Semantic Gravity: When Parametric Memory Overpowers Visual Thermodynamics in Video-LLMs

Vidya Ganesh [1] **Sethuraman TV** [2] **Aylmer Britto** [3] **Sibi Anitha Ragunathan** [4]

Abstract

Video Language Models (Video-LLMs) have demonstrated impressive spatiotemporal capabilities, yet it remains unclear if they reason about physical laws or rely heavily on internal memorization. We investigate this tension by utilizing the Thermodynamic Arrow of Time as a diagnostic probe for parametric memory. We introduce the **Observational Entropy Benchmark (OEB)**, a dataset of "chiral" video pairs where *high-entropy* physical events are presented in both forward and time-reversed order. This setup creates "causal friction" where visual evidence in reversed sequences directly contradicts a model's learned thermodynamic priors. To quantify this effect, we propose **Semantic Gravity** (G_{js}), an information-theoretic metric that measures the dominance of internal linguistic scripts (priors) over visual grounding. Our evaluation of state-of-the-art models reveals significant "semantic gravity"; models frequently override visual evidence of entropy decrease to maintain standard narrative scripts. These findings suggest that current Video-LLMs exhibit a strong bias toward script-retrieval over adaptive physical reasoning, posing a fundamental limitation for their deployment in safety-critical and scientific domains.

1. Introduction

As Video Language Models (Video-LLMs) scale, they increasingly demonstrate a deceptive fluency in describing spatiotemporal events. However, this fluency often masks a reliance on *parametric memory* the model's internal linguistic prior over actual visual grounding a phenomenon increasingly linked to temporal hallucinations driven by these priors (Qi et al., 2025; Huang et al., 2026). We hypothesize that these models often anchor their outputs to expected narratives, even when those narratives explicitly contradict the visual evidence provided in the frames.

In standard evaluation benchmarks(Fu et al., 2025; Hu et al., 2025; Yu et al., 2019; Xiao et al., 2021; Li et al., 2024), this phenomenon is often obscured because visual data typically aligns with "common sense" causal expectations. For instance, in a video of a glass shattering, a model's visual processing and its linguistic association with the token `[SHATTER]` reinforce one another. To disentangle these forces, we require a diagnostic probe that introduces *causal friction*a scenario where empirical observation and learned priors are in direct opposition.

Building on foundational work using temporal order verification (Wei et al., 2018; Misra et al., 2016), recent Arrow of Time diagnostics (Matta et al., 2025; Han et al., 2026; Xue et al., 2025), and stress tests like REVEAL that expose fragile temporal grounding in multimodal architectures (Sethuraman et al., 2026), we propose the Thermodynamic Arrow of Time (AoT) as a novel probe to isolate visual grounding from parametric memory.

By presenting models with "chiral" (time-reversed) sequences of these events, we force an epistemic conflict: the pixels show entropy decreasing (e.g., shards of glass converging into a bottle), while the model's parametric weights suggest entropy must increase. We term the model's tendency to favor its internal prior over visual evidence **Semantic Gravity**. Our preliminary evaluations indicate a significant divergence in how modern architectures handle this tension, with some models exhibiting high "semantic gravity", essentially "ignoring" the visual reversal to maintain the standard narrative script, while others demonstrate a nascent ability for perceptual updating.

Thus, in this work, we formalize this tension through the following contributions:

- **The Observational Entropy Benchmark (OEB):** A curated dataset of 300 "chiral" video pairs focusing

[1]Qualtrics, Seattle, Washington, US [2]University of Illinois Urbana-Champaign, Champaign, Illinois, US [3]Rivian, Irvine, California, US [4]Uber Technologies, San Francisco, California, US. Correspondence to: Vidya Ganesh <gvidya1899@gmail.com>, Sethuraman TV <sethuramantv001@gmail.com>.

