Selective Vision is the Challenge for Visual Reasoning: A Benchmark for Visual Argument Understanding

Anonymous ACL submission

Abstract

Visual arguments, often used in advertising or social causes, rely on images to persuade viewers to do or believe something. Understanding these arguments requires selective vision: only specific visual stimuli within an image are relevant to the argument, and relevance can only be understood within the context of a broader argumentative structure. While visual arguments are readily appreciated by human audiences, we ask: are today's AI capable of similar understanding?

We collect and release VisArgs, an annotated corpus designed to make explicit the (usually implicit) structures underlying visual arguments. VisArgs includes 1,611 images accompanied by three types of textual annotations: 5,112 visual premises (with region annotations), 5,574 commonsense premises, and reasoning trees connecting them to a broader argument. We propose three tasks over VisArgs to probe machine capacity for visual argument understanding: localization of premises, identification of premises, and deduction of conclusions. Experiments demonstrate that 1) machines cannot fully identify the relevant visual cues. The top-performing model, GPT-4-O, achieved an accuracy of only 78.5%, whereas humans reached 98.0%. All models showed a performance drop, with an average decrease in accuracy of 19.5%, when the comparison set was changed from objects outside the image to irrelevant objects within the image. Furthermore, 2) this limitation is the greatest factor impacting their performance in understanding visual arguments. Most models improved the most when given relevant visual premises as additional inputs, compared to other inputs, for deducing the conclusion of the visual argument.

1 Introduction

What we see depends
mainly on what we look for.
– Lubbock (1893)

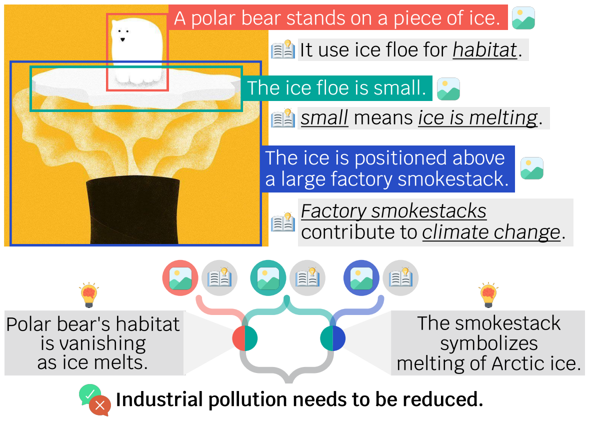


Figure 1: An example from our VisArgs corpus. VisArgs makes the persuasion process in a visual argument explicit by representing it as a reasoning tree. Image credit: Eglė Plytnikaitė

Humans often communicate messages visually. For example, traffic light colors regulate drivers' behavior, while computer icons, such as the trash bin symbol for deleting files or the magnifying glass for searching, guide user actions.

We consider the case of *visual arguments*. Consider Fig. 1, which depicts a polar bear on a shrinking ice floe. Without any text, this image calls attention to climate change: a visual metaphor connects melting ice to industrial emissions from factories. A plausible interpretation of the argument concludes: *industrial pollution needs to be reduced.*

We introduce VisArgs, an annotated dataset of 1,611 images containing visual arguments. VisArgs makes explicit the reasoning process in interpreting a visual argument:[1] each image is annotated with *visual premises* grounded on object bounding boxes, *commonsense premises* eliciting implicit knowledge, and *argument trees* formalizing the connection of these premises to the conclusion. An

[1] We note that our corpus contains just one possible *interpretation* of a visual argument (rather than, e.g., claiming to represent the creator's intent).

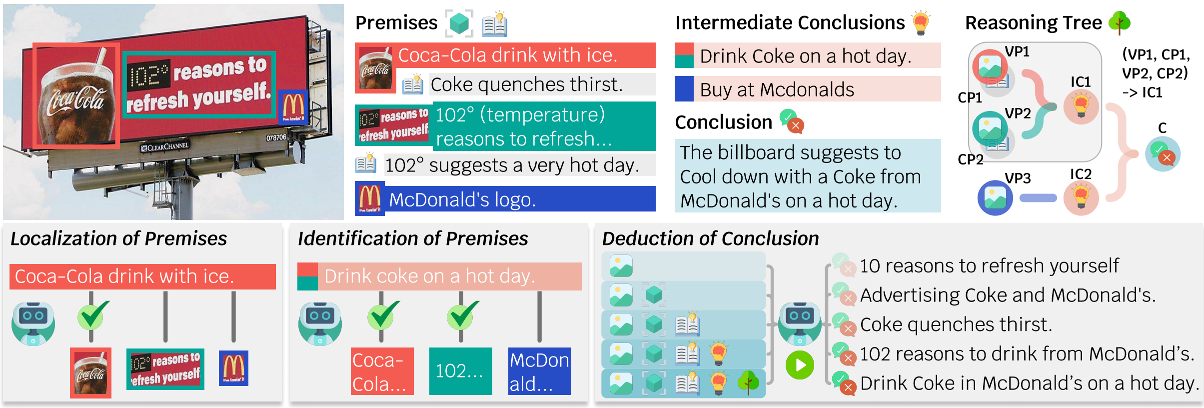


Figure 2: To identify the bottleneck in visual argument understanding, we define three tasks over VisArgs: *Localization of Premises* requires models to ground the visual premises. *Identification of Premises* necessitates models to infer the visual premise relevant to the given intermediate conclusion. *Deduction of Conclusion* studies the ability of models to deduce the argument's conclusion based on different levels of inputs.

argument tree consists of a root node (*conclusion*), some internal nodes (*intermediate conclusion*), and two types of leaf nodes (visual and commonsense premises).

Using VisArgs, we propose three complementary tasks to evaluate different aspects of machine capacity for comprehending visual arguments as illustrated in Fig. 2: 1) *Localization of Premises*: associates the description of a visual premise with a specific region in the image, 2) *Identification of Premises*: Given an image and an (intermediate) conclusion, retrieves the necessary visual premises to support the conclusion, and 3) *Deduction of Conclusion*: generates the conclusion with increasing detail of the annotated visual argument.

Experiments on VisArgs demonstrate that the main bottleneck for machine understanding of visual arguments is *selective vision,* i.e., *Identification of Premises* relevant to a given conclusion (see § 5.2). We show that while machines can identify visual premises within an image (albeit worse than human agreement, see *Localization of premises* § 5.1), they struggle to discern which premises are relevant to the conclusion among them. Results on our final *Deduction of Conclusion* task (§ 5.3) additionally support the hypothesis that difficulties in understanding visual arguments do not stem from deficiencies in raw vision capacity. There, we controlled the level of input to the algorithm, ranging from raw images to explicit reasoning trees. The greatest accuracy gains came from the inclusion of *relevant* visual cues, further supporting our main hypothesis. In all visual argument understanding tasks, machines perform worse than human agreement, providing avenues for future work.

In conclusion, our results suggest that selective attention to visual cues is the main bottleneck for the current AI capacity to understand visual arguments. This finding also establishes visual argument understanding as a distinct area of study in the computational domain: vision does not precede, but works jointly with reasoning in terms of understanding visual arguments. We expect that VisArgs will be utilized as a diagnostic benchmark for selective vision in future multimodal models: even the best current models lag significantly behind human performance in our *Identification of Premises* and *Deduction of Conclusion* tasks.

2 Related Work

Visual arguments are arguments built on visual medium (Boland, 2005). Unlike typical images, a visual argument is intentionally organized to persuade viewers to a certain conclusion (Birdsell and Groarke, 1996; Boland, 2005). This work builds upon to ongoing debates in the human studies literature about the nature of visual arguments (Johnson, 2003; Tseronis, 2018). Our results (§ 5) suggest that understanding visual arguments requires focusing on a subset of the visual context: not all visual cues contribute, and identifying the relevant ones is the key necessity. This task one of *selective vision*: the human capability to focus on behaviorally relevant stimuli. (Desimone and Duncan, 1995). Examples of visual arguments are prevalent in advertisements (Kjeldsen, 2012; Zhang et al., 2018; Ye et al., 2019), cartoons (Birdsell and Groarke, 2007), mathematical educations (Inglis and Mejía-Ramos, 2009), and, arguably, diagrams (Kembhavi et al., 2016; Alikhani and Stone, 2018).

Multimodal reasoning. Recent studies have introduced various multimodal models capable of sophisticated reasoning across different modalities, such as vision and language. Models such as LLaVA (Liu et al., 2023a), Idefics2 (Laurençon et al., 2024), and Qwen-VL (Bai et al., 2023) are built on pretrained large language models (e.g., LLaMA (Touvron et al., 2023)) and integrate vision encoders. Others, including OFA (Wang et al., 2022) and Unified-IO (Lu et al., 2022), are developed from scratch. These models excel in tasks such as localization, image captioning, and commonsense reasoning. Furthermore, models such as Unified-IO-2 (Lu et al., 2023) and GPT-4-O (Achiam et al., 2023) can understand audio, while others (Zellers et al., 2022; Han et al., 2023a) support video understanding, demonstrating broad multimodal reasoning capabilities.

Beyond factual visual understanding. Visual comprehension is moving beyond factual understanding to include various types of writing. These include visual commonsense reasoning (Zellers et al., 2019; Park et al., 2020; Han et al., 2023b; Hessel et al., 2022), humor understanding (Hessel et al., 2023; Hyun et al., 2023), and understanding social interaction (Zadeh et al., 2018). Of particular relevance to our work is visual metaphors (Akula et al., 2023), which express abstract concepts with concrete visual cues. While some overlap exists in the images used, there are clear differences in intention and structure; not all metaphorical images present clear arguments and can be seen as visual arguments. Conversely, not all visual arguments depend on metaphors (Blair, 2012).

Argument structure. An argument is typically understood as a structure that starts from a set of premises (reasons) and ends in a conclusion, often represented symbolically as a tree (Whately, 1863; Freeman, 2011). While there have been extensions, including computational models of arguments (Bench-Capon and Dunne, 2007; Rahwan and Simari, 2009; Atkinson et al., 2017), we use the basic form of trees connecting premises to conclusions, following previous literature (Stab and Gurevych, 2014; Lawrence and Reed, 2020).

3 VisArgs Dataset

VisArgs comprises a total of 1,611 images featuring clear visual arguments. These images are categorized into 914 advertisement images and 697 cartoon images based on their sources. Each image in VisArgs is annotated with descriptions and bounding boxes for the visual premises (VP), descriptions of the commonsense premises (CP), the conclusion, and an argumentation tree (T) detailing the reasoning path from the premises to the conclusion (C). All descriptions are in English, with an average character length of 79, 91, 142, and 105 for VP, CP, C, and T, respectively. On average, each image contains 3.17 visual premises, 3.46 commonsense premises, and 2.88 intermediate conclusions.

