BOTTLEHUMOR: Self-Informed Humor Explanation using the Information Bottleneck Principle

EunJeong Hwang[1,2], **Peter West**[1], **and Vered Shwartz**[1,2]
[1] University of British Columbia [2] Vector Institute for AI
{ejhwang,pwest,vshwartz}@cs.ubc.ca

Abstract

Humor is prevalent in online communications and it often relies on more than one modality (e.g., cartoons and memes). Interpreting humor in multimodal settings requires drawing on diverse types of knowledge, including metaphorical, sociocultural, and commonsense knowledge. However, identifying the most useful knowledge remains an open question. We introduce BOTTLEHUMOR, a method inspired by the information bottleneck principle that elicits relevant world knowledge from vision and language models which is iteratively refined for generating an explanation of the humor in an unsupervised manner. Our experiments on three datasets confirm the advantage of our method over a range of baselines. Our method can further be adapted in the future for additional tasks that can benefit from eliciting and conditioning on relevant world knowledge and open new research avenues in this direction.

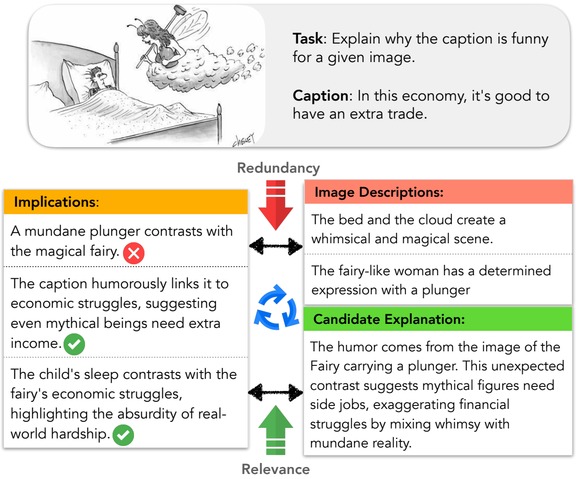

Task: Explain why the caption is funny for a given image.
Caption: In this economy, it's good to have an extra trade.
Redundancy
Implications:
A mundane plunger contrasts with the magical fairy. ✗
The caption humorously links it to economic struggles, suggesting even mythical beings need extra income. ✓
The child's sleep contrasts with the fairy's economic struggles, highlighting the absurdity of real-world hardship. ✓
Image Descriptions:
The bed and the cloud create a whimsical and magical scene.
The fairy-like woman has a determined expression with a plunger
Candidate Explanation:
The humor comes from the image of the Fairy carrying a plunger. This unexpected contrast suggests mythical figures need side jobs, exaggerating financial struggles by mixing whimsy with mundane reality.
Relevance

Figure 1: Humor understanding requires understanding world knowledge. BOTTLEHUMOR aims to reduce redundancy in existing inputs (e.g. image descriptions) while increasing relevance to candidate explanations.

1 Introduction

Humor is an effective communication tool (Stauffer, 1999; Wanzer et al., 2010; Vartabedian, 1993; Kasulis, 1989) that can manifest in various forms, including puns, exaggerated facial expressions, absurd behaviors, and incongruities (Shaw, 2010). It is shaped by multiple factors such as culture, social interactions, societal phenomena, and personal imagination (Warren and Mcgraw, 2015; Warren et al., 2020).

In particular, humor is prevalent in online communications (McCulloch, 2020), often spanning multiple modalities (e.g., cartoons and memes; Shifman, 2013). Interpreting humor across modalities requires “reading between the lines”, connecting textual and visual elements to grasp the meaning (Warren et al., 2020). For example, in Fig. 1, connecting the tooth fairy depicted in the image carrying a plunger to the caption, “In this economy, it’s good to have an extra trade”, creates the humorous interpretation that in this state of the economy, even the imaginary fairy needs a side job as a plumber.

Several datasets for multimodal humor understanding tasks were proposed, where models are tasked with generating free-text humor explanations for an image and a caption (Hwang and Shwartz, 2023; Hessel et al., 2023; Nandy et al., 2024; Hu et al., 2024b). However, they are often overlooked in vision-and-language models (VLMs) evaluations, possibly due to the subjective nature of humor and the challenges in evaluating free-text explanations. With that said, VLMs have demonstrated remarkable visual reasoning capabilities on datasets requiring scientific knowledge (Lu et al., 2022), commonsense knowledge (Schwenk et al., 2022), and spatial reasoning (Liu et al., 2023) and there is a prominent line of work on enhancing multimodal reasoning (Zhang et al., 2024; Mitra et al., 2024; Mondal et al., 2024; Hu et al., 2024a).

In this paper, we introduce BOTTLEHUMOR, a

Findings of the Association for Computational Linguistics: ACL 2025, pages 22611–22632
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

method inspired by the information bottleneck (IB) principle. BottleHumor leverages VLMs to generate and iteratively refine implications and explanations from an image and text, selecting those most relevant for explaining the humor in the image and maximizing information gain. As an off-the-shelf method, it is applicable to any VLM.

We evaluate BottleHumor on three multimodal humor explanation datasets: MemeCap (Hwang and Shwartz, 2023), NewYorker (Hessel et al., 2023), and YesBut (Nandy et al., 2024). Prior work relied on reference-based automatic metrics that overlook lexical variability and the open-endedness of explanations and costly human evaluation. Leveraging the strong text understanding capabilities of LLMs, we propose new automatic evaluation metrics that resemble precision and recall, and better correlate with human judgments. BottleHumor improves F_1 by up to 8.2, 4.3, and 2.8 points on MemeCap, NewYorker, and YesBut, respectively, compared to zero-shot baselines and outperforms existing self-refine methods that merely iterate on and refine the explanation without generating intermediate implications. Our results highlight the importance of incorporating implications, paving the way for future research on incorporating diverse world knowledge in complex reasoning tasks.[1]

2 Related Work

Multimodal Humor Understanding. Earlier works on humor understanding primarily focus on detection in images and videos (Chandrasekaran et al., 2016; Castro et al., 2019; Patro et al., 2021). Recent work shifted to generative tasks, typically explaining humor in an image (Hwang and Shwartz, 2023; Hessel et al., 2023; Nandy et al., 2024) or video (Hyun et al., 2024; Hasan et al., 2019). Understanding and explanation generation remain underexplored due to the complexity of the task and free-text evaluation. The V-Flute dataset (Saakyan et al., 2024) addresses this by re-casting this as predicting whether an image containing humorous elements or visual metaphors *entails* a given description, while providing justification. We focus on the generative version of this task, proposing a method to enhance humor explanation and a framework for automatic evaluation.

Iterative LLM-based Reasoning. Many methods elicit knowledge from the LLM for intermediate reasoning steps. Shwartz et al. (2020) elicited clarification questions and answers, then incorporated these in the input. Modern Few-shot prompting removed the need for supervision for these explanations (Marasovic et al., 2022; Wiegreffe et al., 2022). One popular approach is Chain-of-Thought (CoT; Wei et al., 2022). CoT steers LLMs to generate intermediate reasoning steps towards the final answer, improving multi-step arithmetic, commonsense, and symbolic reasoning tasks. Relevant successor approaches include self-refine (Madaan et al., 2023) which prompts LLMs to iteratively improve their answers with self-generated feedback. Eliciting knowledge from LLMs to improve predictions has been used for opinion understanding (Hwang et al., 2024; Hoyle et al., 2023), factuality (Akyürek et al., 2024), and consistency (Liang et al., 2024).

CoT has been adapted to the vision and language setting (Zhang et al., 2024) by adding external knowledge (Mondal et al., 2024), extracting a scene graph (Mitra et al., 2024), or using visual sketches as intermediate reasoning steps (Hu et al., 2024a). Most existing works focus on benchmarks such as ScienceQA (Lu et al., 2022) and visual commonsense reasoning (Schwenk et al., 2022), with (a) definitive/objective answers; and (b) simple evaluation metrics (e.g., ScienceQA is multiple-choice). We focus on multimodal explanation generation tasks in which the answers are open-ended and nuanced. As in CoT, we elicit intermediate reasoning steps from the models, but propose a novel method using the information bottleneck principle to guide generation and selection of useful knowledge for a correct explanation.

Information Bottleneck Principle. The Information Bottleneck principle (IB; Tishby et al., 1999), based on information theory, extracts relevant information from an input while minimizing redundancy (Sec. 3.1). It has been applied to a wide range of tasks (Ben-Shaul et al., 2023), including representation learning (Wu et al., 2020; Lee et al., 2021), deep learning (Saxe et al., 2018; Kawaguchi et al., 2023), summarization (West et al., 2019; Ju et al., 2021; Li et al., 2021), speech recognition (Hecht et al., 2009), and multimodal learning (Mai et al., 2023; Fang et al., 2024). Most prior works apply the IB principle during training to learn useful feature representations, with the exception of West

[1] Our code and data are available at: `https://github.com/eujhwang/bottle-humor`

et al. (2019); Ju et al. (2021), who use IB for unsupervised summarization. In this work, we extend the IB principle to multimodal humor understanding to identify relevant LLM world knowledge.

3 BottleHumor

Given a humoristic image along with an accompanying text (*caption*), our goal is to generate a descriptive explanation of the humor. For example, in Figure 2, a fairy woman with a plunger looking at a boy can be humorously explained as "The humor comes from a fairy with a plunger, taking a side job because of a tough economy" (from the NewYorker dataset; Hessel et al., 2023).

We propose BottleHumor (Figure 2), a multi-hop reasoning method inspired by the IB principle (Sec. 3.1). We integrate the visual and textual components to generate implications (Sec. 3.2). We then select the most useful implications by employing the IB principle (Sec. 3.3), and add them to the input to generate candidate explanations (Sec. 3.4). This iterative process alternates between refining implications and explanations.

