PRISM: Preconditioned Visual Language Inference and Rationalization using Weak Supervision

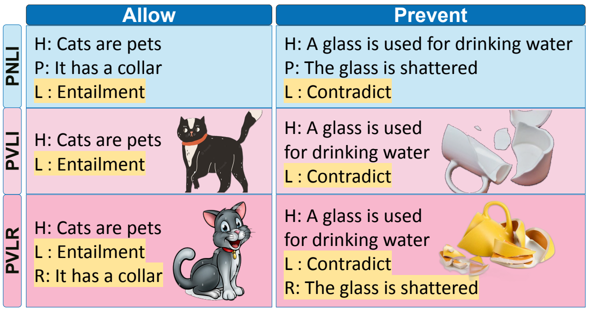
Anonymous submission

Abstract

Humans can infer the affordance of objects by extracting related contextual preconditions for each scenario. For instance, when presented with an image of a shattered cup, we can deduce that this specific condition hinders its suitability for drinking purposes. The process of employing commonsense preconditions for reasoning is extensively studied in NLP, where models explicitly acquire contextual preconditions in textual form. Nonetheless, it remains uncertain whether state-of-the-art visual language models (VLMs) can effectively extract such preconditions and employ them to infer object affordances. In this work, dubbed PRISM, we introduce two tasks: preconditioned visual language inference (PVLI) and rationalization (PVLR). To address these tasks, we propose three strategies for acquiring weak supervision signals and creating a human-validated evaluation resource through crowd-sourcing. Our findings expose the limitations of current state-of-the-art VLM models in these tasks, and we chart a roadmap for overcoming the challenges that lie ahead in their improvement.

1 Introduction

According to the *Theory of Affordance* (Gibson, 2000; Chemero, 2003), understanding the preconditions in which an action or statement is possible or impossible is a key aspect of human intelligence. For example, a glass may be used for drinking water, under an implicit assumption that the water is at normal temperature, but may not be if the glass is shattered. From the cognitive perspective, understanding the affordance of objects, or simply preconditions of actions (Qasemi et al., 2022a), is part of the commonsense knowledge that constitutes what distinguishes humans from a machine to make inference (Lenat, 1998). From an applications perspective, it also has huge implications such as robotics (Ahn et al., 2022), transportations (Prakken, 2017; Seff and Xiao, 2016; Kothawade et al., 2021), and general artificial intelligence (Nguyen et al., 2021).



	Allow	Prevent
PNLI	H: Cats are pets P: It has a collar L : Entailment	H: A glass is used for drinking water P: The glass is shattered L : Contradict
PVLI	H: Cats are pets L : Entailment	H: A glass is used for drinking water L : Contradict
PVLR	H: Cats are pets L : Entailment R: It has a collar	H: A glass is used for drinking water L : Contradict R: The glass is shattered

Figure 1: Preconditioned Visual Language Inference (PVLI) and Reasoning (PVLR) tasks. The “H” and “P” are the input *hypothesis* and *premise*. The outputs, *label* (letter “L”) and *rationale* (letter “R”), are highlighted.

Reasoning with preconditions of commonsense knowledge (i.e. preconditioned inference), is proposed as a benchmarking task for evaluation of the theory of affordance (Qasemi et al., 2022a). Multiple studies have formulated the preconditioned natural language inference (PNLI) as variations of the Natural Language Inference (NLI) (Williams et al., 2018; Bowman et al., 2015a; Condoravdi et al., 2003) task and contributed learning resources that are gathered through crowdsouring (Rudinger et al., 2020; Qasemi et al., 2022a; Hwang et al., 2020; Do and Pavlick, 2021; Jiang et al., 2021b) or weak supervision data (Qasemi et al., 2022b). In PNLI, the models rely on the contextual information (i.e. textual preconditions as *premise*) as input and have to decide whether the *hypothesis* is *allowed* (entailment), *prevented* (contradiction), or undetermined (neutral) given the *premise* (first row in Fig. 1). However, humans reason about affordance using information beyond text (Barsalou, 2010; Andrews et al., 2009) and extract the contextual meaning representations for cognitive tasks (such as PNLI) from the pool of available information in various modalities. For example, upon getting the query “can this person run?” and seeing a picture of a person in a full leg cast, one can imply the contextual information from the image that "the person is injured and incapable of running" and use it to answer the query accordingly. Thus, a visual variation of the PNLI task is cognitively more realistic to benchmark artificial intelligence models.

In this work, we propose *PRISM*, to expand the preconditioned inference and reasoning to the visual-language realm by considering the interaction between linguistic and visual information in common sense. This work presents three contributions. **First**, we introduce the Preconditioned Visual Language Inference (PVLI) and Rationalization (PVLR) tasks (2nd and 3rd rows in Fig. 1), which evaluate the visual-language models' (VLM) capabilities to reason with preconditions associated with commonsense knowledge. In PVLI, the precondition is represented as an image that further constrains the context in which the model has to decide the "*prevent*" or "*allow*" labels. In PVLR, the model has to provide the rationale for the choice between the labels as well. For example, say the model is given a commonsense statement such as "a glass is used for drinking water" as the *hypothesis* and an image of a "broken glass" as the *premise*. Then, in PVLI, the model has to decide whether there is a *prevented*(contradiction) or *allowed*(entailment) relation between them, and in PVLR, it has to provide a rationale for its decision, such as *the glass is broken*. In addition, to foster further research, we created a human-verified evaluation dataset through crowd-sourcing to benchmark models.

Second, we propose three strategies for retrieving a rich amount of cheap and allowably noisy supervision signals for inference and rationalization. Similar to Parcalabescu et al. (2021), *PRISM*'s three strategies rely on the available image captioning datasets (e.g. Changpinyo et al. (2021); Sharma et al. (2018); Gurari et al. (2020a); Lin et al. (2014a)) that are readily available as a result of years of research in the field and maturity of resources. In the first strategy, *Extraction from Captions*, we utilize the PInKS (Qasemi et al., 2022b) method to extract PNLI instances from image captions. PInKS uses a combination of linguistic patterns (e.g. "{action} unless {precondition}") and generative augmentation to extract large quantities of instances from raw text. In the second strategy, *Caption Querying*, we use the existing crowdsourced PNLI instances (e.g. Rudinger et al. (2020); Qasemi et al. (2022a); Hwang et al. (2020); Do and Pavlick (2021); Jiang et al. (2021b)) and find an image caption that is semantically identical to them. The third strategy, *Image Querying*, focuses solely on the PNLI instances and devises queries (such as "you are in a desert") to search directly for corresponding images on the web using image search engines (e.g. Google Images). As post-processing for all three strategies, we use ChatGPT (Brown et al., 2020) to fix the formatting and grammatical issues.

Our **third** contribution is an extensive benchmarking of VLMs based on *PRISM*. We benchmark 4 SOTA VLMs, FLAVA (Singh et al., 2022a), VisualBERT (Li et al., 2019), ViLBERT (Lu et al., 2019) and ViLT (Kim et al., 2021) in inference (§4.1) and Ayyubi et al. (2020) in rationalization (§4.5). Using PABI (He et al., 2021) metric, we provide a theoretical measure of informativeness of both visual and textual modalities to the overall task (§4.3). In addition, we show how an effective rationalization will improve inference in the VLM models (§4.5). We further investigate the fine-tuning (learning) process of VLMs in the inference task (§4.2) and study their exploitation of the spurious correlation in our dataset (§4.4).

2 Construction of *PRISM* and Test Set

This section gives an overview of *PRISM* (summarized in Fig. 2), describing our strategies for obtaining the data, and quality control. For Brevity, details of the human verification through crowdsourcing are moved to Appx. §B, and implementation details of each strategy are discussed in Appx. §A.1.

