal cues, thereby creating the first benchmark that exclusively evaluates a model’s ability to process and understand pure temporal information.

3.1. Dataset Generation

Figure 2 shows our proposed data generation framework. The dataset consists of specially designed videos that encode three types of content - words, images, and videos - using binary noise patterns with specific motion properties. In this approach, content is embedded within noise patterns such that individual frames appear as random noise, while the content becomes perceptible only when viewed as a temporal sequence. Our dataset encodes different types of content (Figure 3) through temporal noise animations in the following categories: **1) Words:** Text rendered as masks in which the background noise and foreground noise move in opposite directions, making the text visible only through temporal movement. **2) Images:** Binary masks generated using SAM2 [67] from single-object images generated using text-to-image model Flux [41], encoded using the same content mask animation approach as words. **3) Dynamic Scenes:** Depth maps extracted from videos in single-object tracking datasets LaSOT [32] and OTB2015 [90] using Video Depth Anything [16]. These are encoded using a technique in which pixels above a brightness threshold move while others remain static as shown in the algorithm 2.

3.2. Temporal Encoding Framework

Our temporal encoding framework implements two distinct motion configurations as detailed in Algorithms 1 and 2. For words, and image masks (Algorithm 1), we employ opposing motion patterns between foreground and background. The content is first converted to a binary mask M where $M(x, y) = 1$ represents foreground pixels and $M(x, y) = 0$ represents background. We generate two separate noise patterns N_{bg} and N_{fg} consisting of random binary values (0 or 255). During animation, foreground pixels sample from N_{fg} with a positive offset that increases with time $(y + vt \bmod h)$, while background pixels sample from N_{bg} with a negative

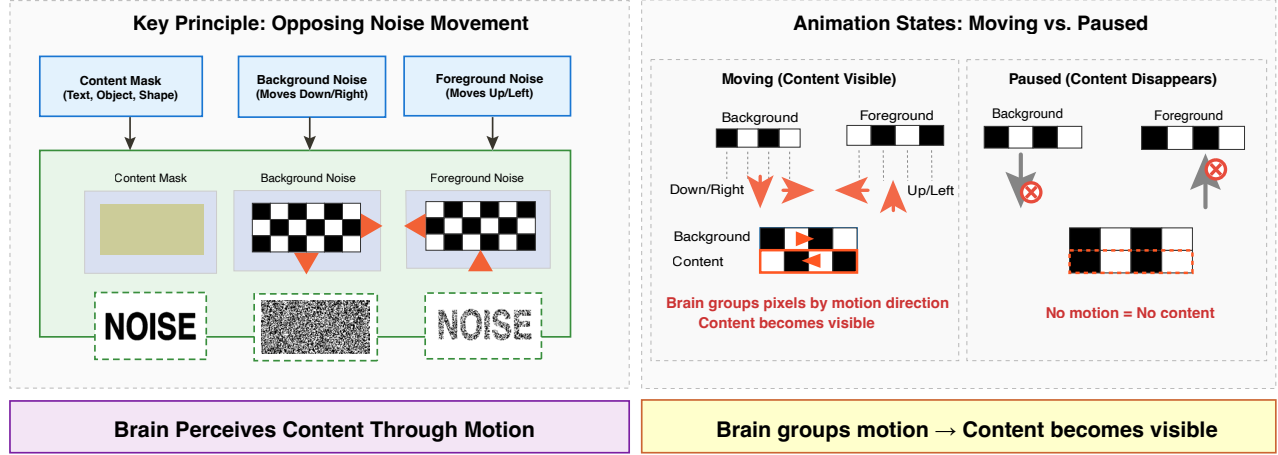


Figure 2. Illustration of the temporal encoding framework used in SpookyBench. **Left:** Core mechanism showing how content becomes visible through opposing motion patterns. A content mask defines regions where foreground noise (moving up/left) and background noise (moving down/right) are applied. When animated, the human visual system groups pixels with similar motion, causing the content to emerge. **Right:** Comparison between moving and paused states, demonstrating how content is only perceptible during animation and disappears when static, as individual frames contain only structured noise.

