VISION LANGUAGE MODELS CANNOT REASON ABOUT PHYSICAL TRANSFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding physical transformations is fundamental for reasoning in dynamic, real-world environments. While Vision Language Models (VLMs) show promises in embodied applications grounded in the physical world, whether they genuinely understand physical transformations remains unclear. To address this gap, we introduce *ConservationBench* to evaluate *conservation*—whether physical quantities remain invariant under transformations despite appearance changes. Spanning four quantitative properties (number, length, volume, size), each task requires integrating visual evidence across time and includes counterfactuals where the targeted quantities are not conserved, forming paired conserving and non-conserving scenarios. With systematic variation in prompts, frame sampling methods, and task design, we generate 13,824 questions evaluating on 34 VLMs. Results reveal consistent failure: none demonstrates systematic conservation. Performance remains marginally above chance, with improvements on conservation tasks often accompanied by severe performance on counterfactual controls. This suggests a dependence on superficial patterns or shortcuts over genuine understanding and reasoning on conservation. Moreover, models show no benefit from higher temporal resolution or prompt design. Together, these findings indicate that current VLMs fail to reason about physical transformation.

1 INTRODUCTION

Recent advances in Vision-Language Models (VLMs) (Zhang et al., 2024b; Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023) have demonstrated remarkable capabilities of perception (Wang et al., 2024b; Chen et al., 2025; Team et al., 2025; Cheng et al., 2024b), reasoning (Zhang et al., 2024a; Xu et al., 2024; Cheng et al., 2024a), and visual commonsense understanding (Zellers et al., 2019; Park et al., 2020). These capabilities hold promise for real-world applications (Brohan et al., 2023), particularly in embodied tasks (Driess et al., 2023; Nasiriany et al., 2024) that demand a genuine understanding of the physical world and its underlying properties (Chow et al., 2025b; Gao et al., 2024). Yet it remains unclear whether VLMs possess a true understanding of physical principles or the capacity to operate reliably in embodied physical environments.

A key factor in human intelligence that enables successful navigation in an embodied, physically grounded world is the ability to understand and reason about physical transformations (Piaget, 1950; 1952; 1965; Baillargeon et al., 1985; 1990; Baillargeon, 1987; 1986; Spelke et al., 1992; Baillargeon & Carey, 2012; Bear et al., 2021; Piloto et al., 2022). This capacity includes tracking objects over time (Spelke et al., 1994; 1995), managing occlusions (Gredebäck & von Hofsten, 2004), and adapting to dynamic environments (Allen et al., 2020). While there are benchmarks evaluating physically plausible video generation (Motamed et al., 2025; Meng et al., 2024a; Yang et al., 2025; Liu et al., 2025; Shi et al., 2024) and physical understanding in VLMs, spanning from everyday scenes (Zheng et al., 2024; Chow et al., 2025a) to high-school physics questions (Wang et al., 2025) and Olympiad-level problems (Qiu et al., 2025; Wang et al., 2025), these efforts focus either on video generation or physical properties in static scenes, leaving underexplored whether VLMs can genuinely reason about physical transformations—where specific properties may or may not remain invariant.

To bridge this gap, we evaluate *conservation* in VLMs—the understanding that physical quantities remain invariant under transformation despite changes in appearance. Here, physical quantity refers to the measurable magnitude of objects along certain dimensions, while spatial transformation de-

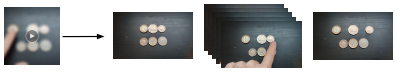
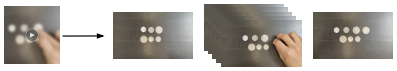
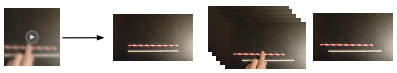
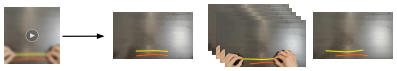
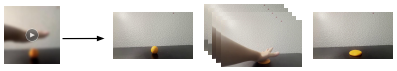
Task	Question	Conserve?	Extracted Frames
Number	*Is the number of coins in the upper row the same as in the lower row in the final image?*	✓	
		✗	
Length	*Is the length of the upper straw the same as the length of the lower straw in the final image?*	✓	
		✗	
Size	*Is the size of the playdough in the first image the same as in the final image?*	✓	
		✗	
Volume	*Is the amount of liquid in the left glass in the first image the same as in the right glass in the final image?*	✓	
		✗	

Figure 1: Illustrative Tasks and Frame Selection Pipeline in *Conservation Bench*.

notes the continuous process through which objects change in appearance. For example, an agent demonstrating conservation would recognize that pouring water into a differently shaped glass does not alter its volume, despite the change in visible form. Achieving conservation thus requires more than linguistic knowledge of quantity: it demands a systematic understanding that is both reversible and grounded in visual as well as conceptual representations. We introduce *ConservationBench*, a cognitively grounded benchmark for evaluating whether VLMs can reason about physical transformations. The benchmark consists of 192 video-based tasks across four core quantitative properties—number, length, volume, and size—each requiring models to judge whether a quantity is conserved despite visual transformations. To control for shortcut exploitation, we include matched non-conserving controls where the target quantity changes while irrelevant features remain constant. We systematically vary frame extraction method, temporal resolution, and prompting strategy, yielding 36 conditions and 13,824 total trials.

Evaluating 34 VLMs (1B–76B parameters), we find that models consistently fail to integrate temporal information to track conserved properties across dynamic scenes. High accuracy on conservation tasks is often driven by default heuristics, which reverse in non-conserving scenarios, revealing brittle, non-generalizable reasoning. Furthermore, prompting with cues encouraging transformation reasoning or providing higher temporal resolution does not help. These findings expose a fundamental limitation in current VLMs and underscore the need for more grounded, temporally-aware models capable of systematic physical inference.

2 Related Works

2.1 Evaluating and Benchmarking VLMs

The evaluation of vision-language models (VLMs) is central to identifying their limitations and shaping future directions. Early efforts relied on single-task benchmarks such as VQA (Antol et al., 2015), OK-VQA (Marino et al., 2019), MSCOCO (Lin et al., 2014), OCR (Liu et al., 2023), and GQA (Hudson & Manning, 2019). However, with the emergence of multi-modal large language models (VLMs) that claim broader perceptual and reasoning abilities, evaluation has shifted toward holistic benchmarks such as LAMM (Yin et al., 2023), MMMU (Yue et al., 2024), SEED-Bench (Li et al., 2024), and MMBench (Liu et al., 2024b). A growing line of benchmarks focuses specifically on quantity understanding (Rane et al., 2024; Paiss et al., 2023; Rahmanzadehgervi et al., 2024; Yuksekgonul et al., 2022). These tasks typically assess a model's ability to individuate and count discrete objects in static scenes. While useful, such evaluations largely reduce to surface-level enu-

meration. They do not test whether models encode numerical invariance—the understanding that quantity persists across spatial or configurational transformations. In contrast, our work examines whether VLMs go beyond perceptual counting to represent quantity as a conserved relational property. Our work also relates to multi-image and video-based benchmarks (Yue et al., 2024; Song et al., 2024; Jiang et al., 2024; Fu et al., 2024; Liu et al., 2024a; Meng et al., 2024b; Wang et al., 2024a; Huang et al., 2023). These evaluations assess logical reasoning, cross-image comparison, temporal dynamics, and context referencing. While conservation tasks share this multi-image nature, they uniquely target whether VLMs can track continuous physical transformations and recognize the stability of invariant properties across them.

