Seeing Cause and Time: a Visually Grounded Evaluation of Multimodal Models

Salvatore Ergoli[1,*], Alessandro Bondielli[1,2,*] and Alessandro Lenci[1]

[1]*CoLing Lab, Department of Philology, Literature and Linguistics, University of Pisa*

[2]*Department of Computer Science, University of Pisa*

Abstract

Reasoning about causal and temporal relationships is fundamental to human intelligence but poses a persistent challenge for AI. Vision-Language Models (VLMs) offer a promising path towards more robust conceptual understanding by grounding language in perception. However, it is unclear if this grounding enables genuine, human-like reasoning. We investigate this question by focusing on the causal and temporal abilities of two leading VLMs using a novel multimodal dataset derived from the ExpliCa dataset. Through a series of carefully designed tasks, we isolate their performance on visual-only input versus combined visual-textual inputs. Our results show that while models exhibit some reasoning capability, they are hindered by a marked "iconicity bias": their performance degrades on relations where the perceptual sequence of images mismatches the logical event order (i.e., anti-iconic). This reliance on simple visual heuristics suggests that their high-level reasoning failures may be symptomatic of a more fundamental, fragile visual understanding.

Keywords

Multimodality, Causal Reasoning, Temporal Reasoning, Vision Language Models

1. Introduction

The ability to comprehend and reason about causal and temporal relationships is a cornerstone of human cognition, underpinning our capacity to understand narratives, predict outcomes and navigate the complexities of the world. We effortlessly discern why an event occurred and the sequence in which events unfolded, integrating information from various modalities. While Large Language Models (LLMs) have demonstrated remarkable fluency in generating text that describes such relationships, a critical question remains: do they possess a genuine, human-like understanding of these fundamental concepts or do they primarily rely on sophisticated pattern matching learned from vast textual corpora [1, 2]? This distinction is crucial, as linguistic proficiency can sometimes obscure deeper cognitive limitations, a phenomenon known as the "fallacy of language as thought" [3].

Recent advancements have led to the development of Vision Language Models (VLMs), which are trained on both textual and visual data [4, 5]. This multimodal grounding offers a potential pathway to richer, more robust representations of concepts, potentially bridging the gap between linguistic competence and conceptual understanding, as human meaning representation itself relies on multiple modalities [6, 7]. However, the extent to which this enriched inputs translate to superior causal and temporal reasoning capabilities remains an area in need of investigation.

This paper contributes to this line of inquiry by conducting a focused analysis of the causal and temporal reasoning abilities of two distinct, current generation multimodal models: `Llama-11b-vision` and `Gemini-flash-2.0`. We explore their performances with a series of carefully designed tasks on a multimodal version of the ExpliCa dataset, which explicitly combines causal and temporal relations [8]. Our objective is twofold: first, we aim to assess the models' capacity to infer these relations from visual input alone; second, we want to address how their performances change when the visual stimuli accompany the textual captions. We do so by comparing models with differing architectures and parameter counts and varying the input modalities.

Our experimental methodology involves i) constructing a novel image dataset, that we name **Visual-ExpliCa**, aligned with the ExpliCa dataset,[1] and ii) evaluating the models on five distinct tasks of increasing difficulty. The tasks range from directly identifying the type of relationship (causal vs. temporal) and specifying the antecedent and consequent from image-only input, to selecting the correct linguistic connective and judging the overall acceptability of an event when both images and textual descriptions are provided. Through this graduated approach, we seek to disentangle the models' visual inferencing capabilities from their ability to integrate multimodal information.

Our findings reveal that while both models demonstrate capabilities beyond chance in interpreting visual

[1]https://github.com/Unipisa/explica

sequences, they exhibit distinct strengths, weaknesses and biases, particularly struggling with anti-iconic relations (i.e., when the sequence of events is inverted compared to their chronological and/or logical-causal order) when relying solely on visual input. This suggests that current VLMs, despite their multimodal training, may still heavily favour direct, sequential interpretations of visual information for complex reasoning tasks.

2. Related works

A growing body of work focuses on assessing the reasoning abilities of pre-trained models, particularly in the domain of causality. LLMs have been evaluated on various causal tasks which reveals that their grasp of formal causality is often superficial and prone to heuristic-based errors. A key development in rigorously probing these limits is the CLADDER dataset [9], which moves beyond commonsense questions by grounding them in symbolic queries derived from an oracle causal inference engine. By evaluating models against the formal rungs of Pearl's Ladder of Causation [10], the authors found that even with bespoke prompting strategies like CAUSALCOT, LLMs struggle significantly with formal, rule-based inference. This concern over the fallibility of LLMs causal understanding is echoed by other research, which shows models are susceptible to inferring causality from simple positional cues or temporal precedence (*post hoc fallacy*) and struggle to infer causal links from counterfactual evidence, suggesting a reliance on memorized heuristics rather than deep reasoning [11]. In another work was proposed a novel architecture (CARE-CA) [12] that integrates explicit causal knowledge from resources like ConceptNet with implicit reasoning patterns from LLMs, enhanced by counterfactual analysis.

