

# Which Way Does Time Flow? A Psychophysics-Grounded Evaluation for Vision-Language Models

Shiho Matta<sup>1,\*</sup>, Lis K. Pereira<sup>2,3,4,\*</sup>, Peitao Han<sup>2,3,4</sup>, Fei Cheng<sup>1</sup>, Shigeru Kitazawa<sup>2,3,4</sup>

<sup>1</sup> Kyoto University, Japan

<sup>2</sup> Center for Information and Neural Networks, Japan

<sup>3</sup> National Institute of Information and Communications Technology, Japan

<sup>4</sup> The University of Osaka, Japan

\* matta@nlp.ist.i.kyoto-u.ac.jp, liskanashiro@nict.go.jp

## Abstract

Modern vision–language models (VLMs) excel at many multimodal tasks, yet their grasp of temporal information in video remains weak and, crucially, under-evaluated. We probe this gap with a deceptively simple but revealing challenge: judging the arrow of time (AoT)—whether a short clip is played forward or backward. We introduce **AoT-PsyPhyBENCH**, a psychophysically validated benchmark that tests whether VLMs can infer temporal direction in natural videos using the same stimuli and behavioral baselines established for humans. Our comprehensive evaluation of open-weight and proprietary, reasoning and non-reasoning VLMs reveals that most models perform near chance, and even the best lag far behind human accuracy on physically irreversible processes (e.g., free fall, diffusion/explosion) and causal manual actions (division/addition) that humans recognize almost instantly. These results highlight a fundamental gap in current multimodal systems: while they capture rich visual–semantic correlations, they lack the inductive biases required for temporal continuity and causal understanding. We release the code and data for **AoT-PsyPhyBENCH** to encourage further progress in the physical and temporal reasoning capabilities of VLMs.

## 1 Introduction

Vision–language models (VLMs) have recently achieved remarkable progress in multimodal understanding, including tasks such as video captioning, retrieval, and question answering (Bai et al., 2025; Qwen Team, 2024; Liu et al., 2024; Wang et al., 2024a; Xiao et al., 2021; Wu et al., 2024; Fu et al., 2025). Yet a fundamental question remains: do these models possess the implicit physical and causal assumptions, i.e. the *inductive biases*, that humans rely on to interpret the physical world? Among the most basic of such biases is the *arrow*

of time (AoT): the implicit assumption that events unfold irreversibly from past to future, constrained by gravity, entropy, and causality. Judging whether a short video plays forward or backward provides a minimal yet powerful diagnostic of such bias: a system that has internalized temporal irreversibility should recognize when physical regularities are violated, whereas a system that relies only on visual correlation will fail despite extensive training.

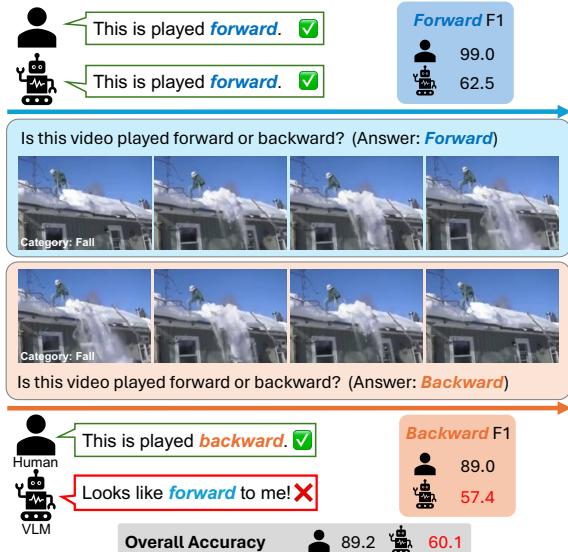


Figure 1: Overview of the arrow-of-time (AoT) task. Humans identify the AoT for both forward and backward playbacks with high accuracy; however, VLMs perform substantially worse and exhibit a label-prediction bias, preferring Forward (see Section 6.2).

To anchor this diagnostic in human performance, Hanyu et al. (2023) conducted a psychophysical study using 360 three-second natural video clips from everyday activities. They found a global forward bias: participants more often judged reversed clips as forward (39% errors on reversed vs. 9% on forward trials). However, humans detected reversals rapidly and almost flawlessly in five specific categories: **Fall** (free fall) and **Dif-**

**fusion** (diffusion/explosion), which involve physically irreversible processes governed by gravity and entropy; **Proceed** (forward locomotion); and **Division** (manual division) and **Put** (manual addition), which involve agent-driven causal sequences. In these cases, reversed playback becomes visibly implausible. For example, when a big chunk of snow ascends back onto the roof, defying gravity (Figure 1). These results demonstrate that humans exploit not only low-level motion cues but prior knowledge of physical regularities and causal structure, providing a psychophysically validated foundation for testing AoT in VLMs.