Category	Basic SNR (dB)	Perceptual SNR	Temporal Coherence	Motion Contrast
Images	-46.95 ± 2.40	-47.28 ± 2.28	8.00 ± 2.08	7.17 ± 5.00
Dynamic Scenes	-48.95 ± 3.64	-63.43 ± 5.74	21.91 ± 5.76	-3.18 ± 10.17
Text	-39.27 ± 1.58	-49.18 ± 3.31	7.84 ± 0.65	8.26 ± 6.44

Table 2. Signal-to-Noise Ratio (SNR) metrics across SpookyBench categories.

offset ($y - vt \bmod h$). This creates the perception of opposing motion within and outside the masked regions. For video depth maps (Algorithm 2), we employ a threshold-based approach. Using depth maps D extracted from videos, pixels with brightness values between lower and upper thresholds ($t_l \leq d \leq t_u$) are animated by sampling a noise pattern N with a time-varying offset ($y + vt \bmod h$), while pixels outside this range remain static. This creates the illusion that brighter regions (typically foreground objects) are moving while darker regions (typically background) remain static. The noise patterns are generated using binary values (0 or 255) in square blocks of variable size. We used different speckle sizes ranging from 1×1 to 3×3 pixels to investigate the effect of noise granularity on perception. For each speckle size, we also varied the noise density - the probability that a block is white versus black - using values of 10%, 30%, 50%, and 90%. These noise patterns arranged in pixel blocks create optimal perceptual conditions for human viewers while remaining challenging for vision language models. To ensure seamless animation, the noise patterns are made tileable by copying edge pixels to the opposite boundaries. All videos maintain consistent technical specifications: 960×540 pixel resolution, with an average duration of 7.11 seconds (ranging from 1.0 to 35.0 seconds) and an average

of 333.5 frames per video. Text videos have a consistent duration of around 4 seconds; however, videos of dynamic scenes are longer, ranging up to 35 seconds. Figure 2 illustrates the structure of the data set and the encoding patterns in categories. We used binary masks for the images using SAM2 [67]. For videos, depth maps are extracted using Depth Anything V2 [94] and Video Depth Anything [16] from the LaSOT [32] and OTB2015 [90] datasets.

3.3. Data Statistics

SpookyBench comprises 451 videos in three distinct categories, each requiring purely temporal reasoning for content identification. The dataset is distributed as follows: Text (46.6%, 210 videos), Object Images (40.8%, 184 videos) and Dynamic Scenes (12.6%, 57 videos). Although Dynamic Scenes comprise only 12.6% of videos, they account for 42.5% of total frames due to their longer duration, balancing the frame-level distribution across categories. This distribution ensures comprehensive coverage of different temporal perception challenges while maintaining a natural frequency distribution that reflects real-world scenarios. Additionally, more dataset can be generated indefinitely through the data generator on our project page, thus the dataset size is essentially unlimited. The “Text” category contains common English words rendered through temporal noise patterns, enabling evaluation of models’ ability to identify linguistic content through purely temporal cues. The “Object Images” category presents single objects extracted from high-quality images using segmentation techniques [67], encoded with the same temporal animation approach. It also contains a synthetic silhouette of simple objects generated using DALL-E 3 [7] and flux [41].



Figure 3. Noise generation process: (Top) masks applied for dynamic noise video generation, (Mid) word-specific mask, and (Bottom) depth map of video frame used for constructing noise-overlaid stimulus.

3.3.1. Analysis of Temporal Metrics

To ensure a rigorous quantification of the temporal information present in each video, we analyzed five key 2. SNR metrics that capture different aspects of the complexity and perceptibility of temporal patterns in SpookyBench, as shown in Table. These metrics provide insight into why temporal patterns might be visible to humans but challenging to detect by computational models.

Basic SNR measures signal-to-noise ratio in decibels:

$$\text{SNR}_B = 10 \log_{10} \left(\frac{P_S}{P_N} \right) \quad (1)$$

where $P_S = \mathbb{E}[\|\nabla \mathbf{F}\|^2]$ is motion boundary energy derived from spatial gradients of optical flow field $\mathbf{F}(x, y) = (F_x, F_y)$, and $P_N = \text{Var}(I_0)$ is variance of the static frame I_0 .