Datasets: The construction of *PRISM* uses existing text-only PNLI and image-captioning datasets as building blocks. For the text-only PNLI datasets, we require that they contain a precondition (e.g. premise, context), an action (e.g. hypothesis, question), and a binary label indicating whether the precondition *allows* or *prevents* the action. To limit the leakage of unclean text from these resources, we instruct ChatGPT (Brown et al., 2020) to fix formatting and grammatical issues. For the image captioning datasets, we simply require images (typically the URL) and their captions. Any datasets that meet these requirements can be used for the following steps.

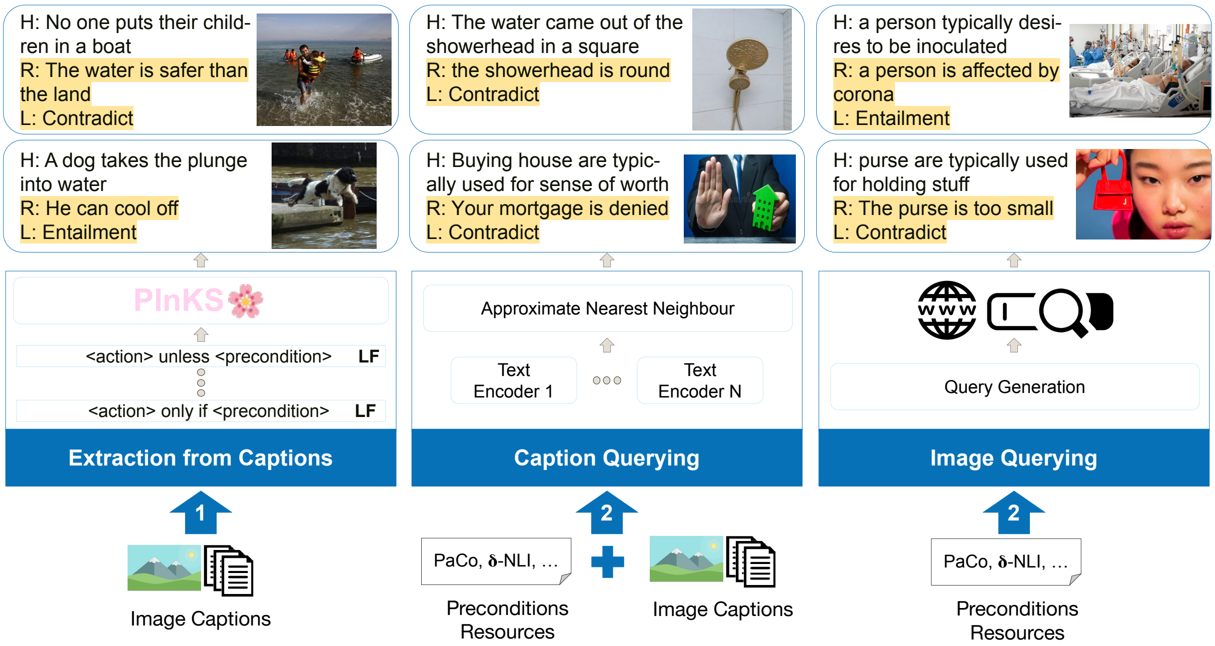

H: No one puts their children in a boat
R: The water is safer than the land
L: Contradict
H: The water came out of the showerhead in a square
R: the showerhead is round
L: Contradict
H: a person typically desires to be inoculated
R: a person is affected by corona
L: Entailment
H: A dog takes the plunge into water
R: He can cool off
L: Entailment
H: Buying house are typically used for sense of worth
R: Your mortgage is denied
L: Contradict
H: purse are typically used for holding stuff
R: The purse is too small
L: Contradict
PInKS
<action> unless <precondition> LF
<action> only if <precondition> LF
Extraction from Captions
Approximate Nearest Neighbour
Text Encoder 1
Text Encoder N
Caption Querying
Query Generation
Image Querying
1
Image Captions
2
PaCo, δ-NLI, ...
Preconditions Resources
Image Captions
2
PaCo, δ-NLI, ...
Preconditions Resources

Figure 2: Overview of weak supervision methods in constructing *PRISM*.

Preprocessing: The PNLI instances often use varying conventions for referring to people. We standardize them by replacing these identifiers with "the person", "another person", "a third person", and so on. For example, the sentence "Alice helps Bob" would become "the person helps another person". This ensures that the specific names or traditionally-associated genders are not mistaken as a focus of the statements. Of the image captions, some may consist of multiple sentences. In these cases, we split the captions into individual examples, pairing each sentence with the original image. Using these preprocessed resources, we then obtain *PRISM* instances with three strategies: extraction from captions, caption querying, and image querying.

Extraction from Captions (EC): Our first strategy focuses solely on the image captions, finding and extracting the few that contain preconditions and actions. By nature, the resulting pairs are already grounded in the images. We use the minimally-supervised approach described in *PInKS* (Qasemi et al., 2022b), where linguistic patterns are used to extract preconditions and actions from raw corpora. This strategy constructs *labeling functions (LF)* based on common conjunctions such as "only if" and "unless". For example, the sentence "Swimming pools have cold water in the winter unless they are heated" is matched by the pattern "{action} unless {precondition}", and therefore we can infer that "they are heated" is a precondition that prevents the action "Swimming pools have cold water in the winter". Such labeling functions can be refined and added to as desired. In cases where the conjunction can be used in multiple senses, part-of-speech tagging is utilized to filter out irrelevant senses. After applying the labeling functions to the image captions, we have a dataset consisting of preconditions and actions, where both are grounded in the associated images. To control for quality, we annotate a sample of matches from each labeling function (precision of each LF) to record whether the relation between precondition and action makes sense. Based on the results, we choose a precision threshold and only include labeling functions that meet this minimum.

Caption Querying (CQ): Our second strategy bridges the PNLI premises (preconditions) and image captions by grounding them in images that have semantically similar captions. We begin by limiting the premises and captions to those whose length is within one standard deviation of the mean (rounded to the nearest integer) in order to remove outliers. We then encode the PNLI premises and image captions in high-dimensional vector embeddings using multiple models. Next, using a PNLI's premise as a query, we find the most similar captions' embedding through approximate nearest neighbors. We then aggregate the top-ranking captions for each model and select the best caption. This strategy of including multiple models in the decision-making process helps make it more robust to model differences and to the approximate nature of the nearest neighbors. The number of models incorporated and the number of top-ranking captions to choose from each depends on balancing the desired robustness and time or computational constraints.

To control for quality, we additionally record two values: *perplexity* score from each model and *model's agreement*. The perplexity is the distance (cosine, dot, etc.) between the query and caption, averaged over the models. In the case when one of the models did not include the chosen caption in its ranking, the distance of the last caption is used for the average. By nature, the perplexity measures how good the models believe the match to be. In contrast, the model agreement is not specific to the chosen caption but instead measures how well-aligned the models' rankings are. Using a ranking similarity metric, we compute the similarity between pairs of rankings and then average the scores for the model agreement. Since a high model agreement indicates that the models agree on which are the closest captions, but does not speak to the actual proximity of the match, it can be thought of as a measure of confidence.

Image Querying (IQ): Our third strategy utilizes image search engines as a source for grounding the PNLI premises into an image. Here we form a search query from the PNLI premise and feed it to a search engine to directly find the relevant images describing the premise from the web. Like the CQ strategy, we limit the premises to those whose length is within one standard deviation of the mean and remove all punctuations from the search query. In order to ensure effective results, we have excluded premises related to abstract concepts, such as the notion of responsibility (the person is responsible) or gratitude (the person will be grateful), as searching for images directly in relation to these concepts is unlikely to yield favorable outcomes. It is important to note that each of the top image results can become an individual example, enabling this strategy to rapidly generate a substantial amount of training data. However, it should be acknowledged that this abundance of data may introduce a certain level of noise and bias.