2.2 Physical Understanding and Conservation

Insights from cognitive science underscore conservation as a critical benchmark for systematic physical reasoning. First proposed by Piaget, success on conservation tasks has long been viewed as evidence of emerging mental operations (Piaget & Inhelder, 1969). Developmental studies show that solving these tasks requires constructing transformation-invariant representations while suppressing misleading perceptual cues (Goldin-Meadow & Beilock, 2010; Houdé et al., 2011; Poirel et al., 2012). Behavioral and neurocognitive research further demonstrates that conservation performance depends on sensorimotor grounding and inhibitory control, highlighting the embodied nature of transformation understanding (Beilock & Goldin-Meadow, 2010; Lozada & Carro, 2016). Conservation also builds on more rudimentary abilities such as object permanence and individuation, revealed through studies exploiting the tunnel effect and violation-of-expectation paradigms (Burke, 1952; Flombaum & Scholl, 2006; Noles et al., 2005; Scholl, 2007), which themselves provide essential foundations for robust physical reasoning. In this light, conservation is widely recognized as a foundational cognitive capacity—providing a scaffold for the higher-level physical reasoning needed to navigate dynamic, embodied environments (Fodor, 1975; Baillargeon & Carey, 2012; Barsalou, 2020; Luo et al., 2025).

Recent studies have examined models' abilities to reason about physical properties, causal interactions, and material dynamics (Chow et al., 2025a; Patel et al., 2022; Zheng et al., 2024). Growing evidence suggests that VLMs struggle with fundamental aspects of visual reasoning and physical understanding (Schulze Buschoff et al., 2025; Campbell et al., 2024; Buschoff et al., 2025), with some work exploring how modular frameworks or synthetic training data might address these limitations (Balazadeh et al., 2024; 2025). However, these efforts largely emphasize outcome prediction or descriptive inference, without testing whether models recognize that certain properties remain invariant under transformation. In many cases, success appears to stem from outcome-based heuristics rather than structured mental operations (Newman et al., 2024; Isola et al., 2015). Consequently, it remains unclear whether current VLMs can genuinely integrate sequential evidence to track physical transformations while maintaining stable representations of underlying properties—a core cognitive capacity directly targeted by conservation tasks (Mitchell & Krakauer, 2023).

3 Experimental Design

3.1 Conservation Tasks

To systematically measure the conservation ability of VLMs–the understanding that specific physical properties remain invariant under transformations despite changes in appearance, we construct a suite of conservation tasks in the form of videos that visually depict physical transformations across four fundamental quantitative properties.

- **Conservation in Number**: Two rows of identical coins are presented in an initial configuration. One row is then spread apart without adding or removing any coins. The task assesses whether the quantity is perceived as unchanged despite the altered spacing.
- **Conservation in Length**: Two straws of identical length are shown in an initial configuration. One of the straws is then repositioned without altering its actual length.
- **Conservation in Volume**: A fixed volume of liquid is poured from one container into another of a different shape. Although the height changes significantly, the volume remains constant throughout the transformation.

- **Conservation in Size**: A lump of playdough is reshaped from one form (e.g., a ball) into another (e.g., a flattened disc). While the shape and surface features change, the total mass remains the same across both states.

Although the four conservation types probe distinct physical properties, the tasks follow a unified structure: a transition from an initial to a final state mediated by an observable transformation. Each video begins with an initial state, proceeds through a continuous transformation (e.g., pouring, spreading, flattening), and ends with a new state where the surface appearance of the object of interest is altered. This design mirrors real-world scenarios where physical reasoning depends on integrating perceptual evidence across time.

Generalization across Task-irrelevant Features To ensure the robustness and generalizability of the conclusions drawn from our benchmark, we systematically vary key visual parameters in each conservation task (Table 3). These parameters include object count, size, color, layout, container shape, and the direction of transformation. Each conservation property consists of 48 unique video instances of different configurations, resulting in a total of 192 videos. This controlled variation guarantees that the core conservation principle is preserved across a wide range of visual contexts, thus preventing models from relying on memorized templates or superficial cues.

Transformation-mandatory and Transformation-helpful Notably, conservation tasks differ in how strongly they depend on observing the transformation. We classify them into two categories: *transformation-mandatory* and *transformation-helpful*. In mandatory tasks (volume and size), witnessing the transformation is essential—for instance, in volume conservation, seeing the liquid poured is necessary, since the final height alone is insufficient for judging quantity. In helpful tasks (number and length), correct judgments can still be made from the initial and final states, as the relevant quantity remains visually accessible despite superficial changes. This distinction enables a more diagnostic evaluation: models that excel on helpful but not on mandatory tasks may rely on static cues rather than forming internal representations of the transformation process.

To this end, we further curated a set of 96 tasks derived solely from the final frame of transformation-helpful tasks. Here, models are prompted to compare numbers and lengths directly based on simple counting and intuitive judgments of spatial extent. This design isolates pre-conceptual, rudimentary forms of quantitative assessment—such as item enumeration and perceptual matching—from the broader representational demands of transformation-based reasoning. By contrasting performance on these static tasks with temporal conservation trials, we can reveal how basic quantitative sensitivity relates to the more systematic representations of quantity that underlie conservation reasoning.

3.2 Non-conserving Tasks

A key limitation of applying conservation tasks to model evaluation is the uniformity of ground-truth labels: since all standard tasks involve quantity preservation, models can appear accurate simply by defaulting their responses to indicate invariance, due to biases from either visual contexts or linguistic patterns in the prompts, without genuinely reasoning about the physical transformation itself (Li et al., 2025b). To address this, we create non-conserving counterfactuals as a set of controlled experiments where the quantity of interest is explicitly altered during the transformation without changing the task-irrelevant features. That is to say, these manipulations are performed within the same environments, using identical object sets and visual contexts, thereby ensuring a controlled comparison. This design enables fine-grained assessment of model sensitivity to actual changes in quantity, rather than reliance on superficial heuristics or distributional priors. We describe the construction of these control tasks across each property below.

- **Non-conserving Transformation of Number**: Two rows of identical coins are presented in an initial configuration. One row is then spread apart, with one coin added to said row.
- **Non-conserving Transformation of Length**: Two straws of identical length are shown in an initial configuration. One of the straws is then repositioned while its actual length is altered (extendable straws are used).
- **Non-conserving Transformation of Volume**: A fixed volume of liquid is poured from one container into another of a different shape. A significant portion of water is left in the original container instead of being completely poured in.

- **Non-conserving Transformation of Size**: A lump of playdough is reshaped from one form (e.g., a ball) into another (e.g., a flattened disc). A significant portion of playdough is left in the experimenter's hand without being integrated into the new shape.

Following this design, we curated a control set in which each non-conserving control is paired with a conservation task under matched configurations, yielding another 192 videos.