3.1 The Information Bottleneck Principle

We use the Information Bottleneck principle (IB; Tishby et al., 1999) to select useful implications in BottleHumor. IB aims to extract the most relevant information from a given input variable while minimizing redundancy. Specifically, IB seeks to compress the input source S into a representation $\hat{S}$ while retaining the information most relevant to predicting the target Y. This objective is formulated as minimizing the following equation:

$$I(S, \hat{S}) - \alpha I(\hat{S}, Y)$$

where I denotes mutual information, and α is a parameter to balance compression term $I(S, \hat{S})$ with relevance term $I(\hat{S}, Y)$.

3.2 Eliciting Multi-Hop Implications

First, we generate a set of natural language implications of the input. The goal of this step is to discover connections across different objects, concepts, and situations described in the input.

Image Descriptions. As a first step, we provide the image I to a VLM to generate a detailed *image description* D, focusing on the scene and objects while ignoring the humoristic meaning behind the image. We limit the description to a maximum of five sentences.

Implications. Using these descriptions, the VLM elicits *implications*: commonsense knowledge, social norms, and possible connections for the objects in the description D and the caption C. Implications generated at hop h are denoted as $P^h = \{p_1^h, p_2^h, \ldots, p_j^h\}$.

In the first hop, the implications are derived from the image I, its caption C, and a subset of two image descriptions D, selected via a sliding window to balance efficiency (i.e., input length and cost) and coverage. From the second hop onward, we provide the VLM with *candidate explanations* (see below) and one of the previously selected top-k implications (Sec. 3.3).

When the number of generated implications exceeds 15, we cluster them using sentence embeddings and select the implications closest to each cluster's centroid. This step reduces redundancy while preserving diversity.

Candidate Explanations. To guide implication selection for generating the correct output, we provide the image I and caption C to the VLM to generate a set of *candidate explanations* at each hop: $R^h = \{r_1^h, r_2^h, \ldots, r_k^h\}$. One candidate explanation acts as an initial hypothesis, refined iteratively when additional information (implications) becomes available. In the first hop, we generate candidate explanations by providing the VLM with the image I, caption C, and descriptions D. From the second hop onward, we condition—in addition to the previous inputs—on each of the k implications selected in the previous hop (§3.3) to generate k candidate explanations. The prompts used for generating image descriptions, implications, and candidate explanations are in Appendix F.

3.3 Selecting and Refining Useful Implications

We aim to select the top k most useful implications at each hop, which should add meaningful information beyond the image and caption while providing relevant context for generating a target response. These requirements lend themselves to the two core IB components: compression and relevance.

Compression. The compression term is used to ensure that new implications provide additional information beyond what is already known. We measure the redundancy of each implication generated in the current hop h, $\{P_j^h\}_{j=1}^J$ with the inputs $X^h = \{C, D, P^{h-1}\}$, which include the image, caption, and implications generated at previous hops (when applicable). We can think of this as

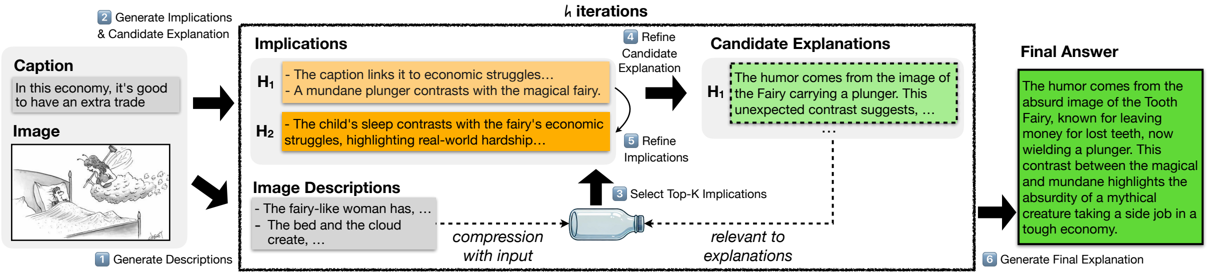

2 Generate Implications & Candidate Explanation
h iterations
Caption
In this economy, it's good to have an extra trade
Image
1 Generate Descriptions
Implications
H₁
- The caption links it to economic struggles…
- A mundane plunger contrasts with the magical fairy.
H₂
- The child's sleep contrasts with the fairy's economic struggles, highlighting real-world hardship…
4 Refine Candidate Explanation
5 Refine Implications
Candidate Explanations
H₁
The humor comes from the image of the Fairy carrying a plunger. This unexpected contrast suggests, …
…
Image Descriptions
- The fairy-like woman has, …
- The bed and the cloud create, …
compression with input
3 Select Top-K Implications
relevant to explanations
Final Answer
The humor comes from the absurd image of the Tooth Fairy, known for leaving money for lost teeth, now wielding a plunger. This contrast between the magical and mundane highlights the absurdity of a mythical creature taking a side job in a tough economy.
6 Generate Final Explanation

Figure 2: Overview of BOTTLEHUMOR. We begin by generating descriptions, implications, and a candidate explanation (steps 1 and 2). Then, we refine the implications and candidate explanations over h iterations using the IB principle (steps 3 to 5), ultimately generating a final explanation from the refined implications and candidate explanations (step 6).

testing whether the new set $X^h + \{P_j^h\}_{j=1}^J$ can be easily compressed back to X^h (redundant). To that end, we embed each of the inputs using sentence embeddings and compute the maximum cosine similarity between the target implication and each input in X^h, representing the maximum redundancy with existing information:

$$\hat{I}(X, P_j^h) = \max_{i \in I}(\cos(X_i, P_j^h))$$

Relevance. The relevance term is used to ensure that implications provide useful information for generating a target explanation. Since our method is unsupervised, we use the VLM to generate candidate explanations at hop $h-1$: $Y = \{R^{h-1}\}_{i=1}^I$, which we use as a proxy for the gold standard answer in the next hop h. We measure the relevance of the target implication P_j^h as the maximum probability (minimum cross entropy loss) for predicting the candidate explanation from the current (textual) inputs $\hat{Z}_j^h = \{C, D, P^{h-1}, P_j^h\}$, which include the caption, image description, implications from previous hops, and the target implication:

$$\hat{I}(P_j^h, Y) = \min_{i \in I}(\text{CE}(R_i^{h-1} \mid \hat{Z}_j^h))$$

Cross-entropy values tend to be lower for short candidate explanations, leading to abnormally low scores for low-quality responses. To address this, we introduce a length penalty to adjust for deviations from the average response length. Responses significantly shorter or longer than the average receive a larger penalty. We incorporate a scaling factor β, defined as the ratio of the average cross-entropy to the average length. The length penalty is then formulated as:

$$LP_i = \beta \cdot |L_i - \bar{L}|, \quad \beta = \frac{\bar{CE}}{\bar{L}}$$

where L_i is a length for i-th candidate explanation, $\bar{L}$ is the mean token length across all candidate explanations, and $\bar{CE}$ is the mean cross-entropy loss across all candidate explanations. The final relevance term for each implication becomes:

$$\hat{I}(P_j^h, Y) = \min_{i \in I}(\text{CE}(R_i^{h-1} \mid \hat{Z}_j^h) + LP)$$

We use the open/efficient Qwen2-1.5B (Yang et al., 2024) LLM to compute cross-entropy values.

Selecting Implications. With these compression and relevance terms, we formulate the final IB-based objective function. Since the goal is to minimize redundancy (maximize compression) and maximize relevance, we select k implications based on the following equation:

$$\min_{\substack{k \\ j \in J}} \quad \hat{I}(X, P_j^h) - \hat{I}(P_j^h, Y) = \qquad (1)$$
$$\min_{\substack{k \\ j \in J}} \left\{ \begin{array}{l} \max_{i \in I}(\cos(X_i, P_j^h)) + \\ \alpha \min_{i \in I}(\text{CE}(R_i^{h-1} \mid \hat{Z}_j^h) + LP) \end{array} \right\}$$

where α is a hyperparameter that controls the trade-off between the compression and relevance terms. In our experiments, we set $\alpha = 0.7$, based on our empirical observation. A detailed analysis of the effect of varying α is provided in Appendix E.

We use the implications in each hop to refine the candidate explanations in the next hop and vice versa. To avoid excessive calculation during the implication refinement step, we keep the number of candidate explanations to a maximum of three based on the cross entropy scores computed using all existing inputs. These inputs, denoted as $\hat{Z}_j^h = \{C, D, P_j^h, R_i^{h-1}\}$, include caption, image descriptions, current hop implications, and previous hop candidate explanations. We then select top-k candidate explanations ($k = 3$) in current hop candidate explanations R_i^h that minimize the cross-entropy:

$$R_{\text{top-}k}^h = \arg min_{i \in I, |I|=k} \mathrm{CE}(R_i^h \mid \hat{Z}_j^h) \quad (2)$$

In our experiments, we set the number of hops H to 2 and the number of reasoning chains k to 3.

3.4 Generating Final Answer

After H iterations of refinement, we generate the final answer. As for candidate explanation generation in earlier hops, we provide the VLM with the image I, its caption C, the k implications selected in the previous hop (Eq. 1), and the k candidate answers selected in the previous hop (Eq. 2), instructing it to generate a response.

We used Sentence Transformer[2] for all sentence embeddings. The prompts for generating multi-hop implications and explanations are in Appendix F.

4 Experimental Setup

4.1 Datasets

We evaluate BottleHumor on three multimodal humor datasets (see examples in Appendix A):

MemeCap (Hwang and Shwartz, 2023). Each instance includes a meme paired with a title (social media post to which the meme was attached). The task is to generate a brief explanation, compared against multiple reference explanations. The task requires interpreting visual metaphors in relation to the text, where models can benefit from reasoning about background knowledge.

New Yorker Cartoon (Hessel et al., 2023). We focus on the explanation generation task: given a New Yorker cartoon and its caption, generate an explanation for why the caption is funny given the cartoon, requiring an understanding of the scene, caption, and commonsense and world knowledge.

YesBut (Nandy et al., 2024). Each instance contains an image with two parts captioned "yes" and "but". The task is to explain why the image is funny or satirical.

