ISO-BENCH: Benchmarking Multimodal Causal Reasoning in Visual–Language Models through Procedural Plans

Ananya Sadana
Stony Brook University
ananyasadana2@gmail.com

Yash Kumar Lal
Ohio State University
lal.77@osu.edu

Jiawei Zhou
Stony Brook University
jiawei.zhou1@stonybrook.edu

Abstract

Understanding causal relationships across modalities is a core challenge for multimodal models operating in real-world environments. We introduce ISO-BENCH, a benchmark for evaluating whether models can infer causal dependencies between visual observations and procedural text. Each example presents an image of a task step and a text snippet from a plan, with the goal of deciding whether the visual step occurs before or after the referenced text step. Evaluation results on ten frontier vision-language models show underwhelming performance: the best zero-shot F1 is only 0.57, and chain-of-thought reasoning yields only modest gains (up to 0.62 F1), largely behind humans (0.98 F1). Our analysis further highlights concrete directions for improving causal understanding in multimodal models.

1 Introduction

Complex visual reasoning is critical for advancing multimodal intelligence, particularly in tasks involving planning and execution in the physical world. In such contexts, visual information plays a central role in how humans interpret, follow, and carry out plans. For instance, assembling furniture with a manual or cooking from a recipe requires understanding the logical and temporal structure of the plan, grounding visual observations within this structure, and reasoning about current and next steps by integrating information across modalities.

As vision-language models (VLMs) continue to improve, a key question emerges: can these models perceive and reason about multimodal plans that unfold in real-world visual contexts? In this work, we focus on a specific and challenging aspect of this problem—*causal reasoning across modalities*. Given a plan in text and a visual observation (e.g., an image showing the status of a task in Figure 1), we investigate whether a model can understand the causal relationship between steps conveyed through different modalities. This line of inquiry forms the foundation for more advanced reasoning capabilities, including video understanding, grounded world modeling, and embodied AI, where agents must interpret visual states to execute plans.

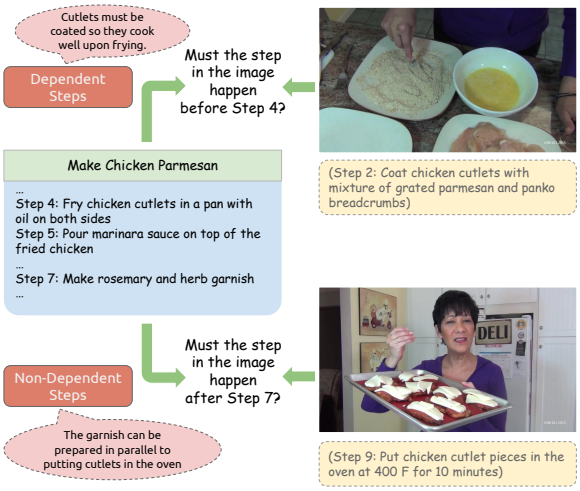

Figure 1: An example plan, visual states, and their causal reasoning. To determine whether the step in the top image must occur before Step 4, one must recognize it as coating chicken cutlets and infer that frying depends on it. In contrast, the bottom image shows cutlets being placed in the oven—a step independent of Step 7 (garnish preparation), which can occur in parallel.

While existing benchmarks in visual intelligence often focus on image or video question answering (Yue et al., 2024; Fu et al., 2025; Fang et al., 2024), they typically lack the procedural and temporal structure needed for physical planning. In contrast, plan reasoning benchmarks are usually text-only (Valmeekam et al., 2023a; Lin et al., 2020), focusing on state changes (Bosselut et al., 2018) or action sequences (Donatelli et al., 2021). Yet real-world plans often rely on visual context and demand reasoning across modalities. Inferring dependencies in such settings requires not only perception and grounding, but also understanding causal relations like preconditions and effects between steps.

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 797–807
July 4, 2026 ©2026 Association for Computational Linguistics

To address this gap, we introduce ISO-BENCH (Image-State Ordering Benchmark), a new evaluation set to study whether models can reason about causal relationships across modalities in procedural plans. Built from two instructional video datasets—YouCook2 (Zhou et al., 2018) and CrossTask (Zhukov et al., 2019)—we manually construct and annotate a diverse set of examples spanning domains like cooking, car maintenance, crafting, and woodworking. Each example includes a text snippet, an image depicting a step, and a question asking whether the image step occurs before or after the textual steps—directly probing multimodal causal understanding.