Table 1: **Overview of Multi-image Task Conditions and Evaluation Scale**

Core Dataset	384 total videos
Conservation Tasks	48 videos $\times$ 4 properties = 192 videos
Non-conserving Control	48 videos $\times$ 4 properties = 192 videos
Multi-frame Curation	36 conditions (factorially combined)
Extraction Method	3 Conditions (Uniform, SEVILA, Human)
Frame Count	3 Conditions (3 frames, 8 frames, 16 frames)
Prompting	4 Conditions (Direct Question, "Sequential", CoT, "Continuous")
Total Task Trials	$384 \times 36 = 13,824$ evaluation trials

3.3 Adaptation to Multi-frame Input

3.3.1 Temporal Resolution

The ability to understand physical transformations critically depends on comprehending dynamic processes over time. Unlike static snapshot reasoning, robust comprehension requires recognizing continuity across successive observations. Human perception benefits from high frame rates (e.g. $\sim$ 30-60 frames per second) that convey rich temporal information, while the architectural and computational limitations of VLMs restrict them to inferring such dynamics from discrete and often sparse inputs.

To investigate the impact of temporal resolution on conservation understanding, we vary the number of frames extracted from each video:

- **3-frame condition**: Only three frames are provided—the first, the last, and one intermediate frame. This condition presents minimal temporal information while retaining just enough cues for humans to solve the task.
- **8-frame condition**: Eight frames are sampled to offer moderate temporal granularity. This condition is designed to contrast qualitatively with the 3-frame condition by enabling multi-frame representations of the temporally continuous scene.
- **16-frame condition**: Sixteen frames are sampled to provide finer-grained temporal information, offering a more detailed depiction of the transformation process, contrasting quantitatively with the 8-frame condition.

In all conditions, the first and last frames are fixed to preserve the initial and final states of the transformation. This manipulation enables us to assess whether increased intermediate visual evidence regarding the transformation process enhances the model's ability to infer conservation.

3.3.2 Sampling Strategy

In studying physical transformations, the sequence and selection of visual inputs are crucial. Due to computational limitations, state-of-the-art VLMs are optimized for sparse, multi-frame reasoning. This raises an important question: do different frame selection strategies influence the model's understanding of dynamic scenes? Additionally, do humans and models rely on different criteria when identifying informative visual moments? To explore these questions, we implement and compare three frame extraction strategies, each reflecting distinct assumptions about what defines a "representative" moment in a physical event.

- **Uniform Sampling**: Frames are sampled uniformly across the timeline, serving as a baseline approach commonly used in prior work, based on the assumption that temporal regularity sufficiently represents informational diversity.

- **Human-based**: To obtain a baseline for human intuition in frame extraction, we recruited N = 18 annotators. Each annotator was randomly assigned a subset of the dataset and asked to manually select the intermediate frames that captured the essential stages of the transformation.
- **Model-based**: We adopt SeViLA (Yu et al., 2023) and leverage a BLIP-2-based Localizer to identify language-aware keyframes. Prompted with the same instruction assigned to humans ("extract the most complete set of frames that capture the entire process"), the Localizer module selects frames with high relevance scores, which are then passed to the Answerer module for inference. This method formalizes a strategy akin to semantic salience: choosing frames that are maximally informative given a specific query.

This design allows us to test whether different frame selection strategies affect model performance on physical transformation reasoning. We hypothesize that optimizing frame selection, rather than merely increasing frame quantity, leads to more effective representations of dynamic events. We detail our data curation process in Appendix A and prompting strategies in Appendix B, redand provide example input in Appendix C.

4 Experiments

4.1 Inference and Evaluation

Inference. We evaluate 34 VLMs spanning diverse model architectures, training data, and parameter scales, covering both mainstream commercial systems and advanced open-source models. To ensure fidelity, comparability, and reproducibility, we strictly adhere to reference configurations and implementations from the official codebases. Refer to Appendix D for further details.

Evaluation. To evaluate free-form outputs of VLMs on multiple-choice questions (MCQs), we follow the two-stage scoring method of Li et al. (2025a). In Stage 1, each VLM output is mapped to a unique choice from the provided options or labeled FAIL when no unambiguous mapping is possible. Mapping follows a hybrid strategy: deterministic template matching is applied first, and unresolved cases are adjudicated by an LLM-as-a-Judge constrained to the option set. Models exhibiting persistently high FAIL rates are excluded from further analyses to avoid bias from nonsensical outputs. In Stage 2, the mapped option is compared against the ground-truth answer, with all FAILs scored as incorrect. Details are provided in Appendix E.

4.2 Human Baseline

Given the large number of questions and the cost of human annotation, we curated a representative subset by randomly selecting one out of every eight task configurations for each quantitative property, counterbalanced across conservation tasks and non-conserving controls. This resulted in a total of 864 questions. We hypothesize that reduced variation in task-irrelevant features is unlikely to compromise the benchmark's validity or generalizability given the robustness of human reasoning. Participants received the same stimuli and three-choice questions as the VLMs, with the exception that they directly selected answers rather than requiring LLM judge parsing. The aggregated human accuracy reaches 95.25%, consistent with decades of developmental research showing that humans from late childhood reliably solve conservation tasks with near-perfect accuracy (Piaget, 1965; Houdé, 1997; Pezzulo et al., 2013; Viarouge et al., 2019). These results validate our benchmark design and its adaptation for evaluating VLMs.

4.3 Main Results

As shown in Figure 2A, model accuracy across 34 VLMs ranges from 27% to 49%, with most performing only marginally above the 33.3% chance level. In contrast, human participants exceed 95% accuracy (Section 4.2), highlighting a clear gap between VLMs and intuitive human reasoning. Even top-performing models (e.g., InternVL-2-8B) fail to generalize across conservation and non-conserving controls. These limitations generalize across all four quantitative domains—number, length, volume, and size (see Appendix G for details). While number and length yield marginally better results, no model demonstrates consistent success across domains. Collectively, these results reveal a core limitation: VLMs struggle to integrate temporal cues or track invariant properties through dynamic transformations, a key requirement for grounded physical reasoning.

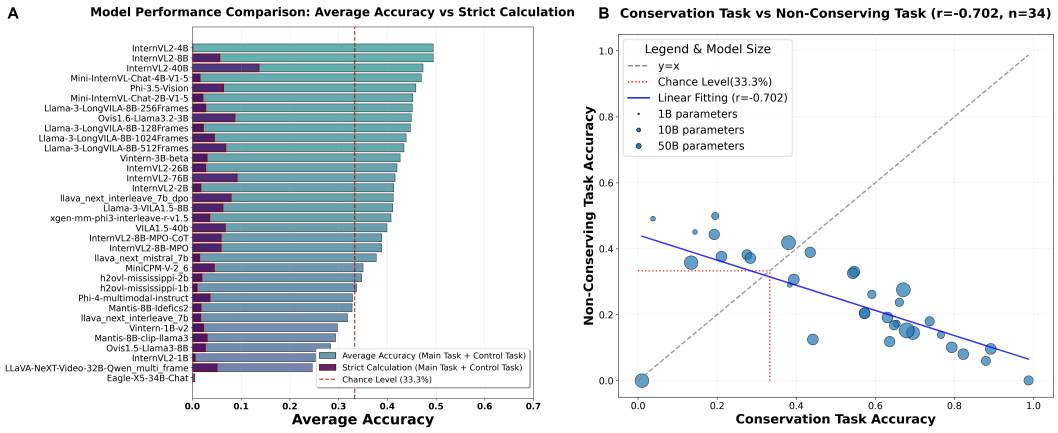


Figure 2: Overall Performance on *ConservationBench*. A. Accuracy averaged across conservation tasks and non-conserving control compared to strict pairwise calculation; B. Performance on non-conserving control in relation to conservation tasks.