Since our method is unsupervised, we use the test set portions of these datasets. Due to resource and cost constraints, we don't evaluate our method on the full test sets. Instead, from each dataset, we randomly sample 100 test instances. We repeat the process three times using different random seeds to obtain three test splits and report average performance and standard deviation.

4.2 Models

We test our method with two closed-source and two open-source VLMs.

GPT-4o (Hurst et al., 2024) is an advanced, closed-source multimodal model processing text, audio, images, and video and generating text, audio, and images. It matches GPT-4's performance in English text tasks with improved vision understanding.

Gemini (Team et al., 2023) is a closed-source multimodal model from Google, available in multiple variants optimized for different tasks. We use `Gemini 1.5 Flash` for evaluation and `Gemini 1.5 Flash-8B` for experiments, a smaller, faster variant with comparable performance.

Qwen2 (Yang et al., 2024) is an open-source multimodal model built on a vision transformer with strong visual reasoning. We use the `Qwen2-VL-7B-Instruct` model, competitive with GPT-4o on several benchmarks.

Phi (Abdin et al., 2024) is a lightweight, open-source 4.2B-parameter multimodal model, trained on synthetic and web data. We use `Phi-3.5-Vision-Instruct`, optimized for precise instruction adherence.

4.3 Baselines

We compare our method to four prompting-based baselines:[3] zero-shot (ZS), Chain-of-Thought (CoT) prompting, and self-refinement with (SR) and without (SR-NoC) a critic.

ZS generates a final explanation directly from the image and caption using VLM. CoT follows a similar setup but instructs the model to produce intermediate reasoning chains (Wei et al., 2022). Additionally, we implement SR, a multimodal variant of self-refinement (Madaan et al., 2023), where a *generator* produces a response, and a *critic* evaluates it based on predefined criteria. The critic's feedback helps refine the output iteratively[4]. Evaluation criteria include correctness, soundness, completeness, faithfulness, and clarity (details in Appendix H). SR-NoC functions identically to SR but without a *critic model*, refining candidate explanations without feedback. This also serves as an ablation of the implications from BottleHumor. Prompts for baselines are in Appendix H.

[2] BAAI/bge-large-en-v1.5

[3] Temperature set to 0.8 for all baselines.

[4] Refinement steps set to 2 for fair comparison.

4.4 Evaluation Metrics

While human evaluation is often the most reliable option for open-ended tasks like ours (Hwang and Shwartz, 2023), it is costly at scale. LLM-based evaluations (e.g., with `Gemini 1.5 Flash`) offer a more affordable alternative but are not always reliable (Ye et al., 2024). Prior research in fact verification has found that modern closed-source LLMs excel at fact checking when the complex facts are decomposed into simpler, atomic facts and verified individually (Gunjal and Durrett, 2024; Samir et al., 2024). Inspired by this approach, we propose LLM-based precision and recall scores.

For recall, we decompose the reference ref into atomic facts: $\{y_1, y_2, ..., y_n\}$ and check whether each appears in the predicted response $pred$.

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big(LLM(y_i, pred) = \text{Yes}\big)$$

where n is the number of atomic facts in ref.

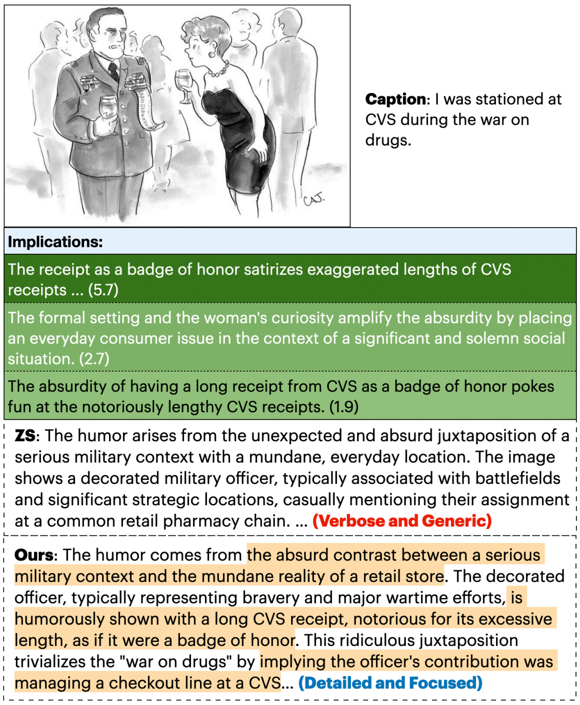
Precision follows the same process in reverse, decomposing $pred$ into a list of atomic facts: $\{x_1, x_2, ..., x_m\}$ and verifying their presence in ref:

$$\text{Precision} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\big(LLM(x_i, ref) = \text{Yes}\big)$$

where m is the number of atomic facts in $pred$. Both decomposition and verification use Gemini-Flash-1.5 with a temperature of 0.2.

In preliminary experiments, we observed that human references tend to omit obvious visual details, whereas model-generated answers are often more complete, referencing visual information. To prevent penalizing the models for these facts, we incorporate literal image descriptions (Sec 3) into the reference by decomposing them and adding them to the atomic facts for fairer evaluation. Based on the precision and recall scores, we report the macro-F_1 score.

To assess the reliability of our metrics, we conducted a human evaluation on 130 random samples across all models and datasets via CloudResearch (details in Appendix D). Human annotators determined whether each atomic sentence appeared in the corresponding text (e.g., reference). The average agreement between the LLM-based evaluator and two human annotators was 77.1% ($\kappa = 54.1$), similar to the agreement between the two annotators: 75.4% ($\kappa = 50.8$), indicating considerable alignment with human judgment. Prompts are in Appendix G.



Caption: I was stationed at CVS during the war on drugs.

Implications:

The receipt as a badge of honor satirizes exaggerated lengths of CVS receipts ... (5.7)

The formal setting and the woman's curiosity amplify the absurdity by placing an everyday consumer issue in the context of a significant and solemn social situation. (2.7)

The absurdity of having a long receipt from CVS as a badge of honor pokes fun at the notoriously lengthy CVS receipts. (1.9)

ZS: The humor arises from the unexpected and absurd juxtaposition of a serious military context with a mundane, everyday location. The image shows a decorated military officer, typically associated with battlefields and significant strategic locations, casually mentioning their assignment at a common retail pharmacy chain. ... **(Verbose and Generic)**

Ours: The humor comes from the absurd contrast between a serious military context and the mundane reality of a retail store. The decorated officer, typically representing bravery and major wartime efforts, is humorously shown with a long CVS receipt, notorious for its excessive length, as if it were a badge of honor. This ridiculous juxtaposition trivializes the "war on drugs" by implying the officer's contribution was managing a checkout line at a CVS... **(Detailed and Focused)**

Figure 3: An example analysis of the explanations of ZS and BOTTLEHUMOR for a New Yorker Cartoon, using SentenceSHAP. Implications are sorted according to their SentenceSHAP score from most to least important.

5 Results

We present the comparison of BOTTLEHUMOR to the baselines (§5.1), look into the contribution of each individual component in our method (§5.2), justify the IB framework (§5.3), and present an error analysis of our method's predictions (§5.3).

5.1 Comparison to the Baselines

Table 1 presents the overall experimental results. Compared to the best of ZS and CoT, BOTTLEHUMOR improves an average of 4.2, 1.6, and 2.1 F_1 points on the MemeCap, NewYorker, and YesBut datasets, respectively, across models. Among all models, GPT-4o performs best, averaging 3.4 F_1 point improvement across datasets. BOTTLEHUMOR significantly boosts recall while maintaining comparable precision. This suggests that our method effectively integrates external knowledge to generate more comprehensive final explanations, with a slight precision drop due to potential noise.

ZS performs reasonably well, likely due to these strong VLMs trained on similar tasks. However,