3 Data Analysis

In this section, we investigate different aspects of the weak supervision data and evaluate the quality of the generated resource. For brevity, additional implementation and experimental details are moved to Appx. §A.1 (for data acquisition results), Appx. §B (for crowd-sourcing results), and Appx. §A.2 (for *Image Querying* results) to conserve space.

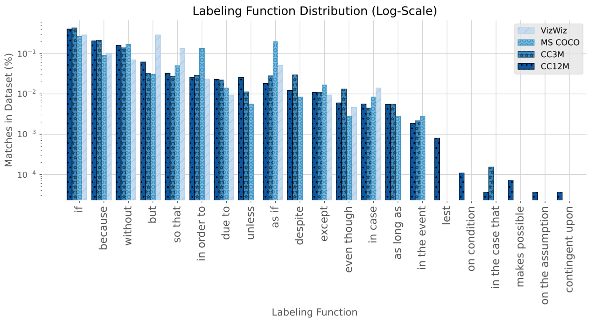


Figure 3: Distribution of instances extracted from captions (log-scale), for each source of the caption.

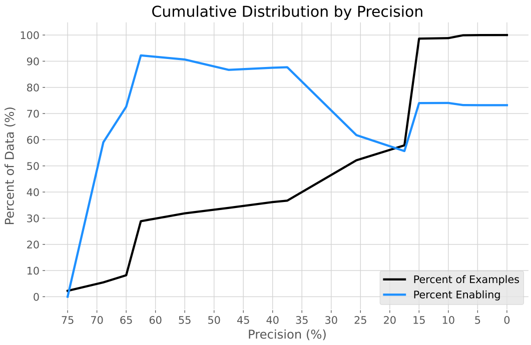


Figure 4: Cumulative distribution of the data with respect to the precision of the generating labeling function.

Extraction from Captions Results: After preprocessing the image-captioning resources we end up with 17 million captions. From this, we utilize the *Extraction from Captions* method that results in 34K extracted *PRISM* instances. Fig. 3 illustrates the percentage of matches that come from each LF, separated by PNLI resouced. General statements such as "if" unsurprisingly make up a large percentage of the data, but interestingly, the PNLI resourced have very different distributions. Among the sources of captions, *VizWiz* has disproportionately high counts of "but" and "so that", while *MS-COCO* is high in "in order to" and "as if". Fig. 4 shows the percent of the data and the percent of "allow" examples for varying precision thresholds. For the results in §4, we use the threshold of 0.6 to have a good balance between the quality and quantity of the final resource.

Caption Querying Results: Figure 8 (moved to Appx. §A.1 to preserve space) summarizes the observed distribution of matches obtain through precondition and image sources in the Caption Querying method. For example, we see that the majority of our precondition matches come from ANION (59.1%) and the majority of captions come from *CC12M* (79.47%), which is unsurprising given their size. To mitigate the effect of size of original source of data, in Fig. 8c and Fig. 8d we take the ratio of the observed percentages to the percentages we would expect based purely on the sizes of the datasets. For example, we observe that *MS COCO* captions are not good matches for *PaCo* actions.

4 Evaluation and Discussion

In this section, we focus on the *PRISM* tasks. We first benchmark state-of-the-art VLMs on the inference (§4.1) and reasoning tasks (§4.5). Then, we focus on evaluating the faithfulness of the VLMs to both modalities of the data through counterfactual analysis in the inference task (§4.4).

4.1 Inference Benchmarking Results

Here, as the main results, we benchmark the SoTA VLM models in the PVLI task.

Experimental Setup: We used 4 SOTA vision-language models: ViLBERT (Lu et al., 2019), ViLT (Kim et al., 2021), FLAVA (Singh et al., 2022a), and CLIP (Radford et al., 2021). For all four models, we start from available pre-trained models and evaluate their performance on the test set in zero-shot and fine-tuned setups. To make sure the models, especially in zero-shot setup, are familiar with the hypothesis-image-label format of the task, we first fine-tune them on Visual Natural Language Inference (VSNLI) task (Vu et al., 2018). VSNLI is a general visual language inference task that has the same format as *PRISM*, however, it does not have any explicit focus on preconditioned inference. So by fine-tuning the models on VSNLI, we make sure we directly evaluate their understanding of preconditions and not their familiarity with the format. We then report the accuracy of the resulting models on the PVLI tasks in zero-shot and fine-tuned, w.r.t. PVLI, setups. For the ViLBERT (Lu et al., 2019) model, we used the pre-trained model provided by the authors[1] that is fine-tuned VSNLI. For the rest of the models, we use the pre-trained weights from the Hugging Face library (Wolf et al., 2020) and fine-tuned them ourselves.

The ViLBERT model, provided by the authors, is originally fine-tuned on the Visual Natural Language Inference (VSNLI) task (Vu et al., 2018). We fine-tune ViLBERT on the PVLI training set with a batch size of 32 for 5 epochs, with the Adam Optimiser to optimize the cross entropy loss between the actual and the predicted labels. For all other hyperparameters, we used the default values by authors.

The Hugging Face library contains the ViLT (Kim et al., 2021) pre-trained on the Visual Question-Answering task, in which the model has to find an answer from a predefined set of tokens including *yes* and *no*. So for zero-shot and fine-tuned results, we format the PVLI dataset into a question-answering format with binary *yes/no* answers. The statement is converted into a question format by appending the phrase “Is this possible?” to the statement. This question is then fed into the model along with the associated image, which acts as the premise. The model then outputs one of the 2 labels - *yes* or *no*, which we use to compute its accuracy on the task.

FLAVA (Singh et al., 2022a) and CLIP (Radford et al., 2021) are multi-modal vision and language models that can be used for tasks such as image-text similarity or zero-shot image classification. Similar to ViLT, the hugging face library does not provide CLIP and FLAVA models that are pre-trained on binary or multi-label classification tasks. For the fine-tuned results of FLAVA model, we extract the multi-modal embeddings it generates and feed them to a classification head. This classification head is fine-tuned on the VSNLI before using in our experiments. For the CLIP model, we utilize the similarity scores between the visual and the textual features. Similarly, we feed the features through a classification head to output the label which indicates whether the precondition “allows” or “prevents” the common sense statement.

From the weak supervision data, we randomly sample 16K for *tuning* and 6K as *noisy test* set. For the *clean test* set we used the 261 human-verified samples obtained through crowd-sourcing on AMT (Discussed in §2 and detailed in Appx. §B). The experiments are conducted on a commodity workstation with an Intel Xeon Gold 5217 CPU and an NVIDIA RTX 8000 GPU.

Discussion Tab. 1 summarizes the results of SoTA VLMs on the PVLI task. In the zero-shot setup, all the models perform below the random

[1] https://github.com/facebookresearch/vilbert-multi-task

Model	0-shot Noisy Test	Clean Test	Finetuned Noisy Test	Clean Test
ViLBERT	52.02	48.48	78.75	55.68
ViLT	50.88	45.83	77.92	55.68
CLIP	30.15	42.80	73.13	56.82
FLAVA	47.38	53.78	80.43	59.47
Random	63.47	56.08		

Table 1: Results of SoTA Visual Language Models on the PVLI task.