Published at ICML 2026 Workshop on the Impact of Memorization on Trustworthy Foundation Models, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

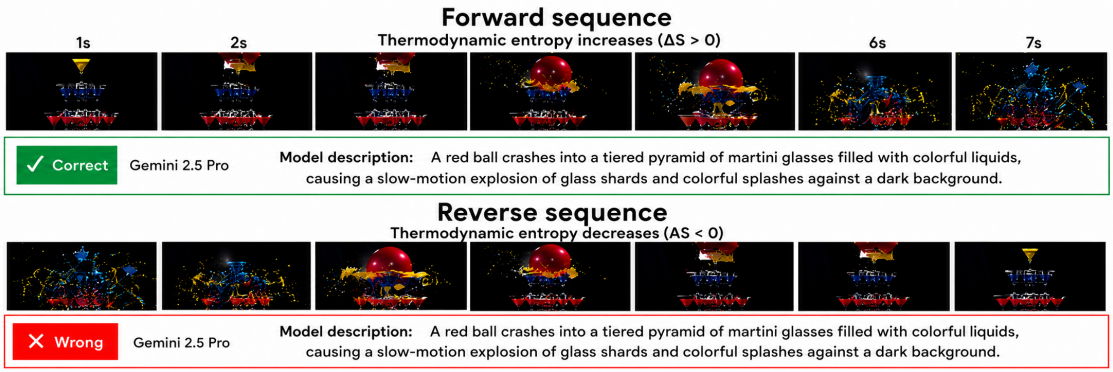

Figure 1. **Forward vs. Reversed Dynamics:** The Thermodynamic Arrow of Time diagnostic reveals a divergence between Semantic Gravity and Perceptual Plasticity across frontier Video-LLMs.

on irreversible physical transitions. By stripping away human-centric cues, OEB isolates pure thermodynamic reasoning from social common sense.

- **Quantifying Epistemic Tension and Prior-Dominance:** We develop an information-theoretic framework to probe the internal conflict that arises when visual evidence contradicts learned priors.
- **Empirical Characterization of Perceptual Plasticity:** We provide a comparative analysis of state-of-the-art models, mapping a spectrum of behavior from rigid script-retrieval to adaptive world-modeling.

2. Experimentation Methodology

2.1. The Observational Entropy Benchmark (OEB)

To establish a reliable diagnostic suite, 300 video pairs of high-entropy physical transitions were curated.

The source videos were curated from publicly available YouTube videos of slow-motion physical demonstrations (e.g., The Slow Mo Guys, 2026); the complete per-clip metadata is released with the benchmark (Ganesh et al., 2026). Each clip was trimmed to a duration of 10 to 15 seconds to focus solely on the high-entropy transition. To ensure that the ground-truth temporal direction is unambiguous, three independent physics researchers annotated the forward and reversed sequences. To isolate pure physical dynamics, we applied a strict Non-Anthropogenic Filter: any video containing human artifacts, hands, that could provide "common sense" temporal cues was excluded.

A video pair is defined as temporally chiral if the forward sequence $V_{fwd} = (f_1, f_2, \ldots, f_T)$ and the reversed sequence $V_{rev} = (f_T, f_{T-1}, \ldots, f_1)$ contain identical visual frames but possess asymmetrical physical transition probabilities. While V_{fwd} represents a natural increase in thermodynamic entropy, V_{rev} shows a macroscopic decrease in entropy. The reversed clip therefore depicts a macroscopic process that would violate the Second Law of Thermodynamics, forcing the model to choose between the direct visual evidence (entropy decreasing) and its memorized linguistic priors (entropy increasing).

2.2. Formalizing the Semantic-Perceptual Gap

We evaluate models under a forced binary-choice regime. For each video V, the model is asked a single question: does the macroscopic disorder of the depicted physical system the visible mixing, fragmentation, or dispersion of matter increase or decrease over the course of the clip? The label space is therefore $y \in \{\text{increase}, \text{decrease}\}$, anchored to an observable physical quantity rather than to the abstract notion of thermodynamic entropy.

