

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

## Supplementary Material

### Contents

<b>A Data Statistics</b>	<b>2</b>
<b>B Prompt Design for Evaluation</b>	<b>2</b>
B.1. Prompt Design Principles	2
B.2. Direct vs. Chain-of-Thought Prompting	2
B.3. Prompt Effectiveness Analysis	2
<b>C LLM-as-a-Judge Evaluation</b>	<b>2</b>
C.1. Evaluation Pipeline	3
C.2. Results	3
<b>D Binary Classifier Experiment Details</b>	<b>3</b>
D.1. Dataset Construction	3
D.2. Experiment Details	3
D.3. Results	5
<b>E Overfitting Experiment with Varying Dataset Sizes</b>	<b>6</b>
E.1. Setup	6
E.2. Additional Data Details	6
E.3. Analysis	6
<b>F. Impact of FPS</b>	<b>6</b>
<b>G Temporal Motion Coherence Analysis</b>	<b>6</b>
G.1. Motion-Based Perception in Noisy Environments	6
G.2. Signal-to-Noise Ratio Analysis	7
<b>H Qualitative Examples of Model Responses</b>	<b>8</b>
H.1. Direct Prompting Responses	8
H.2. Chain-of-Thought Prompting Responses	8
H.3. Analysis of Failure Modes	8
<b>I. Additional Images</b>	<b>8</b>

## A. Data Statistics

Figure 1 shows the data distribution of SpookyBench.

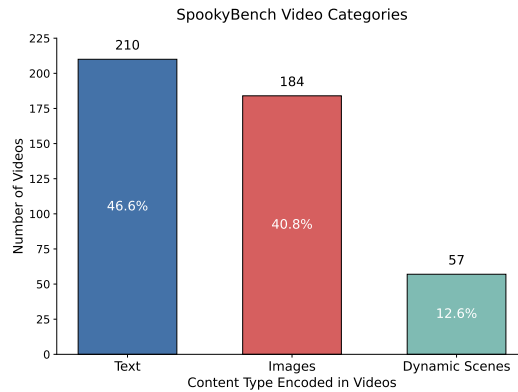


Figure 1. Distribution of the **SpookyBench** dataset across three video categories. Each category represents a different type of content encoded through temporal noise patterns: *Text*, *Object Images*, and *Dynamic Scenes*.

## B. Prompt Design for Evaluation

Prompt design significantly affects the performance of vision-language models [3, 4]. We performed careful prompt engineering to ensure fair and comprehensive evaluation, developing a systematic prompting methodology that builds on established principles while introducing elements specific to temporal pattern recognition.

### B.1. Prompt Design Principles

We designed our prompts based on three key principles:

1. **Specificity**: Each prompt explicitly states that the content is encoded through temporal patterns to direct attention to motion-based cues rather than static frame analysis.
2. **Category targeting**: We created specialized prompts for each content category (text, objects, dynamic scenes) to account for the different perceptual mechanisms involved in each.
3. **Constrained response format**: All prompts request brief, specific answers (1-3 words) to ensure objective evaluation and minimize the influence of language generation capabilities.

### B.2. Direct vs. Chain-of-Thought Prompting

We implemented two distinct prompting strategies shown below. Figure 2 and 3 present our category-specific prompts for both strategies.

- **Direct prompts** test immediate pattern recognition without explicit guidance, similar to how humans naturally perceive temporal patterns without conscious step-by-step processing.
- **Chain-of-Thought (CoT) prompts** provide explicit steps to guide attention and processing, testing whether models could benefit from structured reasoning about temporal patterns.

### B.3. Prompt Effectiveness Analysis

Neither prompt strategy improved model performance on SpookyBench. All tested models achieved 0% accuracy regardless of prompt type, indicating a fundamental architectural limitation rather than a prompt engineering issue. The complete ineffectiveness of even carefully engineered prompts across all tested models further supports our claim that current video-language models lack the mechanisms needed for processing purely temporal patterns.

## C. LLM-as-a-Judge Evaluation

To verify that the observed 0% accuracy is not an artifact of our exact-match evaluation protocol, we additionally employ an LLM-as-a-judge approach, a standard evaluation methodology widely adopted in VLM benchmarks [2]. Specifically,