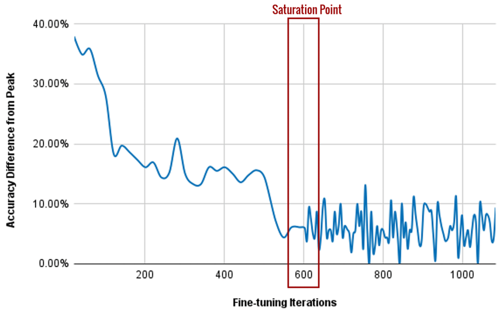


Figure 5: Accuracy difference from the peak value of fine-tuning FLAVA (lower is better) with increasing amounts of tuning data from PVLI. The batch size is 64.

baseline, showing the difficulty and novelty of the task for the models. After fine-tuning, the models' performance improves above the random guess, where the FLAVA's (Singh et al., 2022a) performance elevates by 33.05 points of accuracy to 80.43% on the *noisy-test*. However, it still is not mastering the task. Overall, this shows that SOTA methods generally fall behind human-level performance, therefore indicating the need for further research in order to improve the comprehension of preconditions by commonsense visual reasoners.

4.2 Analysis with Fine-tuning

All 4 models get higher scores on PVLI after a full fine-tuning process, as observed in Tab. 1. Here, we dissect the fine-tuning process to find at what point the model understands the task's requirements.

Experimental Setup Here we focus on FLAVA (Singh et al., 2022a) as the top-performing model in PVLI. We carry the setup from §4.1 and evaluate FLAVA on the noisy test set in fine-grained intervals during fine-tuning.

Discussion Fig. 5 illustrates the progression of the FLAVA model toward its peak accuracy performance (marked as a red box with the label "saturation point"). As illustrated, the model's performance saturates after 600 iterations of fine-tuning or observing 38K instances. The slow saturation of the accuracy score here suggests that the instances in PVLI are not trivial for the model and it has to see a substantial number of instances to be able to perform the task. Considering that the FLAVA has been pre-trained on a vast corpus, our result shows the novelty and uniqueness of the PVLI task. This result is consistent with the similar analysis in Qasemi et al. (2022a), for comparing MNLI task with PNLI (text-only).

4.3 Overall Informativeness of Modalities

To show the necessity of both modalities in *PRISM*, we conducted additional experiments based on the PABI informativeness score (He et al., 2021). Similar to Qasemi et al. (2022b) we use PABI to provide a theoretical justification on the effect of weak supervision data on the target task. In essence, the PABI score quantifies the informativeness of an incidental signal (represented by weak supervision data) in relation to a target task (represented by test data).

Experimental Setup We leverage the PABI score to assess the effectiveness of three variants of the *PRISM* dataset on the test set: utilizing only the text, utilizing only the image, and utilizing the complete dataset encompassing both image and text. In line with the methodology employed in PInKS, we also establish a baseline using the zero-rate classifier (which consistently predicts the majority class) as a point of reference with no task-related information.

Incidental data	PABI on clean test set (100X)
Text+Image	63.23
Text-only	61.95
Image-only	58.12
zero-rate	26.83

Table 2: PABI informativeness scores of three variants of the *PRISM* dataset on the clean test set

Discussion Table 2 summarizes the PABI informativeness scores of three variants of the *PRISM* dataset on the clean test set. The results clearly indicate that the image modality's informativeness on the *PRISM* test set is comparable to that of the text and both are significantly higher than the zero-rate case. This underscores the significance of both the image modality and text modality in addressing the task at hand.

4.4 Faithfulness Evaluation

Large LMs (and by extension VLMs) tend to learn to solve the dataset rather than the task (Bras et al., 2020), by overfitting spurious correlations in the data (Xu et al., 2022). To quantify and eliminate such negative effects, recent studies conduct counterfactual inference used in text classification Qian et al. (2021a) and information extraction tasks (Wang et al., 2022c, 2023). Internally, debiasing through counterfactuals works on the model trained or fine-tuned on the biased classification data. During inference, this technique creates counterfactuals where parts or all of the input are obfuscated to observe what the model would give by seeing only the biasing factors. In this way, bias terms can be distilled from the model, which can be further deducted from the original prediction for debiasing. Specifically, Qian et al. (2021a) design two types of counterfactual variations of the input to produce two counterfactual output distributions that model label bias and keyword bias in the model.

Experimental Setup Since our data contain both images and text, we modified the counterfactuals in Qian et al. (2021a) to fit the task. We create four counterfactual variants of the inputs to consider, visual-token bias, textual-token bias, image bias, and text bias. In the visual-token bias and textual-token bias, we partially mask the input image (50% as in Qian et al. (2021a)) and text (67% as in Qian et al. (2021b)) respectively with no change to the other modality of input. In the image bias and text bias we blind the model in one modality by fully masking their respective modalities. Here, we focus on the FLAVA (Singh et al., 2022a) model and carry over the setup from §4.1 on the noisy test set.

Discussion Our results show that the visually blind FLAVA (Singh et al., 2022a) model is performing on par with the original model (79.88 accuracy on noisy test). This shows that the model may overly rely on the text modality as a shortcut in most of the instances rather than utilizing both image and text. This result further motivates the need for further research in multi-modal faithfulness techniques for models.

4.5 Utilizing Rationale for Inference Task

Here, we try to answer the question "How can the rationales contribute to the inference task?". In other words, we show how the generated rationales can become a piece of useful evidence for inference.

As discussed in §5 (under "Free-Text Rationale Generation"), even though there exists a rich body of literature on the free-text rationale generation models in the text-only tasks, there are limited publicly available models for the visual language tasks. We implement the architecture proposed in Ayyubi et al. (2020) for visually-guided rationale generation[2]. The architecture feeds the visual embeddings from a VLM to the decoder of a LM and jointly trains both in an end-to-end fashion.

Experimental Setup We do an experiment similar to §4.1, except that the VLM model is trained with both the textual *hypothesis* and *rationale* plus the visual *premise* as input. To contain the length of this experiment we only focus on the FLAVA (Singh et al., 2022a) VLM, and evaluate its performance on the noisy test set in a fully fine-tuned setup. We separately experiment with two types of rationales as input: the *FLAVA-rationale-gen* gets the generated rationale, and the *FLAVA-rationale-gold* gets the ground-truth rationale from *PRISM* .

For our implementation of Ayyubi et al. (2020) to generate the rationale, we use a separate FLAVA (Singh et al., 2022a) as the VLM to embed the multi-modal input and use GPT-2 (Radford et al., 2019) as a decoder-only LM to generated the rationale from the multi-modal embeddings. We initialize both models, from pre-trained weights on Hugging Face (Wolf et al., 2020) library and fine-tune them on *PRISM* data for the rationale generation task given the input (text and image).

Discussion The inference accuracy of the *FLAVA-rationale-gold* and *FLAVA-rationale-gen* is 94.2 and 80.56 respectively. First, the significant jump in the performance of *FLAVA-rationale-gold* (from the base of 80.43 in Tab. 1) shows that in the presence of a competent rationalization model, the generated rationales can significantly contribute to the inference task. Second, we observe that a rationale model as simple as *FLAVA-rationale-gen*, can also contribute to the performance (although slightly) of the inference task. This further motivates the need for research in multi-modal rationalization models.

[2] At the time of this writing, the code for Ayyubi et al. (2020) is not public

5 Related Works

Preconditions of Commonsense Knowledge reasoning with preconditions of common sense has been studied in the context of affordance in different fields from cognitive sciences (Garbarini and Adenzato, 2004) to robotics (Ahn et al., 2022) but was recently brought up in natural language understanding. In NLP, the focus has been mainly on proposing human-verified learning resources (Qasemi et al., 2022a; Rudinger et al., 2020; Hwang et al., 2020; Sap et al., 2019; Heindorf et al., 2020; Do and Pavlick, 2021; Jiang et al., 2021a). Among them, Qasemi et al. (2022a) and Rudinger et al. (2020) propose variations of the canonical NLI task for preconditioned inference in common sense. Qasemi et al. (2022b) propose a combination of weak-supervision strategy and biased masking to improve LMs' performance in the task.