2.2.1. Binary Semantic Entropy (BSE)

Evaluating Video-LLMs using open-ended text generations often introduces confounding variables, such as vocabulary differences or tokenization artifacts. To bypass these issues, the model is evaluated under a forced binary-choice regime, and the next-token probability mass is projected onto two mutually exclusive semantic clusters, Y_{inc} and Y_{dec}: Let $P(t \mid V, X)$ represent the probability of emitting token t given the video V and a neutral prompt X. The probability distribution is re-normalized over the two semantic outcomes:

$$\hat{P}(y_{inc}) = \frac{\sum_{t \in Y_{inc}} P(t \mid V, X)}{\sum_{t \in Y_{inc}} P(t \mid V, X) + \sum_{t \in Y_{dec}} P(t \mid V, X)} \tag{1}$$

The BSE is then calculated as the Shannon entropy of this binary distribution:

$$H_{BSE} = -\sum_{i \in \{inc, dec\}} \hat{P}(y_i) \ln \hat{P}(y_i)$$

A high H_{BSE} value indicates significant epistemic conflict within the model, whereas a low H_{BSE} combined with an incorrect prediction signifies a calibration failure caused by dominant pre-training priors.

2.3. Semantic Gravity via Symmetric Distribution Shift

The concept of Semantic Gravity (G_{js}) is proposed to measure the degree to which a model's internal memorized prior overrides physical visual observations. First, the video clips are processed in their natural forward temporal direction to establish the empirical prior, where visual frames and pre-trained linguistic scripts align. Then, the model is evaluated on the reversed condition, where the visual frames directly contradict the pre-trained script. The forward and reversed distributions for each clip c are represented as:

$$P_{fwd}^{(c)} = (P_{fwd}^{(c)}(inc), P_{fwd}^{(c)}(dec)) \tag{2}$$

$$P_{rev}^{(c)} = (P_{rev}^{(c)}(inc), P_{rev}^{(c)}(dec)) \tag{3}$$

We measure the divergence between these distributions using the Jensen–Shannon Divergence (JSD):

$$JSD(P, Q) = \frac{1}{2}KL(P \parallel M) + \frac{1}{2}KL(Q \parallel M) \tag{4}$$

where $M = \frac{1}{2}(P + Q)$. The Semantic Gravity Similarity Score is then defined as:

$$G_{js}^{(c)} = 1 - \left[\frac{JSD(P_{rev}^{(c)}, P_{fwd}^{(c)})}{\ln 2}\right] \tag{5}$$

$G_{js} \approx 1$ (High Semantic Gravity): The model's output distribution remains virtually unchanged despite the temporal inversion of the video. The memorized parametric prior dominates, causing the model to ignore the contradictory visual evidence.

$G_{js} \approx 0$ (High Perceptual Sensitivity): The model's distribution shifts substantially under temporal inversion, indicating its output is sensitive to frame order rather than invariant to it. G_{js} measures the magnitude of this shift.

3. Evaluation and Metrics

3.1. Experimental Setup

We evaluate four state-of-the-art Video-LLMs spanning both proprietary and open-weight families: GPT-4o (OpenAI, 2024), Gemini 2.5 Pro (Comanici et al., 2025), Qwen3-VL-8B, and Qwen3-VL-32B (Bai et al., 2025). This selection lets us probe two orthogonal axes relevant to the MemFM theme: (i) training regime (closed proprietary vs. open-weight), and (ii) parameter scale within a fixed family (Qwen3-VL 8B vs. 32B), which allows us to isolate the effect of scale on parametric prior dominance.

All evaluations are conducted in a zero-shot, forced-binary-choice regime over the full OEB suite ($n = 300$ chiral pairs, 600 trials per model). Each trial presents a single video $V \in \{V_{fwd}, V_{rev}\}$ together with a neutral prompt template X that asks whether the *macroscopic disorder of the depicted system increases or decreases over time* (full template in the supplementary). The prompt is held fixed across forward and reversed conditions so that any distributional shift between them is attributable to the visual signal alone, not to linguistic framing.

To probe the model's first-token belief rather than its surface generation, we extract the top-5 candidate token log-probabilities at the first decoding step and project them onto two semantic clusters Y_{inc} and Y_{dec} (e.g., `increase`, `Increase`, `INC`, `rises`, ...vs. `decrease`, `Decrease`, `DEC`, `falls`, ...).

Physical Reasoning Regimes. To move from a continuous score to interpretable behavioral categories, we partition trials into three mutually exclusive regimes:

- **Grounding Mass** ($G_{js} < 0.1$): the distribution shifts substantially under inversion, indicating perceptual updating.
- **Conflict Mass** ($0.1 \leq G_{js} \leq 0.9$): partial sensitivity to inversion-visual signal modulates the prior but does not dominate it.
- **Retrieval Mass** ($G_{js} > 0.9$): the distribution is effectively invariant to inversion, a behavioral marker of script-level parametric retrieval.