This susceptibility to temporal fallacies underscores a critical prerequisite for robust causal reasoning: a coherent understanding of time itself. However, research demonstrates that LLMs' internal model of time is fragile. Authors in [13] identify several key failure modes, including temporal shifts, invariance and inertia, where models either disregard the specific time in a query or fail to update long-held facts. Recognizing that direct reasoning over unstructured text may be the source of this fragility, some approaches focus on actively mitigating these flaws. The TG-LLM framework, for instance, proposes a two-step process: first translating unstructured text into a formal temporal graph and then fine-tuning the LLM to perform Chain-of-Thought reasoning over this explicit structure [14]. This methodological shift from implicit to explicit representation significantly improves performance, highlighting that the reasoning deficit may lie more in parsing complexity than in logical inability.

The challenge of causal reasoning becomes even more pronounced when extending from the linguistic to the multimodal domain, where models must integrate visual evidence with abstract knowledge. Recent benchmarks reveal that the performance of state-of-the-art VLMs is often no better than random chance. The MuCR benchmark [15], designed to test the inference of cause-and-effect from visual cues alone, found that models either suffer from inadequate visual perception or are biased by their language priors to the point of ignoring contradictory visual evidence. This deficiency is not merely about identifying simple causal chains. The NL-EYE benchmark, which frames abductive reasoning as a visual entailment task, found that VLMs perform at or below random baselines on a task humans find trivial [16]. Crucially, the failure was not one of logic—when given textual descriptions of the scenes, the models succeeded. The breakdown occurs in visual interpretation, where models are distracted by superficial cues and fail to grasp the underlying commonsense relationships. This points to a fundamental gap between a model's linguistic reasoning capabilities and its ability to ground that reasoning in the perceptual world. Similarly, the TemporalVQA benchmark tests models on temporal order understanding and time-lapse estimation between images [17]. Their conclusions reveal that even top-tier models perform at or below random chance, are highly sensitive to image layout and rely on superficial spatial cues rather than genuine temporal comprehension.

3. The Visual-ExpliCa Dataset

The empirical investigation presented in this paper relies on a carefully constructed dataset, specifically created to align visual stimuli with textual ones from the ExpliCa dataset [8]. ExpliCa features 600 unique events, each represented by a pair of sentences. These pairs are linked by an explicit connective that establishes one of three relationship types: causal (so, because), temporal (then, after) or unrelated. The connectives define the nature and directionality of the relationship between the two sentences. Specifically, this directionality distinguishes between iconic relations, where the order of sentences reflects the chronological or causal sequence of events (i.e., with connectives *so* and *then*), and anti-iconic relations, where the presentation order is inverted relative to the logical flow (i.e., with connectives *because* and *after*). Explicit connectives for sentence pairs where selected via crowdsourcing experiments [8]. Additionally, ExpliCa is controlled for potential confounding biases, such as Lexical Association Bias (ensuring that word co-occurrences within sentence pairs do not disproportionately favor certain relationship types) and Frequency Bias (ensuring that the linguistic structures representing different relations are comparably frequent in natural language).

This makes it a robust resource for evaluating genuine reasoning rather than statistical shortcuts.

In building Visual-ExpliCa, we focused exclusively on the causal and temporal relations, excluding the *unrelated* category of the original dataset. In order to collect visuals matching sentences in the dataset, we first conducted some pre-processing steps. These involved i) lemmatisation, to mitigate data sparsity issues and to alleviate issues with VLMs struggling with temporal dimensions encoded in verb conjugations [18], and ii) NER, specifically to replace people NEs with generic placeholders (e.g., "Matteo" is replaced by "[PERSON]"), and prevent image retrieval to focus on specific individuals rather than the core actions and concepts of the sentence. For pre-processing, we used SpaCy.[2]

3.1. Images Collection

Images to match sentences of ExpliCa were mostly collected from the Fondant-CC-25M dataset. [3] It is a large-scale image corpus derived from CommonCrawl, composed exclusively of images with Creative Commons licenses. This choice ensures ethical usage and avoids copyright issues prevalent in many traditional image datasets. To retrieve images, we used the *clip-retrieval* library.[4] This tool leverages CLIP (Contrastive Language-Image Pre-Training) [19] to find images whose embeddings are semantically closest to the text query's embedding. For each sentence, we selected the 10 images with the highest CLIP score. Then, to ensure a reasonable degree of semantic alignment between the visual and textual components, we conducted a further manual review to select the final image for each single sentence.

For a small number of sequences we were not able to retrieve high-quality descriptions. To address these cases, we resorted to text-to-image generation. Specifically, we used the Segmind Stable Diffusion model[5] to create visual representations for captions that were too abstract or specific for the retrieval process. The generative approach was required for 39 individual captions (out of the 778 total captions in the final dataset).