We evaluate 10 state-of-the-art multimodal models, both open-source and proprietary, on ISO-BENCH[1]. Surprisingly, even the strongest models perform near chance. Reasoning-based chain-of-thought (Wei et al., 2022) yields only marginal gains, despite its effectiveness in other reasoning tasks (Camburu et al., 2018; Kumar and Talukdar, 2020). These results reveal a notable gap in current VLM capabilities and position multimodal causal reasoning as a concrete challenge for future research.

2 Related Work

Prior work on plan understanding has examined various aspects such as entity state tracking (Bosselut et al., 2018; Henaff et al., 2017), action prediction (Lin et al., 2020; Donatelli et al., 2021), and next-step inference (Zellers et al., 2019; Zhang et al., 2020). XPAD (Dalvi et al., 2019) focuses on scientific processes with action-dependency prediction, while PizzaCommonsense (Diallo et al., 2024) enriches cooking plans with fine-grained intermediate steps. PlanBench (Valmeekam et al., 2023b) and NaturalPlan (Zheng et al., 2024) explore classical and constraint-based planning, showing that LLMs often struggle to produce valid action sequences.

Several benchmarks evaluate models' ability to reason about procedural dependencies, but primarily in the text modality. For example, CAT-BENCH (Lal et al., 2024) and Kiddon et al. (2015) focus on dependency prediction in cooking recipes, while CREPE (Zhang et al., 2023) probes comparative causal judgments in procedures.

On the multimodal side, RecipeQA (Yagcioglu et al., 2018) aligns text with illustrative images, and TOMATO (Shangguan et al., 2024) evaluates temporal reasoning from videos. COMKitchens (Maeda et al., 2024) and MM-ReS (Pan et al., 2020) provide datasets with structured workflow annotations. However, these benchmarks do not directly assess causal reasoning across modalities. ISO-BENCH fills this gap by testing whether models can integrate visual and textual information to infer temporal dependencies within instructional plans.[2]

3 Benchmark Construction

Understanding and following instructional plans requires the ability to comprehend both textual steps and corresponding visual states, as well as the dependencies between them. We focus on *temporal and causal dependencies* across modalities: if the effects of one step satisfy the preconditions of another, the former must occur before the latter. Recognizing such dependencies involves reasoning over actions shown in images and inferring preconditions, causes, sub-goals, and effects.

To construct ISO-BENCH, we repurpose two instructional video datasets, YouCook2 (Zhou et al., 2018) and CrossTask (Zhukov et al., 2019), covering domains like cooking, woodworking, and more. From each annotated step in the videos, we extract a single frame at the midpoint. For each video and plan, we create text snippets by selecting the first or last $k \in \{2, 3, 4\}$ steps of a plan and pair it with an image from another step in the same plan. When the image comes from a later step, we frame the question as: *Must the step in the image happen after the snippet?* Conversely, if the image is from an earlier step, we ask whether it must happen *before*.

Each (snippet, image) pair is annotated with a binary yes/no label indicating whether there is a causal dependence between the image step and any step in the text. This yields two types of instances: DEP, where such a dependence exists (ground truth is "yes"), and NONDEP, where no causal relationship is present (ground truth is "no").

ISO-BENCH contains 764 examples, roughly evenly split between the two types. The task challenges models to decide whether the visual step occurs before or after the textual steps, probing their ability to reason about cross-modal procedural dependencies. Evaluation is based on per-class precision, recall, and F1 score. Dataset statistics appear in Table 2 (Appendix D), and data processing details are in Appendices B and C.

[1] ISO-BENCH is available at https://huggingface.co/datasets/StonyBrookNLP/ISO-Bench

[2] Please find more detailed related work in Appendix A.