4.3.1 Inverse Performance Across Conservation and Non-Conserving Control

By comparing model performance on non-conserving control tasks against conservation tasks, we observe a strong negative correlation: models that perform better on conservation tasks tend to perform worse on the corresponding control tasks, and vice versa. This trade-off indicates a systematic bias toward different interpretations of the task scenario—models tend to "default" almost randomly across the axis of conserving vs. non-conserving heuristics, regardless of actual transformation evidence (Figure 2B). Crucially, this pattern reveals a diagnostic failure: models are not simply underperforming but are exhibiting mutually offsetting errors across matched task types.

We further validate this pattern using a strict pairwise evaluation across the full set of 6,912 matched conservation and non-conserving control tasks (as shown in Figure 2A; labeled in purple). In this analysis, a model is only marked correct if it answers both tasks in a pair correctly—capturing whether it can jointly recognize quantity preservation and detect meaningful violations under matched visual conditions. We found that all models perform well below chance, uniformly achieving accuracy rates under 10%, indicating that they are unable to reliably distinguish between conserving and non-conserving scenarios. For example, the second best-performing model on the standard benchmark (INTERN-VL2 4B), which reached nearly 50% average accuracy, drops to 0.001% in this joint evaluation, suggesting that its success on conservation tasks is driven almost entirely by a strong bias over quantity invariance rather than genuine reasoning about physical transformations. This finding further supports the conclusion that models fail to internalize structured physical reasoning and instead rely on brittle default strategies of quantity assessment.

4.3.2 Dissociating Sources of Bias.

We show that model performance on conservation tasks is largely driven by shallow heuristics rather than grounded physical reasoning. To disentangle the source of this bias—whether it arises from visual features or linguistic priors—we rerun the same prompts using fully white, content-free images, while keeping all textual inputs constant. Model responses were evaluated as if they were answering standard conservation tasks. If performance were driven by visual cues, models should operate at chance when visual input is removed. Conversely, systematic deviations from chance would indicate a reliance on linguistic biases embedded during pretraining—favoring either conservation (bias toward invariance) or non-conservation (bias toward perceptual change).

The results reveal heterogeneous linguistic biases across models, with no clear relationship to overall performance on the original benchmark (see Appendix H for details). Most models exhibit strong priors, systematically favoring either conservation or non-conservation even in the absence of any visual content, while a smaller subset performs near chance, suggesting greater reliance on visual

features. These divergent patterns underscore the influence of pretraining data and architectural inductive biases, independent of any genuine capacity for physical reasoning.

4.3.3 Does Scaling of Model Size Help?

The advancement of LLMs has been closely tied to the empirical scaling law—predictable power-law improvements in performance with increased compute, parameters, and training data (Kaplan et al., 2020; Henighan et al., 2020; Zhai et al., 2022)—as well as emergence, the abrupt appearance of qualitatively new abilities as models grow larger (Wei et al., 2022; Aghajanyan et al., 2023; Bubeck et al., 2023; Berti et al., 2025). This raises a natural question: *Does the capacity to understand physical transformations and conservation similarly emerge with scale?* Our results suggest not: performance shows no significant relationship to model size, with larger models exhibiting substantial variability in accuracy across both conservation tasks and non-conserving controls (as shown in Figure 2B). These findings indicate that scaling alone is insufficient for current VLMs to acquire the capacity for genuine reasoning about physical transformations.

4.4 Different Prompting Strategies, Frame Numbers, and Sampling Methods

We further analyzed model performance across three experimental factors—prompt type, frame count, and frame sampling method—evaluated separately for Number & Length versus Volume & Size conservation tasks. We highlight the main conclusions as below.

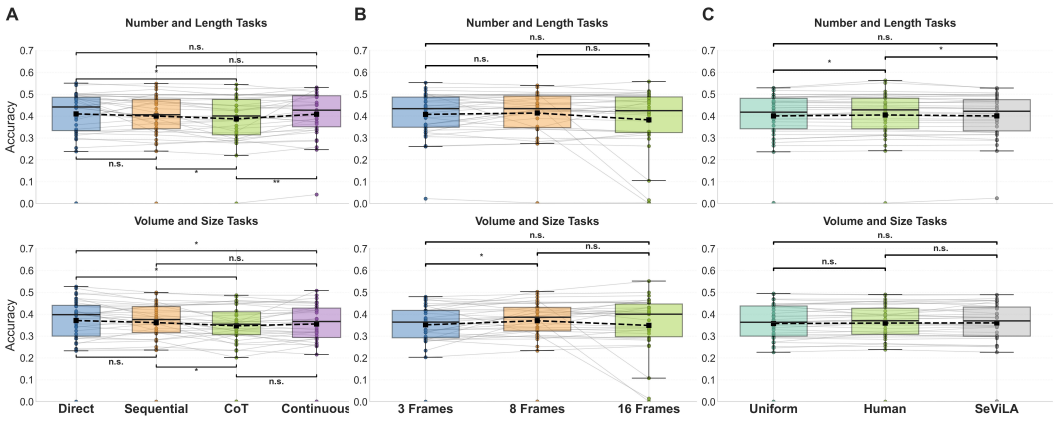


Figure 3: Model performance showing main effects by (A) prompt type, (B) number of frames, and (C) frame sampling method. Each panel averages across the other two factors from the full factorial design (4 prompts × 3 frame counts × 3 extraction methods).

"Continuous" cue fails; CoT makes it worse. We evaluate the impact of different prompting strategies across both conservation tasks and non-conserving controls (Figure 3A). Averaged across quantitative properties, linguistic scaffolding provides limited benefit: none of the three prompting types—Sequential, Chain-of-Thought (CoT), or Continuous—outperform the Direct Question baseline. In fact, performance is significantly worse when using conceptual cues that frame transformations as continuous processes. For Number and Length tasks, Sequential prompts lead to significantly lower accuracy than Direct questions ($t = 2.947$, $p = 0.00585$), as do CoT ($t = 2.701$, $p = 0.01083$) and even Direct vs. Continuous ($t = 2.663$, $p = 0.01188$), all indicating reliable performance drops. CoT prompting performs the worst overall, suggesting that explicitly verbalizing the reasoning process may amplify reliance on brittle heuristics. These results indicate that prompting alone does not improve—and may even impair—transformation reasoning in VLMs.

Increased Temporal Resolution Helps Little. We assessed whether increasing temporal resolution improves model performance by comparing tasks rendered with 3, 8, or 16 visual frames (Figure 3B). Across Number & Length tasks, no statistically significant improvements were observed (all $p > 0.05$), and only a marginal effect was found for Volume & Size: models performed slightly better when given 8 frames compared to 3 frames ($t = -2.075$, $p = 0.04583$). This suggests that

simply providing more visual information over time in general does not enhance models' capacity to track or reason about physical transformations. Despite higher frame counts offering richer depictions of dynamic processes, current VLMs appear unable to integrate temporally extended input. Instead, they continue to rely on static visual heuristics extracted from individual frames, regardless of sequence length or transformation complexity.