Model	Method	MemeCap			NewYorker			YesBut			Avg.
		P	R	F$_1$	P	R	F$_1$	P	R	F$_1$	F$_1$
GPT4o	ZS	81.8$_{2.0}$	34.1$_{2.6}$	48.1$_{2.8}$	75.4$_{1.4}$	42.0$_{2.5}$	53.9$_{1.9}$	73.9$_{2.3}$	47.8$_{6.6}$	58.0$_{5.5}$	53.3$_{1.9}$
	CoT	78.6$_{2.1}$	34.1$_{1.1}$	47.5$_{1.1}$	**76.0**$_{0.7}$	26.1$_{2.7}$	38.8$_{3.1}$	**74.1**$_{0.9}$	26.2$_{2.9}$	38.7$_{3.2}$	41.7$_{1.2}$
	SR	75.3$_{1.3}$	31.8$_{0.9}$	44.8$_{0.9}$	75.1$_{1.2}$	44.6$_{1.7}$	56.0$_{1.7}$	69.4$_{1.3}$	46.7$_{3.9}$	55.8$_{3.1}$	52.2$_{1.0}$
	SR-noC	**81.9**$_{2.6}$	33.5$_{0.5}$	47.5$_{0.7}$	75.2$_{1.2}$	45.8$_{1.0}$	56.9$_{0.5}$	72.4$_{1.6}$	46.3$_{2.6}$	56.5$_{2.4}$	53.6$_{1.1}$
	BottleHumor	79.1$_{2.6}$	**38.2**$_{0.8}$	**51.5**$_{0.3}$	74.5$_{2.2}$	**47.7**$_{0.3}$	**58.2**$_{0.5}$	73.8$_{3.5}$	**51.2**$_{2.9}$	**60.4**$_{2.6}$	**56.7**$_{1.1}$
Flash1.5	ZS	79.2$_{1.7}$	17.7$_{2.3}$	28.9$_{3.0}$	76.6$_{1.2}$	**24.1**$_{3.1}$	36.6$_{3.7}$	74.5$_{1.8}$	28.5$_{3.5}$	41.1$_{3.7}$	35.5$_{0.4}$
	CoT	79.5$_{1.8}$	16.1$_{1.0}$	26.7$_{1.5}$	**76.7**$_{2.6}$	13.1$_{0.5}$	22.3$_{0.8}$	**77.3**$_{2.6}$	16.4$_{1.5}$	27.1$_{2.2}$	25.4$_{0.7}$
	SR	76.2$_{2.1}$	19.4$_{1.2}$	30.9$_{1.6}$	73.7$_{1.7}$	22.9$_{1.5}$	34.9$_{1.8}$	72.7$_{0.9}$	28.9$_{3.5}$	41.3$_{3.5}$	35.7$_{1.1}$
	SR-noC	**80.9**$_{0.7}$	19.4$_{0.5}$	31.3$_{0.7}$	74.0$_{2.0}$	21.2$_{1.1}$	32.9$_{1.5}$	71.3$_{1.1}$	26.5$_{4.7}$	38.5$_{5.0}$	34.3$_{2.3}$
	BottleHumor	79.6$_{0.7}$	**20.8**$_{1.8}$	**32.9**$_{2.2}$	76.2$_{1.0}$	**24.1**$_{1.0}$	**36.7**$_{1.3}$	73.4$_{1.8}$	**30.6**$_{4.6}$	**43.1**$_{4.7}$	**37.6**$_{2.7}$
Qwen2	ZS	74.3$_{2.1}$	22.8$_{1.9}$	34.8$_{2.4}$	66.7$_{1.0}$	**19.4**$_{0.2}$	**30.1**$_{0.3}$	70.0$_{1.8}$	19.7$_{0.5}$	30.7$_{0.4}$	31.9$_{1.2}$
	CoT	71.6$_{4.0}$	22.0$_{1.2}$	33.6$_{1.5}$	**70.9**$_{1.5}$	11.0$_{1.4}$	19.0$_{2.1}$	**72.2**$_{1.6}$	13.6$_{4.4}$	22.7$_{6.3}$	25.1$_{2.6}$
	SR	73.1$_{1.5}$	23.9$_{1.2}$	36.1$_{1.5}$	67.4$_{0.9}$	17.8$_{1.1}$	28.2$_{1.4}$	68.9$_{0.7}$	20.3$_{1.7}$	31.3$_{2.1}$	31.9$_{0.3}$
	SR-noC	**75.0**$_{1.4}$	23.0$_{0.8}$	35.2$_{1.0}$	67.3$_{0.4}$	18.6$_{1.1}$	29.1$_{1.4}$	70.2$_{2.5}$	20.6$_{1.5}$	31.8$_{1.5}$	32.0$_{0.2}$
	BottleHumor	73.5$_{2.3}$	**24.0**$_{0.8}$	**36.2**$_{1.2}$	68.4$_{1.3}$	17.7$_{0.1}$	28.1$_{0.2}$	69.8$_{1.2}$	**22.1**$_{1.2}$	**33.5**$_{1.2}$	**32.6**$_{0.9}$
Phi	ZS	64.2$_{1.2}$	9.8$_{1.1}$	17.0$_{1.6}$	51.2$_{1.0}$	14.8$_{0.9}$	23.0$_{0.9}$	54.4$_{1.4}$	19.3$_{4.4}$	28.3$_{4.7}$	22.7$_{2.0}$
	CoT	59.9$_{0.9}$	11.7$_{1.1}$	19.5$_{1.5}$	**57.4**$_{1.5}$	8.5$_{1.3}$	14.8$_{2.1}$	56.2$_{2.1}$	11.7$_{2.4}$	19.3$_{3.2}$	17.9$_{0.9}$
	SR	56.8$_{0.3}$	15.1$_{3.9}$	23.7$_{4.9}$	49.1$_{0.6}$	13.0$_{0.2}$	20.6$_{0.2}$	52.5$_{1.4}$	17.2$_{3.5}$	25.8$_{4.2}$	23.4$_{2.5}$
	SR-noC	59.5$_{2.5}$	11.8$_{3.0}$	19.6$_{4.3}$	51.2$_{3.7}$	**15.1**$_{2.2}$	23.3$_{2.6}$	54.1$_{2.1}$	17.4$_{4.2}$	26.1$_{4.9}$	23.0$_{1.2}$
	BottleHumor	**65.2**$_{5.2}$	**15.6**$_{2.3}$	**25.2**$_{3.0}$	55.8$_{2.1}$	15.0$_{0.4}$	**23.6**$_{0.4}$	**57.6**$_{1.3}$	**20.0**$_{1.0}$	**29.7**$_{1.1}$	**26.2**$_{1.5}$

Table 1: Precision, Recall, and F1 scores of models and baselines on three multimodal humor benchmarks.

Model	Input	MC	NY	YB
GPT4o	Imp	47.6$_{1.3}$	53.3$_{0.3}$	54.9$_{5.3}$
	Cand	50.0$_{1.9}$	56.5$_{2.7}$	59.7$_{3.1}$
	Ours	**51.5**$_{0.3}$	**58.2**$_{0.5}$	**60.5**$_{2.5}$
Flash1.5	Imp	32.5$_{0.8}$	36.8$_{0.2}$	39.0$_{5.1}$
	Cand	32.8$_{3.2}$	**37.7**$_{1.1}$	**43.7**$_{2.7}$
	Ours	**32.9**$_{2.2}$	36.7$_{1.3}$	43.1$_{4.7}$
Qwen2	Imp	36.2$_{2.1}$	**29.2**$_{0.9}$	**36.2**$_{1.2}$
	Cand	**37.0**$_{1.0}$	**29.2**$_{0.9}$	33.5$_{0.7}$
	Ours	36.2$_{1.2}$	28.1$_{0.2}$	33.5$_{1.2}$
Phi	Imp	23.2$_{4.0}$	23.2$_{1.4}$	26.2$_{4.2}$
	Cand	**27.3**$_{2.2}$	23.1$_{0.8}$	28.1$_{4.9}$
	Ours	25.2$_{3.0}$	**23.6**$_{0.4}$	**29.7**$_{1.1}$

Table 2: F1 score comparison of using a single refined input: implications (Imp) or candidate explanations (Cand) vs. using both.

CoT causes a substantial performance drop. We observe that CoT's reasoning often leads the model to produce more generic explanations and lose focus on explaining the humor.

The self-refine baselines perform similarly to ZS, with SR slightly outperforming SR-noC. This suggests that merely refining the output without adding new information might not be beneficial for these tasks. Furthermore, incorrect feedback from SR could even negatively impact the performance. In contrast, BottleHumor outperforms both self-refinement baselines, improving an average of 2.8, 2.0, and 3.3 F_1 points on the MemeCap, NewYorker, and YesBut datasets, respectively; supporting our hypothesis that humor understanding requires additional world knowledge, which BottleHumor can successfully integrate into the reasoning process.

5.2 Contribution of Individual Components

Since our method introduces several modifications to the standard prompting approach, we assess the contribution of each individual component to the final performance. We conduct ablation tests and employ an explainability technique to point to the features that the model relies on most.

Ablation study. Table 2 presents an ablation study where only a single input is provided after refining implications and candidate explanations. GPT-4o and Phi perform better with both inputs, suggesting they effectively integrate relevant information from both to generate improved explanations. In contrast, Flash-1.5 and Qwen2 models rely more on the candidate explanations, which contain more readily-useful information than the implications, indicating these models are less proficient at ignoring noisy or irrelevant implications.

Feature importance. To further pinpoint the contribution of individual implications to the final explanations, we turn to interpretability methods. We adapt TokenSHAP (Horovicz and Goldshmidt, 2024), which estimates the importance of individual tokens to the model's prediction using Monte

Carlo Shapley value estimation, to a sentence-level variation that we refer to as SentenceSHAP (see Appendix B for details). This approach visualizes each sentence's contribution to the final explanation, as shown in Figure 3. The explanation from ZS misses the humor in the long CVS receipt that the officer is holding as a badge of honor, while BottleHumor is directly informed by the top implication.

5.3 Assessment of the IB Framework

IB component analysis. We focus on GPT4o, the best performing model across all datasets, and analyze the contribution of each IB component in our method through ablation tests. We evaluate four implication selection approaches (iterative refinement; Sec. 3.3): (1) *Random*, where implications are selected randomly; (2) *Cosine*, which selects implications with the lowest cosine similarity to the previous inputs; (3) *CE*, which selects implications that yield the lowest cross-entropy value when we condition on them to generate the candidate explanations; and (4) *Cosine+CE*, our method presented in Sec. 3.3 that combines cosine similarity and cross-entropy based on the IB principle. We conduct the analysis on 100 random instances from each dataset. Figure 4 shows that *Cosine+CE* method outperforms the *Cosine* and *CE* baselines, improving F_1 score by 4.8 and 2.3 points, respectively, confirming the importance of balancing reducing redundancy with increasing the signal.

Quality of intermediate explanations. To analyze whether the candidate explanations improve across iterations, we randomly sample 50 examples from each dataset and their outputs generated by GPT-4o and Flash1.5. Since each iteration generates three candidate explanations, we report the highest F_1 score among them, and the corresponding precision and recall values in Table 3. For GPT-4o, F_1 scores consistently improve across iterations, primarily driven by recall, which increases by an average of 11.4 points at h_2 compared to the initial hop. Precision also improves significantly at h_1, averaging an 8.0 point gain across datasets, then stabilizes. A similar trend is observed in Flash1.5-8B, a considerably smaller model, except for the MemeCap, where F_1 scores peak at h_1 but decrease by 2.5 points at h_2. While precision remains similar at the final hop compared to h_1, recall drops by 2.4 points, suggesting smaller models are more susceptible to noisy information as iterations progress.

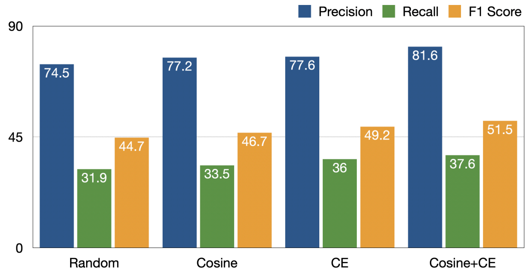


Figure 4: Performance of GPT4o on different IB components.