		DEP			NONDEP			MACRO AVG		
		P	R	F	P	R	F	P	R	F
DeepSeek-VL2	E→A	0.57	0.07	0.12	0.50	0.95	0.66	0.54	0.51	0.39
	A	0.52	0.04	0.07	0.50	0.97	0.66	0.51	0.50	0.36
Llava-1.5	E→A	0.51	0.56	0.53	0.51	0.45	0.47	0.51	0.50	0.50
	A	0.50	0.96	0.66	0.54	0.05	0.09	0.52	0.51	0.52
Qwen2.5-VL	E→A	0.53	0.37	0.43	0.56	0.26	0.35	0.54	0.31	0.39
	A	0.56	0.50	0.53	0.54	0.59	0.57	0.55	0.55	0.55
PaliGemma2	E→A	0.51	0.24	0.33	0.50	0.77	0.61	0.50	0.50	0.47
	A	0.57	0.17	0.26	0.57	0.87	0.64	0.54	0.52	0.53
Llama-3.2	E→A	0.51	0.71	0.59	0.58	0.13	0.22	0.54	0.42	0.41
	A	0.44	0.01	0.02	0.50	0.99	0.66	0.47	0.50	0.34
4o-mini	E→A	0.54	0.79	0.64	0.61	0.33	0.42	0.57	0.56	0.53
	A	0.53	0.85	0.65	0.62	0.24	0.34	0.57	0.55	0.50
4o	E→A	0.58	0.67	0.63	0.61	0.52	0.56	0.60	0.60	0.59
	A	0.56	0.72	0.63	0.61	0.43	0.50	0.58	0.58	0.57
Sonnet	E→A	0.54	0.76	0.63	0.60	0.35	0.44	0.57	0.56	0.54
	A	0.55	0.65	0.60	0.58	0.27	0.37	0.57	0.46	0.48
o3	E→A	0.61	0.61	0.62	0.62	0.61	0.61	0.62	0.62	0.62
o4-mini	E→A	0.58	0.73	0.65	0.63	0.46	0.54	0.61	0.60	0.59
Human*	-	0.97	0.98	0.98	0.98	0.97	0.97	0.98	0.98	0.98

Table 1: Performance of all models on when just providing an answer A and when also explaining that answer E→A. We report per-label as well as weighted macro average precision, recall and F1 score. We report human baseline numbers calculated on a subset of 200 random instances of ISO-BENCH.

4 Benchmarking Models on ISO-BENCH

We benchmark the performance of a variety of state-of-the-art models on ISO-BENCH.

4.1 Models and Setup

We evaluate 10 different vision-language models (VLMs). Open-source models include llava-1.5-7b-hf (`Llava-1.5`), Llama-3.2-11B Vision Instruct (`Llama-3.2`), Qwen2.5-VL-32B (`Qwen2.5-VL`), paligemma2-10b-mix-448 (`PaliGemma2`) and deepseek-VL2-tiny (`DeepSeek-VL2`), and proprietary models include gpt-4o-mini-2024-07-18 (`4o-mini`), gpt-4o-2024-11-20 (`4o`), claude-3-5-sonnet-20241022 (`Sonnet`), OpenAI o3-2025-04-16 (`o3`) and o4-mini-2025-04-16 (`o4-mini`), covering a range of model sizes and capabilities. As described in Appendix E, we also establish how well humans can perform this task.

We test models under two zero-shot settings: (i) direct answer generation (A), and (ii) explanation-augmented reasoning (E→A), where models are prompted to generate an explanation before answering. Full model details, prompts, and evaluation setup are provided in Appendices F and H.

4.2 Results

Table 1 presents the performance of all the models in different settings for ISO-BENCH. We present per-class (DEP and NONDEP) precision, recall and F1 score as well as macro average metrics. Humans are easily able to perform this task, and achieve very high F1 score on ISO-BENCH. We make two main observations on model performance.

Most VLMs perform poorly on ISO-BENCH. We observe that most models perform poorly on the task (A rows in Table 1), with overall F1 scores only ranging from ∼0.3-0.5. Only one model (out of eight) stands out: `4o` (0.57 F1). On the contrary, humans perform near perfectly, achieving 0.98 F1.

Models are clearly better at judging when there is

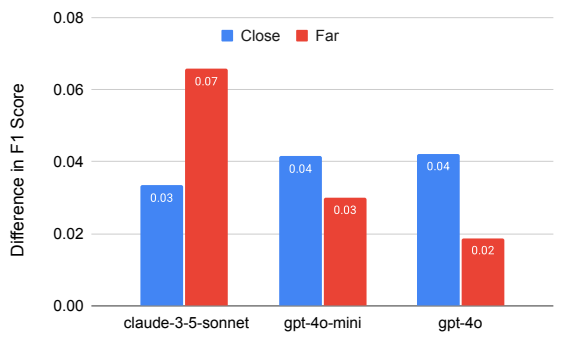

Figure 2: Difference in performance of models between (E→A) and (A) settings split by the distance between the steps being asked about in the question.