Human-aligned Frame Extraction Aids Performance. Across the Number & Length tasks, models achieve significantly higher accuracy when presented with frames selected by humans based on intuitive understanding of the transformation, compared to both uniform temporal sampling ($t = -2.526$, $p = 0.01653$) and LLM-based extraction using SEVILA ($t = 2.523$, $p = 0.01664$). However, this advantage did not extend to the Volume & Size tasks, where no significant performance differences were observed across frame selection strategies (Figure 3C). These findings suggest that human-aligned frame extraction can aid performance in tasks where transformation cues are helpful but not essential, likely by highlighting snapshots that facilitate heuristic-based reasoning or static quantity assessment. In contrast, when success requires tracking continuous transformations, such selection strategies confer no benefit.

4.5 Insights From Transformation-helpful Tasks

Number & Length tasks, categorized as Transformation-helpful, may not necessarily require understanding the physical transformation process, as judging the final frame alone can suffice. For example, the Number task can be solved by directly counting the coins in the last frame. Thus, it is important to evaluate whether conservation plays a significant role in Transformation-helpful tasks or whether they can be solved by simply counting or perceiving length in the last frame. To explore this, we compare model performance on multi-frame versus last-frame-only inputs.

We found no significant difference in mean accuracy between the two conditions across models (both $p > 0.05$). However, the last-frame-only condition exhibited substantially greater variance (Number: $Std = 0.0750$ vs. 0.1861; Length: $Std = 0.0810$ vs. 0.1490), with some models performing notably better than others (see further details in Appendix I). This suggests that certain models may be more attuned to static quantity assessment (e.g., counting or length comparison), effectively solving tasks using cues from the final frame alone. Yet, these same models perform markedly worse on multi-frame tasks involving continuous physical transformations—even when such reasoning is not strictly necessary. This underscores that effective reasoning about physical processes requires more than simple heuristics. In fact, the inability to engage in such reasoning may nullify any advantage in static assessments when tasks demand integration of temporal dynamics.

5 Conclusion

This study introduces a cognitively inspired benchmark that systematically evaluates whether VLMs can reason about physical transformations through both conservation tasks and non-conserving controls. Our findings reveal that current models are consistently incapable of integrating temporal information to track physical properties across dynamic visual inputs. Models achieving high performance on conservation tasks do so by relying on default heuristics over quantitative assessment, leading to inverse performance trade-offs when confronted with non-conserving scenarios. These results underscore a fundamental gap in structured physical reasoning and point to a key challenge for developing more grounded, temporally aware AI systems capable of systematic inference in real-world environments.

Conservation tasks offer a foundational benchmark for evaluating whether models can reason about physical transformations in quantitative domains. Future work may extend this paradigm to more complex physical settings, including richer object dynamics, multimodal inputs, or uncertainty, and evaluate conservation reasoning in goal-directed contexts such as planning or tool use. Such extensions are crucial for testing whether VLMs can support structured physical reasoning in real-world scenarios.

6 REPRODUCIBILITY STATEMENT

Data Availability and Transparency: We provide complete documentation of our data generation process, including precise descriptions of task construction, sampling procedures, and quality control protocols. Representative samples for each quantitative domain are included in the supplementary material. The full dataset—comprising all 13,824 tasks spanning conserving and non-conserving variants—will be released publicly upon publication to ensure transparency and enable direct replication of our experiments.

Evaluation Pipeline Standardization: All evaluations are conducted using publicly released VLMs with default settings from their official implementations. Our pipeline adheres to established evaluation standards and reproduces published results on existing benchmarks such as MMBench using the same configurations. This alignment validates the reliability of our evaluation setup and ensures fair comparisons across all models in our benchmark.

7 ETHICS STATEMENT

This research presents no identifiable ethical concerns. It involves purely computational evaluations of publicly available VLMs using synthetic image-text inputs. No human subjects, private data, or sensitive content are involved at any stage. All models evaluated are publicly released, and all tasks are generated programmatically under controlled conditions to avoid harm or misuse.

8 LLM USAGE STATEMENT

Large Language Models (LLMs) were used to assist with the writing and editing of this manuscript. Specifically, an LLM was employed to refine language, improve clarity, and enhance the overall readability of the text. Tasks included grammar correction, sentence rephrasing, and improving narrative flow across sections. Importantly, the LLM was not involved in any part of the research process, including ideation, methodological design, or data analysis. All scientific content, experimental decisions, and conceptual contributions were developed solely by the authors. The authors retain full responsibility for the content of the manuscript, including any text edited or suggested by the LLM. We confirm that the use of the model adheres to ethical guidelines and does not constitute plagiarism or scientific misconduct.

REFERENCES

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pp. 265–279. PMLR, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Kelsey R Allen, Kevin A Smith, and Joshua B Tenenbaum. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, 117(47):29302–29310, 2020.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Renee Baillargeon. Representing the existence and the location of hidden objects: Object permanence in 6-and 8-month-old infants. *Cognition*, 23(1):21–41, 1986.

Renée Baillargeon. Young infants' reasoning about the physical and spatial properties of a hidden object. *Cognitive development*, 2(3):179–200, 1987.

Renée Baillargeon and Susan Carey. Core cognition and beyond: The acquisition of physical and numerical knowledge. *Early childhood development and later outcome*, 1, 2012.

Renee Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985.

Renée Baillargeon, Marcia Graber, Julia Devos, and James Black. Why do young infants fail to search for hidden objects? *Cognition*, 36(3):255–284, 1990.

Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G Krishnan. Synthetic vision: Training vision-language models to understand physics. *arXiv e-prints*, pp. arXiv–2412, 2024.

Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G Krishnan. Physics context builders: A modular framework for physical reasoning in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7318–7328, 2025.

Lawrence W Barsalou. Challenges and opportunities for grounding cognition. *Journal of Cognition*, 3(1), 2020.

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

Sian Beilock and Susan Goldin-Meadow. Gesture changes thought by grounding it in action. *Psychological Science*, 21(11):1605–1610, 2010.

Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. Emergent abilities in large language models: A survey. *arXiv preprint arXiv:2503.05788*, 2025.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu

Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL https://arxiv.org/abs/2307.15818.

Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Luke Burke. On the tunnel effect. *Quarterly Journal of Experimental Psychology*, 4(3):121–138, 1952.

Luca M Schulze Buschoff, Konstantinos Voudouris, Elif Akata, Matthias Bethge, Joshua B Tenenbaum, and Eric Schulz. Testing the limits of fine-tuning to improve reasoning in vision language models. *arXiv preprint arXiv:2502.15678*, 2025.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Camillo Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven Frankland, Tom Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. *Advances in Neural Information Processing Systems*, 37:113436–113460, 2024.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. Vision-language models can self-improve reasoning via reflection. *arXiv preprint arXiv:2411.00855*, 2024a.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms, 2024b. URL https://arxiv.org/abs/2406.07476.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025a.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding, 2025b. URL https://arxiv.org/abs/2501.16411.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. URL https://arxiv.org/abs/2303.03378.