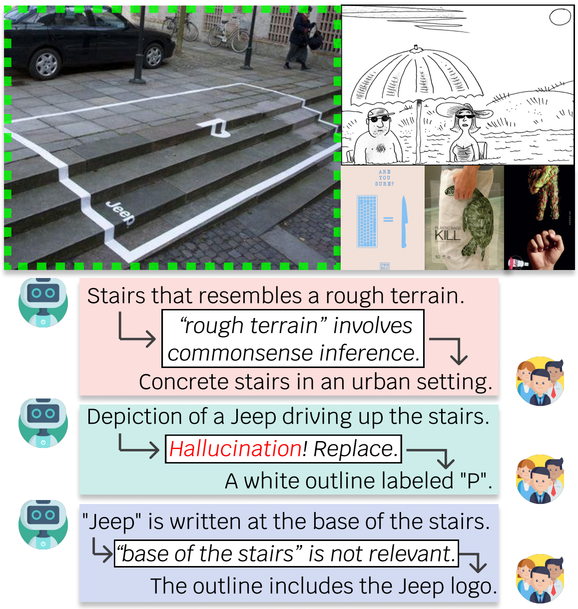


Figure 3: Human workers iteratively refine initial data produced by machines in VisArgs annotation process.

3.1 Annotation Process

We partially rely on GPT-4-O (Achiam et al., 2023) for initial annotations. However, these machine-generated annotations serve only as preliminary seeds, which are then extensively refined by experienced human workers, as illustrated in §3. The machine's role is merely to provide imperfect starting points to facilitate the human annotation process. Below, we detail our annotation procedure.

Collecting Images. We manually collect around 1,600 images from Pinterest.[2] Starting with keyword-based searches (*e.g. creative ads*), we expanded our collection by exploring related images. Cartoons (which often contain visual arguments (Birdsell and Groarke, 1996)) were sourced from a dedicated website.[3] We manually collected around 1,600 cartoons from various categories, in-

[2] www.pinterest.com
[3] www.cartoonmovement.com

cluding politics, education, and environment. For both categories, we followed previous work (Schuhmann et al., 2022; Lee et al., 2021) by including URLs to the images to comply with licensing terms. Refer to Appendix A for details.

Describing Visual Premises. The next step is to explicitly describe the visual argument within each image. However, during the early stages of our annotation process, we discovered that although humans can naturally understand visual arguments, they often find it challenging to articulate their interpretation into structured argumentation trees. Therefore, we used an AI model (GPT-4-O) to generate initial candidates. Human workers then select and modify these initial annotations, as shown in Fig. 3. To facilitate this process, we break down the annotation into two steps: describing the visual premises and specifying the argument structure.

Given an image containing a visual argument, we instructed the model to generate a set of visual premises necessary to support the argument (refer to Appendix J for further details). However, the AI model often fails to fully comprehend the visual argument. To address this, we engaged a pool of experienced human workers to review the machine-generated outputs. They selected the correct visual premises and made necessary modifications to ensure accuracy and coherence. Additionally, we identified that a model-generated visual premise sometimes contains multiple atomic premises. We instructed the reviewers to separate these merged premises into individual atomic premises. Further details are provided in Appendix A.

Specifying Argument Structure. Given the visual premises and the image, we further annotate three components constituting the argumentation structure: commonsense premises, conclusions, and argument trees. As in the previous stage, we first generate initial candidates using an AI model. For this stage, we impose an additional criterion: the set of selected premises should be both necessary and complete (refer to Appendix J). The same pool of human workers then adjust the annotations for greater accuracy. The workers first verify the correctness of the conclusion and discard the image if it is incorrect. They then identify and correct any errors, including semantic and structural mistakes. We discarded 1,593 of the 3,204 images in this process. Details are provided in Appendix A.

Visual Grounding. Lastly, we manually gather bounding box annotations for each visual premise to finalize the multimodal annotations. We assume a one-to-one relationship between each bounding box (vp_i^r) and its corresponding textual description (vp_i^d). Annotators are instructed to ensure accurate matching and precise bounding box tightness, as detailed in Appendix A.

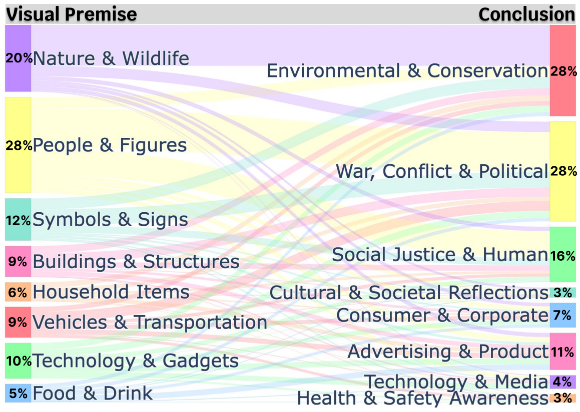


Figure 4: Variety of the topics represented in the visual premises and conclusions in VisArgs.

	Recall	Hit rate
LLaVaNeXT	0.48	0.14
LLaVa-LLaMa3-Docci	0.27	0.02
ShareCaptioner	0.40	0.12

Table 1: Frequency of detailed captions containing visual premises. *Hit rate* denotes how often all visual premises per image are included in the captions.

3.2 Data Analysis

Topic Diversity. To gauge the diversity of topics covered in VisArgs, we run zero-shot categorization using GPT-4-O and LLaMa3 (AI@Meta, 2024) to classify the topics of visual premises and conclusions. The topics cover a wide range of visual objects and argument topics, as shown in Fig. 4. Refer to Appendix B for details.

Visual Cues *vs*. Dense Captioning. In theory, selective attention to visual premises could be collapsed into an NLP problem by describing *everything* in an image. To test this counter-hypothesis, we manually check how often the visual premises are contained in the outputs of detailed captioning models. We include three baselines here: a generalist (LLaVA-Next (Liu et al., 2024b)), a specialist (ShareCaptioner (Chen et al., 2023)), and LLaVA-LLaMa3 (XTuner Contributors, 2023) fine-tuned on a detailed captioning corpus (DOCCI (Onoe et al., 2024))[4]. Tab. 1 summarizes our manual inspection of 100 images, showing that the detailed

[4] huggingface.co/gokaygokay/llava-llama3-docci

	Acc.	Prec.	Rec.	F1	Corr. (ρ)
BLEU-4	67	44	67	53	18
ROUGE	75	76	75	72	35
CIDEr	72	70	72	70	26
GPTEval	75	83	75	76	53
BERTScore	**94**	**94**	**93**	**93**	**59**

Table 2: Correlation of each metric with human decisions in the *Deduction of Conclusion* task.

captions insufficiently capture the visual premises, with the hit rate staying below 15% for all models.

Safety. Since we did not initially filter for safety, we now analyze the safety of VisArgs using standard models. For textual safety, we utilize the Perspective API[5], and for visual domains, we employ LAION-Safety[6]. The toxicity scores for textual descriptions were 0.03 for visual premises and 0.07 for conclusions. Also, given the threshold of 0.7, no descriptions and visual premises were classified as toxic. Furthermore, only 71 among 1611 images are classified as unsafe. Manual inspection reveals that such "unsafe" images were social campaigns advocating *against* the harmful behaviors which presumably triggered the LAION detector.

4 Task Overview

We pose three tasks based on VisArgs for a structured analysis of how machines understand arguments presented in visual form.

An instance of VisArgs consists of an image I, a set of visual premises VP $= \{(vp_0^d, vp_0^r), (vp_1^d, vp_1^r), \ldots\}$ with textual description vp^d along with region grounding with a bounding box $vp^r = \langle x, y, h, w \rangle$, a set of commonsense premises CP $= \{cp_0^d, cp_1^d, \ldots\}$, and the conclusion in textual form C. Further, a single argument tree for each image is built on the premises. Each tree $t \in T$ represents a reasoning path leading to the conclusion C. The nodes N of a tree consist of the following: 1) leaf nodes: subsets of the union of the visual and commonsense premises VP$\cup$CP. 2) internal nodes: elements of the set of intermediate conclusions IC. 3. root node: the conclusion C. An edge e of the tree connects a subset of nodes $\bar{N} \subset$ VP $\cup$ CP $\cup$ IC to either an intermediate conclusion $ic \in$ IC or the final conclusion C.

[5] www.perspectiveapi.com; June 2024 version.
[6] www.github.com/LAION-AI/LAION-SAFETY

4.1 Localization of Premises

The first task focuses on assessing whether machines can accurately align visual premises (VP^d) with the corresponding regions (VP^r) in a given image (I), requiring minimal computational reasoning capabilities. It aims to determine if difficulties in understanding visual arguments originate from basic object detection stages.

We investigate two setups based on the algorithm's ability to output bounding box labels: First, *closed-set grounding* is designed for a broad range of models that lack explicit grounding capabilities. The problem is formulated as a retrieval task where the goal is to match a region in the image (vp_i^r) with an appropriate description (vp_i^d). We adapt standard image-text matching models (*e.g.* CLIP) to perform grounded image-text matching. More details can be found in § 5. Second, *open-set grounding* tests models with explicit grounding capabilities. The task is framed as a visual grounding problem (Yu et al., 2016), where the machine must locate an object in an image based on a natural language expression. Both the ground truth and machine output are represented as bounding box coordinates $\langle x, y, h, w \rangle$. Performance is evaluated using the intersection over union (IoU) ratio, with predictions considered correct if IoU ≥ 0.5.

4.2 Identification of Premises

The second task tests the machines' capabilities to discern visual premises that would better support the given conclusion. Given the image I, the intermediate conclusion ic, and a superset of the gold text descriptions of the visual premises $S \supset$ VP^d, the machine should retrieve a correct visual premise $vp_i^d \in$ VP^d. Note that the candidate set S contains a single ground truth premise vp_i^d and a fixed number $K = 2$ of negative premises.

The complexity of a retrieval task is impacted by the choice of the negative set. We explore four types of *global* samplers and a single *local* sampler for constructing the negative set. The global samplers source the negatives from visual premises that *do not* correspond to the selected image. The only difference is the sample selection strategy: 1. *Random* sampling samples uniformly without replacement. 2. *Visual* sampling samples from the top premise descriptions that are the closest to the given image. We use CLIPScore (Hessel et al., 2021) for the multimodal scoring. 3. *Textual* sampling samples from the top premise descriptions that are the closest to the ground truth premise. We use cosine similarity on the ColBERT (Khattab and Zaharia, 2020) representation space for the textual scoring. 4. *Mixed* sampling combines textual and visual sampling by visually selecting from the top 10 textual retrieval results.

For *local* sampling, we select from the visual premises that *do* correspond to the given image. Relying on our argumentation tree annotation, we can automatically obtain the set of local visual premises that does not help justify the given intermediate conclusion ic. we sample uniformly without duplicates from the local pool and name the method 5. *Semantic* sampling due to its argumentation-dependent nature. Additionally, we report human performance on 100 random samples to mitigate the risk of false negatives.

4.3 Deduction of Conclusion

The final task is to evaluate how each component (I, VP, CP, IC, and T) influences the deduction of the conclusion C. We approach this as a sequence-to-sequence task aimed at generating C. While this allows flexible output formats, it complicates evaluation because the machine-generated text must be compared to the free-form label. Common text comparison practices, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and CIDER (Vedantam et al., 2015) measure surface form similarity, not semantic similarity between conclusions. Alternatively, prompt-based evaluation using general reasoners (*e.g.* GPT-4) (Achiam et al., 2023) can be biased by factors including candidate order (Pezeshkpour and Hruschka, 2023). Human verification, though ideal, is costly and hard to reproduce. We conduct a small-scale comparison study (see Tab. 2) to verify that the model-based metric BERTScore (Zhang* et al., 2020) provides the most stable estimate, making it our primary metric. Details are in Appendix D.