Nevertheless, a smaller subset of captions proved intractable. Specifically, for 12 sentence-pairs of the original dataset, it was not possible to obtain a suitable image for at least one of the two descriptions, either through retrieval or generation. We chose to exclude the entire sentence-pair from the final analysis to ensure the quality and coherence of the dataset. Consequently, the final curated multimodal dataset used for our experiments consists of 388 event pairs. Table 1 shows the distribution of categories in the dataset.

---

[2]spacy.io
[3]https://huggingface.co/datasets/fondant-ai/fondant-cc-25m
[4]https://github.com/rom1504/clip-retrieval
[5]https://huggingface.co/segmind/SSD-1B

Connective	Relation, Direction	Count
so	Caus., Ic.	106
then	Temp., Ic.	105
because	Caus., A-Ic.	99
after	Temp., A-Ic.	78
Total		**388**

Table 1
Distribution of event pairs in the final curated dataset, categorized by connective type. Causal and Temporal are abbreviated with Caus. and Temp. respectively. Iconic and Anti-Iconic are abbreviated with Ic. and A-Ic. respectively.

Figure 1 shows an example of a sentence-pair for a Causal, Iconic (Caus., Ic.) including visuals from the final dataset.

Sentence	Text	Image
A	[PERSON]'s clothes got dirty.	
B	[PERSON] put his clothes in the washing machine.	

Figure 1: An example of a sentence pair with images; the relation in this case is Causal, Iconic.

4. Experimental Setup

4.1. Models

To evaluate the capabilities of current VLMs in causal and temporal reasoning, we selected two prominent models representing distinct architectural families and development origins: `Llama-11b-vision` from Meta AI [20] and `Gemini 2.0 Flash` from Google DeepMind [21].

`Llama-11b-vision` is part of the Llama 3.2-Vision family of models. It was released by Meta in September 2024. These models are designed to be natively multimodal, capable of processing paired image and text inputs to generate textual outputs. Its architecture builds upon the Llama 3.1 LLM family. The instruction-tuned versions of Llama-3.2-Vision, including the variant used here, are optimized through a combination of Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) [20]. Authors argue that this alignment process aims to enhance the model's utility, safety and ability to follow instructions. The vision component was pre-trained on a dataset of 6 billion image-text pairs.

`Gemini 2.0 Flash` is a multimodal large language model (text, image, audio, video) with a 1M-token context

window, positioned as an upgrade over Gemini 1.5 Flash. It is reported to achieve improved efficiency and benchmark performance through a refined Mixture-of-Experts Transformer architecture and supports real-time multimodal interactions [22]. It inherits the general Gemini philosophy of deep interweaving of modalities.

We chose these models to reflect two contrasting trends in multimodal AI: Llama, an open-source and relatively small model accessible for research at modest computational cost, and Gemini Flash, a closed but comparatively compact commercial system optimized for efficiency and lower inference costs. This contrast highlights differences in openness, scale, and resource demands, providing a balanced testbed for evaluating causal and temporal reasoning.

4.2. Tasks design

To systematically probe the models' reasoning capabilities, we designed five distinct experimental tasks grounded in the Visual-ExpliCa dataset. These tasks are structured to progressively increase in complexity and are organized into two primary conditions that directly address our research objectives: assessing reasoning from visual-only input (Tasks 1 to 3) and evaluating multimodal integration (Tasks 4 and 5).

We employ a **Multimodal-Chain-of-Thought (Multimodal-CoT)** strategy for prompting in visual-only tasks. This strategy is inspired by [23], and is aimed at addressing one of the most critical failure modes in prompting VLMs, i.e. their tendency to rely on superficial visual processing and get distracted by irrelevant cues. In contrast, using Multimodal-CoT we structure the prompt to first elicit a description and interpretation of the visual information before attempting further reasoning, to establish a grounded rationale. This visual analysis then serves as the foundation for the reasoning steps needed to derive the final conclusion, effectively creating a reasoning chain [24].[6]

The first three tasks are designed to isolate the models' ability to infer causal and temporal relations relying solely on visual evidence. The model is first prompted to describe the visual content of the two images before being asked to perform the specific reasoning step. The final two tasks assess how performance vary given the support of textual data, thus evaluating the models' capacity to integrate information from both modalities. In the following, we detail each task.

---

[6]We report examples of prompts in the Appendix.

Task 1. Relation identification In the first task, the model's goal is to classify the fundamental relationship between the two visual depictions of events as either *causal* or *temporal*, regardless of the order they are presented in.

Task 2. Directionality Specification In the second task, the model's goal is to determine the logical order of the event, identifying which image represent the *antecedent* and which the *consequent*, regardless of their causal or temporal relation.

Task 3. Connective Selection In the third task, the model's goal is to provide the most appropriate linguistic connective (among *so*, *because*, *then*, and *after*) given the pair of images representing the events, in a specific order. Recall that each connective is directly associated with a Relation (*causal* or *temporal*) and a Direction of such relation (*iconic* or *anti-iconic*).