Jonathan I Flombaum and Brian J Scholl. A temporal same-object advantage in the tunnel effect: facilitated change detection for persisting objects. *Journal of Experimental Psychology: Human Perception and Performance*, 32(4):840, 2006.

Jerry A Fodor. *The Language of Thought*. MIT Press, 1975.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation, 2024. URL https://arxiv.org/abs/2309.02561.

Susan Goldin-Meadow and Sian Beilock. Action's influence on thought: the case of gesture. *Perspectives on psychological science : a journal of the Association for Psychological Science*, 5(6): 664–674, 2010.

Gustaf Gredebäck and Claes von Hofsten. Infants' evolving representations of object motion during occlusion: A longitudinal study of 6- to 12-month-old infants. *Infancy*, 6(2):165–184, 2004. doi: 10.1207/s15327078in0602_2.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Olivier Houdé. Numerical development: From the infant to the child. *Cognitive Development*, 12 (3):373–391, 1997.

Olivier Houdé, Arlette Pineau, Gaëlle Leroux, Nicolas Poirel, Guy Perchey, Céline Lanoë, Amélie Lubin, Marie-Renée Turbelin, Sandrine Rossi, Grégory Simon, Nicolas Delcroix, Franck Lamberton, Mathieu Vigneau, Gabriel Wisniewski, Jean-René Vicet, and Bernard Mazoyer. Functional magnetic resonance imaging study of piaget's conservation-of-number task in preschool and school-age children: a neo-piagetian approach. *Journal of experimental child psychology*, 110(3):332–346, 2011.

Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463*, 2023.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1383–1391, 2015.

Dongfu Jiang, Xuan He, Huaye Zeng, Con Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *CONFERENCE*, 2023.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, and Hokin Deng. Core knowledge deficits in multi-modal language models. *arXiv preprint arXiv:2410.10855*, 2025a.

Yijiang Li, Bingyang Wang, Tianwei Zhao, Qingying Gao, Hokin Deng, and Dezhi Luo. Evaluating multi-modal language models through concept hacking. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*, 2025b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Daochang Liu, Junyu Zhang, Anh-Dung Dinh, Eunbyung Park, Shichao Zhang, Ajmal Mian, Mubarak Shah, and Chang Xu. Generative physical ai in vision: A survey, 2025. URL https://arxiv.org/abs/2501.10928.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. Mibench: Evaluating multimodal large language models over multiple images. *arXiv preprint arXiv:2407.15272*, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

Mariana Lozada and Natalia Carro. Embodied action improves cognition in children: Evidence from a study based on piagetian conservation tasks. *Frontiers in psychology*, 7(393), 2016.

Dezhi Luo, Yijiang Li, and Hokin Deng. The philosophical foundations of growing ai like a child. *arXiv preprint arXiv:2502.10742*, 2025.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation, 2024a. URL https://arxiv.org/abs/2410.05363.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*, 2024b.

Melanie Mitchell and David C Krakauer. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles?, 2025. URL https://arxiv.org/abs/2501.09038.

Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. Pivot: Iterative visual prompting elicits actionable knowledge for vlms, 2024. URL https://arxiv.org/abs/2402.07872.

Kaleb Newman, Shijie Wang, Yuan Zang, David Heffren, and Chen Sun. Do pre-trained vision-language models encode object states? *arXiv preprint arXiv:2409.10488*, 2024.

Nicholaus S Noles, Brian J Scholl, and Stephen R Mitroff. The persistence of object file representations. *Perception & Psychophysics*, 67(2):324–334, 2005.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3170–3180, 2023.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image, 2020. URL https://arxiv.org/abs/2004.10796.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Cripp-vqa: Counterfactual reasoning about implicit physical properties via video question answering. *arXiv preprint arXiv:2211.03779*, 2022.

Giovanni Pezzulo, Lawrence W Barsalou, Angelo Cangelosi, Martin H Fischer, Ken McRae, and Michael J Spivey. Computational grounded cognition: a new alliance between grounded cognition and computational modeling. *Frontiers in psychology*, 3:612, 2013.

Jean Piaget. *The Psychology of Intelligence*. Harcourt, Brace, 1950.

Jean Piaget. *The Origins of Intelligence in Children*. International Universities Press, 1952.

Jean Piaget. *The Child's Conception of Number*. W.W. Norton and Company, 1965.

Jean Piaget and Bärbel Inhelder. *The Psychology of the Child*. Basic Books, New York, 1969.

Luis S Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature human behaviour*, 6(9): 1257–1267, 2022.

Nicolas Poirel, Grégoire Borst, Grégory Simon, Sandrine Rossi, Mathieu Cassotti, Arlette Pineau, and Olivier Houdé. Number conservation is related to children's prefrontal inhibitory control: an fmri study of a piagetian task. *PloS one*, 7(7):e40802, 2012.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, et al. Phybench: Holistic evaluation of physical perception and reasoning in large language models. *arXiv preprint arXiv:2504.16074*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv: 2103.00020*, 2021.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. *arXiv preprint arXiv:2407.06581*, 2024.

Sunayana Rane, Alexander Ku, Jason Baldridge, Ian Tenney, Tom Griffiths, and Been Kim. Can generative multimodal models count to ten? *Proceeding of the Annual Meeting of the Cognitive Science Society*, 46:1235–1241, 2024.

Brian J Scholl. Object persistence in philosophy and psychology. *Mind & Language*, 22(5):563–591, 2007.

Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1):96–106, 2025.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling, 2024. URL https://arxiv.org/abs/2401.15977.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024.

Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological review*, 99(4):605, 1992.

Elizabeth S Spelke, Gary Katz, Susan E Purcell, Sheryl M Ehrlich, and Karen Breinlinger. Early knowledge of object motion: Continuity and inertia. *Cognition*, 51(2):131–176, 1994.

Elizabeth S Spelke, Roberta Kestenbaum, Daniel J Simons, and Debra Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British journal of developmental psychology*, 13(2):113–142, 1995.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Arnaud Viarouge, Olivier Houdé, and Grégoire Borst. The progressive 6-year-old conserver: Numerical saliency and sensitivity as core mechanisms of numerical abstraction in a piaget-like estimation task. *Cognition*, 190:137–142, 2019.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024a.

Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Peng Xia, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, et al. Physunibench: An undergraduate-level physics reasoning benchmark for multimodal models. *arXiv preprint arXiv:2506.17667*, 2025.

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. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024b. URL https://arxiv.org/abs/2409.12191.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, and Xu Jia. Vlipp: Towards physically plausible video generation with vision and language informed physical prior, 2025. URL https://arxiv.org/abs/2503.23368.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36:26650–26685, 2023.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning, 2019. URL https://arxiv.org/abs/1811.10830.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12104–12113, 2022.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024b. URL https://arxiv.org/abs/2410.02713.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B Tenenbaum, and Chuang Gan. Contphy: Continuum physical concept learning and reasoning from videos. *arXiv preprint arXiv:2402.06119*, 2024.

APPENDIX

A DATA CURATION

Curation and Quality Control. *ConservationBench* was curated by three annotators with college-level training in cognitive science or computer science. Each video underwent two independent cross-review passes; items failing to meet design criteria were removed or revised.