5 Experiments

5.1 Localization of Premises

Localization of Premises tests the visual grounding capabilities of machines. Given the image I and description of a visual premise vp^d, the goal is to find a corresponding region vp^r in the image.

Metrics and Models. For *closed-set grounding*, which is an N-way classification task, the goal is to match the given description with the correct bounding box. To evaluate standard image-text matching algorithms (*e.g.* CLIP), we crop the regions accordingly. The models for this task include various CLIP-based models (CLIP (Radford et al., 2021) with different backbones and SigLIP (Zhai et al., 2023)) and a multitask model OFA (Wang et al., 2022). For *open-set grounding*, which is to locate an object in an image based on a natural language expression, we instruct the models to output bounding box coordinates and we compare them to the ground truth region. A predicted coordinate is considered correct if its intersection over union with the gold label is at least (IoU $\geq$ 0.5). We use a diverse set of models that support local region output formats, UNINEXT-H (Yan et al., 2023), LISA (Lai et al., 2023), Unified-IO-2 (Lu et al., 2023), OFA, MM-G-DINO (Liu et al., 2023b).

	Acc. (%)		
	Ads	Cartoon	All
Random	33.33	33.33	33.33
Human	100.00	100.00	100.00
$\text{CLIP}_{\text{RN50}}$	80.83	82.72	81.91
$\text{CLIP}_{\text{ViT-L}}$	82.72	82.96	82.85
$\text{CLIP}_{\text{ViT-L@336}}$	82.09	83.26	82.76
SigLIP	86.10	86.67	86.43
AlphaCLIP	75.15	77.44	76.45
OFA_{Base}	68.75	75.71	72.71
$\text{OFA}_{\text{Large}}$	72.01	79.18	76.10

Table 3: *closed-set* results in *localization of premises*.

	IoU	Acc. (%)
UNINEXT-H	38.75	35.58
LISA	44.25	44.62
Unified-IO-2	48.61	47.15
OFA	50.14	49.13
MM-G-Dino	55.02	54.98

Table 4: *open-set* results in *localization of premises*.

Results. Tab. 3 demonstrates that current models are generally effective in matching descriptions of visual premises to the correct regions in images, thereby meeting the basic vision requirements for understanding visual arguments. However, the results for *open-set* grounding, shown in Tab. 4, are somewhat mixed: the scores are acceptable but not uniformly high. We traced this performance decline to the nature of zero-shot object detectors, which are designed to detect concrete objects and clear segments. In contrast, our bounding boxes are more semantic (Guo et al., 2018). Visual examples can be found in Appendix G.

	Global				Local	
	Random	Visual	Textual	Mixed	Semantic	Semantic
Random	33.33	33.33	33.33	33.33	33.33 (-)	
Human	100.00	99.00	94.00	100.00	98.00 (↑ 4.00)	+ G.T region
OFA	0.00	0.00	0.00	0.00	0.00 (-)	-
Qwen-VL-Chat	86.05	85.77	70.67	75.57	49.74 (↓ **20.93**)	-
CogVLM	97.46	96.39	88.00	92.22	65.31 (↓ **22.69**)	-
Idefics2	98.68	97.83	91.80	95.07	75.01 (↓ **16.79**)	-
InstructBLIP	83.77	79.23	66.95	71.37	61.90 (↓ **5.05**)	78.13 (↑ 16.23)
Unified-IO-2	98.42	96.99	86.87	92.81	34.74 (↓ **52.13**)	84.39 (↑ 49.65)
LLaVA-1.5	98.65	97.91	83.74	89.86	67.43 (↓ **16.31**)	76.67 (↑ 9.24)
LLaVA-NeXT	97.66	96.20	80.90	85.86	78.53 (↓ **2.37**)	82.19 (↑ 3.66)
GPT-4-O	-	-	-	-	79.50 (-)	-

Table 5: Results of the *Identification of Premises* task. Difference between the lowest score in *global* and *local* setup for each model are highlighted.

	Image	+ VP	+ CP	+ Tree
LLaMA3	-	30.2	37.8 (↑**7.6**)	40.8 (↑2.0)
Mistralv0.2	-	18.9	30.2 (↑**11.3**)	36.6 (↑6.4)
Zephyr	-	20.6	28.7 (↑**8.1**)	36.5 (↑7.8)
OFA	-41.3	-24.6 (↑**16.7**)	-16.5 (↑8.1)	-13.9 (↑2.6)
Qwen-VL-Chat	12.8	23.7 (↑**10.9**)	30.2 (↑6.5)	32.7 (↑2.5)
CogVLM	25.7	30.7 (↑**5.0**)	33.6 (↑2.9)	36.3 (↑2.7)
Idefics2	16.4	22.8 (↑6.4)	29.5 (↑6.7)	36.6 (↑**7.2**)
InstructBLIP	-18.4	16.6 (↑**35.0**)	28.9 (↑12.3)	32.2 (↑3.3)
Unified-IO-2	-9.9	-3.4 (↑6.5)	4.2 (↑**7.6**)	8.0 (↑3.8)
LLaVA-1.5	2.2	20.0 (↑**17.8**)	29.6 (↑9.6)	33.7 (↑4.1)
LLaVA-Next	15.1	28.4 (↑**13.3**)	34.3 (↑5.9)	39.5 (↑5.2)
GPT-4-O	25.5	-	34.3 (↑**8.8**)	41.0 (↑6.7)

Table 6: Results of the *Deduction of Conclusion* task, showing how incremental additions of inputs affect the correctness of the conclusion. Scores are presented using BERTScore, with similar trends observed across other metrics as detailed in Appendix F.

5.2 Identification of Premises

Identification of Premises tests the selective attention capabilities, i.e., selecting necessary visual cues to understand an argument. Given the image I and an intermediate conclusion ic, the goal is to select a visual premise vp^d that leads to this intermediate conclusion.

Metrics and Models. For this task, we retain only intermediate conclusions that have at least two unrelated visual premises within the image. We report classification accuracy based on a single gold visual premise and two negative candidates. The negative sets are sourced as described in § 4.2 and are categorized into *random*, *visual*, *textual*, *mixed*, and *semantic* sets. Given the task's requirement for understanding argumentation structure, the models evaluated are primarily multimodal large language models with adequate reasoning capabilities. We experiment with a broad selection of models: OFA (Wang et al., 2022), Qwen-VL-Chat (Bai et al., 2023), CogVLM (Wang et al., 2023), Idefics2 (Laurençon et al., 2024), InstructBLIP (Dai et al., 2024), Unified-IO 2 (Lu et al., 2023), LLaVa-1.5 (Liu et al., 2024a), and LLaVa-Next (Liu et al., 2024b). For the sake of brevity, we do not report per-category results (*Ads* and *Cartoon*) here. Refer to Appendix F for full results.

Results. Tab. 5 highlights a significant trend: models struggle to distinguish negatives within the image (*local*), but excel in identifying *global* negatives. A major challenge for most models was handling *semantic* negatives within the same image, as evidenced by the generally wide margin between models' performance on *global* and *local* setups. Still, the *global* negative samples exhibited more pronounced distinctions based on their sampling scheme. Negatives sampled uniformly were distinguishable by most models with $\geq 90\%$ accuracy. In contrast, retrieval methods proved more challenging across the board, particularly for negatives retrieved using the text-to-text similarity model (*textual*), which increased the problem complexity for most models. Notably, OFA failed to follow zero-shot instructions for multiple-choice answering, scoring close to zero. Finally, we also present results for cropped ground-truth region images. Although cropped images are not lossless representations of the regions, all models exhibited significant improvements, indicating that the ability to infer relevant visual cues is indeed a critical challenge. Thus, we conclude that models struggle to infer which visual cues support the argument.

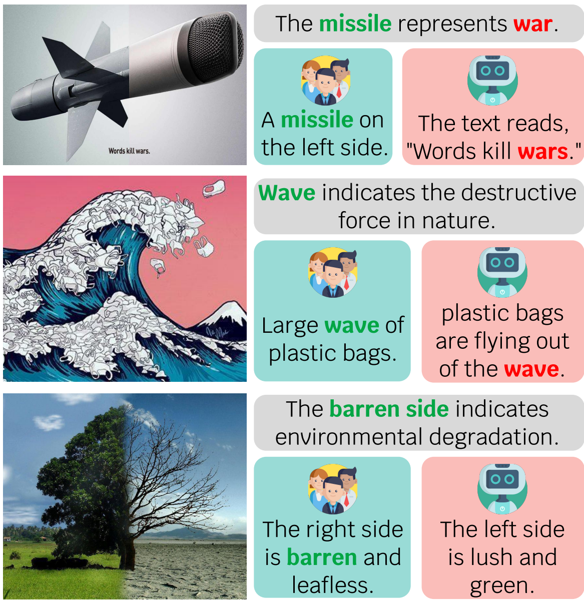


Figure 5: Failure cases of LLaVA-1.5 in *Identification of Premises*. The model incorrectly reasons about relevant objects, relying instead on common words.

	Image	Δ VP	Δ CP	Δ Tree
LLaVA-1.5	3.48	↑ **13.29** (5.42)	↑ 8.57 (4.75)	↑ 4.72 (4.34)
LLaVA-Next	15.04	↑ **11.28** (1.21)	↑ 6.72 (2.78)	↑ 4.14 (4.02)

Table 7: Mean of incremental improvements in BERTScore with each additional input across four different prompts in *Deduction of Conclusion*. Standard deviations are shown in parentheses.

5.3 Deduction of Conclusion

Deduction of Conclusion evaluates the comprehensive ability to deduce the conclusion of an argument. Given a subset of inputs among the image I, the visual premises VP, the commonsense premises CP, and the reasoning tree T, the objective is to generate the conclusion C of an argument.

Metrics and Models. As discussed earlier in § 4.3, we use BERTScore as the primary metric. We supplement this with three additional static metrics (Bleu-4, ROUGE-L, CIDEr) in Appendix F. The models tested in this task include all the multimodal LLMs used in the previous experiment and text-only LLMs (LLaMa-3-Instruct (AI@Meta, 2024), Mistral-Instruct (Jiang et al., 2023), and Zephyr (Tunstall et al., 2023)). All LLMs considered here are the 7 $\sim$ 8b sized variants. The LLMs do not take the image as an input.

Results. Table 6 shows the results for this task. As expected from previous tasks, most models experience the highest gain from the additional information provided by the ground-truth set of visual premises. This supports our hypothesis that selective attention to visual premises is a bottleneck in understanding visual arguments in current models. Also, both multimodal and text-only models benefited from commonsense premises and reasoning trees in most setups, indicating that models cannot yet perfectly understand visual arguments in a text-only format and benefit from explicit reasoning process information. We note that OFA struggled to follow the instruction format, leading to subzero scores. Although rare, BERTScore, based on cosine similarity, can yield negative values. We also clarify that the multimodality of the *deduction of conclusion* task resides in the visual premises, making it solvable by text-only models given them.

5.4 Diagnostics

Prompt Robustness. To ensure the robustness of our empirical results, we differentiated the prompts provided to the models. As shown in Tab. 7, the trend of gains remained stable across four different prompts, confirming the validity of our tests. For detailed prompts, refer to Appendix M, and for results in other tasks, see Appendix E.