Task 4. Connective Selection With Captions The fourth task is analogous to the third task. However, in this case the model is provided with both the images and their corresponding textual description of the events from ExpliCa. This allows for a direct comparison of performance with and without linguistic context.

Task 5. Acceptability rating In the fifth and final task, we replicate one of the experiments conducted on ExpliCa in [8]. Here, the model must perform a holistic evaluation of a complete multimodal input (two images, two captions and a human-provided connective). It is tasked with providing a numerical plausibility rating from 1 to 10, simulating a human-like judgment of coherence. We chose to exclude `Llama-11b-vision` from this specific task, as preliminary tests revealed it was unreliable in consistently generating ratings in the required numerical format. This is a known issue also reported in [8]. We can speculate that it is probably due to the limited model size. Conversely, to robustly assess `Gemini-2.0-Flash` and account for output variability, we prompted it to generate five distinct ratings for each event. This was achieved by querying the model five times, each with a different temperature setting to modulate the randomness of the output. We used the average of these ratings as the final score.

4.3. Evaluation

Our evaluation strategy was designed to measure the multifaceted nature of the models' causal and temporal reasoning across the five experimental tasks. The metrics

Model	Overall Acc.	Caus. Acc.	Temp. Acc.
Gemini	0.72	0.58	**0.87**
LLaMA	0.63	**0.86**	0.40

Table 2
Results for Task 1.

were chosen to reflect the nature of each task, ranging from categorical decisions to graded plausibility judgments.

For tasks requiring a categorical decision (Tasks 1-4), we employed a "cloze test" paradigm, mirroring the evaluation approach often used for the ExpliCa dataset [8]. In this setup, the models were presented with the input (either images-only, or images and partly-hidden captions) and asked to "fill in the blank" by choosing the most suitable option from a predefined list of candidates. A response was considered correct only if it exactly matched the designated ground truth; both incorrect choices and responses that did not conform to one of the choices were marked as an error. The primary evaluation metric for these tasks was **Accuracy**. However, for Tasks 3 (Connective Selection) and 4 (Connective Selection with Captions), which involve a multi-class classification among four connectives, we also computed the F1-score. This metric provides a more balanced assessment than accuracy alone, as it considers both precision and recall for each connective class. This is particularly useful for identifying whether a model's performance is uniform across the different logical relationships or if it excels at some at the expense of others.

For Task 2 (Directionality Specification), correctness was determined by the alignment between the event order identified by the model and the iconicity status (iconic/anti-iconic) of the original pair. For example, if the model identified Image A (presented first) as the antecedent and Image B as the consequent, the answer was deemed correct only if the ground-truth connective for the original pair was iconic (i.e., "so" or "then").

Finally, for Task 5 (Acceptability Rating), evaluation was based on the Pearson correlation between the scores generated by the model and the human-provided acceptability judgments for the highest-rated connective. To ensure the values were comparable on a common scale, both the model ratings and the human judgments were first normalized using min-max technique. This allowed us to quantify the degree of alignment between the plausibility assessment of the model and of humans.

5. Results and Discussion

In this Section, we outline and discuss the results obtained by the models on all tasks. In the presentation of the results, we abbreviate Causal and Temporal Caus. and

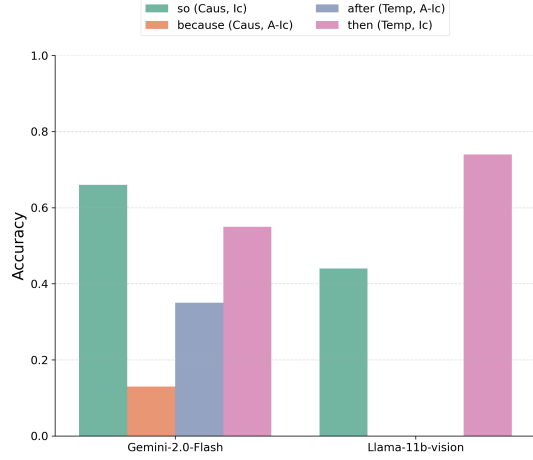Temp. respectively, and abbreviate Iconic and Anti-Iconic with Ic. and A-Ic. respectively.

Figure 2: Results for Task 2 on each connective.

