Content-Specific Humorous Image Captioning Using Incongruity Resolution Chain-of-Thought

Kohtaro Tanaka[1,2], **Kohei Uehara**[1,2], **Lin Gu**[2,1], **Yusuke Mukuta**[1,2], **Tatsuya Harada**[1,2]
[1]The University of Tokyo [2]RIKEN
{k-tanaka, uehara, lingu, mukuta, harada}@mi.t.u-tokyo.ac.jp

Abstract

Although automated image captioning methods have benefited considerably from the development of large language models (LLMs), generating humorous captions is still a challenging task. Humorous captions generated by humans are unique to the image and reflect the content of the image. However, captions generated using previous captioning models tend to be generic. Therefore, we propose incongruity-resolution chain-of-thought (IRCoT) as a novel prompting framework that creates content-specific resolutions from fine details extracted from an image. Furthermore, we integrate logit bias and negative sampling to suppress the output of generic resolutions. The results of experiments with GPT4-V demonstrate that our proposed framework effectively generated humorous captions tailored to the content of specific input images.[1]

1 Introduction

Humorous content comprising an image with an associated caption is universally popular in different communities. For example, Imgflip[2], Bokete[3], and The New Yorker Cartoon Caption Contest[4] all express different tastes in humor using images with text captions. This form of humorous content is important in human communication, such as by providing an effective way to lead others to challenge misinformation (Yeo and McKasy, 2021).

While the topic of humorous image captions remains relatively unexplored, several studies have leveraged large-scale datasets of humorous combinations of images with captions from the Internet to train image captioning models (Peirson V and Tolunay, 2018; Sadasivam et al., 2020; Li et al., 2023). However, previous research has shown that image captioning models trained using cross-entropy loss have a tendency to generate similar captions for different images (Fei and Huang, 2023). LLMs like GPT4 (OpenAI, 2023a) have also been used to generate humorous captions using descriptions of the images provided by humans. However, existing methods are relatively limited, focusing on a single example and lacking in quantitative analysis (Hessel et al., 2023). Furthermore, the capabilities of large multimodal models (LMMs) such as GPT4-V (OpenAI, 2023b) to generate humorous captions have not been previously investigated. In this study, we found that GPT4-V also tends to produce generic captions in attempts at humor, and largely fails to capture the content-specific nuances in images found in humorous captions created by humans, as shown in Figure 1.

Inspired by the incongruity theory of humor, we introduce incongruity-resolution chain-of-thought (IRCoT) as an approach to generate humorous captions related to the content in input images. The incongruity-resolution theory is a well-established framework that describes how humor arises from an unexpected contradiction resolved through a cognitive rule that explains the content's incongruity (Raskin, 1985; Buijzen and Valkenburg, 2004). A study on incongruity in image macro memes, a form of humor comprising an image with an associated caption, suggests that most memes conform to the incongruity-resolution theory (Yus, 2021). IRCoT guides a machine learning model to identify and resolve incongruities in the content of input images as shown in Figure 1. We hypothesized that IRCoT could facilitate the creation of content-specific humorous captions for each image by generating resolutions based on intricate and unique details of the content depicted in the image, which can be recognized in the preceding steps. The results of experiments using GPT4 show that IRCoT improved the specificity of humorous captions generated by GPT4-V compared to prompting

[1]Our project page is available at https://kohtaro246.github.io/publication/IRCoT
[2]https://imgflip.com/
[3]https://bokete.jp/
[4]https://www.newyorker.com/cartoons/contest

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2348–2367
June 16-21, 2024

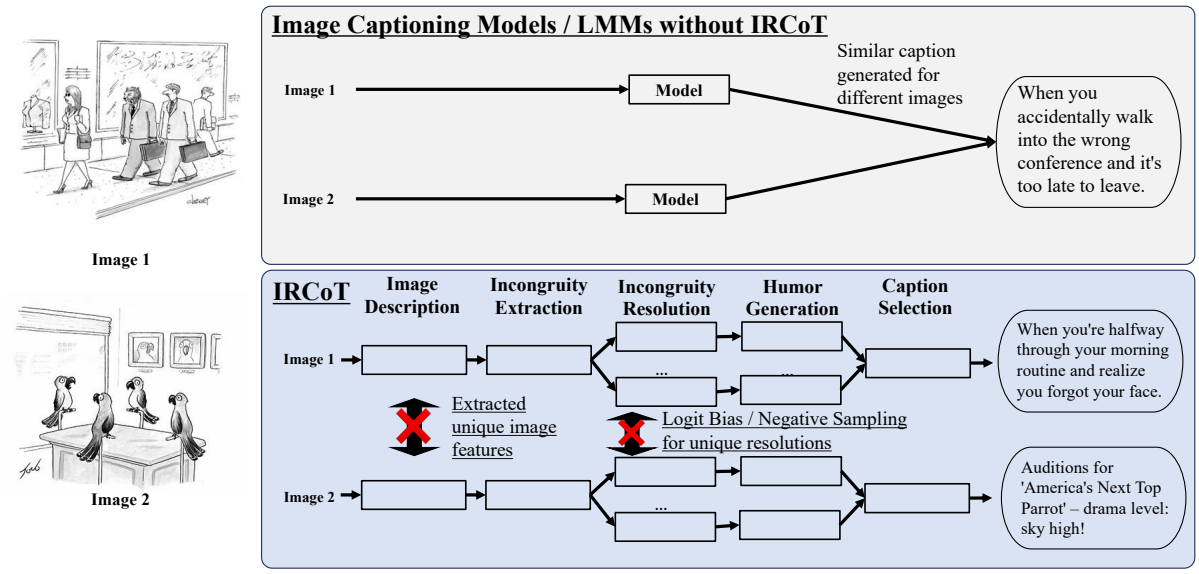

Figure 1: The upper figure shows an example in which GPT4-V without IRCoT generated similar captions for different images. The lower figure shows the proposed IRCoT pipeline to produce content-specific humor using LMMs. The framework is intended to generate humorous captions that are specific to the content in a given image based on the unique details of the image extracted by an "Image Description" module. In addition, we introduce logit bias and negative sampling to generate unique resolutions. This leads to the generation of humorous captions related to the content in a given input image.

the model without using IRCoT.