Building on this, we introduce **AoT-PsyPhyBENCH** (pronounced “AoT-*sci-fi bench*”), an evaluation framework that directly inherits validated stimuli and behavioral baselines from Hanyu et al. (2023). We systematically test modern VLMs, from open-weight to proprietary reasoning/non-reasoning models, on the same AoT judgment task. Our findings reveal that most models perform near chance even in “obvious” irreversible processes, with the best configuration lagging humans by approximately **29 percentage points**, suggesting that they lack the inductive biases underpinning human causal and temporal perception (Figure 1).

**Our contributions are threefold.** **(1) Comprehensive evaluation.** We present the first systematic evaluation of modern VLMs on psychophysically grounded AoT judgment across zero-shot, few-shot, chain-of-thought prompting, and supervised fine-tuning. Despite apparent task simplicity, the gains from elaborate prompting and fine-tuning are minimal and inconsistent, indicating that the bottleneck is temporal and physical understanding, not instruction following. **(2) Psychophysically validated benchmark.** We introduce **AoT-PsyPhyBENCH**, curated from Hanyu et al.’s psychophysical dataset by retaining only clips with clear human consensus on irreversible processes while excluding ambiguous cyclic motions. This yields a low-ambiguity benchmark enabling direct, reliable human–model comparison on temporal direction judgment. **(3) Open resources.** We release our benchmark, evaluation scripts, and model outputs to advance research on temporal reasoning and physical understanding in VLMs.<sup>1</sup>

<sup>1</sup>For anonymity, we release the dataset at <https://anonymous.4open.science/r/AoT-PsyPhyBENCH-222F> (anonymized repository) during the reviewing process. We will update this link in the camera-ready version.

## 2 Related Work

### 2.1 Vision–Language Models (VLMs)

Modern VLMs couple a vision encoder with a large language model (LLM) via a multimodal connector and are typically aligned through instruction tuning on image–/video–text data (Wang et al., 2024b; Qwen Team, 2024; Azzolini et al., 2025; Bai et al., 2025). Despite rapid progress, frontier VLMs often fail on video tasks that depend on event order (Xue et al., 2025).

We broadly categorized the VLMs used in our evaluation along two axes: **(i) proprietary vs. open-weight** and **(ii) reasoning vs. non-reasoning**. Proprietary models are accessible only via APIs with limited transparency, while open-weight models release checkpoints and inference scripts, as well as partial training details. Reasoning models are trained to generate multi-step deliberation (e.g., chain-of-thought reasoning) automatically before producing the final output, often with controllable *reasoning effort*. Non-reasoning models prioritise perceptual understanding (e.g., captioning, visual question-answering) and instruction following, without dedicated reasoning objectives; they typically respond directly unless explicitly prompted for step-by-step reasoning.

Building on this taxonomy, we evaluate the following models. **Proprietary non-reasoning:** GPT-4o and GPT-4.1 (OpenAI, 2024, 2025a). **Proprietary reasoning:** o3, o4-mini, and GPT-5 (OpenAI, 2025c,b); Gemini-2.5-Pro (Comanici et al., 2025). **Open-weight non-reasoning:** Qwen2-VL (Wang et al., 2024b) and Qwen2.5-VL (Bai et al., 2025). **Open-weight reasoning:** CosmosReason1 (Azzolini et al., 2025), which combines explicit AoT supervision and reinforcement learning on ~30k forward/reverse video clip pairs with reasoning traces; and QVQ-72B-Preview (Qwen Team, 2024).

### 2.2 Temporal Reasoning Benchmarks

We evaluate VLMs on a psychophysically validated AoT benchmark to address two key limitations in current temporal reasoning evaluation:

**(1) Lack of temporal dependency.** Several recent benchmarks claim to assess temporal reasoning in VLMs, but many do not *demonstrate* dependence on event order (Liu et al., 2024; Wang et al., 2024a; Fu et al., 2025; Xiao et al., 2021; Wu et al., 2024). Xue et al. (2025) evaluated strong VLMs on standard temporal benchmarks and found

a critical failure: models show minimal or no performance degradation when frames are *shuffled* or *reversed*, indicating these tasks can be solved through scene context without temporal understanding. In contrast, AoT judgment is inherently temporal—reversed videos violate physical laws, making temporal order task-critical rather than incidental.

**(2) Lack of psychophysical validation.** Existing AoT benchmarks simply reverse video datasets (Bagad et al., 2023; Wang et al., 2023; Du et al., 2024; Agarwal et al., 2025; Xue et al., 2025) without *controlled human baselines*, for instance, with quantitative measurements of accuracy, reaction times, and category-specific effects. This creates systematic evaluation problems: difficulty remains uncalibrated (videos ambiguous to humans become spurious failure cases), model–human performance gaps cannot be quantified, systematic biases (forward bias, category effects) go undetected, and bidirectionally plausible cases contaminate test sets. Psychophysical grounding is essential to distinguish genuine deficiencies from artifacts of ambiguous stimuli.

To address these limitations, we introduce **AoT-PsyPhyBENCH**, a psychophysically validated benchmark enabling direct human–model comparison on temporal direction judgment.