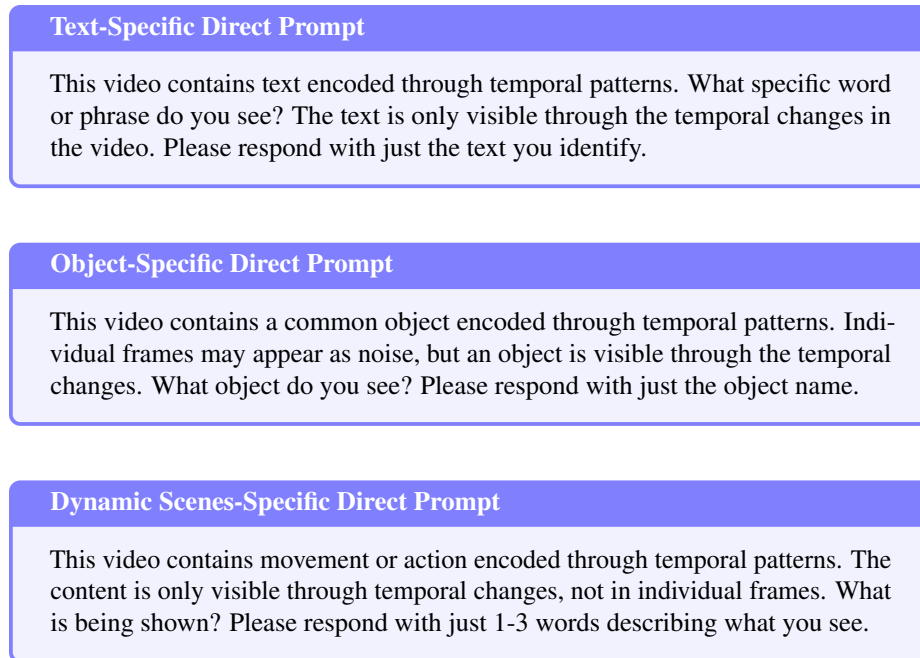


Figure 2. Category-specific direct prompts for the SpookyBench benchmark. These prompts test immediate pattern recognition without step-by-step guidance.

we prompt an LLM to assess whether each model’s output is consistent with the ground-truth label, rather than requiring a verbatim match. This supplementary evaluation addresses the possibility that models may exhibit partial understanding of temporal content yet express it in a form that diverges from the predefined label space.

### C.1. Evaluation Pipeline

Figure 4 illustrates our LLM-as-a-judge evaluation pipeline. For each video, we provide the judge model (GPT-4o) with: (1) the ground-truth label(s), (2) the model’s response, and (3) instructions to determine whether the response correctly identifies the content, allowing for paraphrases, synonyms, and partial matches.

### C.2. Results

The LLM-as-a-judge evaluation yields identical 0% accuracy across all models and categories. This confirms that the failure is not due to strict string matching: models do not produce responses that are even semantically close to the correct answers. Instead, they hallucinate unrelated content (see Appendix H for examples), confirming that the evaluation protocol is robust.

## D. Binary Classifier Experiment Details

To test whether state-of-the-art visual models can detect even the *presence* of foreground objects in SpookyBench videos (without identifying them), we trained binary classifiers using VJEPa-2 [1] and DINOv3 [5].

### D.1. Dataset Construction

We constructed a balanced binary classification dataset as follows:

- **Positive class:** SpookyBench videos containing encoded foreground content (text, objects, or dynamic scenes).
- **Negative class:** Videos generated with the same noise generation pipeline but without any foreground mask applied, i.e., uniform random noise with no embedded temporal pattern.

The dataset was split into 80% training and 20% validation sets, maintaining class balance in both splits.

### D.2. Experiment Details

To evaluate whether pretrained video and image representation models can discriminate foreground noise presence from raw video, we finetune two recent self-supervised models on our binary classification task: VJEPa-2 (ViT-L/16, 64 frames per

### Text-Specific CoT Prompt

This video encodes text through temporal patterns. Let's think step by step to identify it:

1. Observe how noise pixels move differently across regions of the frame
2. Look for areas where opposing motion patterns create visible boundaries forming letters
3. Focus on the overall word or phrase that emerges from these motion boundaries
4. Read the specific text content revealed by the temporal dynamics

Please respond with just the text you identify.

### Object-Specific CoT Prompt

This video encodes an object through temporal patterns. Let's think step by step to identify it:

1. Notice that individual frames appear as random noise
2. Track how groups of pixels move coherently over time in different directions
3. Look for areas where motion patterns reveal object contours and silhouette
4. Focus on the overall form that emerges and determine what specific object is represented

Please respond with just the object name.

### Dynamic Scenes-Specific CoT Prompt

This video encodes movement through temporal patterns. Let's think step by step to identify it:

1. Observe that each frame looks like noise, but regions move differently over time
2. Look for areas where temporal changes reveal motion of a foreground object
3. Focus on the action or activity that emerges from the separation of moving and static regions
4. Identify the specific movement or object in motion

Please respond with just 1-3 words describing what you see.