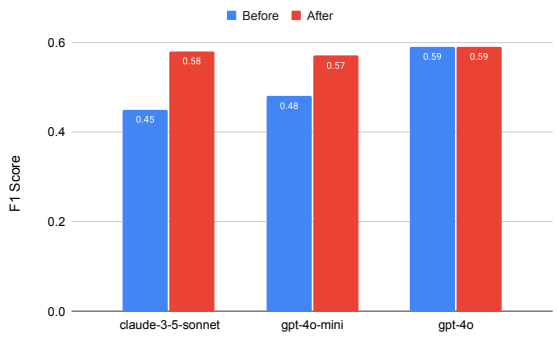

Figure 3: Performance of models (A) split by temporal relation type (*before* and *after*) in the question.

a dependence between the step in the image and the steps in the text snippet. Interestingly, we observe that *most* models have high precision but low recall for NONDEP. Models can reliably identify only some of the NONDEP dependencies. Contrastingly, the best performing models attain high recall and low precision on DEP, indicating a bias towards assessing most data points as having a dependence.

Generating explanations only helps marginally. The E→A rows in Table 1 represent results where models, given the image and text snippet, first perform step-by-step reasoning and then generate a decision on whether there is a dependence between the snippet and the image. Results show that reasoning leads to improvements (∼0.02-0.06) for closed-source models, and for `DeepSeek-VL2` and `Llama-3.2`, but gains are rather small. `o4-mini`, trained to analyze and do reasoning over images, achieves the highest performance (0.62 F1 respectively). For other open-source models, there is a notable drop in performance (∼0.02-0.16) indicating a weakness in their CoT abilities. We hypothesize that this is due to the limitations of instruction tuning in vision-language modeling where models are mainly finetuned to describe or analyze images, not produce reasoning chains across modalities.

5 Analysis

We analyze the performance of the best models by different attributes of the questions and prompts.

5.1 Reasoning as a function of Step Distance

We study how the distance between the step in the image and the closest step in the snippet impacts model (here, `4o-mini`, `4o`, `Sonnet`) performance. A question is said to be about *close* steps if the distance between the step in the image ($step_i$) and the nearest step in the text snippet ($step_j$) is less than 3 ($|j-i| < 3$), and *distant* otherwise. Figure 2 shows the difference in F1 score between explaining then answering (E→A) and just answering (A) with different models as a function of step distance. As step distance increases, it should become harder to reason about whether there is a connection between them. Explanations should provide larger gains for larger distance over smaller distance. Surprisingly, this does not hold true for GPT models.

5.2 Understanding Directional Dependencies

Next, we study how models handle questions about different aspects of the same pair of steps. Typically, reasoning about whether a step must happen *before* another requires reasoning about preconditions and causes, while understanding whether a step must happen *after* another requires understanding the effects of any performed actions. Figure 3 shows model performance (A setting) for these questions. Most models are slightly better at reasoning about the effects of steps. We hypothesize that this is because effects in plans may be more immediate and hence, would be easier to understand.

5.3 Error Analysis

To better understand model failures, we sample and analyze 100 explanations generated by `4o` and `o3` (E→A) where the model produces an incorrect answer. We identify 4 major types of errors:

- Causal Reasoning (62%) – Models fail to understand that the step in the image describes either a precondition or an effect of (one of) the steps in the text snippet.
- Grounding (16%) – Often, models misidentify the step described in the image with another

one not asked about in the question.

- Action Progression (14%) – Models misunderstand an action in progress in the image and assume that the action in question has already been completed when producing their answer.
- Perception (8%) – Models incorrectly identify items in the image leading to wrong reasoning.

6 Conclusion

We introduce ISO-BENCH, a new benchmark to test whether multimodal models can reason about causal dependencies across visual and textual steps in procedural plans. Despite strong general capabilities, current VLMs perform poorly here, with only marginal gains from explanation-based prompting. Our work highlights a specific gap in multimodal causal reasoning and point to clear opportunities to advance model understanding of real-world plans.

Limitations

Our work only investigates English-language documents (plans) and this limits the generalizability of our findings to other languages.

We benchmark a reasonably diverse set of VLMs. Currently, we cover several model families and models of varying sizes. However, due to the current fast-paced landscape of VLM development, we are unable to cover all available VLMs. We will continue to evaluate more VLMs on ISO-BENCH.

Due to difficulty in creating data, we were unable to create a "dev" set which could serve as a source of examples for few-shot experiments. Similarly, we are unable to perform fine-tuning or adapter based experiments on ISO-BENCH. We leave this exploration to future work.

References

Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. 2018. Simulating action dynamics with neural process networks. *ICLR*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1595–1604, New Orleans, Louisiana. Association for Computational Linguistics.