### 3 Experimental Settings

We evaluate the following models (described in Section 2) on AoT-PsyPhyBENCH: **Proprietary non-reasoning**: GPT-4o and GPT-4.1; **proprietary reasoning**: o3, o4-mini, GPT-5, and Gemini-2.5-Pro; **open-weight non-reasoning**: Qwen2-VL and Qwen2.5-VL; **open-weight reasoning**: Cosmos-Reason1-7B and QVQ-72B-Preview. We test these models across multiple experimental settings, which are detailed in the following sections.

#### 3.1 Zero-shot Settings

For the zero-shot experiments, we adopted a simple prompt design. In the system prompt, the model is asked to output only the F (forward) or B (backward) label as follows:

**System prompt** *You will see videos provided from the user, played either forward or backward. Finish your answer with F or B only. F for forward and B for backward.*

In the user prompt, we provide the sampled video frames with the following instruction:

**User prompt** *[Video frames] Detect whether the video plays forward or backward with confidence.*

#### 3.2 Few-shot Settings

For few-shot experiments, we constructed four demonstration examples from AoT-PsyPhyBENCH: two forward and two backward videos exhibiting visually distinctive temporal cues (e.g., explosions, falling objects). Each demonstration consists of sampled video frames, the user prompt from Section 3.1, and the correct label (F for forward, B for backward). These four demonstration videos were excluded from the test set to prevent data leakage.

#### 3.3 Reasoning Effort Ablation Settings

Reasoning-centric models are trained to generate internal reasoning chains automatically before producing outputs, while non-reasoning models usually respond directly unless explicitly prompted for step-by-step reasoning. We ablated the length of thinking of VLMs by either setting the reasoning effort parameter of reasoning models, or by manually prompting chain-of-thought (CoT) reasoning.

**Controlling Reasoning Effort for Proprietary Reasoning Models.** We set the reasoning effort in GPT-5 and Gemini-2.5-Pro to 3 levels: **low**, **medium**, and **high**. To be noted, the low, medium, and high reasoning effort of Gemini-2.5-Pro corresponds to the thinking budget in its native API of 1,024, 8,192, and 24,576 tokens, respectively. These values represent upper bounds; models do not necessarily use the full allocated budget.

**Simulating Reasoning Effort Control with CoT Reasoning for Non-Reasoning Models.** Open-weight models lack explicit reasoning effort control available in proprietary reasoning models. We therefore use few-shot chain-of-thought (CoT) prompting to simulate varying reasoning depths. Using two videos from Section 3.2, we created two CoT variants: **Simple CoT** with brief reasoning and **Multi-Step CoT** with elaborate step-by-step analysis.

**Simple CoT.** The model is instructed to focus on moving elements and identify visual cues indicating playback direction. We manually curated reasoning traces for each demonstration. Due to space limitations, we omitted the example in this version.

**Multi-step CoT.** We designed structured instructions with multi-step reasoning traces divided into three stages: *Observation* (objectively describe visible events without assumptions), *Assumption* (infer plausible physical or causal sequences), and *Conclusion* (determine whether observations align with or contradict *assumptions* to judge playback direction). Due to space limitations, we omitted the prompts in this version. A model output example is shown in Figure 2.

### 3.4 Supervised Fine-Tuning (SFT) Settings

We constructed in-domain fine-tuning datasets to test whether explicit temporal direction supervision improves model performance. Note that all videos in Hanyu et al. (2023), AoT-PsyPhyBENCH, and these fine-tuning datasets are drawn from Moments in Time (Monfort et al., 2019). We randomly sampled 500, 1,500, and 2,500 clips distributed across its 339 action classes in the original dataset. These clips are entirely distinct from the videos in AoT-PsyBench. Each video was used to create two examples: the original (forward, labeled F) and its temporal reverse (backward, labeled B), yielding 1,000, 3,000, and 5,000 training examples.

We fine-tuned Qwen2-VL-7B and Qwen2.5-VL-7B using the zero-shot prompt format (Section 3.1) with standard supervised learning. We then evaluated these models on AoT-PsyPhyBENCH.

### 3.5 Evaluation Metrics

We evaluate models using overall accuracy and F1 scores for Forward and Backward classes separately. Class-specific F1 scores highlight potential directional biases in model predictions.

### 3.6 Hyperparameter Settings

**Frame Sampling and FPS Settings.** Frames were sampled at equal temporal intervals without resizing. Frame rates were set based on model-specific defaults and training configurations: 2 FPS for Qwen2-VL, Qwen2.5-VL, and QVQ-72B-Preview (standard for these models), and 4 FPS for Cosmos-Reason1-7B (matching its inference settings recommendation<sup>2</sup>) and all proprietary models. Preliminary experiments showed that 4 FPS balances performance with API costs (Section 5.5).

**Inference Parameters.** For open-weight models (Qwen2-VL, Qwen2.5-VL, and QVQ-72B-Preview), we set temperature to 0.6 and top-p to

0.95. For Cosmos-Reason1-7B, we used temperature 0.6, top-p 0.95, and repetition penalty 1.05, consistent with settings reported in their paper. For GPT-4o (GPT-4o-2024-11-20) and GPT-4.1 (GPT-4.1-2025-04-14), we applied temperature 0.6, top-p 0.95, and frequency penalty 0. For OpenAI reasoning models (o3, o4-mini, GPT-5), inference parameters such as temperature and top-p are not exposed via the API and thus remained at default values. For Gemini-2.5-Pro (accessed 2025-10-02), we used the OpenAI-compatible API with default values: temperature 1.0 and top-p 0.95.