Figure 3. Category-specific chain-of-thought prompts for the SpookyBench benchmark. These prompts provide explicit step-by-step guidance to test structured temporal reasoning.

clip) [1] and DINOv3 (ConvNeXt-S, pretrained on LV1689M) [5]. For both DINOv3 and VJEPa-2, we load their respective models with a randomly initialized attention pooler and classifier head, and finetuned the full model on 64-frame clips at  $256 \times 256$  resolution. Both models were trained for 100 epochs using cosine-annealed learning rates (DINOv3:  $10^{-3}$  with Adam; VJEPa-2:  $10^{-4}$  with AdamW, weight decay 0.01) on a stratified 90/10 train/validation split of the 1,000-video dataset (500 per class). Training was performed on a single NVIDIA RTX PRO 6000 GPU per model (batch sizes of 64 and 4, respectively). Neither model learned a discriminative representation. DINOv3 training loss converged to  $\sim 0.689$  with validation accuracy fluctuating between 48–52%, barely above the 50% chance baseline. VJEPa-2 exhibited even stronger signs of failure: validation accuracy remained at exactly 50% for the majority of training, with the loss plateauing near  $\ln 2 \approx 0.693$  — the theoretical maximum-entropy baseline for balanced binary classification.

These results confirm that the temporal signal encoding foreground noise presence in our stimuli is inaccessible to architectures that operate on per-frame spatial features, even when those features are aggregated temporally (DINOv3 mean

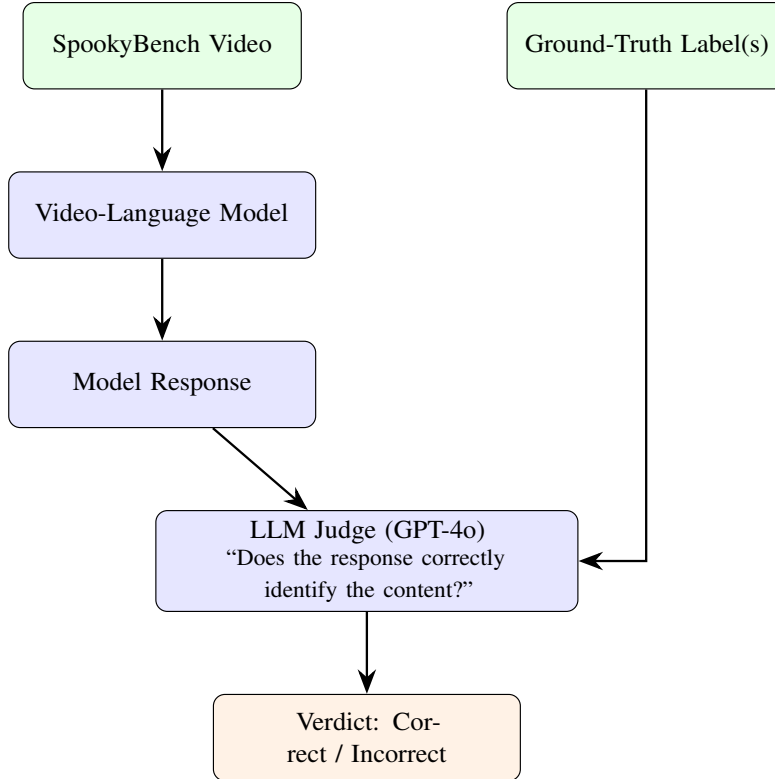


Figure 4. LLM-as-a-Judge evaluation pipeline. The judge receives both the ground-truth label and the model’s response, and determines whether the response correctly identifies the temporal content, allowing for synonyms and paraphrases.

pooling) or processed by a dedicated spatiotemporal encoder (VJEPA-2). The foreground and background noise patterns are visually indistinguishable in any single frame; discrimination requires detecting coherent differential motion across frames, a signal that neither frozen per-frame representations nor end-to-end video encoders could extract from the raw pixel input under these training conditions.

### D.3. Results

Both models converge to approximately 50% validation accuracy (random chance on a balanced dataset), as shown in Figure 4 of the main paper. VJEPA-2 reached 52.8% and DINOv3 reached 53.2% after 30 epochs, with training loss oscillating rather than decreasing. This confirms that the features extracted by these models do not encode the temporal motion patterns present in SpookyBench videos, even at the coarse level of detecting whether a foreground object exists. We also fine-tune Qwen3-VL-8B on the same binary classification task, which similarly achieved 0% meaningful accuracy, further confirming that VLM architectures cannot extract these temporal patterns even with supervised signal. Table 1 reports the training accuracy after 80 epochs across varying dataset sizes. VJEPA-2 and DINOv3 hover near chance level regardless of the number of training videos, while Qwen3-VL-8B fails entirely at 0%.

Model	100 videos	400 videos	1,100 videos
VJEPA-2	52.3%	56.1%	52.8%
DINOv3	52.7%	52.9%	53.2%

Table 1. Training accuracy after 80 epochs for binary classification (foreground presence detection) across different dataset sizes. All models remain near or below chance level.

## E. Overfitting Experiment with Varying Dataset Sizes

A natural question is whether the 0% accuracy of fine-tuned VLMs is simply due to insufficient training data. To address this, we conducted finetuning experiments with varying dataset sizes.