Perceptual SNR incorporates frequency-dependent visual sensitivity:

$$\text{SNR}_P = 10 \log_{10} \left(\frac{\|\mathcal{H}(B) \odot W\|^2}{\|\mathcal{H}(N) \odot W\|^2} \right) \quad (2)$$

where B is the average motion boundary strength, N is the static noise frame, $\mathcal{H}$ is the 2D Fourier transform, $\odot$ denotes element-wise multiplication, and $W(f) = f \cdot e^{-f/f_0}$ is the contrast sensitivity weighting function with peak $f_0 \approx 0.1$ cycles/pixel.

Temporal Coherence SNR quantifies motion consistency:

$$\text{SNR}_T = 10 \log_{10} \left(\frac{\text{Var}(C)}{\mathbb{E}[\text{Var}_{\text{local}}(C)]} \right) \quad (3)$$

where $C = e^{-\text{Var}_{\theta}(\mathbf{F})} \cdot \mathbb{1}(\|\mathbf{F}\| > \tau)$ is the directional coherence map, Var_{θ} computes circular variance of flow direction

angles over time, $\mathbb{1}$ is indicator function, τ is magnitude threshold, and $\text{Var}_{\text{local}}$ computes variance over small spatial neighborhoods.

Motion Contrast SNR measures foreground-background motion differentiation:

$$\text{SNR}_M = 10 \log_{10} \left(\frac{\|\mu_M - \mu_B\|^2}{\frac{1}{2}(\sigma_M^2 + \sigma_B^2)} \right) \quad (4)$$

where $\mu_M = \mathbb{E}[\mathbf{F} | M]$ and $\mu_B = \mathbb{E}[\mathbf{F} | \neg M]$ are mean flow vectors within mask region M and background region $\neg M$ respectively, $\sigma_M^2 = \mathbb{E}[\|\mathbf{F} - \mu_M\|^2 | M]$ and $\sigma_B^2 = \mathbb{E}[\|\mathbf{F} - \mu_B\|^2 | \neg M]$ are corresponding motion variances. The mask M is estimated from the motion boundaries.

The distribution of these metrics reveals why current vision models struggle with **SpookyBench**: they lack mechanisms to leverage temporal coherence (particularly high in Dynamic Scenes, 21.91 ± 5.76 dB) and motion contrast (negative for Dynamic Scenes, -2.20 and -3.18 dB), while text stimuli benefit from higher basic SNR (-39.27 ± 1.58 dB), explaining the observed performance gap.

3.3.2. Binary SNR Threshold Effect in Detection

Our analysis shows a binary threshold phenomenon in detecting text within dynamic noise videos. The words shows negligible detection ($\sim 0\%$) below 2.5dB SNR, but jumped to 85.7% accuracy above this threshold, displaying an abrupt rather than gradual transition as show in figure 5. Similar threshold behavior appears for images (inflection at 6.0 dB) and dynamic scenes (9.0 dB). Prompts performed best (40% accuracy), with Chain-of-Thought reasoning improving general identification tasks compared to direct prompting. This phenomenon parallels medical imaging diagnostics, where pathologies like microcalcifications in mammography become either entirely visible or invisible based on specific SNR thresholds [69]. The implications are significant: unlike perceptual phenomena that degrade gradually with noise, text detection functions as a step function, creating vulnerabilities in safety-critical applications. Just as radiologists cannot diagnose what remains invisible, language models cannot identify text below certain noise thresholds, leading to false certainties and potential catastrophic performance drops with minimal noise increases. This characteristic creates particular concerns for autonomous vehicles reading road signs or medical systems interpreting labels, while also exposing systems to adversarial attacks where slight SNR manipulations could render text completely undetectable.

4. Experiments

4.1. Experimental Setup

Models. We evaluate **SpookyBench** on both open source models (Video-LLaVA [99], LLaVA-NeXT-Video [45], TimeChat [68], InternVL2 [20], Qwen2-VL [80], Qwen2.5-VL [5] etc.) and closed source models (GPT-4o [36], Gemini

Algorithm 1 Content Mask Animation

```

1: Input: Content mask  $M$ , velocity  $v$ 
2: Output: Animated frame  $F_t$ 
3: Generate noise patterns  $N_{bg}, N_{fg}$ 
4: for each pixel  $(x, y)$  do
     $\triangleright$  Check pixel’s mask status
5:   if  $M(x, y)$  then
6:      $F_t(x, y) \leftarrow N_{fg}(x, y + vt \bmod h)$ 
     $\triangleright$  Foreground
7:   else
8:      $F_t(x, y) \leftarrow N_{bg}(x, y - vt \bmod h)$ 
     $\triangleright$  Background
9:   end if
10: end for

```

2.0 Flash [27], and Gemini 1.5 Pro [77]. We design different prompts for each category. All the prompts are included in the Appendix. All prompts instruct models to respond with only 1-5 words identifying the content. We input sequences of multiple video frames simultaneously for models that do not directly support video input.