		GPT4o			Flash1.5		
		$\mathbf{h_0}$	$\mathbf{h_1}$	$\mathbf{h_2}$	$\mathbf{h_0}$	$\mathbf{h_1}$	$\mathbf{h_2}$
MC	P	88.5	**92.7**	**92.7**	81.0	92.2	**92.3**
	R	35.6	47.0	**48.5**	21.0	**35.0**	32.6
	F_1	50.8	62.4	**63.6**	33.3	**50.7**	48.2
NY	P	79.6	**86.5**	84.7	72.2	**83.5**	**83.5**
	R	50.6	57.9	**62.8**	22.9	33.4	**34.7**
	F_1	61.9	69.4	**72.1**	34.8	47.7	**49.0**
YB	P	67.2	**82.3**	82.0	81.0	92.2	**92.3**
	R	48.2	56.2	**57.6**	26.2	36.8	**38.1**
	F_1	56.2	66.8	**67.6**	39.6	52.6	**54.0**

Table 3: Precision, Recall, and F_1 scores on intermediate explanations across hops. h stands for hop. In our experiments, hop $h = 0$ corresponds to $k = 0$ (no implications), $h = 1$ allows up to $k = 3$ implications, and $h = 2$ allows up to $k = 6$ implications.

Error analysis. We manually analyzed 40 randomly sampled explanations across different models where implications negatively impacted performance. The two most common errors are: dilution of focus (81.2%) and introducing irrelevant information (18.7%). Dilution of focus occurs when implications repeat the same concept multiple times or include overly generalized statements that override more specific details. Irrelevant information, such as common phrases unrelated to the humor can also distort the explanation. See Appendix C for examples analyzed using SentenceSHAP.

6 Conclusions

We introduced BottleHumor, an unsupervised method inspired by the information bottleneck principle that addresses humor explanation tasks by eliciting relevant knowledge from VLMs and iteratively refining the explanation. Our experiments show that BottleHumor outperforms a range of baselines on three datasets, underscoring the importance of incorporating relevant world knowledge in humor understanding. Our analysis offers insights into the impact of individual components in our method, and justifies the use of the IB principle. We further propose an LLM-based evaluation framework and an adaptation of an interpretability technique. While we tested our contributions in the context of humor interpretation, future work can adapt them to any task that can benefit from eliciting and reasoning on world knowledge.

Limitations

Subjective nature of humor understanding. Individuals may interpret humor differently based on their personal background knowledge. While we find that the reference in the data is likely the most representative interpretation of the humor in the image and caption, other interpretations can also be valid, which are not captured in our scores.

Evaluation of explanations. Humor explanations are often nuanced and subtle. While breaking down the explanation into atomic sentences helps the model verify the accuracy and relevance of each claim, it may overlook the nuanced meaning that emerges when all the sentences are combined.

Trade-off between interpretability and efficiency. Our method emphasizes interpretable, step-by-step controllable reasoning for the humor explanation tasks, but this comes with increased resource cost. While the computational cost can be managed by limiting the number of implications or image descriptions, the increased cost remains an inherent trade-off for incorporating interpretable reasoning steps. In contrast, less interpretable or controllable approaches may offer greater efficiency. Each call typically involves $\leq$500 input tokens and $\leq$128 output tokens, with up to 20 calls per sample. For 100 samples, this results in an estimated total cost of up to \$4–5 USD using GPT-4o and up to \$1 USD using Gemini-Flash-1.5-8B.

Ethics Statement

Data. All datasets used in our work, MemeCap, NewYorker, and YesBut, are publicly available. The datasets include images, accompanying texts, and humor interpretations collected from humans and may contain offensive content to some people.

Models. The LLMs and VLMs we used for the experiments are trained on a large-scale web corpora and some of them utilize human feedback. Given their training sources, they could potentially generate content (i.e., descriptions, implications, and explanations) that exhibit societal biases.

Data Collection. We use CloudResearch to collect judgments about model-generated explanations in order to validate our proposed automatic evaluation method. To ensure the quality of evaluation, we required that workers were located in English-speaking countries (e.g. US, UK, Canada, Australia, and New Zealand), and had an acceptance rate of at least 93% on 1,000 prior annotations. We paid \$0.20 for the evaluation task, which means that annotators were compensated with an average hourly wage of \$13, which is comparable to the US minimum wage. We did not use any personal information from annotators. We obtained ethics approval from our institution's research ethics board prior to running the study.

Acknowledgements

This work was funded, in part, by the Vector Institute for AI, Canada CIFAR AI Chairs program, Accelerate Foundation Models Research Program Award from Microsoft, an NSERC discovery grant, and a research gift from AI2. We thank Jack Hessel, Benyamin Movassagh, Sahithya Ravi, Aditya Chinchure, and Vasile Negrescu for insightful discussions and feedback.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha

Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Afra Feyza Akyürek, Ekin Akyürek, Leshem Choshen, Derry Wijaya, and Jacob Andreas. 2024. Deductive closure training of language models for coherence, accuracy, and updatability. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9802–9818, Bangkok, Thailand. Association for Computational Linguistics.

Ido Ben-Shaul, Ravid Shwartz-Ziv, Tomer Galanti, Shai Dekel, and Yann LeCun. 2023. Reverse engineering self-supervised learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Arjun Chandrasekaran, Ashwin K. Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2016. We are humor beings: Understanding and predicting visual humor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yingying Fang, Shuang Wu, Sheng Zhang, Chaoyan Huang, Tieyong Zeng, Xiaodan Xing, Simon Walsh, and Guang Yang. 2024. Dynamic multimodal information bottleneck for multimodality classification. In *WACV*, pages 7681–7691.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: Desiderata for decontextualization in LLM fact verification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA. Association for Computational Linguistics.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. 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 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Ron M. Hecht, Elad Noor, and Naftali Tishby. 2009. Speaker recognition by gaussian information bottleneck. In *Interspeech 2009*, pages 1567–1570.

Jack Hessel, Ana Marasovic, 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, Toronto, Canada. Association for Computational Linguistics.

Miriam Horovicz and Roni Goldshmidt. 2024. TokenSHAP: Interpreting large language models with Monte Carlo shapley value estimation. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 1–8, Miami, FL, USA. Association for Computational Linguistics.

Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2023. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13188–13214, Singapore. Association for Computational Linguistics.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. 2024a. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhe Hu, Tuo Liang, Jing Li, Yiren Lu, Yunlai Zhou, Yiran Qiao, Jing Ma, and Yu Yin. 2024b. Cracking the code of juxtaposition: Can AI models understand the humorous contradictions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

EunJeong Hwang, Vered Shwartz, Dan Gutfreund, and Veronika Thost. 2024. A graph per persona: Reasoning about subjective natural language descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1928–1942, Bangkok, Thailand. Association for Computational Linguistics.

Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. 2024. Smile: Multimodal dataset for understanding laughter in video with language models. In *NAACL-HLT (Findings)*, pages 1149–1167.

Jiaxin Ju, Ming Liu, Huan Yee Koh, Yuan Jin, Lan Du, and Shirui Pan. 2021. Leveraging information bottleneck for scientific document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4091–4098, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas P. Kasulis. 1989. Introduction. *Philosophy East and West*, 39(3):239–241.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. 2023. How does information bottleneck help deep learning? In *International Conference on Machine Learning (ICML)*.

Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. 2021. Compressive visual representations. In *Advances in Neural Information Processing Systems*.

Haoran Li, Arash Einolghozati, Srinivasan Iyer, Bhargavi Paranjape, Yashar Mehdad, Sonal Gupta, and Marjan Ghazvininejad. 2021. EASE: Extractive-abstractive summarization end-to-end using the information bottleneck principle. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 85–95, Online and in Dominican Republic. Association for Computational Linguistics.

Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Feiyu Xiong, and Zhiyu Li. 2024. Internal consistency and self-feedback in large language models: A survey. *CoRR*, abs/2407.14507.

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sijie Mai, Ying Zeng, and Haifeng Hu. 2023. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

Gretchen McCulloch. 2020. *Because internet: Understanding the new rules of language*. Penguin.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain of thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kamcot: Knowledge augmented multimodal chain-of-thoughts reasoning. *Preprint*, arXiv:2401.12863.

Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. 2024. ***YesBut***: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16878–16895, Miami, Florida, USA. Association for Computational Linguistics.

Badri N. Patro, Mayank Lunayach, Deepankar Srivastava, Sarvesh Sarvesh, Hunar Singh, and Vinay P. Namboodiri. 2021. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 576–585.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. Understanding figurative meaning through explainable visual entailment. *Preprint*, arXiv:2405.01474.

Farhan Samir, Chan Young Park, Anjalie Field, Vered Shwartz, and Yulia Tsvetkov. 2024. Locating information gaps and narrative inconsistencies across languages: A case study of LGBT people portrayals on Wikipedia. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6747–6762, Miami, Florida, USA. Association for Computational Linguistics.

Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. 2018. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. *A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge*, pages 146–162.

Joshua Shaw. 2010. Philosophy of humor. *Philosophy Compass*, 5(2):112–126.

Limor Shifman. 2013. *Memes in digital culture*. MIT press.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

David Stauffer. 1999. Let the good times roll: Building a fun culture. *Harvard Management Update*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Laurel Klinger Vartabedian, Robert A.; Vartabedian. 1993. *The Annual Meeting of the Speech Communication Association*.

Melissa Wanzer, Ann Frymier, and Jeffrey Irwin. 2010. An explanation of the relationship between instructor humor and student learning: Instructional humor processing theory. *Communication Education - COMMUN EDUC*, 59:1–18.

Caleb Warren, Adam Barsky, and A Peter Mcgraw. 2020. What makes things funny? an integrative review of the antecedents of laughter and amusement. *Personality and Social Psychology Review*, 25:41 – 65.