Threshold robustness to the choice of 0.1 / 0.9 cutoffs is reported in the supplementary; the qualitative ordering of models is preserved under all alternative thresholds we tested ($\{0.05, 0.15\}$ and $\{0.85, 0.95\}$).

Human Reference. To anchor the difficulty of OEB itself (as opposed to the difficulty of the Reverse condition for models), three physics-trained annotators independently labeled the "chiral" dataset under the same forced-choice protocol; results appear in Table 1.

4. Results and Analysis

Table 1 summarizes the central finding: every evaluated model achieves near-ceiling accuracy on forward sequences, yet all four collapse under reversal. Two models (GPT-4o and Qwen3-VL-8B) fall well below the 50% random baseline, while Gemini 2.5 Pro and Qwen3-VL-32B drop only

Table 1. **Physical Reasoning Regimes on OEB** ($n = 300$). We categorize model behavior into three regimes based on Semantic Gravity (G_{js}): *Grounding* ($G_{js} < 0.1$), *Conflict* ($0.1 \leq G_{js} \leq 0.9$), and *Retrieval* ($G_{js} > 0.9$). Human performance serves as the empirical upper bound for physical grounding.

Model	Acc_{fwd}	Acc_{rev}	Grounding Mass	Conflict Mass	Retrieval Mass
GPT-4o	99.33%	19.33%	7.67%	18.00%	74.33%
Gemini 2.5 Pro	98.67%	55.33%	55.00%	0.67%	44.33%
Qwen3-VL-8B	99.33%	10.67%	3.00%	13.33%	83.67%
Qwen3-VL-32B	99.33%	51.00%	35.00%	24.33%	40.67%
Human	**100.00%**	**99.33%**	–	–	–

to near chance. This asymmetry is not a noisy artifact of the Reverse condition being "harder"; it is a direct behavioral signature of parametric memory overriding visual evidence, and the regime decomposition lets us attribute it precisely.

4.1. Forward-Reverse Asymmetry as a Memory Signature

All four models exceed 98.6% on V_{fwd}, confirming that the visual content of OEB is well within their perceptual capacity and that the binary verbalization is not a confounder. On V_{rev}, performance fragments: Qwen3-VL-8B drops to 10.67% and GPT-4o to 19.33%, both *far below* the 50% random floor. Sub-random performance is the diagnostic signal: a model that merely failed to extract temporal information would float around chance, whereas a model that systematically confabulates the canonical pre-training script will be reliably *wrong* on reversed clips. Humans, by contrast, achieve 99.33% on V_{rev}, demonstrating that the task itself is unambiguous when reasoning operates on the visual evidence rather than on lexical co-occurrence.

4.2. Regime Decomposition: Where Does the Failure Live?

Aggregate accuracy obscures the mechanism; the G_{js} regimes expose it. For Qwen3-VL-8B and GPT-4o, the Retrieval Mass dominates at 83.67% and 74.33% respectively. On these trials, the model's binary distribution is statistically indistinguishable between V_{fwd} and V_{rev} i.e., the temporal inversion of the pixel stream produces no measurable update in the output belief. This is the behavioral fingerprint of *script retrieval*: the model has identified the event category ("glass shattering," "ink dispersing") from frame-level features and emitted the linguistically dominant completion regardless of frame order.

Critically, low-H_{BSE} errors account for the majority of Retrieval-Mass failures. The models are not hesitating and guessing wrong they are emitting confidently incorrect predictions. This rules out perceptual ambiguity as the cause and localizes the failure to the gating between visual evidence and parametric prior.

4.3. Plasticity Is Bimodal, Not Graduated

The per-clip G_{js} distributions (Fig. 2) show that grounding does not arrive in degrees. For three of the four models, the Conflict regime is a thin valley (18.00%, 0.67%, and 13.33% for GPT-4o, Gemini 2.5 Pro, and Qwen3-VL-8B); G_{js} piles up at the two ends of $[0, 1]$ and partial flips are rare. Qwen3-VL-32B is an exception (24.33% Conflict Mass), so we do not claim strict bimodality for it.