First, we evaluate the performance of the VLMs on causal and temporal reasoning tasks using only visual inputs. Results from Task 1 (Relation Identification) are reported in Table 2, while results on Task 2 (Directionality Specification) are shown in Figure 2. We observe a two-tiered competency. The models can broadly classify the type of relationship (causal vs temporal) with above-chance accuracy. However, they largely fail to determine its underlying structure and directionality. In Task 1, both models perform significantly better than the random baseline, indicating that they can extract relevant signals from the image pairs. A closer look at the results in Table 2 reveals `Gemini-flash-2.0` shows a clear proficiency on temporal relations (87% accuracy), suggesting a default tendency to interpret visual sequences as a chronological progression. In contrast, `Llama-11b-vision` demonstrates the inverse pattern, excelling at identifying causal relations (86% accuracy), implying a strong prior to infer cause-and-effect. This superficial competence however breaks down when models are required to identify the directionality of the relationship in Task 2 (Figure 2). The performance plummets for both models and this failure is almost entirely attributable to an inability to process anti-iconic relations, thus revealing a noticeable "iconicity bias". This bias manifests as a dependency on the perceptual order of visual events to infer their logical structure. `Llama-11b-vision` excels at identifying the direction for the Temporal Iconic connective *then*, but its performance on the anti-iconic connectives is non-existent. `Gemini-flash-2.0` appear more robust, but displays a similar pattern, with a moderate accuracy on iconic relations but a sharp drop in performance for anti-

Task	Model	Accuracy	Causal Relations (F1)		Temporal Relations (F1)	
			so (Ic.)	*because* (A-Ic.)	*then* (Ic.)	*after* (A-Ic.)
Task 3	Gemini	0.42	0.48	0.28	**0.52**	0.09
	LLaMA	0.31	**0.42**	0.02	0.39	0.04
Task 4	Gemini	0.64	0.66	0.65	**0.70**	0.51
	LLaMA	0.33	0.32	0.06	**0.46**	0.14

Table 3
Model performance with accuracy and F1 Score for connectives on task 3 and task 4
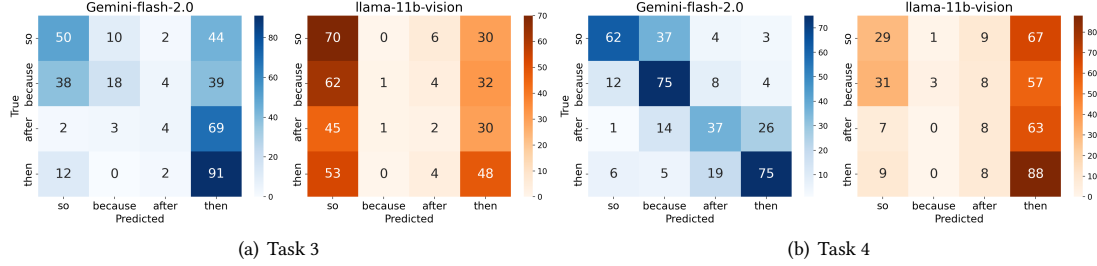

(a) Task 3   (b) Task 4

Figure 3: Comparison of Confusion Matrices for Tasks 3 and 4.

iconic relations (connectives *because* and *after*).

Table 3 reports result on Tasks 3 (Connective Selection) and 4 (Connective Selection With Captions). Task 4, which provides both images and their corresponding textual captions, offers an ideal setting to assess the practical utility of visual grounding in multimodal models. Here, the models receive both images and their corresponding textual captions and their performance can be directly compared to that of the text-only LLMs evaluated on the same cloze task in the original ExpliCa study [8]. The multimodal models, particularly Gemini-flash-2.0 achieve overall comparable or slightly better results (0.64 vs 0.62 accuracy) than strong text-only proprietary models. This suggests that the visual input may actually provide effective grounding, reinforcing or clarifying the relationship expressed via text without being a hindrance. Similarly, Llama-11b-vision's multimodal performance aligns with text-only open-source LLMs (0.33 vs 0.34 accuracy). Nevertheless, if we look at Confusion Matrices in Figure 3 we observe that they reinforce the findings from previous tasks: the models' performance are in general dictated by the iconicity of the underlying relation, even more so than in the original study. This may suggest that, while visual inputs can prove beneficial on a surface level, their order of presentation may strongly affect and bias the models' ability, especially in anti-iconic cases. This may also be taken as indication that the models' training data contained a significantly larger number of "iconic examples".