### E.1. Setup

We fine-tuned InternVL2.5-8B and Qwen2-VL-7B using LlamaFactory [6] on three dataset configurations:

Configuration	Training Videos	Epochs	Test Accuracy
Small	100	30	0%
Medium	400	30	0%
Large	1,100	30	0%

Table 2. Fine-tuning results across different dataset sizes. All configurations yield 0% test accuracy.

### E.2. Additional Data Details

For the large configuration (1,100 videos), we generated additional videos beyond the original 451-video SpookyBench dataset using our data generation pipeline. The additional videos follow the same distribution across categories and use the same noise generation parameters (speckle sizes, noise densities, velocity values) as the original dataset. The test set remained fixed across all configurations to ensure comparability.

### E.3. Analysis

The consistent 0% accuracy across all dataset sizes demonstrates that the failure cannot be attributed to insufficient training data. Even with 1,100 training videos and 30 epochs (providing ample opportunity for overfitting), the models cannot learn to extract temporal patterns from the data. This rules out a simple out-of-distribution explanation and points to an architectural inability to process the temporal information encoded in SpookyBench videos.

## F. Impact of FPS

To test whether frame rate affects the ability to perceive temporally encoded information, we evaluate both human participants and VLMs across frame rates from 1 to 30 FPS. We test three human participants on 120 randomly sampled videos (40 per category) at frame rates of 1, 5, 10, 20, and 30 FPS. For VLMs, we evaluated Qwen2-VL-7B, Qwen2.5-VL-7B, Qwen2.5-VL-3B, and GPT-4o. Human accuracy remains above 95% at 20–30 FPS, degrades to 59.4% at 10 FPS, and drops to 0% at 1 FPS (see Table 3 in the main paper). In contrast, all VLMs achieved 0% accuracy at every frame rate tested (Table 6 in the main paper). Human performance degrades gracefully as frame rate decreases, consistent with the known temporal resolution limits of motion perception. VLMs show no sensitivity to frame rate at all, confirming that they are not engaging with temporal information in any meaningful way, regardless of how much temporal data is provided. This rules out temporal undersampling as an explanation for the performance gap.

## G. Temporal Motion Coherence Analysis

Temporal coherence and motion boundaries provide powerful cues that enable the human visual system to extract shapes from noisy stimuli. We present a comprehensive analysis of these phenomena using our SpookyBench dataset.

### G.1. Motion-Based Perception in Noisy Environments

Even when individual frames have low SNR, temporal integration of motion information allows robust shape perception. Figure 5 shows the motion direction coherence map of an ant silhouette, revealing how consistent motion patterns across frames enable object identification despite significant noise. Figure 7 shows both the average motion boundary strength and its overlay on a noisy frame. Motion boundaries emerge clearly despite extreme noise levels in individual frames (measured at -49.07 dB basic SNR).

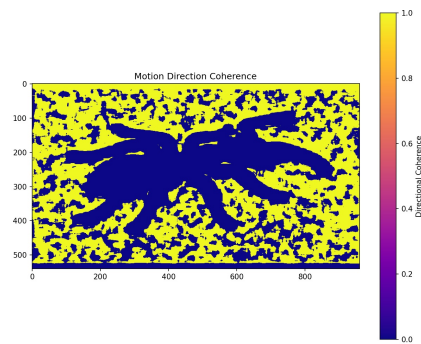
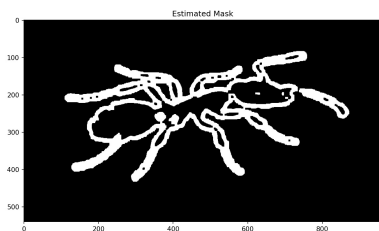
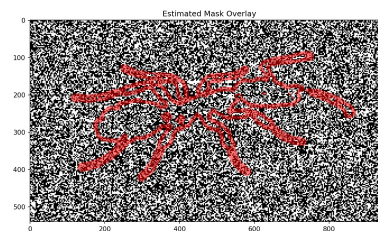


Figure 5. Motion Direction Coherence visualization for the ant silhouette video. Yellow regions (coherence value 1.0) indicate consistent motion direction across frames; blue regions (coherence value 0.0) represent the silhouette itself.

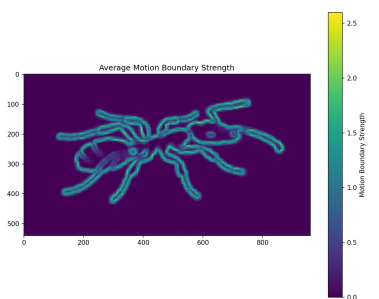


(a) Estimated object mask from temporal motion coherence.

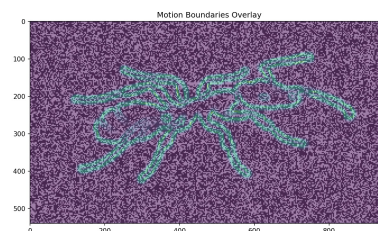


(b) Estimated mask overlay (red) on a single noise frame.