Caleb Warren and A. Peter Mcgraw. 2015. Opinion: What makes things humorous. *Proceedings of the National Academy of Sciences*, 112:7105–7106.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. 2019. BottleSum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. 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 3752–3761, Hong Kong, China. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. 2020. Graph information bottleneck. In *Advances in Neural Information Processing Systems*, volume 33, pages 20437–20448. Curran Associates, Inc.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

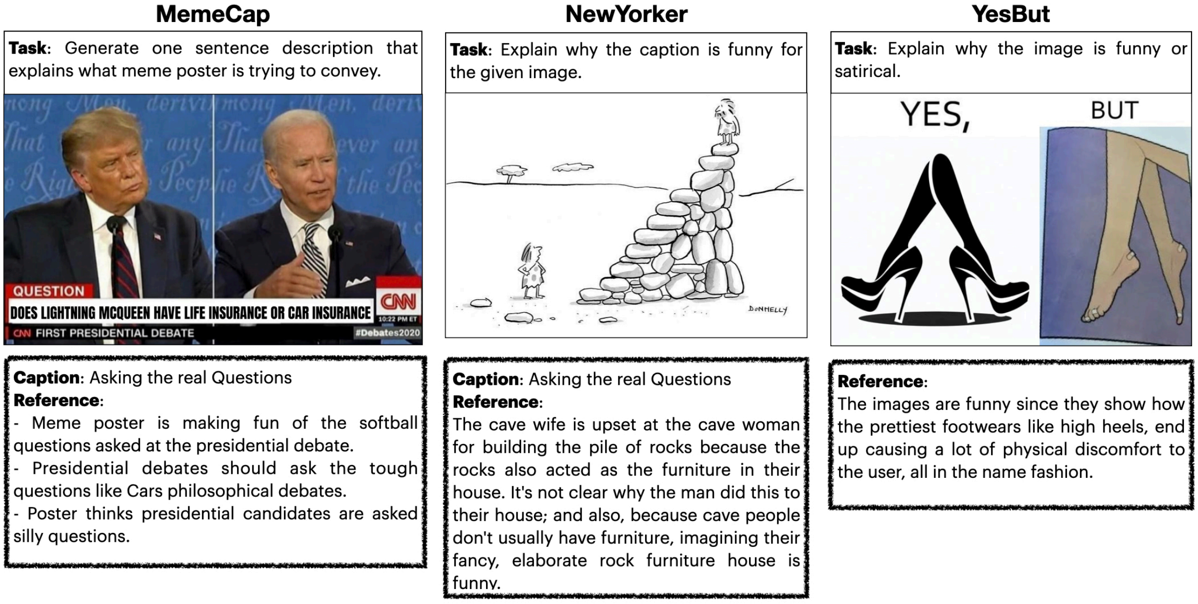
A Dataset Examples

Figure 5 illustrates example data instances from MemeCap, NewYorker, and YesBut.

B SentenceSHAP

In this section, we introduce SentenceSHAP, an adaptation of TokenSHAP (Horovicz and Goldshmidt, 2024). While TokenSHAP calculates the importance of individual tokens, SentenceSHAP estimates the importance of individual sentences in the input prompt. The importance score is calculated using Monte Carlo Shapley Estimation, following the same principles as TokenSHAP.

Given an input prompt $X = \{x_1, x_2, \dots, x_n\}$, where x_i represents a sentence, we generate all possible combinations of X by excluding each sentence x_i (i.e., $X - \{x_i\}$). Let Z represent the set



MemeCap

Task: Generate one sentence description that explains what meme poster is trying to convey.

QUESTION
DOES LIGHTNING MCQUEEN HAVE LIFE INSURANCE OR CAR INSURANCE
CNN
CNN FIRST PRESIDENTIAL DEBATE
#Debates2020

Caption: Asking the real Questions
Reference:
- Meme poster is making fun of the softball questions asked at the presidential debate.
- Presidential debates should ask the tough questions like Cars philosophical debates.
- Poster thinks presidential candidates are asked silly questions.

NewYorker

Task: Explain why the caption is funny for the given image.

Caption: Asking the real Questions
Reference:
The cave wife is upset at the cave woman for building the pile of rocks because the rocks also acted as the furniture in their house. It's not clear why the man did this to their house; and also, because cave people don't usually have furniture, imagining their fancy, elaborate rock furniture house is funny.

YesBut

Task: Explain why the image is funny or satirical.

Reference:
The images are funny since they show how the prettiest footwears like high heels, end up causing a lot of physical discomfort to the user, all in the name fashion.

Figure 5: Dataset Examples on MemeCap, NewYorker, and YesBut.

of all combinations where each x_i is removed. To estimate Shapley values efficiently, we randomly sample from Z with a specified sampling ratio, resulting in a subset $Z_s = \{X_1, X_2, \ldots, X_s\}$, where each $X_i = X - \{x_i\}$.

Next, we generate a base response r_0 using a VLM (or LLM) with the original prompt X, and a set of responses $R_s = \{r_1, r_2, \ldots, r_s\}$, each generated by a prompt from one of the sampled combinations in Z_s.

We then compute the cosine similarity between the base response r_0 and each response in R_s using Sentence Transformer (`BAAI/bge-large-en-v1.5`). The average similarity between combinations with and without x_i is computed, and the difference between these averages gives the Shapley value for sentence x_i. This is expressed as:

$$\phi(x_i) = \frac{1}{s}\sum_{j=1}^{s}\left(\cos(r_0, r_j \mid x_i) - \cos(r_0, r_j \mid \neg x_i)\right)$$

where $\phi(x_i)$ represents the Shapley value for sentence x_i, $\cos(r_0, r_j \mid x_i)$ is the cosine similarity between the base response and the response that includes sentence x_i, $\cos(r_0, r_j \mid \neg x_i)$ is the cosine similarity between the base response and the response that excludes sentence x_i, and s is the number of sampled combinations in Z_s.

C Error Analysis Based on SentenceSHAP

Figure 6 presents two examples of negative impacts from implications: dilution of focus and the introduction of irrelevant information.

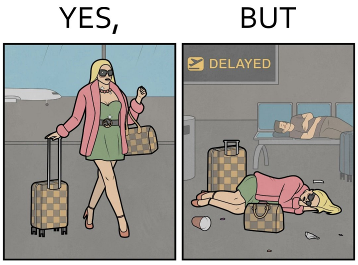
D Details on human anntations

We present the annotation interface on CloudResearch used for human evaluation to validate our evaluation metric in Figure 7. Refer to Sec. 6 for details on annotator selection criteria and compensation.

E Effect of α on explanation quality.

To understand the role of the hyperparameter α in Eq. (1), we conducted an ablation study evaluating F_1 scores from GPT-4o across all three datasets (100 samples each), using a fixed random seed. As shown in Table 4, performance varies with different α values. When $\alpha = 0.0$, the model prioritizes relevance alone; when $\alpha = 1.0$, it focuses exclusively on compression. Across datasets, the best performance tends to occur near $\alpha = 0.7$, indicating that a balanced trade-off between compression and relevance yields the most informative and accurate intermediate explanations.

Error: Dilution of focus (Over-generalization)



Implications:

The 'BUT' panel shows a man lying on the ground near his luggage, which contrasts with the initial impression of a positive outcome. (1.8)

The contrast between the organized, cheerful woman and the chaotic, defeated man underscores the often unpredictable and humorous misfortunes that can occur during travel. (1.3)

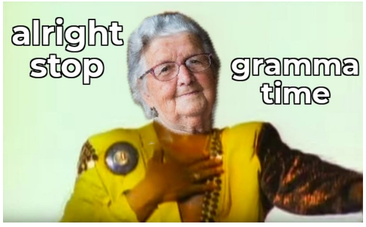
The image uses this scenario to point out the discrepancy between one's expectations upon departure and the reality of potential delays, a universal experience for travelers. (0.6)

Reference: The image is ironical, as having breakfast in bed is a luxury. However, eating while in bed leads to food crumbs, making the bed dirty, along with the need to clean the bed afterwards.

Base: ... On the left, we see an elegant woman confidently striding through an airport, presumably embarking on her journey with her luggage at her side. The right side reveals the disappointing reality of her situation: she is sprawled face down on the floor, appearing exhausted and defeated. The expectation set by the confident woman is humorously subverted by the reality of airport delays, a common travel mishap, resonating with anyone who has experienced similar frustrations. The caption adds a layer of irony to the situation, as it acknowledges that while the journey is underway, it has not gone as planned, a sentiment familiar to anyone who has faced travel delays. ...

Ours: The humor in the image arises from the stark contrast between the woman's confident appearance and the unfortunate reality of the flight delay, coupled with the casual acknowledgment in the caption "Yes, But" which suggests that despite the optimistic beginning, the situation has taken a comically negative turn. This plays on the universal experience of travel delays and the frustration they often cause, making the situation both relatable and exaggeratedly humorous.

Error: Irrelevant information



Implications:

The reference to "gramma time" suggests a lighthearted and endearing celebration of grandmothers, showcasing their unexpected coolness. (7.4)

The use of "alright stop" humorously mirrors the opening lyrics of Vanilla Ice's "Ice Ice Baby" reimagining it with a grandmotherly twist. (2.7)

Using "gramma time" subverts the expectations of the original song, creating a funny and relatable twist for those familiar with both the song and grandparent stereotypes. (1.5)

Caption: from London to the Bay

Reference:
- Meme poster is trying to convey a joke about how old MC Hammer fans are now.
- Meme poster makes a pun about Hammer time about a grandma.

Base: The meme poster is humorously depicting an elderly woman as a cool, hip figure akin to a rap icon.

Ours: The meme poster is humorously reimagining Vanilla Ice's "Ice Ice Baby" with a grandmotherly twist.

Figure 6: Examples of negative impact from implications from Phi (top) and GPT4o (bottom).

Dataset	0.0	0.3	0.7	1.0
MEMECAP	46.7	51.5	51.5	48.2
NEWYORKER	57.5	57.6	57.7	57.1
YESBUT	58.1	56.2	59.0	55.0

Table 4: Ablation study for the hyperparameter α used in Eq. (1).