Bhavana Dalvi, Niket Tandon, Antoine Bosselut, Wen-tau Yih, and Peter Clark. 2019. Everything happens for a reason: Discovering the purpose of actions in procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4496–4505, Hong Kong, China. Association for Computational Linguistics.

Aissatou Diallo, Antonis Bikakis, Luke Dickens, Anthony Hunter, and Rob Miller. 2024. PizzaCommonSense: A dataset for commonsense reasoning about intermediate steps in cooking recipes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12482–12496, Miami, Florida, USA. Association for Computational Linguistics.

Lucia Donatelli, Theresa Schmidt, Debanjali Biswas, Arne Köhn, Fangzhou Zhai, and Alexander Koller. 2021. Aligning actions across recipe graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6930–6942, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. *ICLR*.

Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. 2015. Mise en place: Unsupervised interpretation of instructional recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 982–992, Lisbon, Portugal. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Saksham Lal, Mizuki Morita, Yingya Li, Roberto Rossi, and Peter Lane. 2024. CaT-Bench: Benchmarking language-model understanding of causal and temporal dependencies in plans. *arXiv preprint*, arXiv:2406.15823. To appear in EMNLP 2024.

Angela Lin, Sudha Rao, Asli Celikyilmaz, Elnaz Nouri, Chris Brockett, Debadeepta Dey, and Bill Dolan. 2020. A recipe for creating multimodal aligned datasets for sequential tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4871–4884, Online. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, Yin Li, Jason Baldridge, and H. Sebastian Seung. 2023. LLaVA-1.5: Visual instruction tuning for large language-and-vision assistants. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Koki Maeda, Tosho Hirasawa, Atsushi Hashimoto, Jun Harashima, Leszek Rybicki, Yusuke Fukasawa, and Yoshitaka Ushiku. 2024. Com kitchens: An unedited overhead-view video dataset as a vision-language benchmark. In *Proceedings of the European Conference on Computer Vision*.

Dai Quoc Nguyen, Dat Quoc Nguyen, Cuong Xuan Chu, Stefan Thater, and Manfred Pinkal. 2017. Sequence to sequence learning for event prediction. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 37–42, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Liang-Ming Pan, Jingjing Chen, Jianlong Wu, Shaoteng Liu, Chong-Wah Ngo, Min-Yen Kan, Yugang Jiang, and Tat-Seng Chua. 2020. Multi-modal cooking workflow construction for food recipes. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1132–1141, New York, NY, USA. Association for Computing Machinery.

Paolo Pareti, Benoit Testu, Ryutaro Ichise, Ewan Klein, and Adam Barker. 2014. Integrating know-how into the linked data cloud. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 385–396. Springer.

Jae Sung Park, Ilia Kulikov, Chandra Bhagavatula, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision (ECCV)*.

Ziyao Shangguan, Jiarui Qiu, Andrei Barbu, David Fouhey, and Song-Chang Huang. 2024. TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. *arXiv preprint*, arXiv:2410.23266.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023a. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Preprint*, arXiv:2206.10498.

Karthik Valmeekam, Kartik Talamadupula, and Sanjay Srivastava. 2023b. PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *NeurIPS Datasets & Benchmarks Track*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. 2022. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4525–4542, Dublin, Ireland. Association for Computational Linguistics.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1358–1368.

Jinyoung Yeo, Gyeongbok Lee, Gengyu Wang, Seungtaek Choi, Hyunsouk Cho, Reinald Kim Amplayo, and Seung-won Hwang. 2018. Visual choice of plausible alternatives: An evaluation of image-based commonsense causal reasoning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. 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*, pages 9556–9567.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hongming Zhang, Muhao Chen, Haoyu Wang, Yangqiu Song, and Dan Roth. 2020. Analogous process structure induction for sub-event sequence prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1541–1550, Online. Association for Computational Linguistics.

Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, and Chris Callison-Burch. 2023. Causal reasoning of entities and events in procedural texts. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 415–431,

Dubrovnik, Croatia. Association for Computational Linguistics.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. 2025. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*.

Huaixiu Steven Zheng, Neel Kant, Rui Wang, Peter J. Liu, Julian Michael, and Mohit Iyyer. 2024. NaturalPlan: Benchmarking LLMs on natural-language planning. *arXiv preprint*, arXiv:2406.04520.

Luowei Zhou, Chenliang Xu, Jason J. Corso, Richard Socher, and Caiming Xiong. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5832–5840. Introduces the YouCook2 dataset.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *CVPR*.