Error Analysis. Fig. 5 provides qualitative examples of failure cases. We present straightforward instances to clearly explain the errors. In these cases, the models fail to reason about the relevant object, which is the subject of the given intermediate conclusion, and instead rely on common words, leading to incorrect inference results.

6 Conclusion

We introduce VisArgs, a curated and annotated benchmark for visual argument understanding. Using our benchmark, we affirm a compelling hypothesis: selective vision is a critical bottleneck for visual reasoning in current machines. We aim for our benchmark to serve as a resource for advancing multimodal intelligence beyond passive captioning. Future work includes:

1. Conditional Saliency Analysis: It is demonstrated that the saliency required for visual arguments differs from that needed for passive captioning. Can the varying saliency requirements across different tasks be analyzed?
2. Extending Modalities: In speech recognition, non-conditional selective attention is known as the cocktail party effect. Would conditional selective attention be necessary in modalities other than vision as well?

7 Limitations

VisArgs, which is built on advertisements and cartoons from web sources, does not encompass all forms of visual arguments. Visual arguments also include various forms of media including mathematical diagrams (Inglis and Mejía-Ramos, 2009) and videos, such as films (Alcolea-Banegas, 2009). Consequently, the findings of this study do not represent all forms of visual arguments.

Additionally, the annotations for VisArgs are created by two NLP researchers with similar cultural backgrounds. Although a different group of human evaluators validated these annotations, future research should consider individual variances in the interpretation of visual arguments and the reasoning processes identified by reasoning trees.

Finally, we excluded images containing written text in non-English languages when curating VisArgs, as the annotators were not familiar with other languages. This limitation may confine the cultural context covered by VisArgs, thus representing only a partial depiction of visual arguments. Since the logical relations forming a visual argument can depend on culture-specific elements, this skewed distribution of images can lead to a biased understanding of visual arguments.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, et al. 2023. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23201–23211.

Jesús Alcolea-Banegas. 2009. Visual arguments in film. *Argumentation*, 23:259–275.

Malihe Alikhani and Matthew Stone. 2018. Arrows are the verbs of diagrams. In *COLING*.

Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. 2017. Towards artificial argumentation. *AI magazine*, 38(3):25–36.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Trevor JM Bench-Capon and Paul E Dunne. 2007. Argumentation in artificial intelligence. *Artificial intelligence*, 171(10-15):619–641.

David S Birdsell and Leo Groarke. 1996. Toward a theory of visual argument. *Argumentation and advocacy*, 33(1):1–10.

David S Birdsell and Leo Groarke. 2007. Outlines of a theory of visual argument. *Argumentation and advocacy*, 43(3-4):103–113.

J Anthony Blair. 2012. The possibility and actuality of visual arguments. *Groundwork in the Theory of Argumentation: Selected Papers of J. Anthony Blair*, pages 205–223.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Julie E Boland. 2005. Visual arguments. *Cognition*, 95(3):237–274.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Robert Desimone and John Duncan. 1995. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222.

James B Freeman. 2011. *Dialectics and the macrostructure of arguments: A theory of argument structure*, volume 10. Walter de Gruyter.

Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. 2018. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93.

Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, and Youngjae Yu. 2023a. Champagne: Learning real-world conversation from large-scale web videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15498–15509.

Seungju Han, Junhyeok Kim, Jack Hessel, Liwei Jiang, Jiwan Chung, Yejin Son, Yejin Choi, and Youngjae Yu. 2023b. Reading books is great, but not if you are driving! visually grounded reasoning about defeasible commonsense norms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 894–914.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*.

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *European Conference on Computer Vision*, pages 558–575. Springer.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714.

Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. 2023. Smile: Multimodal dataset for understanding laughter in video with language models. *arXiv preprint arXiv:2312.09818*.

Matthew Inglis and Juan Pablo Mejía-Ramos. 2009. On the persuasiveness of visual arguments in mathematics. *Foundations of Science*, 14:97–110.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ralph H Johnson. 2003. Why "visual arguments" aren't arguments.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Jens E Kjeldsen. 2012. Pictorial argumentation in advertising: Visual tropes and figures as a way of creating visual argumentation. In *Topical themes in argumentation theory: Twenty exploratory studies*, pages 239–255. Springer.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. 2021. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10274–10284.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *NeurIPS*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*.

John Lubbock. 1893. *"The" beauties of nature and the wonders of the world we live in*, volume 2893. Bernhard Tauchnitz.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. 2024. DOCCI: Descriptions of Connected and Contrasting Images. In *arXiv:2404.19753*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 508–524. Springer.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Plotly. 2015. Collaborative data science.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Iyad Rahwan and Guillermo R Simari. 2009. *Argumentation in artificial intelligence*, volume 47. Springer.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 46–56.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Assimakis Tseronis. 2018. Multimodal argumentation: Beyond the verbal/visual divide. *Semiotica*, 2018(220):41–67.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Richard Whately. 1863. *Elements of logic: Comprising the Substance of the Article in the Encyclopaedia Metropolitana*. Sheldon.

XTuner Contributors. 2023. Xtuner: A toolkit for efficiently fine-tuning llm. `https://github.com/InternLM/xtuner`.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. 2023. Universal instance perception as object discovery and retrieval. In *CVPR*.

Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. 2019. Interpreting the rhetoric of visual advertisements. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1308–1323.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *BMVC*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Data Annotation Details

Human Resources. To ensure a comprehensive understanding of the intricate requirements of our setup and maintain consistency across annotations, two of this paper's authors conducted the entire annotation process. Three volunteers from the NLP research community did the human evaluation.

Annotation Interface. We used a custom-built interface for efficient and convenient image annotation. The interface is depicted in Fig. 6 and Fig. 7. Additionally, we provide a snapshot of the human evaluation interface for *Identification of Premises* in Fig. 8. We will open-source this interface along with the dataset.

B Analyzing Topic Diversity

Initially, we considered using the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) method for data visualization, following previous literature (Hessel et al., 2022). However, we found that LDA based on Bag-of-Words representations could not generate meaningful clusters or labels for conclusion topics. As a solution, we developed an adaptive semantic classification technique using multimodal large language models:

Defining Class Labels. We utilize GPT-4-O. We first sample 400 sentences each for VP and C, and then feed them to GPT with the following instructions: For VP: *"Give me well-balanced 10 object type classes for these texts (e.g., eating & dining, environments & landscapes, attire). Just classes."* For C: *"Give me well-balanced 10 classes for these texts. Just classes."* After receiving the 10 classes from the GPT, we manually refine these classes into 8 classes for both VP and C.

Labelling Data. We use a pretrained language model to classify visual premises (VP) and conclusions (C) in a zero-shot manner. We provide the following input to the LLaMA-3[7] LLM:

```
Classes: {}
Your task is to classify a sentence into
the given classes.
Give me just the class.
Sentence: {}
```

Visualization. We use the Plotly (Plotly, 2015) library.

model	dtype	#parameter	version
CLIP	-	623M	RN50x64
CLIP	-	427M	ViT-L/14
CLIP	-	427M	ViT-L/14@336px
SigLIP	-	652M	large-patch16-384
AlphaCLIP	-	428M	clip_l14_336_grit_20m_4xe
UNINEXT-H	-	775M	image_joint_vit_huge_32g
LISA	-	7B	xinlai/LISA-7B-v1
MM-G-DINO	-	343M	grounding_dino_swin-l_pretrain_all
LLaVA-1.5	FP16	7B	llava-1.5-7b-hf
LLaVA-NeXT	FP16	7B	mistral-v0.2
Idefics2	FP16	8B	chatty
OFA	-	470M	vqa-pretrain-large
QwenVLChat	BF16	9B	Qwen-VL-Chat
CogVLM	BF16	17B	cogvlm-chat-hf
InstructBLIP	FP16	7B	instructblip-vicuna-7b
Unified-IO-2	-	3B	uio2-xl

Table 8: Details on the models used in our experiments.

C Experiment Details

C.1 Localization of Premises

For closed-set grounding, we utilized CLIP, SigLIP, AlphaCLIP, and OFA. We measured the alignment between regions and descriptions of visual premises using image-to-text cosine similarity scores. The input regions were provided as cropped images. A model output was considered correct (True) if the similarity between the ground-truth region and the given description was the highest among all candidates; otherwise, it was marked incorrect (False).

For open-set grounding, we employed object grounding models such as MM-GDINO, UNINEXT-H, LISA, OFA, and Unified-IO-2 to directly generate bounding box coordinates. We applied a threshold of 0.35 to the outputs, merging the selected regions into the tightest rectangle union. For LISA, we converted the output segmentation mask into bounding boxes. We then calculated the Intersection over Union (IoU) score for each bounding box. To compute the accuracy metric, we used a threshold of 0.5 for binary classification over the IoU. We calculated the local mean, which is the mean per visual premise in an image, and the mean per image.

C.2 Identification of Premises

We utilized OFA, Qwen-VL-Chat, CogVLM, Idefics2, InstructBLIP, Unified-IO-2, LLaVA-1.5, LLaVA-Next, and GPT-4-O for our experiments. We created multiple-choice questions with three possible answers: one correct answer and two incorrect answers. Five conditions were set for sampling the negatives for incorrect answers:

[7] meta-llama/Meta-Llama-3-8B-Instruct

- *Random Sampling:* This global sampler selects samples uniformly without duplication.
- *Visual Sampling:* This global sampler chooses the top 2 premise descriptions most similar to the image, using CLIP to score the cosine similarity between the image and text. We set the CLIP similarity threshold to 0.24 to ensure negative premises do not accurately describe the image.
- *Textual Sampling:* This sampler selects the top 2 premise descriptions most similar to the ground truth premise, using ColBERT to score the cosine similarity between texts. We set the ColBERT similarity threshold to 25 to prevent negative premises from accurately describing the image.
- *Mixed Sampling:* This approach combines visual and textual sampling, visually selecting from the top 10 textual retrieval results.

To ensure a fair comparison across various negative sampling methods, we use only intermediate conclusions that have three or more related visual premises. This results in 1,775 visual premises for the advertisement category and 1,774 for the cartoon category, totaling 3,549 visual premises, which is 62.34% of the overall visual premises.

Human Evaluation. We randomly selected 100 images from each data category and had human annotators perform the same tests as the machines across all negative set setups. The results demonstrated that humans achieved nearly perfect accuracy in this task, as shown in Tab. 12.

C.3 Deduction of Conclusion

We conducted experiments on both Multi-Modal Large Language Models (MLLM) and Large Language Models (LLMs). The MLLMs used in our experiments include LLaVA-1.5, LLaVA-NeXT, Idefics2, OFA, InstructBLIP, Qwen-VL-Chat, CogVLM, and Unified-IO-2. The LLMs include LLaMA-3, Mistral, and Zephyr.

Prompting. Before conducting the experiments, we established a set of instructions to be applied to all models to elicit appropriate responses. During this process, we encountered several issues with prompt engineering, such as model refusal to address controversial or unsafe questions, the inclusion of unnecessary tokens, multiple sentences, and the positioning of image tokens. Ultimately, we decided on the following prompt: *"<image> <information> Your task is to answer what the image wants to convey. You should respond in only one sentence without any unnecessary prefixes. ANSWER:"*

C.4 Resource & Hyperparameters

Computation. We utilized RTX-4090 and A6000 GPUs for our experiments. All models, except for CogVLM, were implemented using RTX-4090 GPUs. Due to the size of its model weights, CogVLM was implemented on an A6000 GPU. Each model required up to 8 RTX-4090 GPU-hours per task. In total, conducting all tasks demanded 200 RTX-4090 GPU-hours.