Finally, results for Task 5 are shown in Figures 4 and 5 and Table 4. Recall that the objective of the task 5 is is to provide a numerical plausibility rating from 1 (completely incoherent) to 10 (perfectly coherent) for the complete multimodal event: both images, their corresponding textual captions, and the human-provided connective linking them. Also recall that Task 5 was evaluated only on Gemini-flash-2.0. To enable a direct comparison between the model's output and the human judgments, both sets of scores were first normalized to a common scale using a *min-max scaler*. The density plots in Figure 4 reveal both a promising alignment and critical divergences. For the iconic connectives, the model's scores show a distribution that closely resembles the human distribution of the connective with the highest rating. Both distributions are heavily skewed towards higher values (0.8-1.0), indicating that the model, like humans, find these iconic constructions highly plausible. Conversely, a significant discrepancy emerges for the anti-iconic connectives. For *because* and especially *after*, the human ratings show a much broader distribution with a notable peak in the mid-to-low range, indicating greater uncertainty and lower acceptability in general. To quantify this alignment, we computed the Pearson correlation between the model's ratings and human judgments (see Table 4). The results confirm the visual trend: We observe a moderate and statistically significant correlation for the iconic connectives *so* and *then*. The correlation is weaker for the anti-iconic connective *because*, and becomes statistically insignificant for *after*.
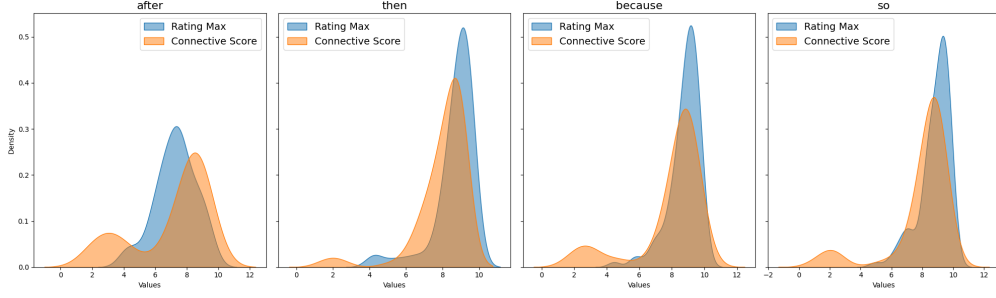
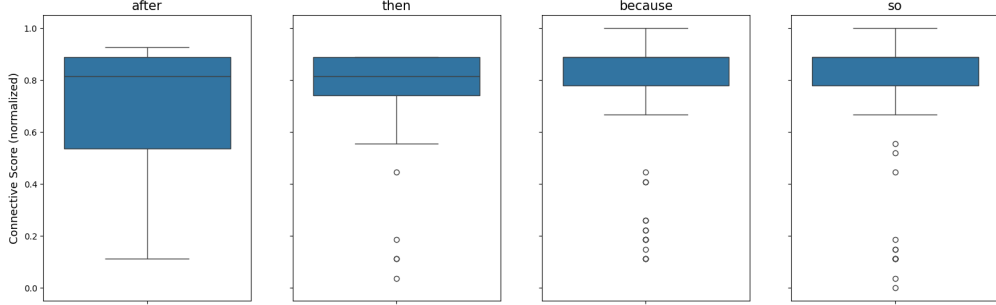**Figure 4:** Density plots comparing model-generated acceptability ratings with human plausibility judgments

Figure 5: Box plots illustrating the distribution of model-generated acceptability ratings for each connective.

Connective	Pearson ρ
so (Caus., Ic.)	0.55*
then (Temp., Ic.)	0.53*
because (Caus., A-Ic.)	0.39*
after (Temp., A-Ic.)	0.21

Table 4

Pearson correlation between model acceptability ratings and human judgments, grouped by connective type. *Indicates a statistically significant correlation ($p < 0.05$).

To better understand the sources of divergence between the model's and human judgments, particularly for the cases that the model rated as highly implausible, we performed an outlier analysis. We specifically focused on low-scoring outliers, which we formally identified using the interquartile range (IQR) rule: any data point falling below the first quartile (Q1) minus 1.5 times the IQR was flagged. As noted in the original ExpliCa dataset, a subset of sentences were intentionally designed to be *socially challenging*, touching on sensitive topics like religion, immigration, drug abuse or sex. Our analysis (Figure 5) reveals that a significant portion of the outliers are directly attributable to this subset. Specifically, 13 out of the 31 most prominent low-scoring outliers correspond to these socially challenging sentences. This finding sug-

gests that the model's performance may be influenced by its internal bias-mitigation and safety alignment protocols. When confronted with sensitive content, the model appears to override its linguistic and logical assessment, assigning a very low acceptability score regardless of the sentence's grammatical or causal coherence. This highlights a potential conflict where safety-driven heuristics can interfere with and ultimately degrade the model's core reasoning capabilities on specific types of content.

6. Conclusion and future works

This paper investigated the capacity of modern Vision-Language Models to reason about the structure of events. We augmented a curated dataset on causal reasoning with visual stimuli, and designed five tasks of increasing difficulty to asses how well the evaluated systems handle causal and temporal relationships, particularly when the logical flow of events diverges from their visual presentation. The central finding of our experiments is a profound vulnerability of the tested VLMs to an "iconicity bias." This manifests as a sharp decline in accuracy for anti-iconic relations, revealing a dependency on perceptual order over abstract logic. This weakness in abstract reasoning is likely rooted in an equally fragile foundational visual understanding. Recent studies using controlled evaluation frameworks [25], have in fact shown

that VLMs struggle to robustly identify even fundamental object properties (like color or shape) and their basic spatial relations. Indeed, their performance is heavily dependent on positional biases, with objects at the center of an image being recognized more reliably than those at the periphery. If models fail to build a stable and reliable representation of a single scene, their ability to infer complex causal and temporal relationships across multiple scenes becomes inherently compromised. The macroscopic failures we observed (e.g., the iconicity bias) can therefore be seen as a direct consequence of these microscopic weaknesses. Furthermore, our analysis indicates that this reasoning is not purely logical; it may also be modulated by the models' safety training, which can produce inconsistent evaluations of causally coherent but sensitive content. Taken together, these results challenge the notion that scaling and multimodal pre-training are sufficient for achieving robust, human-like reasoning. The models' reliance on perceptual heuristics points to a fundamental gap between their pattern-matching prowess and their ability to model the more complex, non-sequential nature of real-world events.