A Detailed Related Work

Plan understanding tasks span testing knowledge and reasoning about multiple aspects such as entity states (Bosselut et al., 2018; Henaff et al., 2017), actions (Pareti et al., 2014; Lin et al., 2020; Donatelli et al., 2021), next events (Nguyen et al., 2017; Zellers et al., 2019; Zhang et al., 2020) and more. XPAD (Dalvi et al., 2019) extend ProPara (Dalvi et al., 2018), originally focused on understanding scientific processes, by adding the new task of explaining actions by predicting their dependencies. PizzaCommonsense (Diallo et al., 2024) contains commonsense knowledge about intermediate and implicit steps for cooking recipes, providing explicit input/output pairs of each action with fine-grained annotations. PlanBench (Valmeekam et al., 2023b) focuses on classical AI planning domains, such as BlocksWorld, and shows that LLMs fail to produce valid, executable action sequences. NaturalPlan (Zheng et al., 2024) evaluate real-world constraint optimization based planning tasks. CREPE (Zhang et al., 2023) measures how well LLMs understand the comparative likelihood of two events occurring in a procedure. CAT-BENCH (Lal et al., 2024) and Kiddon et al. (2015) explore predicting dependencies in cooking recipes. Most work has evaluated how well models understand aspects of plans described only in a single (text) modality.

VCOPA (Yeo et al., 2018) posit the task of identifying the correct next step given a premise step, with all the steps described in images. VisualComet (Park et al., 2020) present actions in stories as images along with summaries of what happened before and after, and show that integration between visual and textual commonsense reasoning is required to holistically reason about actions. RecipeQA (Yagcioglu et al., 2018) align text instructions in cooking recipes with images corresponding to the central action in each step. TOMATO (Shangguan et al., 2024) evaluates how multimodal foundational models perform temporal reasoning about continuous actions in videos. Pan et al. (2020) build MM-ReS, the first large-scale dataset for cooking workflow construction, consisting of 9,850 recipes with human-labeled workflow graphs. Similarly, the COMKitchens (Maeda et al., 2024) dataset provides cooking videos annotated with a structured visual action graph. Wu et al. (2022) test whether models can correctly sequence images of misordered actions described in instruction manuals. COMFORT (Zhang et al., 2025) assess VLMs' spatial reasoning capabilities of along the lines of frames of references.

Existing datasets evaluate different aspects of plans, but only a few focus on assessing whether models understand temporal ordering constraints on the steps of an instructional plan. However, all these benchmarks only consider plans written in text. ISO-BENCH is created to test using and integrating visual and textual understanding capabilities required to understand plans.

B Processing YouCook2

YouCook2 (Zhou et al., 2018) contains 2,000 cooking videos with clean stepwise captions and gives us YouTube links plus the exact time spans and step-by-step instructions for each recipe. We start by taking every span, jumping to its midpoint, and saving one JPEG frame; that leaves us with a clean image for every recipe step.

After we extracted a JPEG for every recipe step, we built a labeled pool of text-image pairs. We filter for videos with atleast 5 steps. For each such video it randomly picked an integer k between 2 and 4. If the task was "before", the text snippet is formed from the last k steps of the recipe; if the task was "after", the snippet uses the first k steps. We define two candidate frame groups: (1) Positive pool – frames whose step comes just outside the snippet, within three steps: For the "before" task these are the steps that immediately precede the snippet (i.e., recipe positions total-k-3 ... total-k-1). For the "after" task they are the steps that immediately follow the snippet (k ... k+2). These frames genuinely belong before or after the snippet, so pairing them with the snippet should be logically correct. (2) Negative pool – frames taken inside the snippet itself (for "before", positions inside the last k steps; for "after", positions inside the first k steps). Because they come from the same span that forms the snippet, showing them separately violates the required temporal relation and therefore constitutes an incorrect match.

We built a Streamlit app to annotate each text-image pair with a gold label. The app shows the snippet, the frame, and whether the sample is tagged POS (the frame should come just before or after the snippet, as specified) or NEG (the frame is taken from inside the snippet and therefore should not match the requested relation). We look at the pair: if the frame truly matches (for a POS) or truly conflicts (for a NEG), we click Accept; otherwise

we click Reject. We provide a screenshot of the annotation app in Figure 4.

C Processing CrossTask

Starting from the 18 primary tasks, we capped ourselves at ten annotated clips per task (180 videos total). For every step span in each clip we jumped to the midpoint and saved one JPEG, giving us a clean image for every unique step.