Data Acquisition. All videos were captured under standardized recording conditions using a fixed camera setup, with consistent lighting and background held constant within each property category. Each transformation was carefully scripted to ensure visual clarity, reproducibility, and minimal ambiguity.

Design Principles. To ensure conceptual integrity and interdisciplinary rigor, we adopt three design criteria for each item: (i) *Discriminativeness*—tasks are constructed such that models lacking the targeted knowledge are systematically driven toward incorrect responses; (ii) *Minimal confounding*—instances are designed to minimize reliance on ancillary skills (e.g., object recognition); and (iii) *Minimal textual shortcuts*—tasks cannot be solved using textual cues alone and instead require genuine multimodal reasoning.

B PROMPTING STRATEGY

Reasoning about conservation often requires interpreting the transformation as a continuous process across the videos or sequence of frames. To examine how prompts influence temporal integration and transformation-based reasoning, we design four prompt types, each progressively enhancing the model's awareness of the underlying continuous process, as summarized in Table 2.

Table 2: **Four different prompt format used in our benchmark.**

Prompt Type	Prompt Example
Direct Question	Is the number of coins in the upper row the same as in the lower row in the final image?
"Sequential" Prompt	Please process the images below sequentially, and then answer: Is the number of coins in the upper row the same as in the lower row in the final image?
CoT Prompt	Please process the images below sequentially. First describe what happens across the images, then answer: Is the number of coins in the upper row the same as in the lower row in the final image?
"Continuous" Prompt	The above images represent a continuous process. Please answer: Is the number of coins in the upper row the same as in the lower row in the final image?

Together, these prompting strategies enable us to evaluate how different forms of linguistic scaffolding shape model engagement with visual dynamics. The "Sequential" and CoT prompts encourage frame-by-frame perception with step-by-step reasoning, directing attention to frame-wise visual evidence. In contrast, the "Continuous" prompt explicitly presents the multi-frame input as a continuous process, offering a conceptual cue to support conservation reasoning.

C EXAMPLE INPUT

To provide clarity on the exact format of inputs provided to models, we present a complete example task below, including both the visual frames and the full textual prompt.

Task: Conservation of Number (Conserving condition)

Task Configuration: This example demonstrates a Number conservation task using Uniform extraction method with 8 frames and Direct Question prompt format.

Visual Input: The model receives a sequence of frames extracted from the video in temporal order (Frame 1 through Frame 8), ensuring that the transformation process is presented chronologically without any frame order disruption. Figure 4 shows an example with 8 frames, where frames are sampled uniformly across the video timeline. The first frame shows the initial state (two rows of coins with equal numbers), intermediate frames capture the transformation process (spreading one row), and the final frame shows the end state (one row spread out while maintaining the same number of coins).

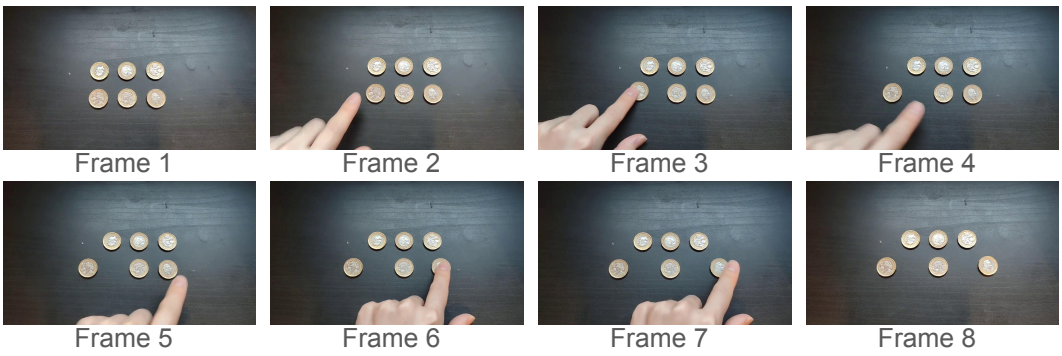


Figure 4: **Example visual input:** A sequence of 8 frames from a number conservation task, showing the initial state, transformation process, and final state.

Textual Input: Below is the structure of the prompt provided to the model (using the "Direct Question" format). The [Image] placeholders indicate where the corresponding frames from Figure 4 are embedded in the actual input:

```
Frame 1: [Image]
Frame 2: [Image]
Frame 3: [Image]
Frame 4: [Image]
Frame 5: [Image]
Frame 6: [Image]
Frame 7: [Image]
Frame 8: [Image]

Is the number of coins in the upper row the same as in the lower row
in the final image?

Please choose one of the following options:
(A) No, the lower row has more coins.
(B) No, the upper row has more coins.
(C) Yes, they are the same.
```

Ground Truth: Option (C) - Yes, they are the same.

Alternative Prompt Formats: For other prompt types, the question is prefixed with additional instructions. For example, the "Sequential" format would begin with "Please process the images below sequentially, and then answer: [question]", while the CoT format would include "Please process the images below sequentially. First describe what happens across the images, then answer: [question]". See Table 2 for details on all four prompt formats.

D MODEL INFERENCE

We evaluate 34 VLMs spanning diverse architectures, training regimes, and parameter scales, including mainstream commercial systems and advanced open-source models ranging from 1B to 76B parameters. Inference is conducted on a cluster equipped with 8× NVIDIA H100 (80 GB) GPUs. As a practical policy, models of 1–13B parameters typically run on a single GPU; 13–32B on two GPUs; 32–70B on four GPUs; and $>$70B on all eight GPUs.

To preserve fidelity and reproducibility, we adhere to configurations and reference implementations from the official codebases, avoiding unnecessary modifications. We build a scalable evaluation framework supporting parallel execution and compartmentalized environments. Inference jobs are distributed across GPUs via a dynamic scheduler that maximizes utilization and minimizes idle time. We additionally develop a lightweight modality-verification suite that prompts each model to summarize the media information it receives, and then the responses are checked by human reviewers to verify correct input routing and modality handling in our inference pipelines.

E EVALUATION

Rule-based matching degrades when model outputs are complex (e.g., chain-of-thought), yielding elevated false positives/negatives and requiring continual template optimization to cover corner cases. LLM-based matching better identifies the intended choice within complex free-form text, but it can hallucinate—especially when a brief answer is embedded within extensive context. To balance these trade-offs, we introduce Hybrid Matching, which prioritizes deterministic template matching to extract answers from VLM responses and, on failure, falls back to an ensemble of four LLM judges (Qwen2.5-72B-Instruct, Mixtral-8x7B-Instruct-v0.1, DeepSeek-R1-Distill-Llama-70B, and Llama-3.1-70B). The ensemble decision is accepted only if at least three of four models return a consistent extraction; otherwise, the mapping is deemed unsuccessful. By coupling the precision of regular-format extraction with the semantic flexibility of LLM adjudication, Hybrid Matching delivers more robust and reliable mappings across diverse response styles.

F COUNTERBALANCING CONDITIONS

Table 3: **Counterbalanced variations of task-irrelevant features** Each unique combination of parameter values yields 48 distinct task instances per domain.