Visual Language Inference With the advent of visual language models (VLMs; Li et al. 2022b; Liu et al. 2021; Li et al. 2019; Cho et al. 2021; Huang et al. 2022) that can simultaneously process visual and linguistic information, there is growing attention to enrich text-only tasks with visual context (Parcalabescu et al., 2021; Xie et al., 2018; Vu et al., 2018). Vu et al. (2018) propose a visually-grounded version of the textual entailment task, supported by the cognitive science view of enriching meaning representations with multiple modalities. According to how Visual Language Inference (VLI; Xie et al. 2018; Vu et al. 2018) is defined, the task is regarded as a visual extension of the NLI task. In VLI, the *premise* is substituted with an image with visual context instead of the text in NLI (Xie et al., 2018). Instead of relying on crowdsourcing, both works augment the Stanford NLI (SNLI) dataset (Bowman et al., 2015b). Since the textual *premise*s of SNLI are extracted from image captions on Flickr, each *premise* can be easily replaced with its respective image. Our proposed PVLI task is a variation of the VLI that focuses on the preconditions (affordance) of tasks/objects.

Weak Supervision Instead of using direct supervision from annotated data, weak supervision in NLP tasks typically use linguistic patterns to infer large-scale "noisy" or "imperfect" labels on unlabelled corpora (Rekatsinas et al., 2017; Zhang et al., 2017; Dehghani et al., 2017; Singh et al., 2022b), e.g. using heuristic rules. Models fine-tuned on weak supervision data have shown considerable improvements across NLU tasks lacking direct supervision, including temporal commonsense reasoning (Zhou et al., 2020), rationale generation (Brahman et al., 2020), document ranking (Dehghani et al., 2017), ultra-fine entity typing (Dai et al., 2021; Choi et al., 2018), and preconditioned inference (Qasemi et al., 2022b).

Free-Text Rationale Generation There is a large body of research on free-text rationale generation toward faithful and explainable NLP. Work like this typically fine-tunes a single LM to generate the task output and rationale (Narang et al., 2020; Marasović et al., 2021; Zelikman et al., 2022), or uses a separate LM to generate the rationale that another LM uses to generate the output (Wang et al., 2022a; Wei et al., 2022; Kumar and Talukdar, 2020; Rajani et al., 2019). In the visual-language realm, free-text rationale generation is limited, where based on our observation it can be due to the lack of large-scale learning resources. Dua et al. (2021) and Ayyubi et al. (2020) repurpose the VCR (Zellers et al., 2019) data and propose VL models to generate free-text rationale (instead of picking one as is in the VCR) for it. Other works, e.g. Su et al. (2022); Li et al. (2022a), use visual inputs for text generation, but they are not focused on the rationale generation.

6 Conclusion and Future Work

We present the Preconditioned Visual Language Inference (PVLI) and Rationalization (PVLR) tasks as novel approach for assessing the capabilities of Visual Language Models (VLMs) in extracting preconditions and deducing object affordance. To establish a reliable benchmark, we introduce the PVLIR dataset, which has been meticulously evaluated by human experts through crowd sourcing. Our findings reveal a substantial performance gap between SoTA VLMs and human performance in the proposed tasks. Moreover, we demonstrate the beneficial impact of incorporating preconditioned rationalization into the inference process. To enhance the performance of VLMs in the inference task, we propose three effective strategies for acquiring and retrieving a substantial volume of cost-effective and tolerably noisy supervision signals. By conducting counterfactual analysis, we quantitatively assess the influence of spurious correlation on VLMs' performance and outline a road map for addressing the inherent challenges in their improvement.

Limitations

The quality of data produced from our weak-supervision strategies is dependent on the range of concepts covered by the image caption datasets, hence it benefits from a very large corpus of captions. Image captioning datasets we used are limited both in breadth and depth. We have not investigated the use of automatically generated captions, e.g. Wang et al. (2022b), in our weak-supervised pipeline, but it is a viable path for future extensions of this work. Alternatively, automatic text-to-image generation techniques, e.g. stable diffusion (Rombach et al., 2022) or Dall-E (Ramesh et al., 2022), are gaining a lot of attention and are promising but require a lot of prompt engineering that is challenging on a large scale. In addition, the lack of access to a large number of free-text rationale generation models (through libraries such as Huggingface (Wolf et al., 2020)) limited the evaluation of our PVLR tasks. We hope the availability of resources, such as ours, elicits more research effort in the field.

Ethical Concerns

We started from publicly available data that is both crowd-verified and neutralized, however, multiple studies have shown the existence of bias and ethical issues in such resources, e.g. Mehrabi et al. (2021). Since our work is based on weak supervision, we have no additional filter on the acquired instances, hence our resource exacerbates the bias in models by reinforcing it with biased evidence, e.g. results from the query “fat person” will only return images of obese white males. In addition, there is a combination of well-studied biases in the large models trained on raw text, e.g. Bender et al. (2021).

Finally, in this work, we have only relied on English resources. In addition, we have only used English-speaking annotators. Hence the judgments and design decisions are heavily skewed culturally which will aggravate the bias issues of our work.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116(3):463.

Hammad A Ayyubi, Md Tanjim, Julian J McAuley, Garrison W Cottrell, et al. 2020. Generating rationales in visual question answering. *arXiv preprint arXiv:2004.02032*.

Lawrence W Barsalou. 2010. Grounded cognition: Past, present, and future. *Topics in cognitive science*, 2(4):716–724.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015b. A large annotated corpus for learning natural language inference.

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2020. Learning to rationalize for nonmonotonic reasoning with distant supervision. *arXiv preprint arXiv:2012.08012*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts.

Anthony Chemero. 2003. An outline of a theory of affordances. *Ecological psychology*, 15(2):181–195.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. *arXiv preprint arXiv:1807.04905*.

Cleo Condoravdi, Dick Crouch, Valeria De Paiva, Reinhard Stolle, and Daniel Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning*, pages 38–45.

Arthur H Copeland. 1951. A reasonable social welfare function. Technical report, Mimeo, University of Michigan USA.

Kevin Crowston. 2012. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches*, pages 210–221. Springer.

Hongliang Dai, Yangqiu Song, and Haixun Wang. 2021. Ultra-fine entity typing with weak supervision from a masked language model. *arXiv preprint arXiv:2106.04098*.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74.

Nam Do and Ellie Pavlick. 2021. Are rotten apples edible? challenging commonsense inference ability with exceptions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2061–2073.

Radhika Dua, Sai Srinivas Kancheti, and Vineeth N Balasubramanian. 2021. Beyond vqa: Generating multi-word answers and rationales to visual questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1623–1632.

Francesca Garbarini and Mauro Adenzato. 2004. At the root of embodied cognition: Cognitive science meets neurophysiology. *Brain and cognition*, 56(1):100–106.

Eleanor J Gibson. 2000. Where is the information for affordances? *Ecological Psychology*, 12(1):53–56.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020a. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pages 417–434. Springer.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020b. Captioning images taken by people who are blind.

Hangfeng He, Mingyuan Zhang, Qiang Ning, and Dan Roth. 2021. Foreseeing the Benefits of Incidental Supervision. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 3023–3030. ACM.