We perform the same procedure on data from CrossTask as we did in Appendix B.

Using 180 videos we ended up with 139 human-vetted instances: 72 "before" and 67 "after", spanning 73 dependent/positive and 66 non-dependent/negative examples.

D Dataset Statistics

Relation	DEP	NONDEP	ISO-BENCH
Before	187	185	372
After	196	196	392
Total	**383**	**381**	**764**

Table 2: ISO-BENCH statistics with data points spanning 480 different instructional plan videos. Each data point contains a text snippet of the plan, an image description of a step not in the snippet, a binary question and answer about the causal dependence between them.

E Human Baseline for ISO-BENCH

To establish how well humans perform on this task, we presented the goal of the plan, the plan snippet and the image of a step and asked one annotator to answer the associated question. We randomly sampled 200 data points from ISO-BENCH for this, and we use the same task formulation that we use with models. The last row in Table 1 presents the per-class as well as macro average precision, recall and F1 scores for humans. This establishes a possible upper bound for the ISO-BENCH task. We find that humans are able to identify implicit causal dependencies between steps represented across modalities fairly easily. In fact, they rarely make mistakes.

F Benchmarking Models

We provide details of each model we evaluate on ISO-BENCH.

`gpt-4o-2024-11-20` accepts as input any combination of text, audio, image, and video and generates any combination of text, audio, and image outputs. It is especially better at vision and audio understanding compared to existing models.

`gpt-4o-mini-2024-07-18` has a context window of 128K tokens, supports up to 16K output tokens per request. It surpasses other small models released to that date on academic benchmarks across both textual intelligence and multimodal reasoning, and supports the same range of languages as 4o.

`claude-3-5-sonnet-20241022` sets new industry benchmarks for graduate-level reasoning (GPQA), undergraduate-level knowledge (MMLU), and coding proficiency (HumanEval). It shows marked improvement in grasping nuance, humor, and complex instructions, and is exceptional at writing high-quality content with a natural, relatable tone.

`o3-2025-04-16` excels at solving complex math, coding, and scientific challenges while demonstrating strong visual perception and analysis. It uses tools in its chains of thought to augment its capabilities; for example, cropping or transforming images, searching the web, or using Python to analyze data during the thought process.

`o4-mini-2025-04-16` is a smaller model optimized for fast, cost-efficient reasoning—it achieves remarkable performance for its size and cost, particularly in math, coding, and visual tasks. It is the best-performing benchmarked model on AIME 2024 and 2025. It performs especially strongly at visual tasks like analyzing images, charts, and graphics.

`llava-1.5-7b-hf` (Liu et al., 2023) is an open-source chatbot trained by fine-tuning LLaMA/Vicuna on GPT-generated multimodal instruction-following data. It is an auto-regressive language model, based on the transformer architecture. It combines a frozen CLIP/Vicuna vision encoder with a 7B-parameter language backbone, and is instruction-tuned on about 600k image–text pairs and is open-source.

`Llama-3.2-11B-Vision-Instruct` is the 11B version of the Llama 3.2-Vision set of multimodal LLMs which have been instruction tuned for image reasoning. It is built on top of the pretrained