Figure 6. Shape extraction through temporal integration, demonstrating how object shape can be recovered from noisy video sequences.



(a) Average motion boundary strength across frames.



(b) Motion boundaries (teal) overlaid on a single noise frame.

Figure 7. Motion boundary analysis showing how temporal integration extracts object boundaries despite extremely noisy individual frames.

## G.2. Signal-to-Noise Ratio Analysis

Table 3 presents detailed SNR results for the ant silhouette video. While basic and perceptual SNR are extremely low (-49.07 dB and -55.02 dB), temporal coherence SNR and motion contrast SNR are positive (7.18 dB and 14.24 dB), showing

SNR Metric	Value (dB)
Basic SNR	-49.07
Perceptual SNR	-55.02
Temporal Coherence SNR	7.18
Motion Contrast SNR	14.24
Combined SNR	-20.61

Table 3. Signal-to-Noise Ratio Analysis for Ant Silhouette Video.

that temporal information provides signal enhancement that supports human perception. The contrast between negative frame-based SNR values and positive temporal SNR metrics directly supports our hypothesis: temporal integration is essential for perceiving content in SpookyBench stimuli, and current architectures lack this capability.

## H. Qualitative Examples of Model Responses

We present representative model responses to illustrate the failure modes observed across VLMs on SpookyBench. These examples show that models do not produce partially correct or semantically related answers; instead, they hallucinate entirely unrelated content.

### H.1. Direct Prompting Responses

Figure 8 presents four representative model responses to the text-specific direct prompt. These responses illustrate distinct failure modes, including timestamp hallucination, prompt echoing, outright refusal, and confident hallucination of unrelated content.

### H.2. Chain-of-Thought Prompting Responses

Figure 9 shows model responses when given explicit step-by-step reasoning guidance. Despite the structured CoT prompt, models generate elaborate but entirely fabricated scene descriptions that bear no relation to the actual video content.

### H.3. Analysis of Failure Modes

Across all tested models, we observe four distinct failure patterns:

1. **Timestamp hallucination:** Models output numerical timestamps instead of content, suggesting they attempt to process temporal metadata rather than temporal visual patterns.
2. **Prompt echoing:** Models repeat parts of the prompt as their answer, indicating no actual visual processing occurred.
3. **Confident hallucination:** Models produce specific but incorrect answers with high confidence. For instance, Qwen-series models consistently predict “clock” or “coffee cup” regardless of input, indicating pattern collapse to high-frequency training concepts.
4. **Scene fabrication:** When given CoT prompts, models generate elaborate descriptions of imaginary scenes (cats, buildings, paper textures) that have no correspondence to the actual video content, suggesting they are generating plausible-sounding narratives from their language prior rather than processing visual input.

None of these failure modes produce responses that are even partially related to the actual content encoded in the videos, which is consistent with the 0% accuracy observed under both exact-match and LLM-as-a-judge evaluation.

## I. Additional Images

We present additional images from the analysis of temporal motion coherence across all categories in SpookyBench. Figures 10, 11, 12, and 13 show motion boundaries, boundary overlays, estimated masks, and mask overlays for different examples in our dataset, demonstrating the varying effectiveness of temporal integration across different content types. In Figure 13, the temporal motion coherence method does not perform effectively for dynamic scenes. Since this category contains real-life videos with complex motion patterns, several factors reduce clarity:

1. **Distributed Motion Patterns:** Human movement involves multiple articulated body parts moving in different directions simultaneously, creating competing motion signals that fragment coherent boundaries.
2. **Non-rigid Deformation:** Dynamic content involves continuous shape changes throughout motion sequences, making consistent boundary extraction more challenging than for static objects.