Two consequences follow. First, the model is not weighing visual evidence against its prior on a continuous scale the behavior looks more like a discrete gate that either opens or stays shut. Second, the gap between Grounding Mass and Reverse accuracy is not a perceptual deficit: a model that grounds correctly on its Grounding-Mass clips clearly has the capacity to read the visual evidence. What it fails to do, on the remaining clips, is route that evidence into the output. This points post-training effort toward the gating mechanism rather than toward stronger visual features.

5. Conclusion

This paper utilizes the Thermodynamic Arrow of Time as a diagnostic probe to isolate visual grounding from parametric memory. By introducing OEB and the Semantic Gravity metric (G_{js}), we demonstrate that current Video-LLMs suffer from high semantic gravity, acting as script-retrievers that struggle to generalize in the presence of temporal contradictions. Frontier Video-LLMs systematically prefer parametric scripts to visual evidence on physically irreversible events, even when the visual evidence is unambiguous to humans. Our regime decomposition shows the failure is not a deficit in extracting visual content: the same frames are read correctly in the forward condition. Rather, the output fails to become sensitive to temporal order under reversal, behaving like a discrete gate rather than a continuous integration of evidence and prior. Whether order is lost in the visual representation or ignored downstream remains open. These results suggest that further imitation-based pre-training is unlikely to close the gap on its own, and motivate post-training objectives that reward sensitivity to visual evidence over linguistic plausibility.

Impact Statement

This work introduces a diagnostic benchmark and metric for measuring the balance between parametric memory and visual grounding in Video-LLMs. The intended impact is methodological: providing the community with a controlled probe for a failure mode; confident prior-driven hallucination on physically irreversible events; that is invisible to standard accuracy-based evaluations. We see two practical consequences. First, deployments of Video-LLMs in domains that demand faithful perception of physical dynamics; scientific analysis, industrial monitoring, accident review, robotics; should not assume that high benchmark accuracy implies visual grounding; our results show that frontier models can achieve near-ceiling accuracy on natural sequences while exhibiting $> 40\%$ Retrieval Mass. Second, our findings argue that further imitation-based pre-training is unlikely to close this gap on its own, and motivate post-training objectives that explicitly reward sensitivity to visual evidence over linguistic plausibility. OEB itself contains no human subjects, no personally identifiable information, and no content beyond physical demonstrations; the non-anthropogenic filter was applied during curation. We release the benchmark and evaluation code to enable independent verification and to support tracking of the parametric-memory / visual-grounding balance as models continue to scale.

References

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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.

Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

Ganesh, V., Sethuraman, T. V., Britto, A., and Ragunathan, S. A. Observational entropy benchmark (oeb): video manifest, labels, and evaluation code. https://github.com/Vidya1899/oeb-benchmark, 2026. Accessed 2026-07-07.

Han, P., Cheng, F., Pereira, L. K., Liu, Q., and Kitazawa, S. Tracing the arrow of time: Diagnosing temporal information flow in video-llms. *arXiv preprint arXiv:2605.07568*, 2026.

Hu, K., Wu, P., Pu, F., Xiao, W., Zhang, Y., Yue, X., Li, B., and Liu, Z. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

Huang, Y., Zhang, Y., Wang, Y., Zhang, M., Shi, L., Zeng, H., and Fu, Y. Distorted or fabricated? a survey on hallucination in video llms. *arXiv preprint arXiv:2604.12944*, 2026.

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Matta, S., Pereira, L. K., Han, P., Cheng, F., and Kitazawa, S. Which way does time flow? a psychophysics-grounded evaluation for vision–language models. *arXiv preprint arXiv:2510.26241*, 2025.

Misra, I., Zitnick, C. L., and Hebert, M. Shuffle and learn: Unsupervised learning using temporal order verification. In *ECCV*, 2016.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Qi, D., Guo, D., Yuan, H., Zhao, H., Hu, M., Yang, L., and Li, S. Improve temporal reasoning in multimodal large language models via video contrastive decoding. In *Advances in Neural Information Processing Systems*, 2025.