**SFT Parameters.** We froze the vision encoder and applied LoRA (Hu et al., 2021) (rank 8) to both the language model and the vision-language bridging MLP layer. We used a maximum sequence length of 2,048 tokens, a batch size of 2 per device with gradient accumulation over 2 steps (effective batch size of 4), a learning rate of  $1 \times 10^{-5}$ , and trained for 5 epochs with a cosine learning rate scheduler and 10% warmup.

## 4 AoT-PsyPhyBENCH: A Psychophysically Validated Arrow-of-Time Benchmark

We evaluate whether vision–language models (VLMs) can infer the arrow of time—that is, distinguish *forward* and *backward* playbacks in everyday videos. Our work builds on the psychophysical study of Hanyu et al. (2023), enabling direct comparison with human performance. The source dataset comprises 360 three-second clips at 29.97 FPS drawn from the Moments in Time dataset (Monfort et al., 2019), covering a broad range of daily dynamics. Ten participants viewed every clip twice—once forward and once reversed—across two sessions on different days (each session: 180 forward, 180 reversed; order randomized per participant). Three raters annotated six motion categories: (1) **Proceed**: forward locomotion of people, animals, or vehicles; (2) **Fall**: free-fall/ballistic motion under gravity; (3) **Diffusion**: centrifugal diffusion or small-particle explosions; (4) **Division**: division of material by hand or tool; (5) **Put**: addition/construction of material by hand; and (6) **Reciprocal**: reciprocating (cyclic) motion (Table 1). The first five categories reflect *irreversible* processes that, when reversed, tend to violate fundamental physical regularities (entropy increase, gravity, causal ordering), whereas reciprocating motion is *bidirectional* and often appears plausi-

<sup>2</sup><https://huggingface.co/nvidia/Cosmos-Reason1-7B>

ble in both directions. Categories are not mutually exclusive. Human performance exhibited a pronounced *forward bias*: participants made errors on **39%** of reversed trials versus **9%** of forward trials, with accuracy strongly modulated by the presence of motions from the first five categories. Performance was significantly lower for the reciprocating motion category.

In this work, we introduce **AoT-PsyPhyBENCH**, a curated subset of [Hanyu et al. \(2023\)](#) that *excludes the reciprocating (cyclic) motion category* (category 6) and retains only high-consensus clips from categories 1–5. This exclusion is justified by human performance data: the cyclic motion category achieved only 61.1% overall accuracy (accuracy with forward and backward videos combined) compared to 83.8–93.5% for the other motion categories, substantially degrading overall benchmark reliability (77.6% with cyclic motion vs. 89.2% without). From the original 720 videos (360 clips with their reversed counterparts), our filtering yields 212 clips (totaling **424** videos with reversals). This produces a low-ambiguity, psychophysically validated AoT evaluation benchmark that enables direct and reliable human–model comparison. Table 1 provides an overview of the motion categories with comparative statistics illustrating how AoT-PsyPhyBENCH filters the original dataset.

## 5 Results

### 5.1 Zero-shot Performance

Table 2 presents zero-shot results, revealing two key gaps: **open vs. proprietary** and **VLM vs. human** performance gaps. Additionally, we did not find advantages of reasoning models over non-reasoning models.

**Open vs. Proprietary.** **Open-weight models** cluster around 50% accuracy (random baseline), with Cosmos-Reason1-7B slightly higher at 52.1%. QVQ-72B-Preview, designed for mathematical reasoning over single images rather than physical dynamics, performs at chance with extreme label bias (Section 6.2). Cosmos-Reason1-7B, despite being trained on an undisclosed AoT dataset where it reportedly achieved 60% accuracy, shows minimal improvement on our benchmark. **Proprietary models** consistently exceed random guessing, with GPT-4.1 achieving the highest zero-shot performance (60.1%).

**VLM vs. Human.** All VLMs lag substantially behind humans (89.2%), with a gap of 29.1 percentage points even for the best zero-shot model, GPT-4.1.

**Reasoning vs. Non-Reasoning:** Explicit reasoning capability provides no clear advantage. The non-reasoning model GPT-4.1 (60.1%) outperforms reasoning models o3 (55.2%), o4-mini (56.1%), and GPT-5 (56.1%).

### 5.2 Few-shot Performance

We test whether few-shot examples elicit in-context learning for AoT (Table 3). Overall, few-shot prompting does *not* consistently improve performance and often amplifies label prediction bias. GPT-5 shows modest, stable gains (56.1→57.4% accuracy; Backward F1: 26.8→30.9 with 4 shots). In contrast, GPT-4.1 degrades with few-shot prompts (60.0→56.0/55.5% accuracy; Backward F1: 57.4→51.7/49.9). GPT-4o shows severe degradation: accuracy drops (52.6→50.7%) while Backward F1 collapses catastrophically (24.9→4.6/5.1), indicating near-complete failure to detect reversed videos. The open model Qwen2.5-VL-72B exhibits minimal accuracy change (49.5→51.9%) but substantially worsened bias (Backward F1: 38.0→18.7/21.1). These results demonstrate that few-shot prompting provides limited benefits for temporal reasoning: only one model (GPT-5) shows modest improvement, while others either decline or amplify their existing forward bias, suggesting that in-context learning is insufficient to elicit robust AoT understanding.