Hyperparameters. Our experiments are deterministic, given the pretrained model weights, the greedy decoding scheme, and the instruction prompts. We explore prompt diversification in § 5.4 and Appendix E.

C.5 Model Details

We specify all exact model identifiers and sizes in Tab. 8.

D Comparison of Metrics for *Deduction of Conclusion*

Here, we describe details for human evaluation of goodness per each metric illustrated in Tab. 2.

Human Evaluation. We sampled 200 target images and collected responses from three models: LLaVA-Next, Qwen-VL-Chat, and GPT-4o. Human annotators then determined whether each model's conclusion was semantically similar to the reference conclusion.

Metrics. To evaluate accuracy, precision, recall, and F1-score, we first converted each metric into binary decisions using derived thresholds. We established these thresholds by training a logistic regression model on 100 pairs of metric scores and human decisions. Subsequently, we inferred binary decision labels on the remaining 100 pairs. The results are presented in Tab. 9. Additionally, the correlation between the metrics and human decisions is reported using Pearson's coefficient (Cohen et al., 2009).

E Prompt Robustness in *Identification of Premises*

Extending the robustness study in Tab. 7, we conducted a similar prompt diversification experiment for the task of *Identification of Premises*. By paraphrasing the original prompt as described in Appendix L, we performed the same evaluation. The results, presented in Tab. 10, demonstrate that our

	Accuracy	Precision	Recall	F1-score	Pearson Corr. (ρ)
BLEU-4	0.67	0.44	0.67	0.53	0.18
ROUGE	0.75	0.76	0.75	0.72	0.35
CIDEr	0.72	0.70	0.72	0.70	0.26
GPTEval	0.75	0.83	0.75	0.76	0.53
BERTScore	**0.94**	**0.94**	**0.93**	**0.93**	**0.59**

Table 9: Comparison of metrics with human decision on *Deduction of Conclusion*

		Global				Local
	Prompt	Random	Visual	Textual	Mixed	Semantic
LLaVA-NeXT	Original	97.10	96.14	80.53	84.70	77.51
InstructBLIP		90.65	84.53	71.54	74.75	58.21
LLaVA-NeXT	Paraphrase 1	97.24	96.59	80.98	85.63	77.60
InstructBLIP		90.93	84.73	72.78	74.98	59.68
LLaVA-NeXT	Paraphrase 2	97.60	96.25	81.46	86.00	76.67
InstructBLIP		93.32	89.83	79.01	81.71	64.50

Table 10: Assessment of prompt robustness with different paraphrases in *Identification of Premises*. Accuracy is measured as a percentage.

	IoU			Acc. (%)		
	Ads	Cartoon	All	Ads	Cartoon	All
UNINEXT-H	34.50	44.33	38.75	31.67	40.71	35.58
LISA	40.05	49.17	44.25	40.52	50.01	44.62
Unified-IO-2	45.81	52.29	48.61	44.66	50.43	47.15
OFA	49.10	51.49	50.14	49.06	49.22	49.13
MM-G-Dino	52.70	58.06	55.02	52.39	58.37	54.98

Table 11: *Open-set* grounding results in *localization of premises*.

experimental outcomes remain stable for *Identification of Premises* across different prompt paraphrases.

F Full Results

This section presents the comprehensive versions of the results summarized in the main paper. Tab. 11 displays the *open-set* grounding results for *Localization of Premises*, while Tab. 12 provides the results for *Identification of Premises*. The results for the task of *Deduction of Conclusion* are detailed by category: advertisements are shown in Tab. 13, cartoons in Tab. 14, and the average across both categories in Tab. 15.

G Qualitative Samples on *Open-Set* Grounding

To identify the cause of low performance in the *open-set* evaluation of the *Localization of Premises* task, we examine qualitative samples shown in Fig. 9. Traditional object detection models are typically trained on single object labels, whereas our semantic region labels may encompass multiple objects with similar meanings. Consequently, although the models may detect the correct target, the intersection over union (IoU) scores are lower, resulting in reduced accuracy.

H Qualitative Samples on *Deduction of Conclusion*

Inference results of different models with varying inputs are shown in Fig. 10 and Fig. 11. The outputs of the models display discrepancies; for instance, CogVLM exhibits weak conditioning on additional inputs, producing similar outputs despite the incremental increase in information provided through different inputs.

I Credits

We do not claim any rights to the images included in our dataset. Therefore, we provide only the URLs to the corresponding images instead of distributing the raw files. For usage outside of an academic context, please contact the copyright holders directly.

Figures. All icons used in the figures are from `www.flaticon.com`.

- Figure 1: `www.art-vibes.com/design/egle-plytnikaite-environmental-issues`
- Figure 2: `www.commarts.com/project/24399/mcdonald-s-refresh`
- Figure 3: `www.aisleone.net/2007/10/30/jeep/` | `www.nextml.github.io/caption-c`

ontest-data/dashboards/630.html | www.i.pinimg.com/originals/ac/32/16/ac321665c9e8f5feccc62eb3f6d09d37.jpg | www.ipnoze.com/publicite-sociale/ | www.adsoftheworld.com/campaigns/scissors-1e569372-d5e7-488b-9e06-8bf46580801e

- Figure 5: www.adsoftheworld.com/campaigns/words-1c383606-d2b3-4aea-9f19-c0627b6fb4ff | www.behance.net/gallery/68747547/The-Great-Plastic-Wave | www.fanpop.com/clubs/global-warming-prevention/images/33088666/title/global-warming-photo

J Prompts for Annotation

- **Annotation for Visual Premises**

```
Your task is to identify visual premises from the image. These are visual cues that support or illustrate the conclusion, enhancing the overall understanding and clarity of the image.

Example
Visual Premises (VP):
1. The image depicts a maze with entry point and exit.
2. At the entry point of a maze labeled "Start," there is a cigarette.
3. The exit of the maze is labeled "Lung Cancer."
4. There's a text saying, "Or you can start here," with an arrow pointing to another text that reads, "Make the right choice. DON'T SMOKE."
```

- **Annotation for Constructing Arguments**

```
Visual Premises (VP):
1. VP1
2. VP2
3. VP3

Given the visual premises of the image, your task is to generate the necessary commonsense premises and conclusion of the image. The conclusion should be one simple sentence. Then show the reasoning steps to reach the conclusion. The reasoning steps should include all visual premises and commonsense premises. You can refer to the following example.

Example
Visual Premises (VP):
1. The image depicts a maze with entry point and exit.
2. At the entry point of a maze labeled "Start," there is a cigarette.
3. The exit of the maze is labeled "Lung Cancer."
4. There's a text saying, "Or you can start here," with an arrow pointing to another text that reads, "Make the right choice. DON'T SMOKE."

Commonsense Premises (CP):
1. Mazes are often used to represent complex journeys or paths one must navigate.
2. Cigarettes are known to be harmful to health and a major cause of lung cancer.
3. The phrase "Make the right choice" implies that there is a decision to be made that can impact one's health .
4. Public health messages often use strong visuals to convey the importance of making healthy choices.

Conclusion (C):
The image is a public health message that illustrates the dangerous path from smoking to lung cancer while encouraging individuals to choose not to smoke for their health.

Reasoning Steps:
(VP1, CP1 -> IC1): The maze represents the difficult and potentially harmful journey.
(VP2, CP2 -> IC2): The presence of a cigarette at the maze's entry point indicates the start of this hazardous journey.
(VP3, CP2 -> IC3): Labeling the maze's exit as "Lung Cancer" directly links smoking to this deadly disease.
(VP4, CP3, CP4 -> IC4): The additional text offers an alternative choice to avoid smoking, emphasizing the importance of preventive health measures.
(IC1, IC2, IC3, IC4 -> C): The image is a public health message that warns about the risks of smoking and encourages making the right choice for one's health.

Answer
```

K Prompts for Evaluation

- **GPTEval**

```
Task Description: You will be given a ground truth sentence that describes an image and a model-generated sentence. Your task is to evaluate the semantic similarity between the model-generated sentence and the ground truth sentence. You don't need to give me any description. Just score should be answered.
```

```
Evaluation Criteria: T/F. False means the sentences are completely different. True means they mean exactly the same thing.

Ground Truth: {}
Generated: {}
```

L Prompts for *Identification of Retrieval*

- **Original Prompt**

```
<image>
The following are multiple choice questions (with answers) about image understanding.

When given an image, a conclusion, and several visual cue options, you need to identify the visual cue that best relates to the conclusion. To do this effectively, carefully analyze how each visual cue connects to the key elements of the conclusion. Select the visual cue that most directly supports or illustrates the conclusion, ensuring that it enhances the overall understanding and clarity of the message. Answer A), B), or C) with no additional explanation. Conclusion: {conclusion}
{vp_options}
ANSWER:
```

- **Paraphrase 1**

```
<image>
The following are multiple choice questions (with answers) about image understanding.

When given an image, a conclusion, and several visual cue options, identify the visual cue that best relates to the conclusion. Select the visual cue that most directly supports or illustrates the conclusion, ensuring that it enhances the overall understanding and clarity of the message. To do this effectively, carefully analyze how each visual cue connects to the key elements of the conclusion. Answer A), B), or C) with no additional explanation. Conclusion: {conclusion}
{vp_options}
ANSWER:
```

- **Paraphrase 2**

```
<image>
The following are multiple choice questions (with answers) about image understanding.
```

```
Given an image, what is the visual cue most related to the given conclusion? Answer A), B), or C) with no additional explanation. Conclusion: {conclusion}
{vp_options}
ANSWER:
```

M Prompts for *Deduction of Conclusion*

- **Image -> C**

```
<image>
Your task is to answer what the image wants to say. You should answer in only one sentence without an unnecessary prefix. ANSWER:
```

- **Image, VP -> C**

```
<image>
"Visual Premises (VP)" are the important features presented in the images.

Visual Premises (VP):
1. VP1
2. VP2
3. VP3

Your task is to answer what the image wants to say. You should answer in only one sentence without an unnecessary prefix. ANSWER:
```

- **Image, VP, CP -> C**

```
<image>
"Visual Premises (VP)" are the important features presented in the images. " Commonsense Premises (CP)" are not visually depicted in the image but are commonly understood by people.

Visual Premises (VP):
1. VP1
2. VP2
3. VP3

Commonsense Premises (CP):
1. CP1
2. CP2
3. CP3

Your task is to answer what the image wants to say. You should answer in only one sentence without an unnecessary prefix.
ANSWER:
```

- **Image, VP, CP, Tree -> C**

```
<image>
"Visual Premises (VP)" are the important features presented in the images. " Commonsense Premises (CP)" are not visually depicted in the image but are commonly understood by people. " Reasoning Steps" are the structure of explanation of how we came up to the "Intermediate Conclusion(IC) and "Conclusion".

Visual Premises (VP):
1. VP1
2. VP2
3. VP3

Commonsense Premises (CP):
1. CP1
2. CP2
3. CP3

Reasoning Step:
(VP1, CP1 -> IC1): IC1
(VP2, CP2 -> IC2): IC2
(VP3, CP3 -> IC3): IC3
(IC1, IC2, IC3 -> C):

Your task is to answer what the image wants to say. You should answer in only one sentence without an unnecessary prefix. ANSWER:
```