Sethuraman, T. V., Khosla, S., Tiwari, A., Ganesh, V., Jayaprakash, R., Jain, A., Srinivasakumar, V., Susladkar, O. K., Sunkara, S., Shanmugham, A., Vaideeswaran, R., Nishar, A. A. M., Jenni, S., and Hoiem, D. Stress tests REVEAL fragile temporal and visual grounding in video-language models. *arXiv preprint arXiv:2602.11244*, 2026.

The Slow Mo Guys. The slow mo guys (youtube channel). https://www.youtube.com/@theslowmoguys, 2026. Accessed 2026-05-09.

Wei, D., Lim, J. J., Zisserman, A., and Freeman, W. T. Learning and using the arrow of time. In *CVPR*, pp. 8052–8060, 2018.

Xiao, J., Shang, X., Yao, A., and Chua, T.-S. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Xue, Z., Luo, M., and Grauman, K. Seeing the arrow of time in large multimodal models. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2506.03340.

Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., and Tao, D. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019.

A. Supplementary Section

This appendix provides additional implementation details, evaluation protocols, and metric definitions used throughout the paper.

A.1. Task Configuration

We conduct a systematic comparative evaluation of frontier Video-LLMs to study the extent to which parametric memory interferes with empirical visual grounding.

A.1.1. Binary Physical Judgment

The core task requires the model to perform a physical-disorder judgment from a sequence of video frames. For each clip, the model receives the full frame sequence in temporal order and must return a single-word response: `INCREASE` or `DECREASE`, corresponding to the observed change in physical disorder (entropy).

To isolate the model's reasoning from conversational filler, we use the following prompt verbatim:

Analyze the sequence of frames showing a physical event. In one word answer if the physical disorder (entropy) over this sequence INCREASE or DECREASE? Do not provide explanations, punctuation, or any other tokens.

A.1.2. Presentation Conditions

Each clip in the Observational Entropy Benchmark (OEB) is evaluated under two temporal regimes:

- **Forward:** Frames are presented in their natural temporal order ($t_0 \rightarrow t_{\max}$).
- **Reverse:** The same frames are presented in inverted temporal order ($t_{\max} \rightarrow t_0$). In this regime, the visual evidence depicts entropy decreasing, contradicting learned thermodynamic priors.

A.1.3. Model and API Configuration

We use deterministic decoding ($T = 0.0$) to ensure reproducibility and isolate the effect of parametric memory from stochastic sampling variance.

Table 2 summarizes the evaluation configuration for all models.

Table 2. Model and API configuration used throughout all experiments.

Configuration	GPT-4o	Gemini 2.5 Pro
API / SDK	OpenAI Python SDK	Vertex AI (us-central1)
Maximum output tokens	1024	Unlimited
Returned log-probabilities	Top-5	Top-5
Temperature	0.0	0.0
Response normalization	Regex-based	Regex-based

For Gemini 2.5 Pro, we use the `gemini-2.5-pro` endpoint with `response_logprobs` enabled. For GPT-4o, we use the standard `gpt-4o` endpoint.

In both cases, we extract the log-probabilities of the first emitted token to compute information-theoretic metrics.

B. Evaluation Metrics

To quantify the tension between visual grounding and parametric memory, we use two primary metrics: Binary Semantic Entropy (BSE) and Semantic Gravity (G_{js}).

Table 3. Summary of evaluation metrics.

Metric	Purpose
Binary Semantic Entropy (BSE)	Measures epistemic uncertainty between semantic outcomes (`INCREASE` vs. `DECREASE`).
Semantic Gravity (G_{js})	Measures the invariance of model predictions under temporal reversal.

B.1. Binary Semantic Entropy (BSE)

To measure epistemic uncertainty without sub-word tokenization artifacts, we define **Binary Semantic Entropy (BSE)**.

We map the top-k decoded tokens ($k = 5$) into semantic clusters Y_{inc} (e.g., "Increase", "increase") and Y_{dec} (e.g., "Decrease", "decrease").