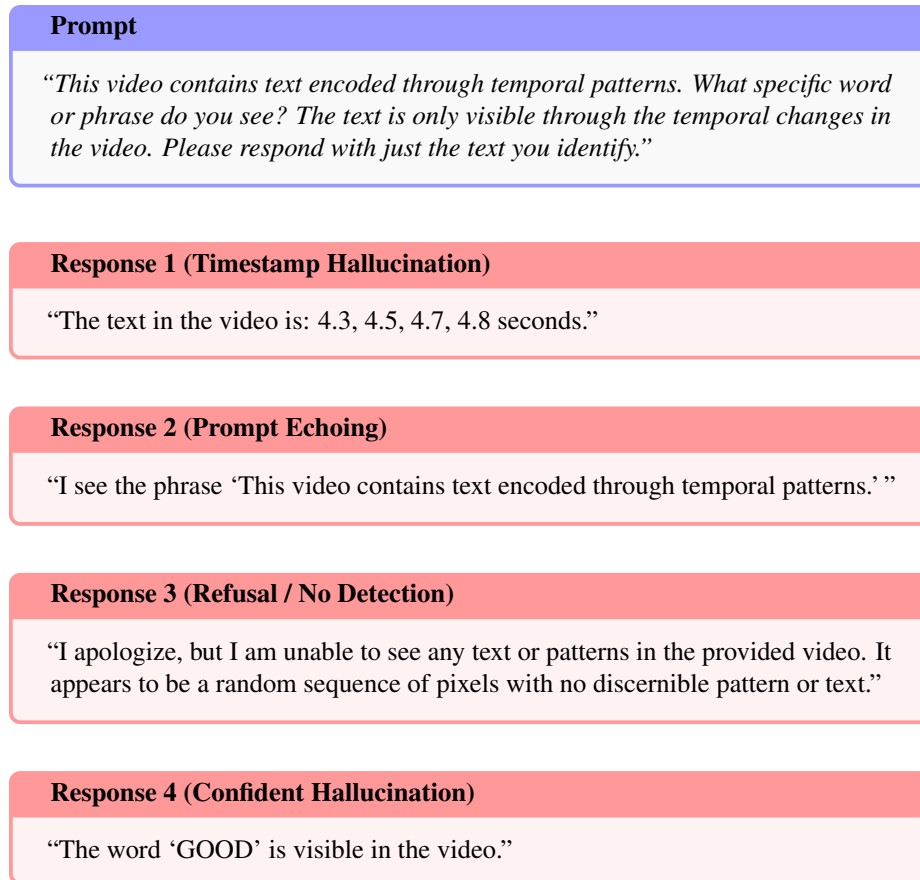


Figure 8. Direct prompting responses on SpookyBench. The prompt (top, blue) is shown together with four representative model responses (red), illustrating distinct failure modes.

3. **Complex Temporal Dynamics:** Mechanical motions in examples like Plane\_2 and Bicycle\_10 create temporal discontinuities that disrupt the motion coherence essential for shape perception.

### Prompt

*“This video encodes text through temporal patterns. Let’s think step by step to identify it: 1. Observe how noise pixels move differently across regions of the frame. 2. Look for areas where opposing motion patterns create visible boundaries forming letters. 3. Focus on the overall word or phrase that emerges from these motion boundaries. 4. Read the specific text content revealed by the temporal dynamics. Please respond with just the text you identify.”*

### CoT Response 1 (Scene Fabrication)

“In the video, a black and white photo of a cat is shown first. Next, the scene shifts to a black and white texture, followed by a black and white photo of a building. Then, a black and white photo of a cat is shown again. The scene changes to a black and white photo of a building and a black and white photo of a cat again. Finally, a black and white texture of a television screen appears. The video continues to show a black and white photo of a cat. Throughout the video, the dense captions describe the appearance of various objects in each scene, including trees, buildings, people, and various textures. The video is black and white, and the scenes appear to be sequenced in a linear fashion.”

### CoT Response 2 (Texture Description)

“First, the video shows a texture of an old sheet of paper in black and white. Then, it shows a series of lines and textures on a black and white background. Next, the video shows a black and white texture of old paper. After that, it shows a texture of a piece of paper on a black and white background. Then, it shows a black and white background of textured material. Finally, the video shows a black and white textured pattern. Throughout the video, there are also various objects, such as a dog, a chair, and a handwritten note, but they are not the main focus of the video.”

Figure 9. Chain-of-thought prompting responses on SpookyBench. The CoT prompt (top, green) is shown together with two representative model responses (red), illustrating elaborate but entirely fabricated descriptions.