### 5.3 Reasoning Effort Ablation

**Controlling reasoning effort in proprietary reasoning VLMs.** We varied the reasoning effort parameter (low/medium/high) for GPT-5 and Gemini-2.5-Pro (Table 4). Surprisingly, increasing reasoning effort does not improve performance and often degrades it. Gemini-2.5-Pro with low effort achieves 61.1% accuracy, the highest performance across all our experimental settings, yet medium and high efforts decrease to 59.9% and 58.7%, with Backward F1 declining from 52.1 to 51.4 to 49.0. From the low effort setting, we highlight one correctly predicted example, where we asked Gemini-2.5-Pro to explain its reasoning in a follow-up dialogue turn (bottom-right of Figure 2). GPT-5 showed an even more pronounced pattern: low effort achieves 57.8% accuracy, while medium and

| Category & Description   | Reversal is easy for humans? | Human F1 (F/B)    | # samples | Included in AoT-PsyPhyBENCH? | Example 1  | Example 2  |
|--|------------------------------|-------------------|-----------|------------------------------|--|--|
| (1) <b>Proceed:</b> forward locomotion of people, animals, or vehicles   | ✓                            | 86.5/82.5         | 82        | Yes                          |   |   |
| (2) <b>Fall:</b> free-fall/ballistic motion under gravity                | ✓                            | 86.9/82.8         | 84        | Yes                          |   |   |
| (3) <b>Diffusion:</b> centrifugal diffusion or small-particle explosions | ✓                            | 84.6/78.7         | 56        | Yes                          |   |   |
| (4) <b>Division:</b> division of material by hand or tool                | ✓                            | 86.0/80.6         | 37        | Yes                          |   |   |
| (5) <b>Put:</b> addition/construction of material by hand                | ✓                            | 84.1/77.4         | 67        | Yes                          |   |   |
| (6) <b>Reciprocal:</b> reciprocating (cyclic) motion                     | ✗                            | 71.6/ <b>38.5</b> | 148       | No                           |  |  |

Table 1: Motion categories following Hanyu et al. (2023), with human performance, counts, and visual examples. Human F1 scores are reported as forward (F)/backward (B) for each motion category. AoT-PsyPhyBENCH includes categories 1–5 (all clips) and excludes category 6. ✓ marks categories where reversal is typically easy for humans (irreversible processes); ✗ marks cyclic, bidirectional motions where reversal is challenging.

high efforts drop to 56.1% and 56.4%. More critically, Backward F1 collapses with increased effort (34.4 to 26.8 to 25.7), indicating amplification of forward bias. These results suggest that extended reasoning does not elicit better temporal reasoning and instead reinforces existing directional biases.

**Chain-of-thought (CoT) prompting in an open VLM.** We tested explicit step-by-step reasoning with Qwen2.5-VL-72B using two CoT variants (Simple and Multi-step; Section 3.3). As shown in Table 5, CoT prompting fails to improve performance and amplifies label prediction bias, most notably with Multi-step CoT (2-shot), where Backward F1 drops by 13.2 points (Section 6.2). In the model output example using the Multi-step CoT prompt (top-right in Figure 2), we saw that the model usually could identify the event in the video, yet it strongly believed that no physical law was violated and the video was played forward. Combined with the reasoning effort results, these findings demonstrate that additional deliberation, whether through explicit effort controls or prompted CoT,

does not compensate for the lack of robust temporal and physical understanding, and often exacerbates existing biases.

## 5.4 SFT Performance

We fine-tuned Qwen2-7B and Qwen2.5-VL-7B on 1,000, 3,000, and 5,000 training examples to test whether explicit supervision improves AoT performance (Table 6). Fine-tuning failed to improve accuracy, which remained near chance level (~50%) across all training set sizes. While fine-tuning reduced the extreme forward bias present in vanilla models, leading to a more balanced forward/backward F1 scores, it did not yield genuine task competence. Performance plateaued after 1,000 examples, with no gains from additional data. These results suggest that AoT task might require different training approaches beyond conventional SFT.

## 5.5 FPS Ablation

We evaluated GPT-4.1 across frame rates from 2 to 30 FPS and compared results to human per-



**An incorrect output from Qwen2.5VL-72B with Multi-step CoT prompting:**  
**Observation:** The video shows two men engaged in arm-wrestling at a desk. (.....) Their hands are clasped together on the desk, and they are pulling against each other.  
**Assumption:** In a typical arm-wrestling match, the participants start with their hands clasped and then exert force to try to pin the opponent's arm to the table. (.....)  
**Conclusion:** The video shows the men in the middle of the struggle, with their hands clasped and both leaning forward in an effort to win. This matches the expected sequence of a forward-playing arm-wrestling match. There are no signs of reversed motion, such as the hands separating or the men leaning backward unexpectedly. Therefore, the video is played forward.