Luyang Huang, Guocheng Niu, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Du-vlg: Unifying vision-and-language generation via dual sequence-to-sequence pre-training. *arXiv preprint arXiv:2203.09052*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *arXiv preprint arXiv:2010.05953*.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021a. " i'm not mad": Commonsense implications of negation and contradiction. *arXiv preprint arXiv:2104.06511*.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021b. "i'm not mad": Commonsense implications of negation and contradiction.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Suraj Kothawade, Vinaya Khandelwal, Kinjal Basu, Huaduo Wang, and Gopal Gupta. 2021. Autodiscern: Autonomous driving using common sense reasoning. *arXiv preprint arXiv:2110.13606*.

Sawan Kumar and Partha Talukdar. 2020. Nile: Natural language inference with faithful natural language explanations. *arXiv preprint arXiv:2005.12116*.

Doug Lenat. 1998. The dimensions of context-space.

Bin Li, Yixuan Weng, Ziyu Ma, Bin Sun, and Shutao Li. 2022a. Scene-aware prompt for multi-modal dialogue understanding and generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 179–191. Springer.

Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. 2022b. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014a. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014b. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Alexander H Liu, SouYoung Jin, Cheng-I Jeff Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. 2021. Cross-modal discrete representation learning. *arXiv preprint arXiv:2106.05438*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. 2021. Few-shot self-rationalization with natural language prompts. *arXiv preprint arXiv:2111.08284*.

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. *arXiv preprint arXiv:2103.11320*.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2021. Advanced semantics for commonsense knowledge extraction. In *Proceedings of the Web Conference 2021*, pages 2636–2647.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena.

Henry Prakken. 2017. On the problem of making autonomous vehicles conform to traffic law. *Artificial Intelligence and Law*, 25(3):341–363.

Ehsan Qasemi, Filip Ilievski, Muhao Chen, and Pedro Szekely. 2022a. Paco: Preconditions attributed to commonsense knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing-Findings*, page 6781–6796, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ehsan Qasemi, Piyush Khanna, Qiang Ning, and Muhao Chen. 2022b. Pinks: Preconditioned commonsense inference with minimal supervision. In *AACL-IJCNLP 2022*.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021a. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021b. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445, Online. Association for Computational Linguistics.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *Proc. VLDB Endow.*, 11(3):269–282.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Theodoros Rekatsinas, Xu Chu, Ihab F Ilyas, and Christopher Ré. 2017. Holoclean: Holistic data repairs with probabilistic inference. *arXiv preprint arXiv:1702.00820*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI 2019*, pages 3027–3035. AAAI Press.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2018. Atomic: An atlas of machine commonsense for if-then reasoning.

Ari Seff and Jianxiong Xiao. 2016. Learning from maps: Visual common sense for autonomous driving. *arXiv preprint arXiv:1611.08583*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022a. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Shikhar Singh, Ehsan Qasemi, and Muhao Chen. 2022b. Viphy: Probing" visible" physical commonsense knowledge. *arXiv preprint arXiv:2209.07000*.

Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. 2022. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.

Hoa Trong Vu, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt. 2018. Grounded textual entailment. *arXiv preprint arXiv:1806.05645*.

Haoyu Wang, Hongming Zhang, Yuqian Deng, Jacob R Gardner, Muhao Chen, and Dan Roth. 2023. Extracting or guessing? improving faithfulness of event temporal relation extraction. In *EACL*.

Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022a. Pinto: Faithful language reasoning using prompt-generated rationales. *arXiv preprint arXiv:2211.01562*.

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

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022c. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *NAACL*.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP 2020: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2018. Visual entailment task for visually-grounded language learning. *arXiv preprint arXiv:1811.10582*.

Nan Xu, Fei Wang, Bangzheng Li, Mingtao Dong, and Muhao Chen. 2022. Does your model classify entities reasonably? diagnosing and mitigating spurious correlations in entity typing. *arXiv preprint arXiv:2205.12640*.

Eric Zelikman, Yuhuai Wu, and Noah D Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ce Zhang, Christopher Ré, Michael Cafarella, Christopher De Sa, Alex Ratner, Jaeho Shin, Feiran Wang, and Sen Wu. 2017. Deepdive: Declarative knowledge base construction. *Communications of the ACM*, 60(5):93–102.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. *arXiv preprint arXiv:2005.04304*.

Appendices

A Weak Supervision Methods

A.1 Implementation Details and Experimental Setup

This section discusses the experimental setup and implementation details for the results in §3.

Precondition Resources: For our P-NLI datasets, we pull from ANION (Jiang et al., 2021b), ATOMIC (Sap et al., 2018), *PaCo* (Qasemi et al., 2022a), δ-NLI (Rudinger et al., 2020) and WINOVENTI (Do and Pavlick, 2021). For the image caption datasets, we use CC12M (Changpinyo et al., 2021), CC3M (Sharma et al., 2018), MS COCO (Lin et al., 2014b), and VizWiz (Gurari et al., 2020b).

Preprocessing Setup: Since ANION and ATOMIC use fixed identifiers (Alice/Bob, PersonX/PersonY), we rely on regex rules to replace them with "the person" and "another person". WINOVENTI uses random first names, and so we utilize Flair's *ner-english-fast* model (Akbik et al., 2018) to identify and replace the spans that are identified as people with greater than 90% confidence. *PaCo* does not have any such identifiers to replace.

For the image caption data, we break the captions into multiple lines using Natural Language Toolkit's sentence tokenizer (Bird et al., 2009) in combination with splitting on newline characters. We also notice that some contained "<PERSON>" tags and use regex to replace them.

As the last step, we leverage regex to fix whitespace issues and replace "the person's" with "their" to increase fluidity. We also found that datasets were easier to clean after lowercasing, particularly as some contained inconsistent capitalization.

Extraction from Captions Setup: We modify some of the original labeling functions from *PInKS* (Qasemi et al., 2022b) and add eight new ones after inspecting our caption corpus. In *PInKS*, the authors also calculate precision values for each of their labeling functions by sampling 20 examples from each function. The samples are then marked as *relevant* (score of 1) or *irrelevant* (score of 0) to the task by two human annotators. The average score of each labeling function provides an estimate of the quality that each labeling function returns and is used for tie-breaking matches or filtering out low-quality functions. We follow their lead and do the same, computing these precision values specifically on our caption corpus. Tab. 3 summarizes all the labeling functions, patterns, their precision, and other details associated with them. Balancing quality and quantity in our data, we select a threshold of 0.60 and only use the labeling functions that meet this minimum. The labeling functions are applied using Snorkel (Ratner et al., 2017), a SOTA framework for algorithmically labeling data—see the original *PInKS* paper for more detail on the setup for Snorkel. Finally, We noticed that not all of our patterns were being used in the sentence as conjunctions, and utilized Flair's *pos-english-fast* model to remove some examples for select patterns.

Caption Querying Setup: For our models, we use the sentence transformers (Reimers and Gurevych, 2019) all-distilroberta-v1, all-MiniLM-L12-v2, and all-mpnet-base-v2 from HuggingFace (Wolf et al., 2020). When forming the rankings, we retrieve the 50 closest captions, as that provides a decent overlap and completes within a reasonable amount of time. To aggregate our rankings and select the best caption, we use Copeland's method (Copeland, 1951). To compare our rankings for model agreement, we utilize the extrapolated form of rank-biased overlap (Webber et al., 2010).

Furthermore, we annotate a subset of our data to assess the ability of perplexity and model agreement to separate good training examples from poor ones. When we rate examples for quality of match between the statement (precondition/action) and the fetched image caption, on a scale from 1 (worst) to 4 (best), we find that our measures are reasonably successful (See Fig. 6). However, when we ask Amazon Mechanical Turk workers to vote on examples for overall quality, requiring that both statements and images be cohesive, the measures are unable to isolate better examples (See Fig. 7). Given that the measures are based purely on the textual match, it makes sense that it would perform better without the incorporation of the image. Unfortunately, matching with a caption is not always sufficient for matching with the associated image. Further work is needed to develop useful heuristics for the overall quality of a training example.