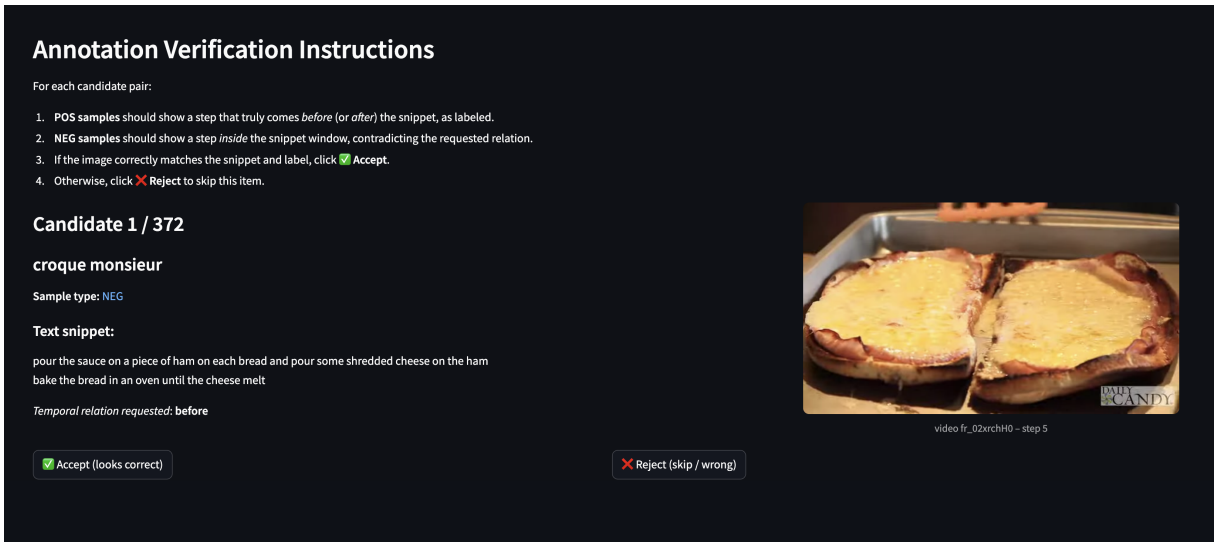

Figure 4: Screenshot with instructions provided to annotators from whom we collected gold labels for ISO-BENCH.

Llama 3.1 text only LLM by combining a seperately trained vision adapter module. Using a combination of supervised fine-tuning and reinforcement learning from human feedback, the model has been optimized to do a variety of vision tasks like image recognition, reasoning, captioning, and question answering on images.

`Qwen2.5-VL-32B-Instruct` is a 32 billion parameter vision language model. It is created on top of the Qwen-2.5 7 billion language model by following the ViT architecture. It has been extensively instruction tuned on (image,text) tuples to so that the model understands all things visual, is agentic, can comprehend long videos and events, can do visual localization, and generate structured outputs.

`paligemma2-10b-mix-448` checkpoints are fine-tuned on a diverse set of tasks and are ready to use out of the box while pt checkpoints are pre-trained and intended for further fine-tuning. These tasks include short and long captioning, optical character recognition, question answering, object detection and segmentation, and more. The model is available in the bfloat16 format for research purposes only.

`deepseek-vl2-tiny` is a 3 billion parameters mixture-of-experts vision language model. It has shown been to demonstrate enhanced performance across multiple tasks like visual question answering, optical character recognition, document/table/chart understanding, and visual grounding. It improves upon its predecessor, DeepSeek-VL, by using an improved high-resoultion vision encoder for better visual comprehension and an optimized language model backbone for training and test time efficiency. It is trained on a data that boosts performance and gives new capabilities to the model such as precise visual grounding.

G Model Setup

Python was the main scripting language for data collection and experimentation. For experiments using closed source models, we used OpenAI[3] and Anthropic[4] APIs. The total cost for OpenAI was $\sim$ 75 USD and $\sim$ 20 USD for Claude. The open-source experiments were conducted on 4 A6000 GPUs, each having 48 GB. The total GPU hours for all the experiments was $\sim$15. The models were downloaded from Huggingface and hosted for inference using Huggingface transformers module and vLLM. We use greedy decoding and temperature set to 0.0 when the option is available. We use GitHub Co-Pilot to help with writing code but verify it manually before running any experiments.

H Prompts Used

For each ISO instance, we provide the following prompt:

```
GOAL: {Goal Name}
Plan:
{step_1}
{step_2}
. . .
```

[3] https://openai.com/api/pricing/
[4] https://://www.anthropic.com/pricing

```
QUESTION: {Does this image show a step that must
come <before|after> the <first|last> step in the
plan?}
```

The prompt mirrors the reasoning process we want the model to follow. The `GOAL` line primes topical knowledge about the dish (e.g., Indian curries usually add tomatoes after onions). The `Recipe excerpt` gives an ordered context without revealing the full plan, so the model must place the image relative to these steps—exactly the causal inference ISO is designed to test. Finally, an explicit `QUESTION` asks about *before* or *after*, limiting the required output to a binary choice and removing ambiguity about expected format.

We test two modes:

1. **Answer-Only**: We append "Answer only with YES or NO." and use the first "YES" or "NO" as the prediction.

2. **Explain-Then-Answer**: We request a short chain-of-thought enclosed by `<think>...</think>`, then a final answer enclosed by `<answer>YES</answer>` or `<answer>NO</answer>`. We evaluate only the final tag, similar to CoT prompting.

The image is provided via the model's vision interface alongside this prompt. We limit generation to 120 tokens, which suffices for both reasoning and answer.

Metric. Since each instance is a binary question, we report *accuracy*, *precision*, *recall*, and *F1* on the subset of responses containing an unambiguous "YES" or "NO."[5]

[5]We discard outputs lacking either token.