Setup. We evaluate model performance using exact match accuracy. For Text, each video has a single correct label y_i . For *Object Images* and *Dynamic Scenes*, we define a set of acceptable labels $Y_i = \{y_{i1}, y_{i2}, \dots, y_{in}\}$ to account for semantic ambiguity (e.g., a video showing “a man playing basketball” accepts “playing basketball”, “man”, “human”, or “woman playing basketball”). Accuracy is computed as $\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(r_i \in L_i)$, where $L_i = y_i$ for Text and $L_i = Y_i$ for the other categories. We also verified all responses manually and with LLM-as-a-judge evaluation, a standard practice in VLM evaluation pipelines [30]; both yield identical 0% performance (details in Appendix). Despite this flexible protocol, no model produced responses matching any acceptable option.

4.2. Human Evaluation

To evaluate human performance against our benchmark, we designed and conducted a controlled experiment involving human participants. We recruited a total of six human participants for this study, each independently evaluating all videos. Participants were instructed to view each video carefully and subsequently record their responses on an anonymized website in the following structured form: 1) **Perceptibility Rating (1-5):** Participants rated how perceptible the presented word, shape, or object was, ranging from 1 (very difficult to perceive) to 5 (very clearly perceptible). This measure provided insights into the clarity and ease of visual grouping. 2) **Words/Objects Identification:** Participants typed out exactly what they identified in the video. This response directly tested the accuracy of their visual perception.

Algorithm 2 Video Depth Map Animation

```

1: Input: Depth map  $D$ , thresholds  $(t_l, t_u)$ , velocity  $v$ 
2: Output: Animated frame  $F_t$ 
3: Generate noise pattern  $N$ 
4: for each pixel  $(x, y)$  do
5:    $d \leftarrow$  brightness from  $D(x, y)$ 
6:   if  $t_l \leq d \leq t_u$  then
7:      $F_t(x, y) \leftarrow N(x, y + vt \bmod h)$ 
     $\triangleright$  Moving noise
8:   else
9:      $F_t(x, y) \leftarrow N(x, y)$ 
     $\triangleright$  Static noise
10:  end if
11: end for

```

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

Table 3. Human evaluation results showing accuracy and perceptibility ratings across different visual categories in **SpookyBench**.

We collect and evaluate participant responses using exact match criteria based on our predefined labels. Similar to the evaluation accuracy of the video language models for the categories of Object Images and Dynamic Scenes, we accepted multiple correct responses to avoid ambiguity. Table 3 shows the average precision and the perception rating of different annotators for different categories. The results show high human performance across all categories: participants correctly identified Words with 98% accuracy, while Object Images had 92% accuracy. We also observe a very high perceptibility rating (4.8 for texts and 4.3 and 4.0 for Object images and Dynamic scenes, respectively) across all three categories. This shows that the human brain can easily extract coherent information in videos, which seems to be very difficult for video language models.

4.3. Impact of Frame Rates

To examine whether temporal sampling affects performance, we evaluate both humans and VLMs across frame rates from 1 to 30 FPS. Three human participants tested 120 randomly sampled videos (40 per category) at 1, 5, 10, 20, and 30 FPS, while four VLMs (Qwen2-VL-7B, Qwen2.5-VL-7B, Qwen2.5-VL-3B, and GPT-4o) were evaluate using identical temporal downsampling. As shown in Tables 4 and 6, human accuracy remains above 95% at 20-30 FPS, degrades to 59.4% at 10 FPS, and drops to 0% at 1 FPS. In contrast,

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

Table 4. Human accuracy (%) across different content categories at varying frame rates. Results are averaged across 3 participants on 120 videos (40 per category).