**Gemini-2.5-Pro Self-Explained Reasoning:**  
The video shows two men engaged in an arm wrestling match.  
1. **Normal Progression of Arm Wrestling:** In a standard arm wrestling match, the contest begins with both participants' arms in an upright, near-vertical position. The goal is to force the opponent's arm down onto the table. (.....)  
2. **Observed Motion in the Video:** The video clip starts with the man in the white shirt in a clear losing position, with his arm almost completely pinned to the desk. As the video plays, his arm moves upward and away from the desk, (.....)  
3. **Conclusion:** The action shown is the exact reverse of how an arm wrestling match progresses. A match ends with an arm going down, it doesn't start there and move up. Because the video shows the conclusion of the match first and then moves back toward the starting position, it is playing in reverse.

Figure 2: **Left:** a backward video clip (category: Put). **Top-Right:** Qwen2.5VL-72B Multi-step CoT reasoning. **Bottom-Right:** Gemini-2.5-Pro’s self-explained reasoning trace. Qwen2.5VL-72B correctly identified the event in the scene and made a valid assumption, but failed to observe that the event was reversed which led to an incorrect conclusion. In contrast, Gemini-2.5-Pro correctly detected the reversal of the event in Step 2 based on a valid assumption it made in Step 1.

| Model                       | F. F1 | B. F1 | Acc.        |
|-----------------------------|-------|-------|-------------|
| <b>Baselines</b>            |       |       |             |
| Random                      | —     | —     | 50.0        |
| Human                       | 99.0  | 89.0  | 89.2        |
| <b>Open Models</b>          |       |       |             |
| <b>Non-reasoning Models</b> |       |       |             |
| Qwen2-VL-7B                 | 66.7  | 0.0   | 50.0        |
| Qwen2.5VL-7B                | 63.0  | 19.5  | 49.3        |
| Qwen2.5VL-72B               | 57.4  | 38.2  | 49.5        |
| <b>Reasoning Models</b>     |       |       |             |
| QVQ-72B-Preview             | 66.1  | 0.0   | 49.4        |
| cosmos-reason1 7B           | 31.2  | 63.3  | <b>52.1</b> |
| <b>Proprietary Models</b>   |       |       |             |
| <b>Non-reasoning Models</b> |       |       |             |
| GPT-4o                      | 65.4  | 24.9  | 52.6        |
| GPT-4.1                     | 62.5  | 57.4  | <b>60.1</b> |
| <b>Reasoning Models</b>     |       |       |             |
| o3                          | 67.2  | 29.1  | 55.2        |
| o4-mini                     | 67.4  | 33.1  | 56.1        |
| GPT-5                       | 68.7  | 26.8  | 56.1        |
| Gemini-2.5-pro              | 65.9  | 51.4  | 59.9        |

Table 2: Zero-shot performance.

| Model                     | #Shots | F. F1 | B. F1 | Acc.        |
|---------------------------|--------|-------|-------|-------------|
| <b>Open Model</b>         |        |       |       |             |
| Qwen2.5 VL-72B            | 0      | 57.4  | 38.0  | 49.5        |
|                           | 2      | 64.2  | 18.7  | 50.2        |
|                           | 4      | 65.4  | 21.1  | 51.9        |
| <b>Proprietary Models</b> |        |       |       |             |
| GPT-4.1                   | 0      | 62.3  | 57.4  | <b>60.0</b> |
|                           | 2      | 59.5  | 51.7  | 56.0        |
|                           | 4      | 60.0  | 49.9  | 55.5        |
| GPT-4o                    | 0      | 65.4  | 24.9  | 52.6        |
|                           | 2      | 66.8  | 4.61  | 50.7        |
|                           | 4      | 66.5  | 5.07  | 50.7        |
| GPT-5                     | 0      | 68.7  | 26.8  | 56.1        |
|                           | 2      | 69.5  | 29.2  | 57.4        |
|                           | 4      | 69.2  | 30.9  | 57.4        |

Table 3: Few-shot performance.

| <b>Model</b>              | <b>Effort</b> | <b>F. F1</b> | <b>B. F1</b> | <b>Acc.</b> |
|---------------------------|---------------|--------------|--------------|-------------|
| <b>Proprietary Models</b> |               |              |              |             |
| GPT-5                     | Low           | 68.9         | 34.4         | <b>57.8</b> |
|                           | Medium        | 68.7         | 26.8         | 56.1        |
|                           | High          | 69.1         | 25.7         | 56.4        |
| Gemini-2.5-pro            | Low           | 67.2         | 52.1         | <b>61.1</b> |
|                           | Medium        | 65.9         | 51.4         | 59.9        |
|                           | High          | 65.3         | 49.0         | 58.7        |

Table 4: Reasoning effort ablation.