Domain	Parameter	Variations
Number	P1: Object Type	2 variants (Uniform, Mixed)
	P2: Mapping Shift	2 variants (Lower vs. Upper row moved)
	P3: Distance Spread	2 variants (Near, Far)
	P4: Number of Objects	6 variants (3–8 coins)
	Total combinations:	$2 \times 2 \times 2 \times 6 = 48$
Length	P1: Object Type	2 variants (Uniform, Mixed)
	P2: Mapping Shift	2 variants (Lower vs. Upper straw moved)
	P3: Distance Moved	2 variants (Near, Far)
	P4: Direction	2 variants (Left, Right)
	P5: Transformation Action	3 variants (Slide, Rotate, Vertical)
	Total combinations:	$2 \times 2 \times 2 \times 2 \times 3 = 48$
Volume	P1: Liquid Color	8 variants
	P2: Glass Transaction	2 variants (Tall $\rightarrow$ Short, Short $\rightarrow$ Tall)
	P3: Liquid Volume	3 variants (Small, Medium, Large)
	Total combinations:	$2 \times 8 \times 3 = 48$
Size	P1: Object Color	8 variants
	P2: Shape Transformation	6 variants (Crossing Sphere, Cylinder, Plane)
	Total combinations:	$6 \times 8 = 48$

G COMBINED MODEL PERFORMANCE BY QUANTITATIVE PROPERTIES

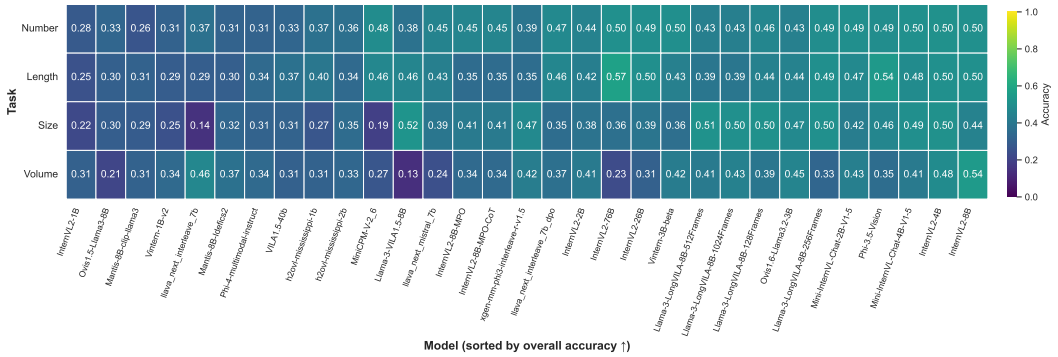


Figure 5: **Model performance by conservation task properties.** Models consistently perform around chance accuracy and significantly below humans, with no significant differences across quantitative properties.

H MODEL RESPONSE ON WHITE CONTROL AS COMPARED TO CONSERVATION TASK SCENARIOS.

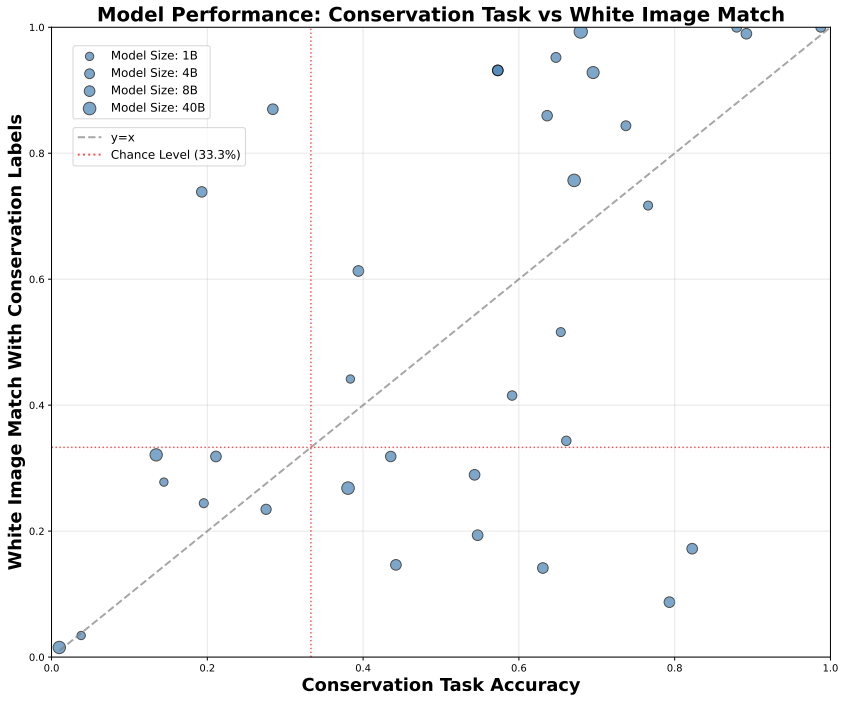


Figure 6: **Model response on white control as compared to conservation task scenarios.** We assess model response to prompts asking about physical transformation, match them in accordance to conservation labels, and compared their accuracy with conservation task performances, revealing distinct sources of bias across linguistic and visual features.

I COMPARING MULTI-FRAMES VS LAST FRAME PEROFRMANCE ON TRANSFORMATION-HELPFUL TASKS

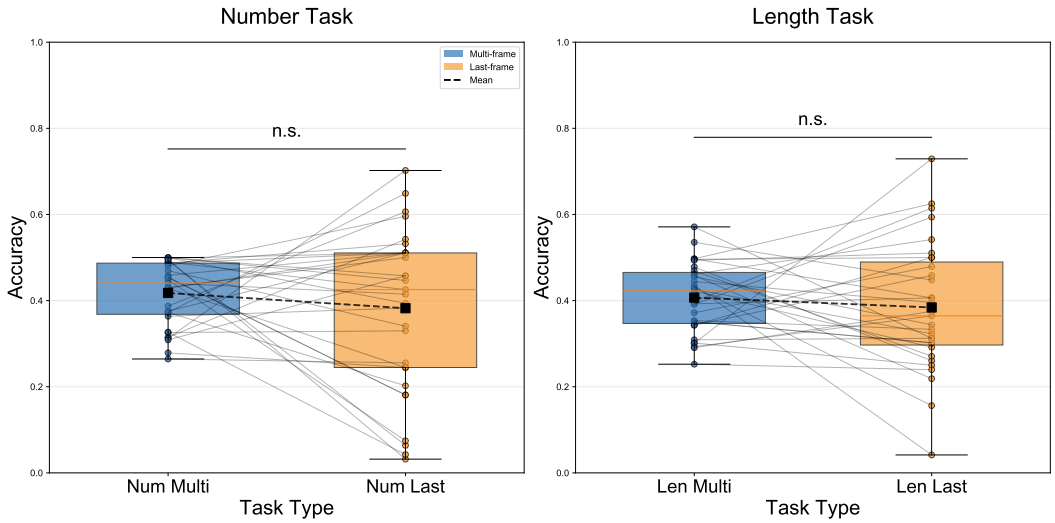


Figure 7: **Model performance on Multi-frames vs Last Frame Conditions in Transformation-helpful Tasks.** While mean accuracy does not differ significantly between the two conditions, variance is substantially higher in the last-frame-only setting.