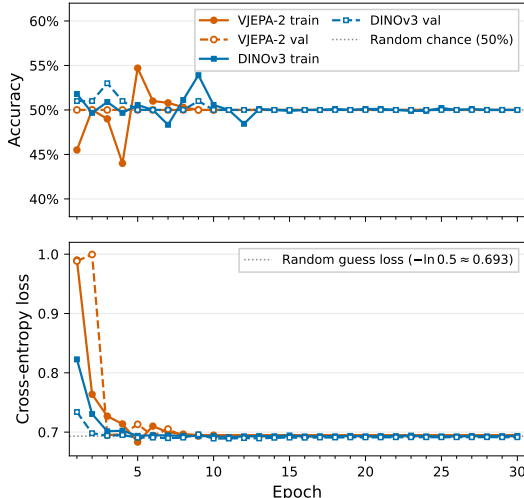


Figure 4. Training and validation loss over all batch steps for a random binary classifier tasked with detecting foreground noise in video clips. The loss oscillates around 0.7, consistent with random prediction on a balanced dataset.

all VLMs achieved 0% accuracy across all frame rates. This demonstrates that temporal sampling frequency does not explain the performance gap between humans and current video-language models, indicating that VLMs lack the architectural mechanisms to process information conveyed through temporal patterns regardless of temporal resolution.

4.4. Impact of Finetuning

To investigate whether the performance gap stems from out-of-distribution data rather than architectural limitations, we finetuned two state-of-the-art video-language models on SpookyBench: InternVL2.5-8B and Qwen2-VL-7B. Both models were trained on 400 SpookyBench videos for up to 30 epochs using LlamaFactory [101]. Despite this targeted training on the exact task and data distribution, both models maintained 0% accuracy on the test set. This result demonstrates that the failure to decode temporal patterns is not attributable to domain mismatch or insufficient exposure to the task, but rather indicates a fundamental architectural inability to process information conveyed purely through motion without relying on spatial content. We refer the reader

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%

Table 5. Performance comparison with and without motion boundary augmentation. Pre-computing and highlighting motion boundaries dramatically improves accuracy, demonstrating that temporal information is computationally extractable but inaccessible to current VLM architectures.

to the supplementary material for detailed results.

Can SOTA Visual Models Tell There’s an Object? We finetuned two video classification models, VJEPa-2[2] and DINOv3[72], on the task of predicting foreground noise presence. After 30 epochs, training loss saturated near chance-level accuracy (50%) on the validation set (Figure 4), indicating that neither model could learn a discriminative representation from the frame-level features. The inability to overfit confirms that temporal patterns in this data are inaccessible to architectures operating on individual frames.

4.5. Impact of Motion Boundary Augmentation

To further validate our hypothesis that VLM failure stems from an inability to extract temporal patterns rather than task impossibility or OOD, we pre-computed motion boundaries using classical optical flow (Farneback method [33]) and overlaid them onto the noisy frames before feeding them to models. *This converts implicit temporal information into explicit spatial cues.* As shown in Table 5, Qwen2-VL-7B jumps from 0% to 51.54% overall (50.95% on text, 70.51% on images), and GPT-4o reaches 59.10%. The substantial performance gain when temporal information is converted to spatial representations provides computational evidence that: (1) the benchmark tasks are solvable when appropriate preprocessing is applied, (2) current VLM architectures can process motion information if presented spatially, and (3) the fundamental limitation lies in the absence of temporal integration mechanisms that perform frame differencing and motion extraction, operations that classical computer vision methods accomplish routinely.

The Videos category shows minimal improvement (1.75%), consistent with our optical flow analysis in the supplementary material: complex articulated motion in dynamic scenes produces fragmented boundaries that remain hard to interpret even when spatially highlighted.

5. Results and Discussion

Table 1 presents the accuracy scores on the **SpookyBench** benchmark. Human participants achieved 98% accuracy under all test conditions. In contrast, all Video-VLMs scored 0% regardless of the type, size, or origin of the model. This pattern was held across all three task categories in our bench-

Model	Qwen2-VL-7B	Qwen2.5-VL-7B	Qwen2.5-VL-3B	GPT-4o
Accuracy (%)	0.0	0.0	0.0	0.0

Table 6. VLM accuracy (%) averaged across all tested frame rates (1-30 FPS).

mark: temporal symbol recognition, temporal sequence understanding, and temporal pattern reasoning. We test two different prompting strategies to determine if performance limitations could be overcome through interface modifications. First, we used direct prompts with basic instructions asking the models to identify content in the videos. Next, we implemented chain-of-thought prompts with explicit guidance to focus on temporal patterns rather than individual frames. As shown in Table 1, none of these approaches yielded improvements. All models maintained 0% accuracy across all prompting conditions, suggesting that the limitation is inherent in the model architectures rather than a matter of optimization or prompt design. Examination of model output revealed consistent failure modes when processing **SpookyBench** videos.