| <b>Setting</b>                | <b>Shots</b> | <b>F. F1</b> | <b>B. F1</b> | <b>Acc.</b> |
|-------------------------------|--------------|--------------|--------------|-------------|
| <b>Qwen2.5VL-72B-instruct</b> |              |              |              |             |
| No CoT                        | 2            | 64.2         | 18.7         | 50.2        |
| Simple                        | 2            | 67.3         | 3.76         | 51.2        |
| Multi-step                    | 2            | 66.7         | 5.48         | <b>50.7</b> |

Table 5: Chain-of-Thought performance on the Qwen2.5VL-72B-instruct model.

| <b>Data Size</b>             | <b>F. F1</b> | <b>B. F1</b> | <b>Acc.</b> |
|------------------------------|--------------|--------------|-------------|
| <b>Qwen2-7B Training</b>     |              |              |             |
| Vanilla                      | 66.7         | 0.0          | 50.0        |
| 1000                         | 64.1         | 12.2         | 49.1        |
| 3000                         | 54.2         | 48.2         | 51.4        |
| 5000                         | 59.4         | 31.6         | 49.1        |
| <b>Qwen2.5VL-7B Training</b> |              |              |             |
| Vanilla                      | 63.0         | 19.5         | 49.3        |
| 1000                         | 54.8         | 43.5         | 49.8        |
| 3000                         | 54.5         | 43.4         | 49.5        |
| 5000                         | 54.5         | 43.4         | 49.5        |

Table 6: SFT performance.

| <b>Model</b> | <b>FPS</b> | <b>F. F1</b> | <b>B. F1</b> | <b>Acc.</b> |
|--------------|------------|--------------|--------------|-------------|
| Human        | 30         | 99.0         | 89.0         | <b>89.2</b> |
|              | 2          | 62.1         | 51.2         | 57.3        |
|              | 4          | 62.5         | 57.4         | 60.1        |
|              | 8          | 62.9         | 55.8         | 59.7        |
|              | 16         | 62.4         | 53.7         | 58.5        |
|              | 30         | 64.4         | 55.9         | <b>60.6</b> |

Table 7: FPS ablation results.

formance (Table 7). Notably, at 30 FPS, corresponding to the complete video with all frames, the model achieves only 60.6% accuracy, falling far short of human performance at 89.2%. This 28.6 percentage point gap demonstrates that the performance deficit is not due to insufficient temporal information. Across tested frame rates, model accuracy varies minimally (57.3–60.6%). Given the marginal improvement from 4 FPS (60.1%) to 30 FPS and API cost considerations, we selected 4 FPS as the default for all proprietary models.

## 6 Analyses

### 6.1 Motion Category Analysis

We analyze per-category performance across the five motion classes in AoT-PsyPhyBENCH: Proceed, Fall, Diffusion, Division, and Put. Figure 3 compares humans with three VLMs: Cosmos-Reason1 7B (best open-weight model in the zero-shot setting, 52.1%), GPT-4.1 (best proprietary and overall model in the zero-shot setting, 60.1%), and Gemini-2.5-Pro (best overall model, 61.1%). Humans maintain ≈80-100% accuracy across all categories and directions. VLMs show substantial gaps with striking asymmetry: relatively better on forward videos but dramatically worse on backward videos, revealing strong *label prediction bias*.

**Category-specific patterns.** The largest F1 deficits appear in Proceed and Fall, where all models struggle to capture the causal and gravitational cues defining temporal direction. Diffusion and Put show moderate F1 scores but remain below human performance even in forward playback, suggesting incomplete modeling of dispersion and goal-directed manipulation. Division is comparatively easier for both humans and models, likely due to clear temporal discontinuities (e.g., splitting). Notably, Cosmos-Reason1-7B achieves relatively strong backward F1, particularly on Fall and Diffusion, consistent with its explicit AoT-oriented training, whereas Gemini-2.5-Pro and GPT-4.1 perform better in forward playback but drop sharply when the temporal order is reversed. Overall, these results suggest that current VLMs rely primarily on directional visual priors and dataset correlations rather than robust physical causality understanding.

### 6.2 Label Prediction Bias

We observed that most VLMs exhibit **label prediction bias**: models strongly favor one label over the other despite balanced forward/backward video