- **Prompt Style 1**

```
<image>
"Visual Premises (VP)" are the important features presented in the images. " Commonsense Premises (CP)" are not visually depicted in the image but are commonly understood by people. " Reasoning Steps" are the structure of explanation of how we came up to the "Intermediate Conclusion(IC) and "Conclusion".

Visual Premises (VP):
...

Commonsense Premises (CP):
...

Reasoning Step:
...

Answer in one sentence what the image wants to convey. ANSWER:
```

- **Prompt Style 2**

```
<image>
"Visual Premises (VP)" are the key visual elements in the image. "Commonsense Premises (CP)" are elements based on general common sense. "Reasoning Steps" are the process of reaching the "Intermediate Conclusion (IC)" and "Conclusion".

Visual Premises (VP):
...

Commonsense Premises (CP):
...

Reasoning Step:
...

Write the message of the image in one sentence. You should answer in only one sentence without an unnecessary prefix. RESPONSE:
```

- **Prompt Style 3**

```
<image>
"Visual Premises (VP)" represent the important features observed in the image. "Commonsense Premises (CP)" are things not visually depicted but generally understood. "Reasoning Steps" are the explanation process leading to the "Intermediate Conclusion (IC)" and "Conclusion".

Visual Premises (VP):
...

Commonsense Premises (CP):
...

Reasoning Step:
...

Write the main message of the image in one sentence. RESPONSE:
```

- **Prompt Style 4**

```
<image>
"Visual Premises (VP)" are the key features observed in the image. "Commonsense Premises (CP)" are not visually depicted but can be understood through general knowledge. " Reasoning Steps" are the logical explanation process leading to the " Intermediate Conclusion (IC)" and " Conclusion".

Visual Premises (VP):
...

Commonsense Premises (CP):
...

Reasoning Step:
...

Write the meaning the image wants to convey in one sentence. RESPONSE:
```

	Global												Local		
	Random			Visual			Textual			Mixed			Semantic		
	Ads	Cartoon	All	Ads	Cartoon	All	Ads	Cartoon	All	Ads	Cartoon	All	Ads	Cartoon	All
Random	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33
Human	100.00	100.00	100.00	100.00	98.00	99.00	96.00	92.00	94.00	100.00	100.00	100.00	98.00	98.00	98.00
OFA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen-VL-Chat	88.90	83.21	86.05	88.67	82.87	85.77	73.73	67.61	70.67	77.00	74.14	49.73	53.21	46.25	75.57
CogVLM	97.58	97.35	97.46	96.45	96.34	96.39	88.78	87.21	88.00	91.66	92.79	92.22	69.28	61.35	65.31
Idefics2	98.59	98.76	98.68	97.91	97.75	97.83	93.18	90.42	91.80	95.15	94.99	95.07	77.40	72.62	75.01
InstructBLIP	82.41	85.13	83.77	78.07	80.39	79.23	68.55	65.35	66.95	71.87	70.87	71.37	66.91	56.90	61.90
Unified-IO-2	98.31	98.54	98.42	97.29	96.68	96.99	88.78	84.96	86.87	92.28	93.35	92.81	34.67	34.82	34.74
LLaVA-1.5	98.82	98.48	98.65	98.08	97.75	97.91	84.44	83.04	83.74	89.23	90.48	89.86	73.34	61.52	67.43
LLaVA-NeXT	97.35	97.97	97.66	96.05	96.34	96.20	81.17	80.62	80.90	84.33	87.38	85.86	82.69	74.37	78.53
GPT-4-O	-	-	-	-	-	-	-	-	-	-	-	-	75.22	82.56	79.50

Table 12: Results on *Identification of Premises*.

	Inputs				Automatic			Semantic
	I	VP	CP	RS	BLEU-4	ROUGE	CIDEr	BERT
LLaMA3		✓			7.07	28.41	33.08	43.00
		✓	✓		8.65 (↑ 1.58)	31.44 (↑ 3.03)	40.87 (↑ 7.79)	59.58 (↑ 16.58)
		✓	✓	✓	8.34 (↓ 0.31)	31.18 (↓ 0.26)	41.94 (↑ 1.07)	56.70 (↓ 2.88)
Mistral		✓			2.95	19.84	23.28	24.86
		✓	✓		4.95 (↑ 2.00)	25.13 (↑ 5.29)	33.90 (↑ 10.62)	39.92 (↑ 16.05)
		✓	✓	✓	6.15 (↑ 1.20)	27.06 (↑ 1.93)	38.34 (↑ 4.43)	49.54 (↑ 9.62)
Zephyr		✓			2.78	16.35	24.06	16.15
		✓	✓		3.25 (↑ 0.47)	17.44 (↑ 1.08)	30.41 (↑ 6.35)	31.38 (↑ 6.52)
		✓	✓	✓	5.20 (↑ 1.94)	22.70 (↑ 5.26)	36.29 (↑ 5.88)	45.23 (↑ 13.85)
OFA	✓				0.00	0.13	0.01	-41.26
	✓	✓			0.00 (-)	5.24 (↑ 5.10)	0.47 (↑ 0.47)	-22.52 (↑ 18.75)
	✓	✓	✓		0.00 (-)	5.79 (↑ 0.55)	0.37 (↓ 0.10)	-15.87 (↑ 6.65)
	✓	✓	✓	✓	0.00 (-)	6.53 (↑ 0.75)	0.70 (↑ 0.33)	-12.51 (↑ 3.36)
QwenVLChat	✓				0.72	13.12	8.41	14.32
	✓	✓			4.02 (↑ 3.30)	24.73 (↑ 11.61)	30.58 (↑ 22.17)	28.74 (↑ 14.41)
	✓	✓	✓		4.85 (↑ 0.83)	26.67 (↑ 1.94)	35.30 (↑ 4.72)	34.05 (↑ 5.31)
	✓	✓	✓	✓	4.89 (↑ 0.03)	26.89 (↑ 0.23)	38.30 (↑ 3.00)	35.11 (↑ 1.06)
CogVLM	✓				4.96	24.40	25.56	27.38
	✓	✓			6.06 (↑ 1.09)	27.37 (↑ 2.97)	39.19 (↑ 13.63)	33.24 (↑ 5.87)
	✓	✓	✓		7.18 (↑ 1.13)	29.25 (↑ 1.88)	47.21 (↑ 8.02)	36.41 (↑ 3.17)
	✓	✓	✓	✓	7.68 (↑ 0.50)	30.03 (↑ 0.78)	51.44 (↑ 4.23)	37.71 (↑ 1.29)
Idefics2	✓				4.00	21.97	18.56	21.27
	✓	✓			4.53 (↑ 0.53)	24.13 (↑ 2.17)	28.79 (↑ 10.24)	27.39 (↑ 6.12)
	✓	✓	✓		5.56 (↑ 1.03)	25.17 (↑ 1.04)	38.21 (↑ 9.42)	33.22 (↑ 5.83)
	✓	✓	✓	✓	7.48 (↑ 1.92)	27.73 (↑ 2.56)	53.22 (↑ 15.01)	38.40 (↑ 5.17)
InstructBLIP	✓				0.00	4.22	1.01	-15.92
	✓	✓			3.27 (↑ 3.26)	18.94 (↑ 16.26)	22.16 (↑ 21.15)	23.15 (↑ 39.07)
	✓	✓	✓		6.00 (↑ 2.73)	26.91 (↑ 7.32)	44.23 (↑ 22.07)	35.20 (↑ 12.05)
	✓	✓	✓	✓	6.27 (↑ 0.27)	28.29 (↑ 0.37)	45.53 (↑ 1.29)	35.52 (↑ 0.32)
Unified-io 2	✓				0.07	10.02	0.89	-8.49
	✓	✓			0.65 (↑ 0.58)	14.81 (↑ 4.79)	5.18 (↑ 4.29)	-0.04 (↑ 8.45)
	✓	✓	✓		0.82 (↑ 0.17)	15.27 (↑ 0.46)	7.73 (↑ 2.55)	5.63 (↑ 5.68)
	✓	✓	✓	✓	0.93 (↑ 0.11)	16.10 (↑ 0.83)	10.12 (↑ 2.39)	8.43 (↑ 2.79)
LLaVA	✓				1.38	14.93	3.73	3.25
	✓	✓			3.92 (↑ 2.54)	22.93 (↑ 8.01)	21.49 (↑ 17.76)	22.21 (↑ 18.96)
	✓	✓	✓		5.49 (↑ 1.58)	26.40 (↑ 3.47)	39.29 (↑ 17.80)	32.48 (↑ 10.28)
	✓	✓	✓	✓	5.68 (↑ 0.19)	26.46 (↑ 0.06)	42.65 (↑ 3.36)	34.22 (↑ 1.74)
LLaVA-NeXT	✓				3.62	21.08	15.23	18.05
	✓	✓			6.78 (↑ 3.16)	28.31 (↑ 7.23)	42.17 (↑ 26.94)	32.98 (↑ 14.94)
	✓	✓	✓		7.51 (↑ 0.73)	30.03 (↑ 1.72)	50.54 (↑ 8.37)	37.93 (↑ 4.95)
	✓	✓	✓	✓	8.51 (↑ 1.00)	31.19 (↑ 1.15)	61.70 (↑ 11.16)	40.96 (↑ 3.02)
GPT-4-O	✓				2.38	17.20	23.08	25.96
	✓	✓			5.44 (↑ 3.07)	24.47 (↑ 7.27)	46.05 (↑ 22.98)	36.09 (↑ 10.13)
	✓	✓	✓	✓	6.82 (↑ 1.38)	26.36 (↑ 1.88)	61.20 (↑ 15.15)	41.08 (↑ 4.99)

Table 13: Results on *Deduction of Conclusion* in the Advertisement category.