Across all models tested, we observed attempts to extract information from individual frames rather than temporal patterns. When explicitly prompted to consider temporal changes, the models acknowledged the instruction but still failed to identify the patterns. **For instance, Qwen-series models consistently predict “clock” or “coffee cup” regardless of input, indicating pattern collapse rather than meaningful temporal processing.** Fine-tuned models produced outputs that mimicked training examples without correctly identifying test patterns. In particular, specialized temporal models like TimeChat [68], which were specifically designed for fine-grained temporal understanding, failed at the same rate as general-purpose models. This suggests that the limitation extends beyond general Video-VLMs to models explicitly optimized for temporal tasks.

5.1. Architectural Implications for Vision Models.

Distinctive signal profiles in **SpookyBench** demonstrate a fundamental gap between human and machine perception of temporal information. Current vision models struggle with **SpookyBench** stimuli primarily because they: (1) lack robust temporal integration mechanisms that could leverage high temporal coherence, (2) process information primarily through spatial rather than temporal channels, and (3) fail to perform motion-based figure-ground segregation effectively. The consistently high temporal coherence values in Dynamic Scenes, coupled with their poor static-frame metrics, suggest that successful models must implement recurrent processing or attention mechanisms that operate across extended temporal windows rather than focusing on frame-level feature extraction. The negative motion contrast observed in Dynamic Scenes further indicates that models require more so-

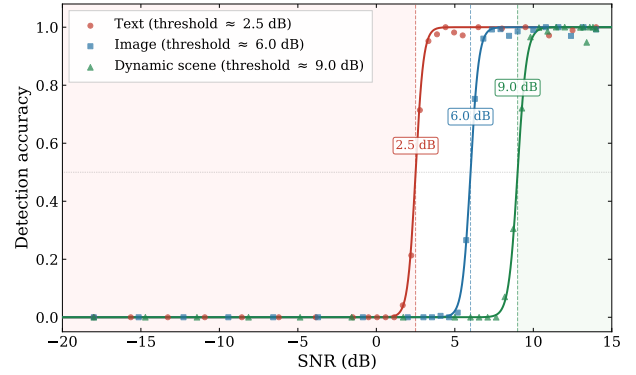


Figure 5. Analysis of effects of SNR on detecting text, image and dynamic scenes with direct prompting and chain of thought prompting.

phisticated motion segregation capabilities to match human perceptual abilities in dynamic visual environments. These findings highlight the need for architectural innovations that specifically address temporal processing limitations. Future models should incorporate dedicated temporal coherence pathways, motion contrast analysis, and longer temporal integration windows to bridge the perception gap demonstrated by **SpookyBench**.

6. Conclusion

In this paper, we introduced **SpookyBench**, a novel benchmark designed to evaluate the temporal reasoning capabilities of video-language models by isolating temporal understanding from spatial comprehension. Our experiments revealed a striking performance gap: while humans effortlessly achieve 98% accuracy on tasks requiring pure temporal pattern recognition, all tested models, including state-of-the-art open and closed-source systems, fail completely with 0% accuracy. This consistent failure across different model architectures, scales, and prompting strategies highlights a fundamental limitation in current video understanding approaches, which typically process spatial features first and then establish temporal connections, rather than integrating spatio-temporal information simultaneously. Importantly, our additional analyses with foreground noise detection and providing explicit motion cues to the model further validate that the deficit arises not from task impossibility but from architectural inability to internally extract temporal structure. The benchmark effectively exposes the *time blindness* of current architectures that remain hidden in conventional evaluation settings where spatial features can provide shortcuts to correct answers. We hope that **SpookyBench** will inspire the development of next-generation temporal-connected models.