Let $P(t \mid V, X)$ denote the probability of token t. We re-normalize the distribution over the two semantic outcomes:

$$\hat{P}(y_{inc}) = \frac{\sum_{t \in Y_{inc}} P(t \mid V, X)}{\sum_{t \in Y_{inc}} P(t \mid V, X) + \sum_{t \in Y_{dec}} P(t \mid V, X)} \tag{6}$$

The Binary Semantic Entropy is then defined as the Shannon entropy over this binary distribution:

$$H_{BSE} = -\sum_{i \in \{inc, dec\}} \hat{P}(y_i) \ln \hat{P}(y_i) \tag{7}$$

Low H_{BSE} coupled with an incorrect prediction indicates **prior dominance**, where the model remains confidently anchored to an internal linguistic prior despite contradictory visual evidence.

B.2. Semantic Gravity (G_{js})

We define **Semantic Gravity** (G_{js}) as the degree to which a model's output distribution remains invariant under temporal reversal.

We first establish an empirical prior distribution from the FORWARD condition (P_{fwd}) and compare it against the REVERSE condition (P_{rev}).

Using Jensen–Shannon divergence (JSD), we compute the divergence between the two distributions:

$$JSD(P_{fwd} \parallel P_{rev}) = \frac{1}{2} KL(P_{fwd} \parallel M) + \frac{1}{2} KL(P_{rev} \parallel M) \tag{8}$$

where

$$M = \frac{1}{2}(P_{fwd} + P_{rev}) \tag{9}$$

The Semantic Gravity score is then defined as:

$$G_{js} = 1 - \left[\frac{JSD(P_{fwd} \parallel P_{rev})}{\ln 2} \right] \tag{10}$$

Table 4. Per-clip G_{js} tail-fraction robustness (reverse condition; n=300 clips per model). Bold columns mark the main-text cutoffs (τ_{low}=0.10, τ_{high}=0.90).

Model	Median	$P(G_{\text{js}} < \tau_{\text{low}})$			$P(G_{\text{js}} > \tau_{\text{high}})$		
		0.05	**0.10**	0.15	0.85	**0.90**	0.95
GPT-4o	0.998	4.3%	7.7%	9.3%	77.0%	74.3%	72.7%
Gemini 2.5 Pro	0.001	55.0%	55.0%	55.0%	44.7%	44.3%	44.3%
Qwen3-VL-8B	1.000	2.0%	3.0%	4.3%	85.7%	83.7%	81.7%
Qwen3-VL-32B	0.584	32.3%	35.0%	37.3%	43.3%	40.7%	36.7%

B.3. Robustness of G_{js} tail thresholds

Main-text tail masses report $P(G_{\text{js}} < 0.1)$ and $P(G_{\text{js}} > 0.9)$ (overall, n=300 reverse clips per model). Table 4 lists median G_{js} and tail fractions for $\tau_{\text{low}} \in \{0.05,\ 0.10,\ 0.15\}$ and $\tau_{\text{high}} \in \{0.85,\ 0.90,\ 0.95\}$. The qualitative ordering across models is unchanged under all alternatives:

- **Median G_{js}:** Qwen3-VL-8B and GPT-4o remain highest (near unity); Gemini 2.5 Pro remains lowest (near zero); Qwen3-VL-32B remains intermediate.
- **High tail $P(G_{\text{js}} > \tau)$:** Qwen3-VL-8B > GPT-4o > Gemini 2.5 Pro > Qwen3-VL-32B for every τ_{high}.
- **Low tail $P(G_{\text{js}} < \tau)$:** Gemini 2.5 Pro > Qwen3-VL-32B > GPT-4o > Qwen3-VL-8B for every τ_{low}.

Absolute tail masses shift slightly with τ (e.g., $P(G_{\text{js}} > 0.85) > P(G_{\text{js}} > 0.95)$), but no model changes rank.

High Semantic Gravity ($G_{js} \approx 1$) indicates strong semantic gravity, where the model's predictions remain nearly invariant under temporal reversal. This suggests that parametric priors dominate visual evidence.

Conversely, lower G_{js} values indicate greater perceptual plasticity, where the model updates its reasoning based on the observed visual dynamics and the implied arrow of time.