	Inputs				Automatic			Semantic
	I	VP	CP	RS	BLEU-4	ROUGE	CIDEr	BERT
LLaMA3		✓			5.51	27.67	31.28	26.46
		✓	✓		6.65 (↑ 1.13)	29.95 (↑ 2.28)	48.01 (↑ 16.73)	33.71 (↑ 7.25)
		✓	✓	✓	8.81 (↑ 2.17)	32.17 (↑ 2.22)	61.74 (↑ 13.73)	39.20 (↑ 5.49)
Mistral		✓			1.87	16.29	11.24	13.23
		✓	✓		4.01 (↑ 2.14)	22.24 (↑ 5.95)	28.15 (↑ 16.91)	25.22 (↑ 11.99)
		✓	✓	✓	6.45 (↑ 2.44)	27.82 (↑ 5.59)	45.00 (↑ 16.85)	34.39 (↑ 9.17)
Zephyr		✓			1.70	13.88	11.54	16.15
		✓	✓		3.28 (↑ 1.58)	17.51 (↑ 3.63)	24.92 (↑ 13.38)	26.40 (↑ 10.25)
		✓	✓	✓	6.91 (↑ 3.63)	27.20 (↑ 9.69)	46.63 (↑ 21.71)	36.70 (↑ 10.29)
OFA	✓				0.00	0.45	0.01	-41.35
	✓	✓			0.00 (-)	5.30 (↑ 4.84)	0.27 (↑ 0.26)	-27.23 (↑ 14.12)
	✓	✓	✓		0.00 (-)	8.15 (↑ 2.85)	0.26 (↓ 0.01)	-17.30 (↑ 9.93)
	✓	✓	✓	✓	0.00 (-)	8.45 (↑ 0.30)	0.58 (↑ 0.32)	-15.74 (↑ 1.56)
QwenVLChat	✓				0.49	13.98	4.23	10.79
	✓	✓			3.18 (↑ 2.69)	23.80 (↑ 9.81)	18.02 (↑ 13.79)	17.18 (↑ 6.39)
	✓	✓	✓		4.23 (↑ 1.05)	26.40 (↑ 2.60)	26.76 (↑ 8.74)	25.03 (↑ 7.85)
	✓	✓	✓	✓	4.79 (↑ 0.56)	27.87 (↑ 1.47)	32.10 (↑ 5.34)	29.49 (↑ 4.46)
CogVLM	✓				4.89	27.48	24.56	23.51
	✓	✓			5.49 (↑ 0.60)	28.50 (↑ 1.02)	32.11 (↑ 7.54)	27.24 (↑ 3.73)
	✓	✓	✓		6.43 (↑ 0.94)	29.64 (↑ 1.14)	38.02 (↑ 5.91)	29.88 (↑ 2.64)
	✓	✓	✓	✓	7.89 (↑ 1.46)	31.72 (↑ 2.08)	53.05 (↑ 15.04)	34.45 (↑ 4.56)
Idefics2	✓				3.50	21.74	16.07	16.36
	✓	✓			3.61 (↑ 0.78)	23.47 (↑ 2.04)	20.02 (↑ 7.20)	16.70 (↑ 6.78)
	✓	✓	✓		5.42 (↑ 1.81)	26.58 (↑ 3.11)	29.96 (↑ 9.94)	24.50 (↑ 7.81)
	✓	✓	✓	✓	8.28 (↑ 2.86)	31.01 (↑ 4.43)	54.64 (↑ 24.68)	34.27 (↑ 9.76)
InstructBLIP	✓				0.00	3.24	0.40	-21.55
	✓	✓			2.53 (↑ 2.53)	18.94 (↑ 15.14)	16.35 (↑ 15.60)	16.62 (↑ 34.98)
	✓	✓	✓		5.33 (↑ 2.80)	26.91 (↑ 7.98)	36.87 (↑ 20.52)	28.92 (↑ 12.30)
	✓	✓	✓	✓	6.18 (↑ 0.85)	28.29 (↑ 1.38)	42.53 (↑ 5.65)	32.17 (↑ 3.25)
Unified-io 2	✓				0.00	9.07	0.40	-11.68
	✓	✓			0.56 (↑ 0.56)	11.32 (↑ 2.25)	2.60 (↑ 2.20)	-7.80 (↑ 3.88)
	✓	✓	✓		0.62 (↑ 0.06)	13.79 (↑ 2.48)	5.40 (↑ 2.80)	2.39 (↑ 10.19)
	✓	✓	✓	✓	1.33 (↑ 0.71)	16.50 (↑ 2.70)	12.63 (↑ 7.24)	7.47 (↑ 5.08)
LLaVA	✓				1.46	17.84	3.55	0.92
	✓	✓			3.73 (↑ 2.27)	23.22 (↑ 5.38)	12.56 (↑ 9.01)	14.92 (↑ 14.01)
	✓	✓	✓		5.61 (↑ 1.87)	27.63 (↑ 4.41)	28.99 (↑ 16.44)	25.89 (↑ 10.97)
	✓	✓	✓	✓	7.87 (↑ 2.26)	30.46 (↑ 2.84)	47.25 (↑ 18.25)	33.11 (↑ 7.22)
LLaVA-NeXT	✓				2.71	22.32	11.26	11.26
	✓	✓			5.78 (↑ 3.07)	27.79 (↑ 5.48)	28.12 (↑ 16.86)	22.46 (↑ 11.20)
	✓	✓	✓		7.31 (↑ 1.52)	30.44 (↑ 2.65)	43.84 (↑ 15.73)	29.57 (↑ 7.11)
	✓	✓	✓	✓	9.16 (↑ 1.85)	33.09 (↑ 2.65)	61.44 (↑ 17.60)	35.88 (↑ 6.32)
GPT-4-O	✓				4.07	23.37	23.15	24.96
	✓	✓			6.40 (↑ 2.34)	27.39 (↑ 4.02)	40.64 (↑ 17.50)	32.05 (↑ 7.09)
	✓	✓	✓	✓	8.69 (↑ 2.29)	31.32 (↑ 3.93)	63.13 (↑ 22.48)	40.78 (↑ 8.73)

Table 14: Results on *Deduction of Conclusion* in the Cartoon category.

	Inputs				Automatic			Semantic
	I	VP	CP	RS	BLEU-4	ROUGE	CIDEr	BERT
LLaMA3		✓			6.40	28.09	37.93	30.22
		✓	✓		7.78 (↑ 1.39)	30.80 (↑ 2.71)	54.57 (↑ 16.65)	37.77 (↑ 7.55)
		✓	✓	✓	8.54 (↑ 0.76)	31.61 (↑ 0.82)	58.88 (↑ 4.31)	40.75 (↑ 2.98)
Mistral		✓			2.48	18.30	18.41	18.93
		✓	✓		4.54 (↑ 2.06)	23.88 (↑ 5.57)	34.83 (↑ 16.42)	30.15 (↑ 11.21)
		✓	✓	✓	6.28 (↑ 1.74)	27.39 (↑ 3.51)	47.58 (↑ 12.75)	36.63 (↑ 6.48)
Zephyr		✓			2.31	15.28	19.10	20.64
		✓	✓		3.26 (↑ 0.95)	17.47 (↑ 2.18)	28.59 (↑ 9.49)	28.67 (↑ 8.04)
		✓	✓	✓	5.94 (↑ 2.67)	24.65 (↑ 7.18)	45.84 (↑ 17.25)	36.47 (↑ 7.79)
OFA	✓				0.00	0.27	0.01	-41.30
	✓	✓			0.00 (-)	5.26 (↑ 4.99)	0.39 (↑ 0.38)	-24.55 (↑ 5.68)
	✓	✓	✓		0.00 (-)	6.81 (↑ 1.55)	0.32 (↓ 0.06)	-16.49 (↑ 1.73)
	✓	✓	✓	✓	0.00 (-)	7.36 (↑ 0.55)	0.65 (↑ 0.32)	-13.91 (↑ 1.22)
QwenVLChat	✓				0.62	13.50	6.60	12.79
	✓	✓			3.66 (↑ 3.04)	24.33 (↑ 10.83)	25.15 (↑ 18.54)	23.74 (↑ 10.94)
	✓	✓	✓		4.58 (↑ 0.92)	26.55 (↑ 2.23)	31.61 (↑ 6.46)	30.15 (↑ 6.41)
	✓	✓	✓	✓	4.85 (↑ 0.26)	27.31 (↑ 0.76)	35.62 (↑ 4.01)	32.68 (↑ 2.53)
Idefics2	✓				3.50	21.74	16.07	16.36
	✓	✓			4.13 (↑ 0.64)	23.84 (↑ 2.11)	25.00 (↑ 8.92)	22.76 (↑ 6.41)
	✓	✓	✓		5.50 (↑ 1.37)	25.78 (↑ 1.93)	34.64 (↑ 9.64)	29.45 (↑ 6.69)
	✓	✓	✓	✓	7.82 (↑ 2.32)	29.15 (↑ 3.37)	53.84 (↑ 19.20)	36.61 (↑ 7.16)
InstructBLIP	✓				0.00	3.80	0.75	-18.36
	✓	✓			2.53 (↑ 2.53)	18.94 (↑ 15.14)	16.35 (↑ 15.60)	16.62 (↑ 34.98)
	✓	✓	✓		5.33 (↑ 2.80)	26.91 (↑ 7.98)	36.87 (↑ 20.52)	28.92 (↑ 12.30)
	✓	✓	✓	✓	6.18 (↑ 0.85)	28.29 (↑ 1.38)	42.53 (↑ 5.65)	32.17 (↑ 3.25)
CogVLM	✓				4.93	25.73	25.13	25.53
	✓	✓			5.81 (↑ 0.88)	27.86 (↑ 2.13)	36.13 (↑ 11.00)	30.65 (↑ 4.94)
	✓	✓	✓		6.86 (↑ 1.04)	29.42 (↑ 1.56)	43.23 (↑ 7.10)	33.59 (↑ 2.94)
	✓	✓	✓	✓	7.77 (↑ 0.92)	30.76 (↑ 1.34)	52.14 (↑ 8.90)	36.30 (↑ 2.71)
Unified-io 2	✓				0.04	9.61	0.68	-9.87
	✓	✓			0.61 (↑ 0.57)	13.30 (↑ 3.69)	4.07 (↑ 3.39)	-3.40 (↑ 6.47)
	✓	✓	✓		0.74 (↑ 0.12)	14.63 (↑ 1.33)	6.72 (↑ 2.66)	4.23 (↑ 7.63)
	✓	✓	✓	✓	1.10 (↑ 0.37)	16.27 (↑ 1.64)	11.21 (↑ 4.48)	8.01 (↑ 3.78)
LLaVA	✓				1.50	16.01	3.65	2.24
	✓	✓			3.86 (↑ 2.36)	22.88 (↑ 6.87)	18.69 (↑ 15.04)	19.98 (↑ 17.74)
	✓	✓	✓		5.54 (↑ 1.69)	26.93 (↑ 4.05)	34.84 (↑ 16.14)	29.63 (↑ 9.66)
	✓	✓	✓	✓	6.63 (↑ 1.09)	28.19 (↑ 1.26)	44.64 (↑ 9.80)	33.74 (↑ 4.11)
LLaVA-NeXT	✓				3.23	21.62	13.51	15.11
	✓	✓			6.35 (↑ 3.12)	28.09 (↑ 6.47)	36.09 (↑ 22.58)	28.43 (↑ 16.75)
	✓	✓	✓		7.42 (↑ 1.07)	30.21 (↑ 2.12)	47.64 (↑ 11.55)	34.31 (↑ 8.07)
	✓	✓	✓	✓	8.46 (↑ 1.04)	31.69 (↑ 1.49)	61.14 (↑ 13.50)	39.50 (↑ 2.58)
GPT-4-O	✓				3.11	19.87	23.11	24.50
	✓	✓			5.86 (↑ 2.75)	25.74 (↑ 5.87)	43.71 (↑ 20.61)	31.89 (↑ 7.39)
	✓	✓	✓	✓	7.63 (↑ 1.77)	28.51 (↑ 2.77)	62.03 (↑ 18.32)	38.42 (↑ 6.53)

Table 15: Results for the *Deduction of Conclusion* averaged across the two categories.