References

- [1] Valtteri Arstila. Theories of apparent motion. *Phenomenology and the Cognitive Sciences*, 15(3):337–358, 2016. 1
- [2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. 7
- [3] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024. 1, 2, 3
- [4] Hammad Ayyubi, Junzhang Liu, Ali Asgarov, Zaber Ibn Abdul Hakim, Najibul Haque Sarker, Zhecan Wang, Chia-Wei Tang, Hani Alomari, Md Atabuzzaman, Xudong Lin, et al. Enter: Event based interpretable reasoning for videoqa. *arXiv preprint arXiv:2501.14194*, 2025. 1, 2
- [5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 3, 5
- [6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 1
- [7] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023. 4
- [8] Domenica Bueti and Vincent Walsh. The parietal cortex and the representation of time, space, number and other magnitudes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1525):1831–1840, 2009. 1, 2, 3
- [9] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13–23, 2025. 2
- [10] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024. 1, 2, 3
- [11] Albert D Carlson and Jonathan Copeland. Flash communication in fireflies. *The Quarterly review of biology*, 60(4): 415–436, 1985. 1
- [12] Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. *arXiv preprint arXiv:2504.03072*, 2025. 2
- [13] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 1, 2, 3
- [14] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motion-llm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024. 3
- [15] Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024. 1, 2
- [16] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv preprint arXiv:2501.12375*, 2025. 3, 4
- [17] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 1
- [18] Xinlong Chen, Yuanxing Zhang, Chongling Rao, Yushuo Guan, Jiaheng Liu, Fuzheng Zhang, Chengru Song, Qiang Liu, Di Zhang, and Tieniu Tan. Vidcapbench: A comprehensive benchmark of video captioning for controllable text-to-video generation. *arXiv preprint arXiv:2502.12782*, 2025. 1
- [19] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3
- [20] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 1, 2, 3, 5
- [21] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2
- [22] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 3
- [23] François Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. Arc prize 2025: Technical report, 2026. 1
- [24] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blis-

- tein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3
- [25] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024. 1, 2, 3
- [26] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeibi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024. 1
- [27] Google DeepMind. Gemini flash, 2025. Accessed: 2025-02-24. 2, 3, 6
- [28] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1
- [29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1
- [30] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024. 6
- [31] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1
- [32] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 3, 4
- [33] Gunnar Farneback. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003. 7
- [34] Fawad Javed Fateh, Umer Ahmed, Hamza Khan, M Zeeshan Zia, and Quoc-Huy Tran. Video llms for temporal reasoning in long videos. *arXiv preprint arXiv:2412.02930*, 2024. 2
- [35] Lishuai Gao, Yujie Zhong, Yingsen Zeng, Haoxian Tan, Dengjie Li, and Zheng Zhao. Linvt: Empower your image-level large language model to understand videos. *arXiv preprint arXiv:2412.05185*, 2024. 2
- [36] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 3, 5
- [37] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2):201–211, 1973. 1, 2
- [38] Kumara Kahatapitiya, Anurag Arnab, Arsha Nagrani, and Michael S Ryoo. Victr: Video-conditioned text representations for activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18547–18558, 2024. 1, 2
- [39] Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. Do you remember? dense video captioning with cross-modal memory retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13894–13904, 2024. 1, 2
- [40] Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. Large language models are temporal and causal reasoners for video question answering. *arXiv preprint arXiv:2310.15747*, 2023. 2
- [41] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 3, 4
- [42] Ji Soo Lee, Jongha Kim, Jeehye Na, Jinyoung Park, and Hyunwoo J Kim. Vidchain: Chain-of-tasks with metric-based direct preference optimization for dense video captioning. *arXiv preprint arXiv:2501.06761*, 2025. 2
- [43] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2
- [44] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. *arXiv preprint arXiv:2412.03735*, 2024. 1, 2, 3
- [45] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 1, 2, 3, 5
- [46] Kunchang Li, Yali Wang, Yanan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 1
- [47] Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024. 1, 2
- [48] Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*, 2024. 2
- [49] Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, et al. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2(3), 2024. 2
- [50] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeibi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF confer-*