Label	Conjunction	Precision	Regex Pattern
enables	so that	0.689	{P} so that {A}
	in order to	0.650	{P} in order to {A}
	because	0.625	{A} because (?!of\b){P}
	due to	0.550	{A} due to {P}
	in case	0.475	{A} in case (?!of\b){P}
	as if	0.400	{A} as if {P}
	as long as	0.375	{A} as long as {P}
	if	0.150	{A}(?<!\bas) if (?!not\b){P}
	in the event	0.100	{A} in the event {P}
	on condition	0.045	{A} on condition (?!of anonymity\b){P}
	supposing	0.000*	{A} supposing {P}
	on the assumption	0.000*	{A} on the assumption {P}
	in the case that	0.000*	{A} in the case that {P}
	contingent upon	0.000*	{A} contingent upon {P}
	with the proviso	—	{A} with the proviso {P}
	to understand event	—	to understand the event "{E}", it is important to know that {P}\.
	statement is true	—	the statement "{E}" is true because {P}\.
	only if	—	{A} only if {P}
	on these terms	—	{A} on these terms {P}
	makes possible	—	{P} makes {A} possible\.
disables	unless	0.750	{A} unless {P}
	even though	0.550	{A} even though {P}
	despite	0.475	{A} despite {P}
	if not	0.300	{A}(?<!\bas) if not (?!(more|most|many|all)\b){P}
	without	0.257	{A} without {P}
	but	0.175	{A} but {NP}
	except	0.075	{A} except {P}
	lest	0.045*	{A} lest {P}
	excepting that	—	{A} excepting that {P}
	except for	—	{A} except for {P}

Table 3: Regex patterns for the labeling functions. A=action, E=event, P=precondition, NP=negative precondition. Patterns with fewer than 20 examples in the corpora are marked with asterisks, and those with no examples are left empty. Bolded conjunctions were followed with part-of-speech tagging to confirm that they were used as conjunctions.

Image Querying Setup: To find the top images on the internet, we use *Google Images Download*[3] to retrieve the URLs of images. We obtain the top 10 images for each query as it is large enough to generate lots of data while keeping them relevant to the query.

Caption Querying Results

A.2 Image Search Results

While less information is available for the *Image Querying* data, we can look at the websites most frequently drawn from for the matches. Tab. 4 and Tab. 5 display the top 10 websites for each NLI dataset for preconditions and actions, respectively. If desired, it is possible to remove images from unwanted websites.

A.3 Model Sizes and Run-times

For results in Tab. 1, the runtimes are FLAVA=4hr, VilBERT=4hr, Clip=5hr, ViLT=4hr; the model sizes for VLMs and LMs are identical to their respective implementations from the source (e.g. gpt2 has 1.5B parameters on Wolf et al. (2020)). The classification head added to VLMs (e.g. FLAVA) has $1.4k$ parameters.

B Data Annotation Details

We used Amazon Mechanical Turk (AMT) (Crowston, 2012) to evaluate the quality of extracted PVLIR instances through our proposed weak supervision methods. This enabled us to coordinate the study and access a large pool of English-speaking participants as our study population. The AMT is especially suitable for this study as it can facilitate accessing a diverse population of participants which is necessary for any notion of common sense. Our study on AMT consists of two parts: a tutorial, which also serves as a qualification test, and

[3] https://github.com/Joeclinton1/google-images-download

WinoVenti	*PaCo*	Anion
m.media-amazon.com (89)	quotefancy.com (102)	quotefancy.com (4721)
i.ytimg.com (85)	i.ytimg.com (80)	thumbs.dreamstime.com (2668)
cdn.shopify.com (67)	thumbs.dreamstime.com (73)	i0.wp.com (2074)
upload.wikimedia.org (64)	i0.wp.com (72)	i.pinimg.com (1662)
media.istockphoto.com (52)	media.istockphoto.com (46)	www.wikihow.com (1597)
thumbs.dreamstime.com (50)	m.media-amazon.com (46)	www.verywellmind.com (1546)
i0.wp.com (47)	i.pinimg.com (39)	media.istockphoto.com (1251)
images.squarespace-cdn.com (35)	c8.alamy.com (36)	miro.medium.com (997)
i5.walmartimages.com (34)	upload.wikimedia.org (35)	www.incimages.com (967)
c8.alamy.com (33)	www.wikihow.com (33)	previews.123rf.com (900)

Table 4: Top 10 websites for preconditions by NLI dataset. There are a total of 10,975 unique websites for 50,729 unique images belonging to 82,740 examples.

WinoVenti	*PaCo*	Anion
m.media-amazon.com (54)	i0.wp.com (79)	thumbs.dreamstime.com (3164)
i.ytimg.com (25)	www.verywellmind.com (62)	quotefancy.com (2943)
i5.walmartimages.com (24)	upload.wikimedia.org (45)	i0.wp.com (2050)
thumbs.dreamstime.com (21)	post.healthline.com (34)	c8.alamy.com (1964)
cdn.shopify.com (21)	media.cheggcdn.com (24)	media.istockphoto.com (1962)
i0.wp.com (20)	quotefancy.com (19)	www.wikihow.com (1469)
c8.alamy.com (20)	qph.cf2.quoracdn.net (19)	www.verywellmind.com (1262)
i.etsystatic.com (19)	www.helpguide.org (17)	i.insider.com (1213)
upload.wikimedia.org (18)	media.self.com (15)	previews.123rf.com (1075)
media.istockphoto.com (17)	images.squarespace-cdn.com (15)	i.ytimg.com (1050)

Table 5: Top 10 websites for actions by NLI dataset. There are a total of 9,700 unique websites for 48,305 unique images belonging to 80,170 examples.

the main survey. In addition, we implemented two levels of quality control: in the first one we use a response checker code and in the second we use human annotators to ensure only high-quality responses wind up in the final data.

B.1 Main AMT Survey

In the main survey, the participants are given a set of question units each consisting of a prompt question, an image, and the radio buttons with three options. We then ask participants to select their responses for each prompt question from the available options in the unit (e.g. “true” “false” “not sure” sample in Fig. 9). We create a question until through the PVLI instances with image and text, that was discussed in §2.

Since our annotated images are not perfect, there are a lot of possible points of failure that can render the question units to be impossible to understand. For example, some of the annotations may not be correct, the automatic conversion of meta-data to a sentence can be wrong in corner cases, or the image links be corrupted. Hence some of the question units may have odd grammar (e.g. "An net is used for catch fish"). Consequently, some of the question units may be hard to understand or just be wrong. To help us find those question units and ignore them in future iterations, each question unit has a checkbox in front of it with the label "not sure/does not make sense". The participant may choose to select the option and skip answering that prompt. To make the payment structure fair for the participants, they will get paid regardless of their responses. We keep the right reserved to block the participants who abuse this option using the annotator agreement metric.