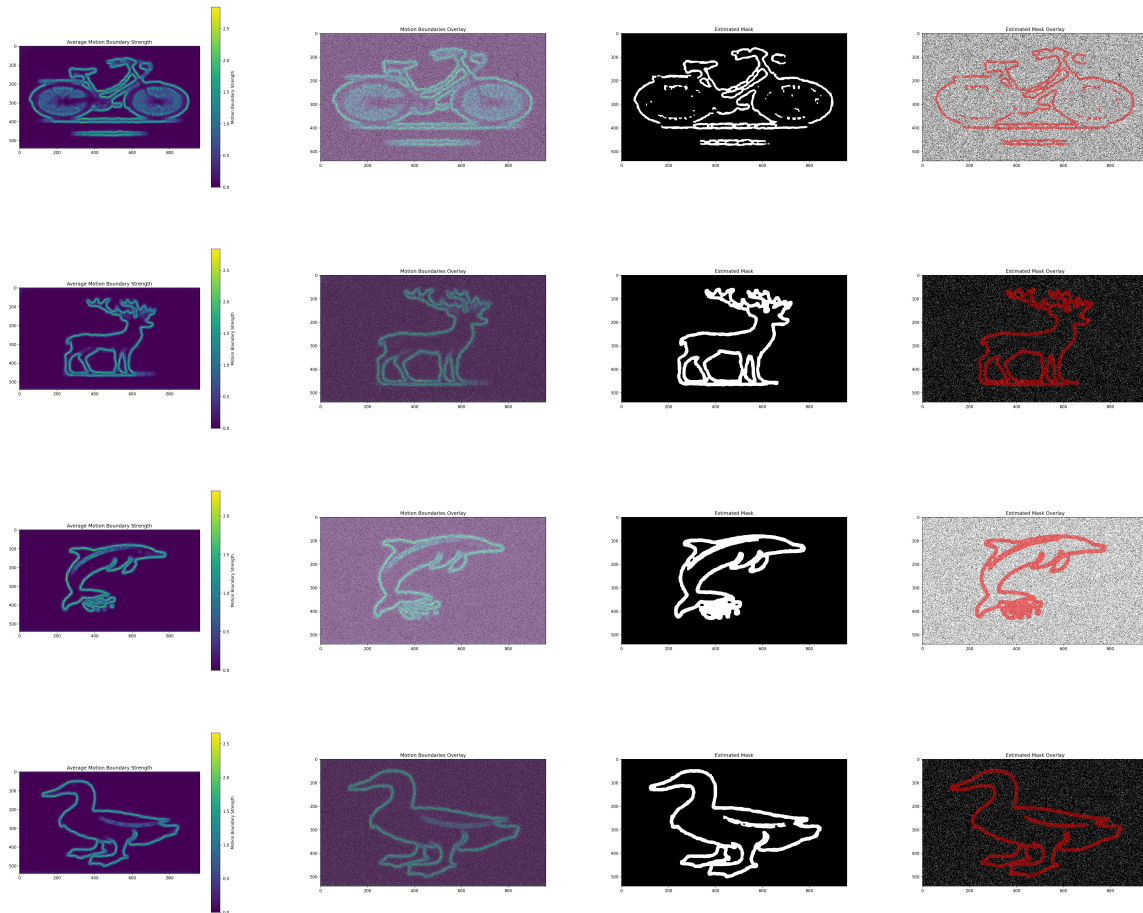


Figure 10. Temporal motion coherence analysis for Images category (Part 1). Each row shows motion boundaries, boundary overlay, estimated mask, and mask overlay for: Cycle, Deer, Dolphin, and Duck (top to bottom).