F Generation Prompts for Selection and Refinement

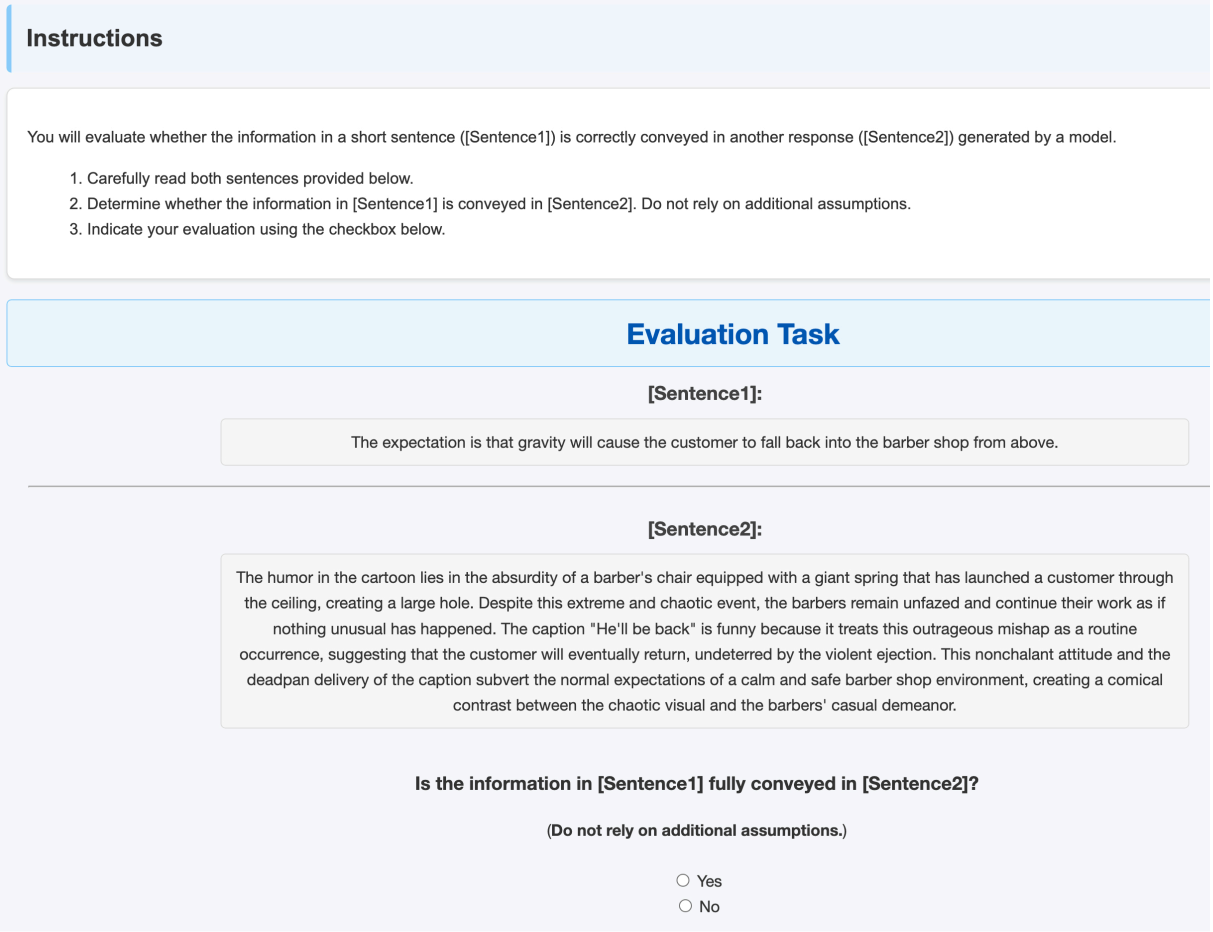
Figures 8, 9, and 10 show the prompts used for generating image descriptions, seed implications (1st hop), and non-seed implications (2nd hop onward). Figure 11 displays the prompt used to generate candidate and final explanations. Image descriptions are used for candidate explanations when existing data is insufficient but are not used for final explanations. For calculating Cross Entropy values (used as a relevance term), we use the prompt in Figure 11, substituting the image with image descriptions, as LLM is used to calculate the cross entropies.

G Evaluation Prompts

Figures 12 and 13 present the prompts used to calculate recall and precision scores in our LLM-based evaluation, respectively.

H Prompts for Baselines

Figure 14 presents the prompt used for the ZS, CoT, and SR Generator methods. While the format remains largely the same, we adjust it based on the



Instructions

You will evaluate whether the information in a short sentence ([Sentence1]) is correctly conveyed in another response ([Sentence2]) generated by a model.

1. Carefully read both sentences provided below.
2. Determine whether the information in [Sentence1] is conveyed in [Sentence2]. Do not rely on additional assumptions.
3. Indicate your evaluation using the checkbox below.

Evaluation Task

[Sentence1]:

The expectation is that gravity will cause the customer to fall back into the barber shop from above.

[Sentence2]:

The humor in the cartoon lies in the absurdity of a barber's chair equipped with a giant spring that has launched a customer through the ceiling, creating a large hole. Despite this extreme and chaotic event, the barbers remain unfazed and continue their work as if nothing unusual has happened. The caption "He'll be back" is funny because it treats this outrageous mishap as a routine occurrence, suggesting that the customer will eventually return, undeterred by the violent ejection. This nonchalant attitude and the deadpan delivery of the caption subvert the normal expectations of a calm and safe barber shop environment, creating a comical contrast between the chaotic visual and the barbers' casual demeanor.

Is the information in [Sentence1] fully conveyed in [Sentence2]?

(Do not rely on additional assumptions.)

○ Yes
○ No

Figure 7: Annotation interface on CloudResearch used for human evaluation to validate our evaluation metric.

baseline being tested (e.g., CoT requires generating intermediate reasoning, so we add extra instructions for that). Figure 15 shows the prompt used in the SR critic model. The critic's criteria include: (1) *correctness*, measuring whether the explanation directly addresses why the caption is humorous in relation to the image and its caption; (2) *soundness*, evaluating whether the explanation provides a well-reasoned interpretation of the humor; (3) *completeness*, ensuring all important aspects in the caption and image contributing to the humor are considered; (4) *faithfulness*, verifying that the explanation is factually consistency with the image and caption; and (5) *clarity*, ensuring the explanation is clear, concise, and free from unnecessary ambiguity.

Prompt for Image Descriptions

Describe the image by focusing on the noun phrases that highlight the actions, expressions, and interactions of the main visible objects, facial expressions, and people.

Here are some guidelines when generating image descriptions:
* Provide specific and detailed references to the objects, their actions, and expressions. Avoid using pronouns in the description.
* Do not include trivial details such as artist signatures, autographs, copyright marks, or any unrelated background information.
* Focus only on elements that directly contribute to the meaning, context, or main action of the scene.
* If you are unsure about any object, action, or expression, do not make guesses or generate made-up elements.
* Write each sentence on a new line.
* Limit the description to a maximum of 5 sentences, with each focusing on a distinct and relevant aspect that directly contribute to the meaning, context, or main action of the scene.

Here are some examples of desired output: —
[Description] (example of newyorker cartoon image):
Through a window, two women with surprised expressions gaze at a snowman with human arms.
—
[Description] (example of newyorker cartoon image):
A man and a woman are in a room with a regular looking bookshelf and regular sized books on the wall.
In the middle of the room the man is pointing to text written on a giant open book which covers the entire floor.
He is talking while the woman with worried expression watches from the doorway.
—
[Description] (example of meme):
The left side shows a woman angrily pointing with a distressed expression, yelling "You said memes would work!".
The right side shows a white cat sitting at a table with a plate of food in front of it, looking indifferent or smug with the text above the cat reads, "I said good memes would work".
—
[Description] (example of yesbut image):
The left side shows a hand holding a blue plane ticket marked with a price of "$50", featuring an airplane icon and a barcode, indicating it's a flight ticket.
The right side shows a hand holding a smartphone displaying a taxi app, showing a route map labeled "Airport" and a price of "$65".
—

Proceed to generate the description.
[Description]:

Figure 8: A prompt used to generate image descriptions.

Prompt for Seed Implications

You are provided with the following inputs:
- [Image]: An image (e.g. meme, new yorker cartoon, yes-but image)
- [Caption]: A caption written by a human.
- [Descriptions]: Literal descriptions that detail the image.

Your Task:
`[ One-sentence description of the ultimate goal of your task. Customize based on the task. ]`
Infer implicit meanings, cultural references, commonsense knowledge, social norms, or contrasts that connect the caption to the described objects, concepts, situations, or facial expressions.

Guidelines:
- If you are unsure about any details in the caption, description, or implication, refer to the original image for clarification.
- Identify connections between the objects, actions, or concepts described in the inputs.
- Explore possible interpretations, contrasts, or relationships that arise naturally from the scene, while staying grounded in the provided details.
- Avoid repeating or rephrasing existing implications. Ensure each new implication introduces fresh insights or perspectives.
- Each implication should be concise (one sentence) and avoid being overly generic or vague.
- Be specific in making connections, ensuring they align with the details provided in the caption and descriptions.
- Generate up to 3 meaningful implications.

Example Outputs:
Example 1 (example of newyorker cartoon image):
[Caption]: “This is the most advanced case of Surrealism I’ve seen.”
[Descriptions]: A body in three parts is on an exam table in a doctor’s office with the body’s arms crossed as though annoyed.
[Connections]:
1. The dismembered body is illogical and impossible, much like Surrealist art, which often explores the absurd.
2. The body’s angry posture adds a human emotion to an otherwise bizarre scenario, highlighting the strange contrast.