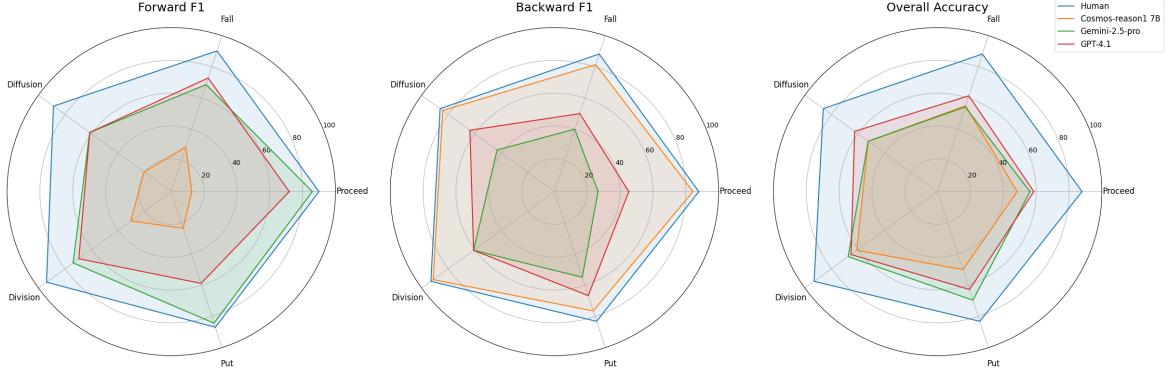


Figure 3: Per-category comparison on AoT-PsyPhyBENCH across three representative models. Cosmos-reason1-7B (zero-shot; best open-weight in this setting), GPT-4.1 (zero-shot; best proprietary and best overall model in this setting), and Gemini-2.5-Pro (zero-shot, low-reasoning effort; best model across all settings), and humans: (a) forward F1 (left), (b) backward F1 (middle), and (c) overall accuracy (right). Humans remain consistently high across all categories and both directions ( $\approx 80\text{-}100\%$ ). In contrast, VLMs show substantial gaps. Backward detection is the most challenging, revealing a forward-direction bias (with Cosmos-reason1 as a notable exception, showing comparatively strong backward F1).

distribution in AoT-PsyPhyBENCH. In zero-shot evaluation (Table 2), GPT-4o predicted 87% of clips as Forward (only 13% as Backward), while QVQ-72B-Preview predicted 100% Forward, causing minority-class F1 scores to drop below 0.4 due to low recall. Most models exhibit Forward bias (Qwen2.5-VL series, QVQ-72B-Preview, GPT-4o, o3, o4-mini, GPT-5), while Cosmos-Reason1-7B shows Backward bias, likely due to extensive reversed-clip exposure during reinforcement learning.

Furthermore, **reasoning amplifies bias**: when models engage in step-by-step reasoning, biases intensify rather than improve. QVQ-72B-Preview often correctly identified scene events but failed to detect reversed motion, with its reasoning process reinforcing the idea that the video is played forward. Similarly, increasing reasoning effort (GPT-5 in Table 4) or enforcing chain-of-thought (Table 5, Figure 2) further amplified label prediction bias rather than correcting it.

## 7 Conclusion

We present **AoT-PsyPhyBENCH**, a psychophysically validated benchmark designed to assess whether modern vision–language models (VLMs) possess a human-like inductive bias for inferring the *arrow of time*—the implicit understanding that physical events unfold irreversibly from past to future. Our experiments reveal that current VLMs perform far below human baselines, even on intuitive scenarios such as falling objects, which hu-

mans find easy to interpret. This striking shortfall cannot be attributed to limited data or lack of reasoning depth. Instead, it highlights a fundamental absence of inductive biases for temporal continuity, causality, and physical irreversibility—principles that humans internalize effortlessly through interaction with the physical world. We release AoT-PsyPhyBENCH and its evaluation code to foster the development of multimodal systems that move beyond statistical pattern recognition toward genuine physical understanding.

## 8 Limitations

Our work has two limitations. First, for proprietary models, even though we report the specific model names and versions used, exact replication may be affected by undisclosed updates or API changes beyond our control. Second, we haven’t been able to offer a conclusive explanation for why SFT did not work on AoT. One possible reason is that open-source VLMs may not effectively encode temporal information along the frame sequence. Cosmos-Reason1-7B saw improvement with reinforcement learning in their work, however, they did not publish their dataset and we did not see improvement on our benchmark, either.

## References

- Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, and 1 others.

2025. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*.
- Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, and 1 others. 2025. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*.
- Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. 2023. Test of time: Instilling video-language models with a sense of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2503–2516.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Yang Du, Yuqi Liu, and Qin Jin. 2024. Reversed in time: A novel temporal-emphasized benchmark for cross-modal video-text retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5260–5269.
- Chaoyou Fu, Yuhua Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.
- Nao Hanyu, Kei Watanabe, and Shigeru Kitazawa. 2023. Ready to detect a reversal of time’s arrow: a psychophysical study using short video clips in daily scenes. *Royal Society open science*, 10(4):230036.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, and Weiming Hu. 2024. MIBench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22417–22428, Miami, Florida, USA. Association for Computational Linguistics.
- Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and 1 others. 2019. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508.
- OpenAI. 2024. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>.
- OpenAI. 2025a. Introducing gpt-4.1 in the api.
- OpenAI. 2025b. Introducing gpt-5.
- OpenAI. 2025c. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- Qwen Team. 2024. Qvq: Qwen2-vl with visual question answering. Technical report, Alibaba Cloud.
- Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, and 2 others. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *Preprint*, arXiv:2406.09411.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. 2023. Paxion: Patching action knowledge in video-language foundation models. *Advances in Neural Information Processing Systems*, 36:20729–20749.
- Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2024. Star: A benchmark for situated reasoning in real-world videos. *Preprint*, arXiv:2405.09711.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa:next phase of question-answering to explaining temporal actions. *Preprint*, arXiv:2105.08276.
- Zihui Xue, Mi Luo, and Kristen Grauman. 2025. Seeing the arrow of time in large multimodal models. *arXiv preprint arXiv:2506.03340*.