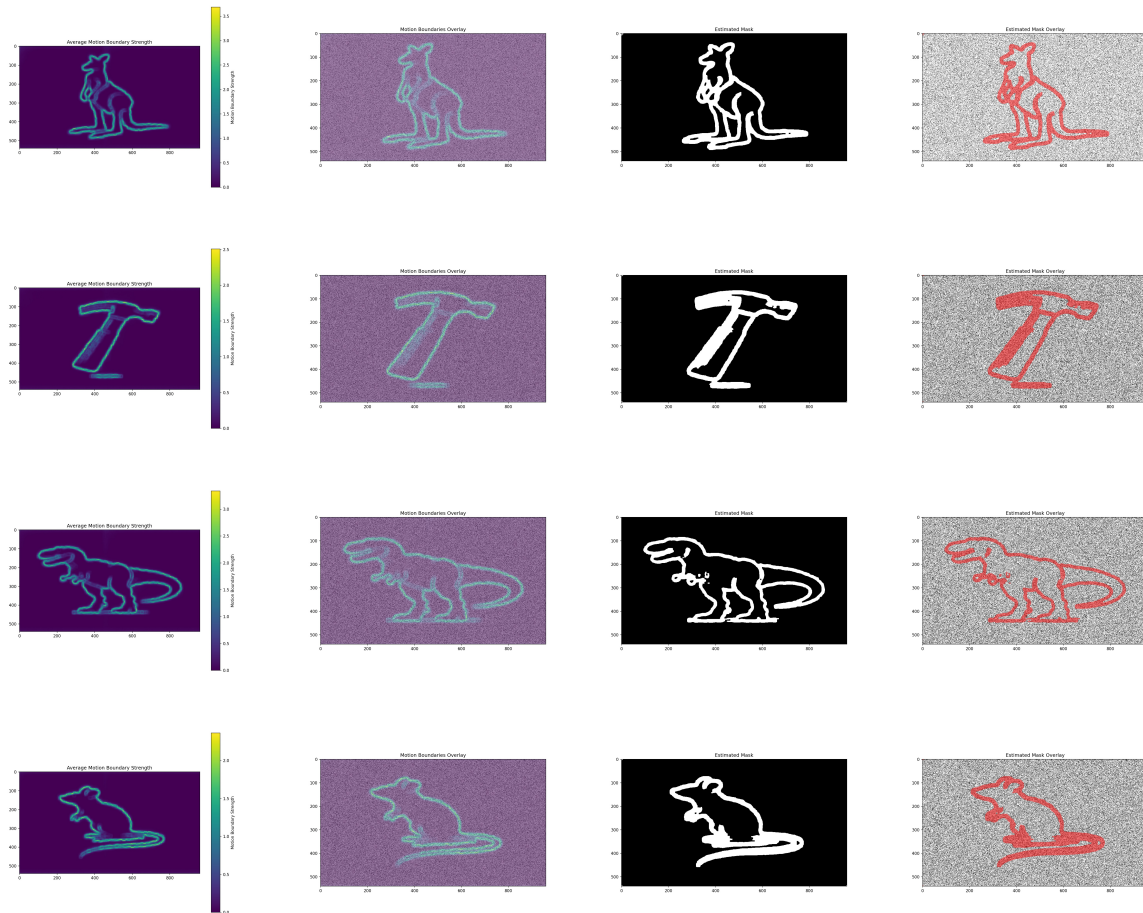


Figure 11. Temporal motion coherence analysis for Images category (Part 2). Each row shows motion boundaries, boundary overlay, estimated mask, and mask overlay for: Kangaroo, Hammer, T-Rex, and Mouse (top to bottom).



Figure 12. Temporal motion coherence analysis for Words category. Each row shows motion boundaries, boundary overlay, estimated mask, and mask overlay for: Gold, Laser Beams Cross, Ancient Olive Trees, and Artificial Minds Think (top to bottom).

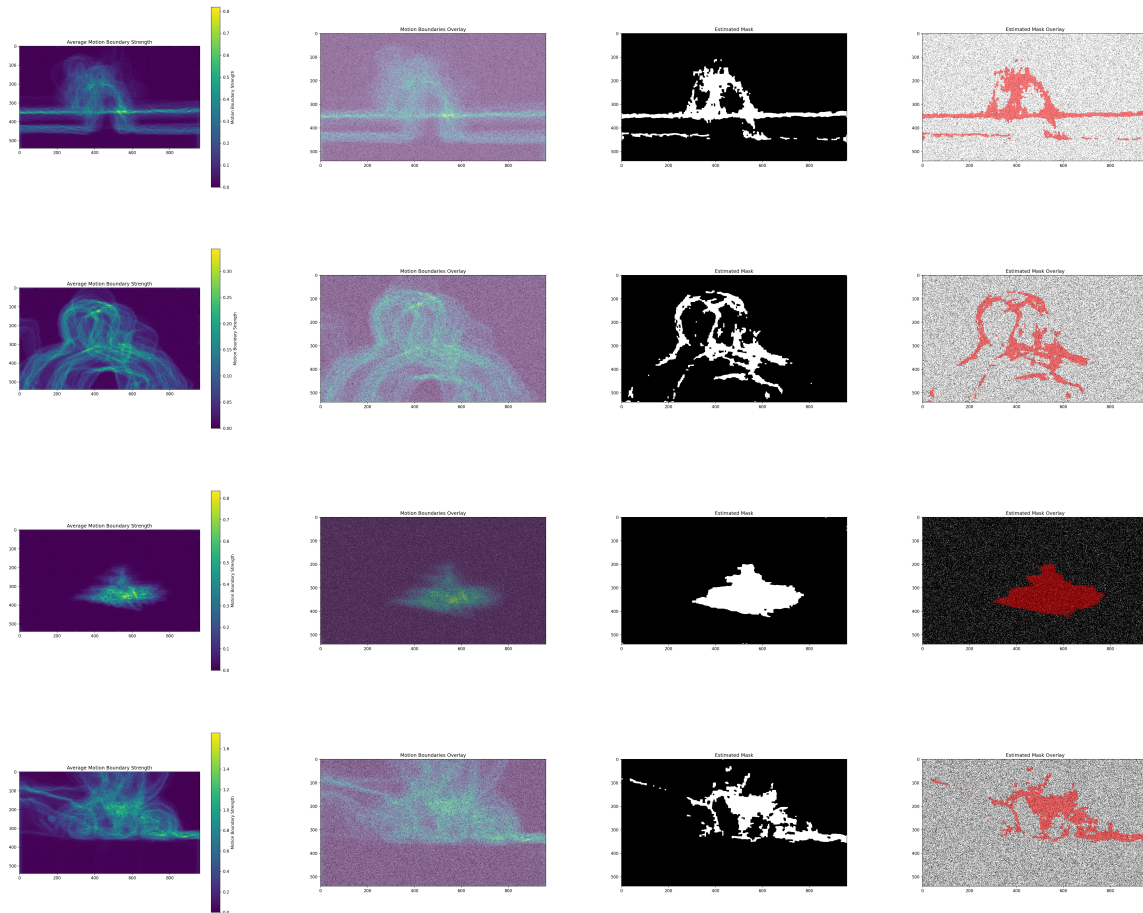


Figure 13. Temporal motion coherence analysis for Videos category. Each row shows motion boundaries, boundary overlay, estimated mask, and mask overlay for: Human\_6, Man\_1, Plane\_2, and Bicycle\_10 video sequence (top to bottom).

## References

- [1] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zhohus, 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. [3](#), [4](#)
- [2] 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. [2](#)
- [3] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023. [2](#)
- [4] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*, 2021. [2](#)
- [5] 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. [3](#), [4](#)
- [6] 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. [6](#)