A crucial next step is to investigate whether these behavioral failures reflect a deeper deficit in the models' underlying competence. A more direct evaluation, drawing on the framework of Hu and Levy [26], would involve measuring the log-likelihood that models assign to different event structures. However, this approach faces a significant technical barrier: the public APIs for state-of-the-art multimodal models, including Gemini 2.0 Flash, do not currently provide access to token-level log-likelihoods. This constraint makes it impossible to directly probe their internal probability distributions. Future work should therefore seek to replicate this study using open-source VLMs where such access is possible.

Limitations

While the present work provide some interesting insights, it is fundamental to point out several of its limitations. First, the two models chosen for the analysis can be considered as good representatives of open-weights and closed-weights models in the small to medium-sized model range; we purposely avoided using larger VLMs as they tipically come with a high computational (or monetary) cost. However, we must acknowledge that the paper's results may not hold for other VLMs.

Second, we leverage CoT prompting, but do not present here an analysis of the results from the CoT; these could be point to additional insights. In addition to this, we must note that we did not perform any prompt-level optimization to improve the performances of each model individually.

Third, we do not account for the abstractness of the stimuli. While the ExpliCa dataset contains mostly concrete, everyday scenarios, searching for relations between their abstractness and the performances of the model may yield more robust findings.

References

[1] A. Lenci, Understanding natural language understanding systems. a critical analysis, 2023. URL: https://arxiv.org/abs/2303.04229. arXiv:2303.04229.

[2] C. D. Manning, Human language understanding & reasoning, Daedalus 151 (2022) 127–138. URL: https://api.semanticscholar.org/CorpusID:248377870.

[3] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, E. Fedorenko, Dissociating language and thought in large language models, 2024. URL: https://arxiv.org/abs/2301.06627. arXiv:2301.06627.

[4] Y. Du, Z. Liu, J. Li, W. X. Zhao, A survey of vision-language pre-trained models, 2022. URL: https://arxiv.org/abs/2202.10936. arXiv:2202.10936.

[5] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao, Vision-language pre-training: Basics, recent advances, and future trends, 2022. URL: https://arxiv.org/abs/2210.09263. arXiv:2210.09263.

[6] L. W. Barsalou, Grounded cognition: Past, present, and future, Topics in Cognitive Science 2 (2010) 716–724. doi:10.1111/j.1756-8765.2010.01115.x.

[7] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, N. Pinto, J. Turian, Experience grounds language, 2020. URL: https://arxiv.org/abs/2004.10151. arXiv:2004.10151.

[8] M. Miliani, S. Auriemma, A. Bondielli, E. Chersoni, L. Passaro, I. Sucameli, A. Lenci, ExpliCa: Evaluating explicit causal reasoning in large language models, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 17335–17355. URL: https://aclanthology.org/2025.findings-acl.891/. doi:10.18653/v1/2025.findings-acl.891.

[9] Z. Jin, Y. Chen, F. Leeb, L. Gresele, O. Kamal, Z. Lyu, K. Blin, F. G. Adauto, M. Kleiman-Weiner, M. Sachan, B. Schölkopf, Cladder: Assessing causal reasoning in language models, 2024. URL: https://arxiv.org/abs/2312.04350. arXiv:2312.04350.

[10] J. Pearl, D. Mackenzie, The Book of Why: The New Science of Cause and Effect, 1st ed., Basic Books, Inc., USA, 2018.

[11] N. Joshi, A. Saparov, Y. Wang, H. He, Llms are prone

to fallacies in causal inference, 2024. URL: https://arxiv.org/abs/2406.12158. arXiv:2406.12158.

[12] S. Ashwani, K. Hegde, N. R. Mannuru, M. Jindal, D. S. Sengar, K. C. R. Kathala, D. Banga, V. Jain, A. Chadha, Cause and effect: Can large language models truly understand causality?, 2024. URL: https://arxiv.org/abs/2402.18139. arXiv:2402.18139.

[13] J. Wallat, A. Jatowt, A. Anand, Temporal blind spots in large language models, 2024. URL: https://arxiv.org/abs/2401.12078. arXiv:2401.12078.

[14] S. Xiong, A. Payani, R. Kompella, F. Fekri, Large language models can learn temporal reasoning, 2024. URL: https://arxiv.org/abs/2401.06853. arXiv:2401.06853.