- ence on computer vision and pattern recognition, pages 26689–26699, 2024. 3
- [51] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2
- [52] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2
- [53] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 2
- [54] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1, 2, 3
- [55] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024. 2, 3
- [56] Michael D Mauk and Dean V Buonomano. The neural basis of temporal processing. *Annual review of neuroscience*, 27: 307–340, 2004. 1, 2, 3
- [57] Hugo Merchant, Deborah L Harrington, and Warren H Meck. Neural basis of the perception and estimation of time. *Annual review of neuroscience*, 36:313–336, 2013. 2, 3
- [58] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245, 2024. 1, 2
- [59] Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. Video-language understanding: A survey from model architecture, model training, and data perspectives. *arXiv preprint arXiv:2406.05615*, 2024. 1, 2
- [60] Avalon CS Owens, Mira Van den Broeck, Raphaël De Cock, and Sara M Lewis. Behavioral responses of bioluminescent fireflies to artificial light at night. *Frontiers in Ecology and Evolution*, 10:946640, 2022. 1
- [61] Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryoo, Donghyun Kim, and Michael S Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa. *arXiv preprint arXiv:2406.09396*, 2024. 1
- [62] Joseph J Paton and Dean V Buonomano. The neural basis of timing: distributed mechanisms for diverse functions. *Neuron*, 98(4):687–705, 2018. 1, 2, 3
- [63] Iqra Qasim, Alexander Horsch, and Dilip Prasad. Dense video captioning: A survey of techniques, datasets and evaluation protocols. *ACM Computing Surveys*, 57(6):1–36, 2025. 1
- [64] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*, 2024. 2
- [65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. Pmlr, 2021. 1
- [66] GM Ramírez-Ávila, J Kurths, and Jean-Louis Deneubourg. Fireflies: A paradigm in synchronization. *Chaotic, Fractional, and Complex Dynamics: New Insights and Perspectives*, pages 35–64, 2018. 1
- [67] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 4
- [68] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 2, 3, 5, 8
- [69] Berkman Sahiner, Heang-Ping Chan, Lubomir M Hadjiiski, Mark A Helvie, Jun Wei, Chuan Zhou, and Yao Lu. Computer-aided detection of clustered microcalcifications in digital breast tomosynthesis: a 3d approach. *Medical physics*, 39(1):28–39, 2012. 5
- [70] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 2
- [71] Eero P Simoncelli and David J Heeger. A model of neuronal responses in visual area mt. *Vision research*, 38(5):743–761, 1998. 2
- [72] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. 7
- [73] Itaf Ben Slimen, Larbi Boubchir, and Hassene Seddik. Epileptic seizure prediction based on eeg spikes detection of ictal-preictal states. *Journal of biomedical research*, 34(3): 162, 2020. 1
- [74] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 3
- [75] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023. 1, 2
- [76] Matthias Tangemann, Matthias Kümmerer, and Matthias Bethge. Object segmentation from common fate: Motion

energy processing enables human-like zero-shot generalization to random dot stimuli. *Advances in Neural Information Processing Systems*, 37:137135–137160, 2024. 1, 2

- [77] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 3, 6
- [78] Qwen Team. Qwen3 technical report, 2025. 3
- [79] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024. 1, 2
- [80] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 3, 5
- [81] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 1, 2
- [82] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 3
- [83] Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. Videollm knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format. *arXiv preprint arXiv:2411.17991*, 2024. 2
- [84] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2.5: Empowering video mlms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 1, 2, 3
- [85] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024. 1, 2, 3
- [86] Max Wertheimer. *On perceived motion and figural organization*. Mit Press, 2012. 2
- [87] Marc Wittmann. The experience of time: neural mechanisms and the interplay of emotion, cognition and embodiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1525):1809–1813, 2009. 3
- [88] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 1, 2
- [89] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2847–2855, 2023. 1, 2
- [90] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015. 3, 4
- [91] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pillava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 1
- [92] Yifang Xu, Yunzhuo Sun, Zien Xie, Benxiang Zhai, and Sidan Du. Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt. *Applied Sciences*, 14(5):1894, 2024. 1
- [93] Antoine Yang, Arsha Nagrai, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023. 1, 2
- [94] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2025. 4
- [95] Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding, 2025. 1, 2
- [96] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023. 1, 2
- [97] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2, 3
- [98] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufer, Nina Wenzel, Forrest Huang, Dhruvi Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mml. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024. 1, 2
- [99] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1, 2, 3, 5
- [100] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 1, 2
- [101] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. 7
- [102] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answer-

ing: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022. [1](#)

- [103] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [2](#)