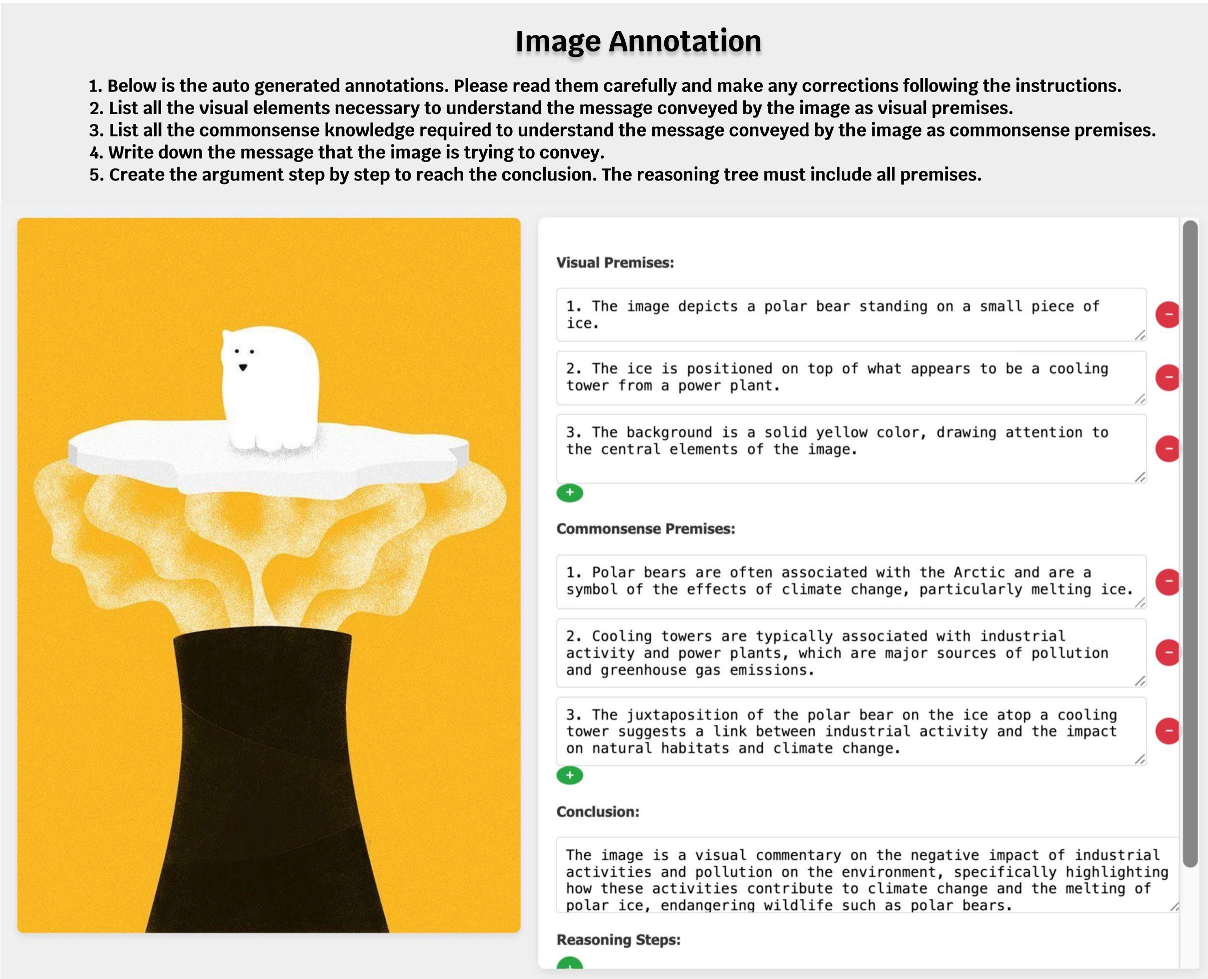
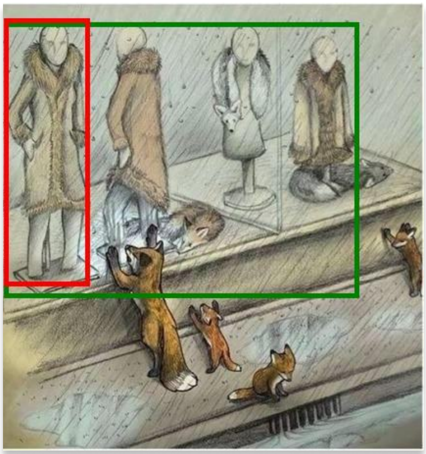


Figure 6: Human annotation interface for collecting textual annotations.

Figure 7: Human annotation interface for collecting bounding boxes of visual premises.

Figure 8: Human evaluation interface for *Identification of Premises*.



Target Visual Premise:
Mannequins in a store window wearing fur coats.

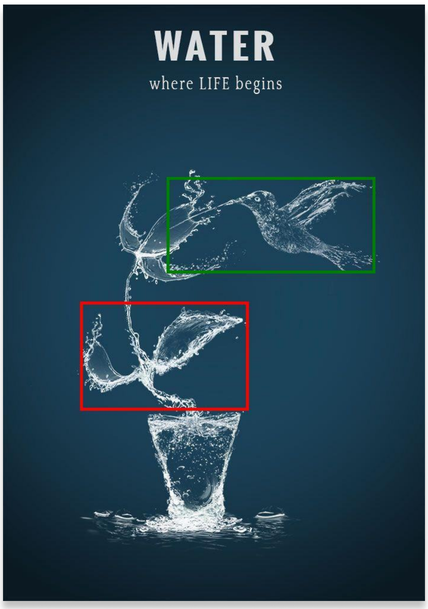
IOU : 0.23

Explanation:
The model seems to be detecting the correct target, but only for a single object.

Target Visual Premise:
A garlic bulb and a chili pepper, both smaller than onion.

IOU : 0.29

Explanation:
The model seems to be detecting the correct target, but only for a single object.



Target Visual Premise:
A bird, made of water, is drinking from the water plant.

IOU : 0.00

Explanation:
The model detecting on object related to the target, but on wrong object.

Figure 9: Qualitative samples of *open-set* grounding results.

LLaVANeXT

I -> C

The image shows a man trying to fix a broken wall with a hammer and a chisel, but the elephant is causing chaos by knocking down the bricks.

I, VP -> C

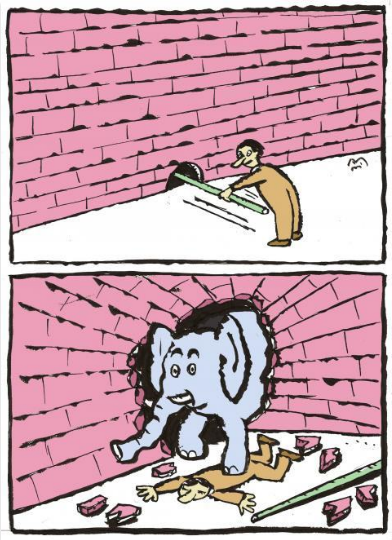
The image illustrates the consequences of curiosity and the unexpected consequences of one's actions.

I, VP, CP -> C

The image humorously illustrates the consequences of curiosity and the unexpected power of an elephant.

I, VP, CP, T -> C

The image warns against curiosity or investigation that may lead to dangerous consequences.



InstructBLIP

I -> C

elephant

I, VP -> C

The image wants to say that the man was knocked down by the elephant as it burst through the wall.

I, VP, CP -> C

The image depicts a man using a stick to probe a small hole in the wall, and an elephant bursts through the wall, creating a large hole and knocking the man down.

I, VP, CP, T -> C

The image suggests that curiosity can lead to unexpected and potentially harmful consequences.

LLaVANeXT

I → C

The image wants to say that mcdonald's is everywhere.

I, VP → C

The image wants to say that McDonald's is a ubiquitous presence in urban environments, even at night.

I, VP, CP → C

The image showcases the convenience of McDonald's delivery service in a city setting, emphasizing the brand's presence and the availability of nighttime delivery.

I, VP, CP, T → C

The image showcases McDonald's delivery service in a city environment, highlighting the convenience of nighttime food delivery.

InstructBLIP

I → C

mcdonalds

I, VP → C

The image is advertising McDonald's delivery service.

I, VP, CP → C

The image wants to showcase the convenience of McDonald's delivery service in an urban European setting, emphasized by the illuminated golden arches logos on various building facades.

I, VP, CP, T → C

The image is advertising McDonald's delivery service in a nighttime urban setting with illuminated golden arches logos on various building facades.

Figure 10: Qualitative samples of *Deduction of Conclusion*, with inference results from LLaVA-NeXT and InstructBLIP models.

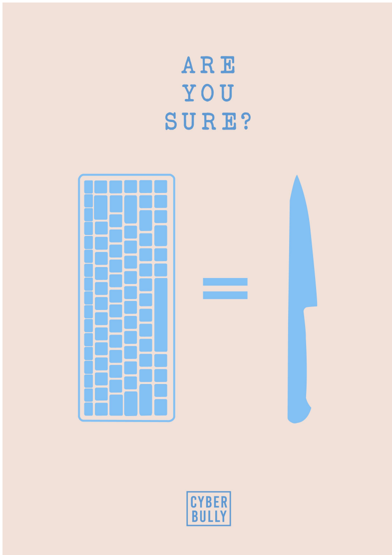
CogVLM

I -> C
The image is a metaphorical representation of the dangers of cyberbullying, suggesting that it is as harmful as physical violence.

I, VP -> C
The image is conveying a message about the potential dangers of cyberbullying, suggesting that one's online actions can have real-world consequences.

I, VP, CP -> C
The image is emphasizing the severity of cyberbullying by comparing it to physical violence, suggesting that both can have harmful consequences.

I, VP, CP, T -> C
The image is emphasizing the harmful nature of cyberbullying by comparing it to physical violence, suggesting that one should be cautious and sure about their online actions.



QwenVLChat

I -> C
the image wants to say that cyberbullying is not a joke.

I, VP -> C
The image wants to say that cyberbullying is a serious issue and should not be taken lightly.

I, VP, CP -> C
The image wants to say that cyberbullying can be as harmful as physical violence, and it uses a keyboard and a knife to represent these concepts.

I, VP, CP, T -> C
The image wants to say that cyberbullying can be as harmful as physical violence, and it is a serious issue that should not be underestimated.

CogVLM

I → C
The image conveys the challenges and dangers faced by journalists in conflict zones.

I, VP → C
The image conveys the dangers and challenges faced by journalists in conflict zones.

I, VP, CP → C
The image conveys the dangers and challenges faced by journalists in conflict zones, emphasizing the sacrifices they make to report the truth.

I, VP, CP, T → C
The image conveys the dangers and challenges faced by journalists in conflict zones, highlighting the vulnerability of press freedom and the sacrifices made for the sake of truth.

QwenVLChat

I → C
the image wants to say that journalists are often threatened and attacked while doing their jobs.

I, VP → C
The image depicts a journalist kneeling in front of armed soldiers, symbolizing the danger and sacrifice of reporting in conflict situations.

I, VP, CP → C
The image portrays a journalist kneeling in front of armed soldiers, conveying the dangers and challenges faced by journalists in reporting from conflict zones.

I, VP, CP, T → C
Journalists often face danger and violence while reporting news, even in conflict zones where they are supposed to be protected by the "PRESS" label.

Figure 11: Qualitative samples of *Deduction of Conclusion*, with inference results from CogVLM and Qwen-VL-Chat models.