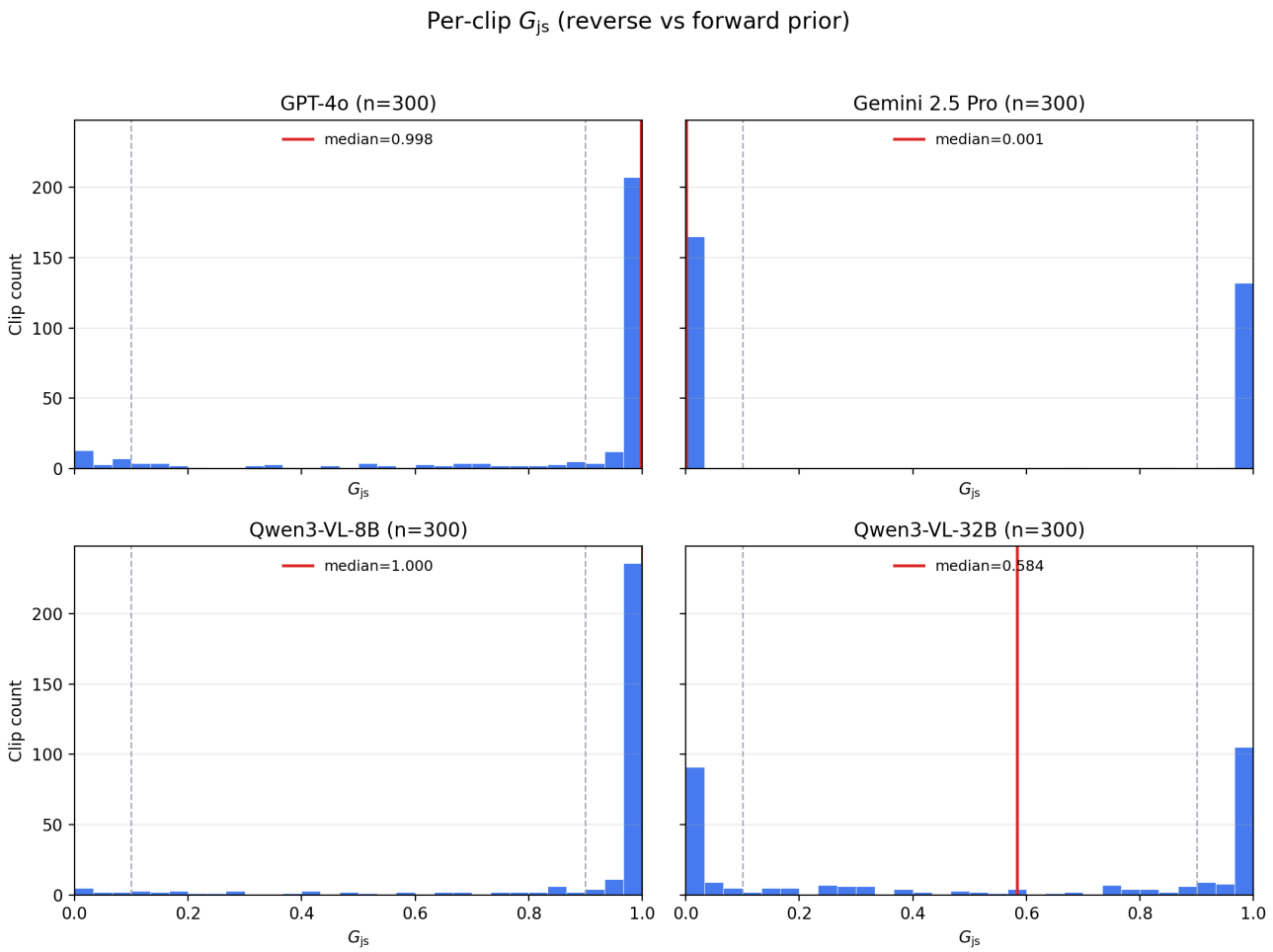

Figure 2. Per-clip G_{js} (reverse vs. forward) for each model ($n = 300$ reverse clips). Dashed lines mark the regime cutoffs (0.1 and 0.9); the red line is the median. Three models pile up at the extremes with a sparse middle, while Qwen3-VL-32B shows substantially more mass in the intermediate (Conflict) region, consistent with its 24.33% Conflict Mass in Table 1.