B.2 Qualifying Participants

In the tutorial, first, we have prepared detailed instructions that explain to the participants what

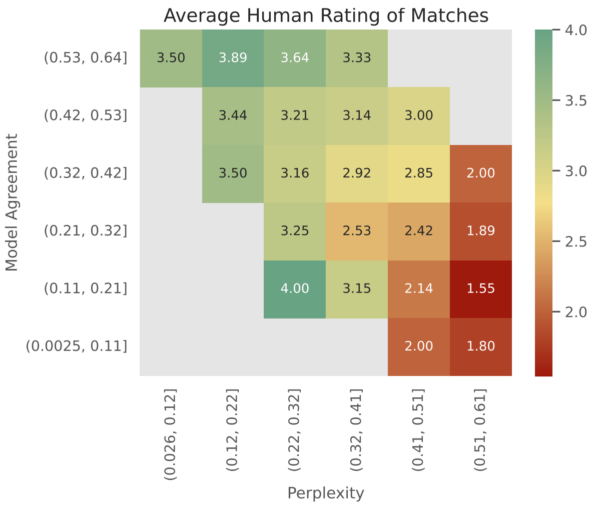


Figure 6: Heatmap graph comparing the measures of perplexity and model agreement with expert human evaluation in the caption querying method. Bins are computed using 6-quantiles for each axis.

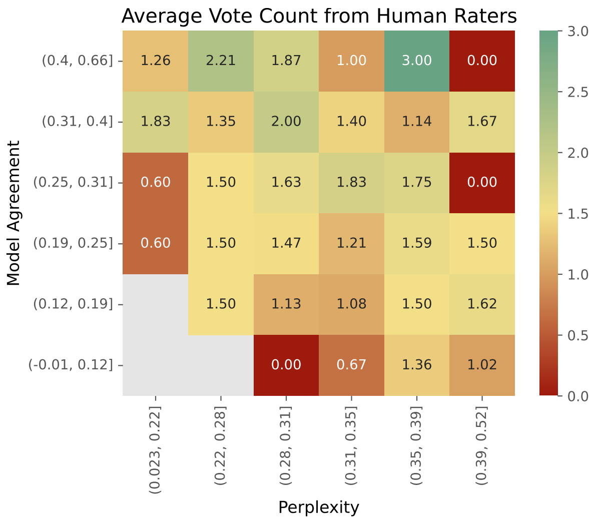


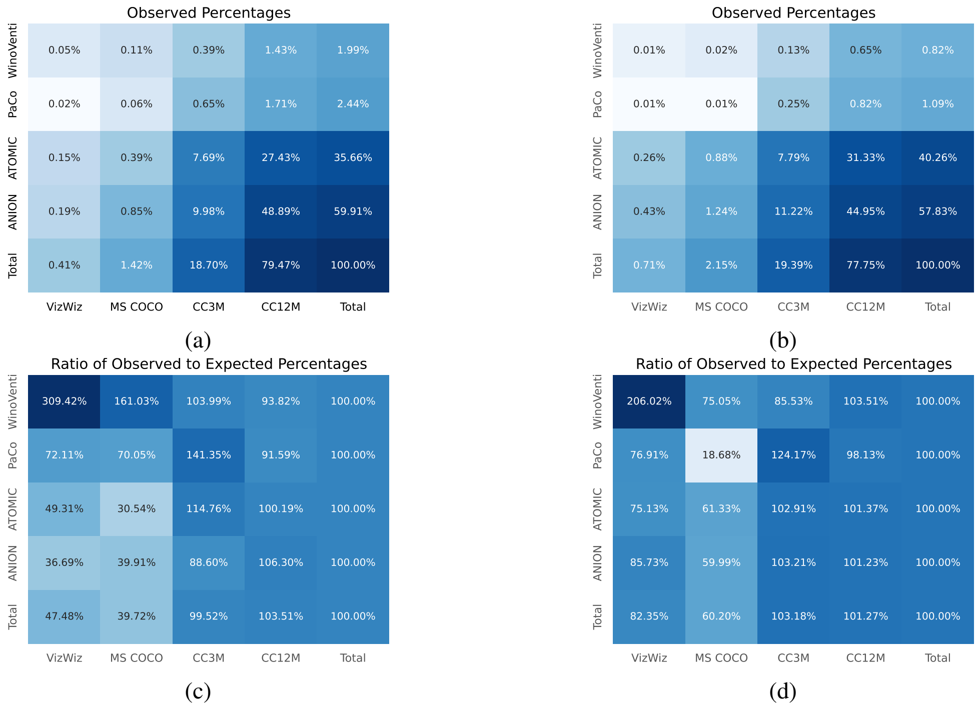
Figure 7: Heatmap graph comparing the measures of perplexity and model agreement with human evaluation from Amazon Mechanical Turk in the caption querying method. Bins are computed using 6-quantiles for each axis.

they need to do and what are the criteria for a good vs bad response. For example, in the instructions, we ask participants to avoid answering “correct”/“incorrect” when they are not sure or when there is something wrong with the image or text of the question unit. The instruction is <500 words with an expected reading time of <7 mins. Additionally, we have prepared a set of good/bad examples associated with each rule that can also be accessed in the tutorial. Each one of the good/bad examples comes with a short explanation that discusses the reason for the good/bad rating of the response. The participants are then asked to give the qualification test as a check on whether they have read and understood the instructions. The qualification test contains ∼10 question units similar to the ones they will see in the original survey (due to AMT limitations the qualification question units have a different visual layout but contain the same information). We have carefully designed each qualification question unit such that it tests the participants’ understanding of the rules individually and give them feedback on their wrong answers. For example, for the rule discouraging the use of “correct”/“incorrect” when the question unit is invalid, we have two question units where first the image is not visible, and second, the text is gibberish. After successfully passing the test, participants with acceptable scores are granted a qualification badge that allows them to engage in the main survey. It must be noted that the detailed instructions and the good/bad examples are both available in the main survey as a memory refresher for the participants.

To coordinate the study and access a large pool of participants we use Amazon’s Mechanical Turk (AMT) service to hire English-speaking people with no specific background as our study population. As part of AMT’s service design, the main survey can be divided into thousands of micro-tasks that each is related to a handful of unique question units. In this setup, the participants may choose their amount of participation in the study by accepting micro-task jobs whenever they want or fits with their schedule. Our goal is that each micro-task takes a short time to complete (less than 1 min) so we can attract a larger group of participants. It must be noted that participants can quit at any time and they will be compensated for their submitted work up until that point. To ensure the quality of the responses, the AMT service allows us to review and accept the responses from each participant individually, this allows us to pinpoint workers with low-quality responses (e.g. disagreement on more than 50 percent of the tasks with other participants) and ban them from future participation. Even after being banned, the participants with low-quality responses will be compensated for their previous accepted works.

B.3 Mechanical Turk Results

Asked annotators to go through 500 instances of PVLI. Each instance was annotated by 3 randomly-



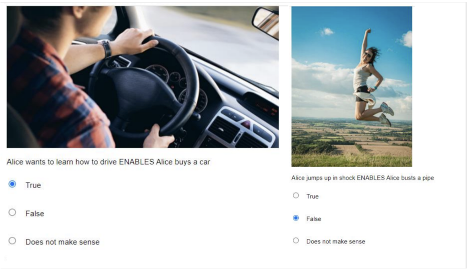
(a)

(b)

(c)

(d)

Figure 8: a) Observed distribution of matches for preconditions. b) Observed distribution of matches for actions. c) Deviation from the expected distribution of matches for preconditions. d) Deviation from the expected distribution of matches for actions.



Alice wants to learn how to drive ENABLES Alice buys a car

● True

○ False

○ Does not make sense

Alice jumps up in shock ENABLES Alice busts a pipe

○ True

● False

○ Does not make sense

Figure 9: A sample question-unit used in the main survey on the AMT.

selected workers from mainly English-speaking countries: U.S., Canada, England, India, and Australia. We selected the instances that are found correct by at least 2 annotators and use them as the *clean-test* set. The final *clean-test* set, consists of 261 human-verified instances with 151 allow labels. The inter-annotator agreement (Fleiss' Kappa) measure between our annotators is 0.78, showing good agreement among them.