[15] Z. Li, H. Wang, D. Liu, C. Zhang, A. Ma, J. Long, W. Cai, Multimodal causal reasoning benchmark: Challenging vision large language models to discern causal links across modalities, 2025. URL: https://arxiv.org/abs/2408.08105. arXiv:2408.08105.

[16] M. Ventura, M. Toker, N. Calderon, Z. Gekhman, Y. Bitton, R. Reichart, Nl-eye: Abductive nli for images, 2024. URL: https://arxiv.org/abs/2410.02613. arXiv:2410.02613.

[17] M. F. Imam, C. Lyu, A. F. Aji, Can multimodal llms do visual temporal understanding and reasoning? the answer is no!, 2025. URL: https://arxiv.org/abs/2501.10674. arXiv:2501.10674.

[18] L. A. Hendricks, A. Nematzadeh, Probing image-language transformers for verb understanding, 2021. URL: https://arxiv.org/abs/2106.09141. arXiv:2106.09141.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020. arXiv:2103.00020.

[20] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. A.-D. et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[21] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. S. et al., Gemini: A family of highly capable multimodal models, 2025. URL: https://arxiv.org/abs/2312.11805. arXiv:2312.11805.

[22] Google DeepMind, Gemini 2.0 flash – model card, 2025. URL: https://ai.google.dev/gemini-api/docs/models, model card published April 15, 2025.

[23] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, A. Smola, Multimodal chain-of-thought reasoning in language models, 2024. URL: https://arxiv.org/abs/2302.00923. arXiv:2302.00923.

[24] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.

[25] M. Rizzoli, S. Alghisi, O. Khomyn, G. Roccabruna, S. M. Mousavi, G. Riccardi, Civet: Systematic evaluation of understanding in vlms, 2025. URL: https://arxiv.org/abs/2506.05146. arXiv:2506.05146.

[26] J. Hu, R. Levy, Prompting is not a substitute for probability measurements in large language models, 2023. URL: https://arxiv.org/abs/2305.13264. arXiv:2305.13264.

A. Prompts

Task 1. Relation Identification (Causal)

```
The image above contain two separated
    images: Image a (on the left) and
    Image b
(on the right). Describe the elements
    in both images. Now, think
    abstractly
about the relationship between the two
    images. Focus on the general
    cause-
and-effect pattern rather than
    specific details. The antecedent
    is the event that
happens first and directly causes
    another event (the cause). The
    consequent is
the event that happens as a result of
    the antecedent (the effect). If
    Image a is
the consequent and Image b is the
    antecedent, respond with Image a.
    If Image b
is the consequent and Image a is the
    antecedent, respond with Image b.
    Do not
provide explanations, additional text
    or commentary.
```

Task 1. Relation Identification (Temporal)

The image above contain two separated images: Image a (on the left) and Image b
(on the right). Describe the elements in both images. Now, think about the
temporal relationship between the two images. Focus on the sequence of events
rather than specific details. If Image a follows Image b, respond with Image a.
If Image b follows Image a, respond with Image b. Do not provide explanations,
additional text or commentary.

Task 2. Directionality Specification

Analyze the relationship between Image A (left) and Image B (right). Determine
whether the connection is temporal (one event happens before or after the other)
or causal (one event directly causes the other). Respond with only one word,
either 'temporal' or 'causal'. Do not provide explanations, additional text or
commentary.

Task 3. Connective Selection

Your task is to select the most appropriate word to connect the two images. There
are four words:
- So: causal relation in which IMAGE A causes IMAGE B;
- Because: causal relation in which the IMAGE B causes IMAGE A;
- Then: temporal relation in which IMAGE A precedes IMAGE B;
- After: temporal relation in which IMAGE A follows IMAGE B;
Answer only with the connective that best expresses the relationship between the
two images. Do not provide explanations or additional details. Your answer has
to be coherent with your previous reasoning.

Task 4. Connective Selection With Captions

You are given two sentences: Sentence A and Sentence B and a couple of images
(Image A refers to Sentence A and Image B refers to Sentence B). Your task is
to select the most appropriate word to connect the two sentences logically and
coherently. The chosen word should fit grammatically and contextually
Format:
Sentence A: Sentence A
Sentence B: Sentence B
There are four words:
- Then;
- After;
- So;
- Because;
Thinks about the two sentences and answer only with the word that best expresses
the relationship between the two sentences.

Task 5. Acceptability Rating

```
Evaluate the acceptability of
    sentences that describe two
    events linked by connectives: 'so
    ', 'because', 'after', and 'then
    '. Rate each sentence on a scale
    from
1 to 10 based on how well the
    connective expresses the
    relationship between
the events. Each event is also
    visually represented by an image:
     the left image
corresponds to the first sentence and
    the right image corresponds to
    the second
sentence. Sentence: sentence a
    connective sentence b. Provide
    only a numerical
rating between 1 and 10, without
    explanations.
```

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI), Gemini (Google), and Grammarly in order to: Generate images, Paraphrase and reword, Improve writing style, and Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.