Example 2 (example of newyorker cartoon image):
[Caption]: “He has a summer job as a scarecrow.”
[Descriptions]: A snowman with human arms stands in a field.
[Connections]:
1. The snowman, an emblem of winter, represents something out of place in a summer setting, much like a scarecrow’s seasonal function.
2. The human arms on the snowman suggest that the role of a scarecrow is being played by something unexpected and seasonal.

Example 3 (example of yesbut image):
[Caption]: “The left side shows a hand holding a blue plane ticket marked with a price of ‘$50’.”
[Descriptions]: The screen on the right side shows a route map labeled “Airport” and a price of ‘$65’.
[Connections]:
1. The discrepancy between the ticket price and the taxi fare highlights the often-overlooked costs of travel beyond just booking a flight.
2. The image shows the hidden costs of air travel, with the extra fare representing the added complexity of budgeting for transportation.

Example 4 (example of meme):
[Caption]: “You said memes would work!”
[Descriptions]: A cat smirks with the text “I said good memes would work.”
[Connections]:
1. The woman’s frustration reflects a common tendency to blame concepts (memes) instead of the quality of execution, as implied by the cat’s response.
2. The contrast between the angry human and the smug cat highlights how people often misinterpret success as simple, rather than a matter of quality.

Now, proceed to generate output:
[Caption]: `[ Caption ]`

[Descriptions]:
`[ Descriptions ]`

[Connections]:

Figure 9: A prompt used to generate seed implications.

Prompt for Non-Seed Implications (2nd hop onward)

You are provided with the following inputs:
- [Image]: An image (e.g. meme, new yorker cartoon, yes-but image)
- [Caption]: A caption written by a human.
- [Descriptions]: Literal descriptions that detail the image.
- [Implication]: A previously generated implication that suggests a possible connection between the objects or concepts in the caption and description.

Your Task:
`[ One-sentence description of the ultimate goal of your task. Customize based on the task. ]`
Infer implicit meanings across the objects, concepts, situations, or facial expressions found in the caption, description, and implication. Focus on identifying relevant commonsense knowledge, social norms, or underlying connections.

Guidelines:
- If you are unsure about any details in the caption, description, or implication, refer to the original image for clarification.
- Identify potential connections between the objects, actions, or concepts described in the inputs.
- Explore interpretations, contrasts, or relationships that naturally arise from the scene while remaining grounded in the inputs.
- Avoid repeating or rephrasing existing implications. Ensure each new implication provides fresh insights or perspectives.
- Each implication should be concise (one sentence) and avoid overly generic or vague statements.
- Be specific in the connections you make, ensuring they align closely with the details provided.
- Generate up to 3 meaningful implications that expand on the implicit meaning of the scene.

Example Outputs:
Example 1 (example of newyorker cartoon image):
[Caption]: "This is the most advanced case of Surrealism I've seen."
[Descriptions]: A body in three parts is on an exam table in a doctor's office with the body's arms crossed as though annoyed.
[Implication]: Surrealism is an art style that emphasizes strange, impossible, or unsettling scenes.
[Connections]:
1. A body in three parts creates an unsettling juxtaposition with the clinical setting, which aligns with Surrealist themes.
2. The body's crossed arms add humor by assigning human emotion to an impossible scenario, reflecting Surrealist absurdity.
...
`[ We used sample examples from the prompt for generating seed implications (see Figure 9), following the above format, which includes [Implication]:. ]`
—

Proceed to Generate Output:
[Caption]: `[ Caption ]`

[Descriptions]:
`[ Descriptions ]`

[Implication]:
`[ Implication ]`

[Connections]:

Figure 10: A prompt used to generate non-seed implications.

Prompt for Candidate and Final Explanations

```
You are provided with the following inputs:
- **[Image]:** A New Yorker cartoon image.
- **[Caption]:** A caption written by a human to accompany the image.
- **[Image Descriptions]:** Literal descriptions of the visual elements in the image.
- **[Implications]:** Possible connections or relationships between objects, concepts, or the caption and the image.
- **[Candidate Answers]:** Example answers generated in a previous step to provide guidance and context.

### Your Task:
Generate **one concise, specific explanation** that clearly captures why the caption is funny in the context of the image. Your explanation must provide detailed justification and address how the humor arises from the interplay of the caption, image, and associated norms or expectations.

### Guidelines for Generating Your Explanation:
1. **Clarity and Specificity:**
- Avoid generic or ambiguous phrases.
- Provide specific details that connect the roles, contexts, or expectations associated with the elements in the image and its caption.

2. **Explain the Humor:**
- Clearly connect the humor to the caption, image, and any cultural, social, or situational norms being subverted or referenced.
- Highlight why the combination of these elements creates an unexpected or amusing contrast.

3. **Prioritize Clarity Over Brevity:**
- Justify the humor by explaining all important components clearly and in detail.
- Aim to keep your response concise and under 150 words while ensuring no critical details are omitted.

4. **Use Additional Inputs Effectively:**
- **[Image Descriptions]:** Provide a foundation for understanding the visual elements."
- **[Implications]:** Assist in understanding relationships and connections but do not allow them to dominate or significantly alter the central idea.
- **[Candidate Answers]:** Adapt your reasoning by leveraging strengths or improving upon weaknesses in the candidate answers.

Now, proceed to generate your response based on the provided inputs.

### Inputs:
[Caption]: [ Caption ]

[Descriptions]:
[ Top-K Implications ]

[Implications]:
[ Top-K Implications ]

[Candidate Anwers]:
[ Top-K Candidate Explanations ]

[Output]:
```

Figure 11: A prompt used to generate candidate and final explanations.

Prompt for Evaluating Recall Score

Your task is to assess whether [Sentence1] is conveyed in [Sentence2]. [Sentence2] may consist of multiple sentences.

Here are the evaluation guidelines:
1. Mark ’Yes’ if [Sentence1] is conveyed in [Sentence2].
2. Mark ’No’ if [Sentence2] does not convey the information in [Sentence1].

Proceed to evaluate.

[Sentence1]: `[ One Atomic Sentence from Decomposed Reference Explanation ]`

[Sentence2]: `[ Predicted Explanation ]`

[Output]:

Figure 12: Prompt for evaluating recall score.

Prompt for Evaluating Precision Score

Your task is to assess whether [Sentence1] is inferable from [Sentence2]. [Sentence2] may consist of multiple sentences.

Here are the evaluation guidelines:
1. Mark "Yes" if [Sentence1] can be inferred from [Sentence2] — whether explicitly stated, implicitly conveyed, reworded, or serving as supporting information.
2. Mark ’No’ if [Sentence1] is absent from [Sentence2], cannot be inferred, or contradicts it.

Proceed to evaluate.

[Sentence1]: `[ One Atomic Sentence from Decomposed Predicted Explanation ]`

[Sentence2]: `[ Reference Explanation ]`

[Output]:

Figure 13: Prompt for evaluating precision score.

Prompt for Baselines

```
You are provided with the following inputs:
- **[Image]:** A New Yorker cartoon image.
- **[Caption]:** A caption written by a human to accompany the image.
[ if Self-Refine with Critic is True: ]
- **[Feedback for Candidate Answer]:** Feedback that points out some weakness in the current candidate responses.
[ if Self-Refine is True: ]
- **[Candidate Answers]:** Example answers generated in a previous step to provide guidance and context.

### Your Task:
Generate **one concise, specific explanation** that clearly captures why the caption is funny in the context of the image. Your explanation must provide detailed justification and address how the humor arises from the interplay of the caption, image, and associated norms or expectations.

### Guidelines for Generating Your Explanation:
1. **Clarity and Specificity:**
- Avoid generic or ambiguous phrases.
- Provide specific details that connect the roles, contexts, or expectations associated with the elements in the image and its caption.

2. **Explain the Humor:**
- Clearly connect the humor to the caption, image, and any cultural, social, or situational norms being subverted or referenced.
- Highlight why the combination of these elements creates an unexpected or amusing contrast.

3. **Prioritize Clarity Over Brevity:**
- Justify the humor by explaining all important components clearly and in detail.
- Aim to keep your response concise and under 150 words while ensuring no critical details are omitted.

[ if Self-Refine is True: ]
4. **Use Additional Inputs Effectively:**
- **[Candidate Answers]:** Adapt your reasoning by leveraging strengths or improving upon weaknesses in candidate answers.
[ if Self-Refine with Critic is True: ]
- **[Feedback for Candidate Answer]:** Feedback that points out some weaknesses in the current candidate responses.

[ if CoT is True: ]
Begin by analyzing the image and the given context, and explain your reasoning briefly before generating your final response.

Here is an example format of the output:
{{
"Reasoning": "...",
"Explanation": "..."
}}

Now, proceed to generate your response based on the provided inputs.

### Inputs:
[Caption]: [ Caption ]

[Candidate Answers]: [ Candidate Explanations ]

[[Feedback for Candidate Answer]:]: [ Feedback for Candidate Explanations ]

[Output]:
```

Figure 14: A prompt used for baseline methods, with conditions added based on the specific baseline being experimented with.

Prompt for Self-Refine Critic

`[ Customize goal text here: ]`
`MemeCap:` You will be given a meme along with its caption, and a candidate response that describes what meme poster is trying to convey.
`NewYorker:` You will be given an image along with its caption, and a candidate response that explains why the caption is funny for the given image.
`YesBut:` You will be given an image and a candidate response that describes why the image is funny or satirical.

Your task is to criticize the candidate response based on the following evaluation criteria:
- Correctness: Does the explanation directly address why the caption is funny, considering both the image and its caption?
- Soundness: Does the explanation provide a meaningful and well-reasoned interpretation of the humor?
- Completeness: Does the explanation address all relevant aspects of the caption and image (e.g., visual details, text) that contribute to the humor?
- Faithfulness: Is the explanation factually consistent with the details in the image and caption?
- Clarity: Is the explanation clear, concise, and free from unnecessary ambiguity?

Proceed to criticize the candidate response ideally using less than 5 sentences:

[Caption]: `[ caption ]`

[Candidate Response]:
`[ Candidate Response ]`

[Output]:

Figure 15: A prompt used in SR critic model.