Furthermore, we show that using logit bias and negative sampling fine-tuning during the resolution step enhanced the specificity of the generated captions without the need for training data provided by humans. These techniques penalize the model for generating generic resolutions that can resolve any incongruities.

In addition to achieving content-specificity, we argue that IRCoT may reduce the risk of models generating offensive content compared to other data-driven approaches that use image-caption humor datasets that contain offensive content (Kiela et al., 2020). Intermediate explanations generated by IRCoT may improve an offensiveness detection model with challenging samples. This issue has become particularly pressing in light of increasing social demands for generative models to avoid harmful content, as outlined in the EU Artificial Intelligence Act (European Parliament, 2023).

The contributions of this study are summarized as follows:

- We discovered that GPT4-V typically produces generic captions lacking content-specificity.
- We propose a novel prompting framework called IRCoT that enables GPT4-V to generate content-specific humorous captions by inducing the model to generate resolutions based on fine details of an input image.
- We established that incorporating logit bias and negative sampling fine-tuning improved the content-specificity of humorous captions.

2 Related Work

2.1 Humorous Image Captioning

Computational tasks involving image-text humor can be broadly classified into three categories, including detecting, evaluating, and generating humor. Although several studies have focused on humor detection and evaluation (Sharma et al., 2020; Kiela et al., 2020; Bejan, 2020; Tanaka et al., 2022), the topic of generating humorous captions has received comparatively less attention.

Previous studies on humorous captioning using neural networks trained popular image captioning models such as LSTM (Graves and Graves, 2012) and Transformer (Vaswani et al., 2017) models using large-scale humor datasets. These datasets were either created through manual annotation of a large number of images via crowdsourcing (Gan et al., 2017) or by scraping humorous content from meme-sharing websites (Peirson V and Tolunay, 2018; Sadasivam et al., 2020; Li et al., 2023). While this approach enabled the generation of captions in a humorous style, the content-specificity of the gener-

ated captions is not guaranteed due to the inherent problem of image captioning models generating generic captions (Fei and Huang, 2023). To address the lack of diversity in generated captions using trained image captioning models, Li et al. (2023) proposed the position-conditioned loss. In addition, a model trained on data that include offensive content often prevalent in the Internet carries the risk of the model generating offensive content (Kiela et al., 2020).

With the advent of LLMs like GPT-4, their potential for generating humor was explored in an appendix of research focused on the capabilities of these models to understand humor (Hessel et al., 2023). This research tested the few-shot ability of GPT4 to generate a humorous caption when prompted with several human-generated captions and explanations of the images. However, this was tested for only a single example image and no quantitative analysis was conducted.

2.2 Chain-of-Thought for Zero-shot Reasoning

LLMs that are pretrained with extensive datasets demonstrate impressive zero-shot capabilities across a range of tasks (OpenAI, 2023a; Liu et al., 2023a). However, for certain complex reasoning tasks, such as solving mathematical problems or puzzles, simple prompts have proven insufficient to fully leverage the capabilities of these models (Rae et al., 2021). The chain-of-thought (CoT) prompting method was introduced to address this by enhancing the zero-shot performance of LLMs in complex reasoning scenarios (Wei et al., 2022). CoT prompting encourages a model to generate intermediate steps that mimic human thought processes to enable it to arrive at accurate solutions for previously unseen problems. Various adaptations of CoT have since been developed to further augment the zero-shot capabilities of LLMs (Wang et al., 2023; Long, 2023; Besta et al., 2023).

2.3 Content-specific Image Captioning

Recent LLMs, which are based on pretrained transformers, employ next-token prediction during their pre-training phase (Brown et al., 2020). However, previous research indicates that this training approach focusing on minimizing cross-entropy loss for generated tokens often results in a model producing generic captions (Fei and Huang, 2023). Several methods have been proposed to address this issue. One such approach involves using a negative sampling loss, which trains the model to avoid outputting certain words (Welleck et al., 2019). Another method involves training a "teacher" model using generic captions and then training a "student" model to avoid generating tokens that the teacher model produces (Fei and Huang, 2023). While these methods have successfully produced more discriminative captions in smaller-scale transformer models, they all require extensive training data, which poses a significant challenge for LLMs due to the high associated computational costs.

3 Content-Specificity of Humorous Captions Generated by LMMs

In this section, we describe our analysis of the content-specificity of humorous captions generated by an LMM using GPT4-V. We selected GPT4-V for this analysis because it is considered a benchmark in LMMs, and is commonly used to create training data for other models and evaluate their performance (Liu et al., 2023b). Additionally, prior research on the capabilities of computational models to assess humor identified GPT4 as the most proficient model among the three tested (Hessel et al., 2023).

3.1 Metric

To quantitatively evaluate the content-specificity of the generated captions, we employed Self-CIDEr (Wang and Chan, 2019), mBLEU, and Div-1 (Li et al., 2016). These metrics are designed to measure the differences in captions associated with different images at a token level. We concluded that evaluation metrics relying on pretrained feature extractors are unsuitable for this task, primarily because feature extractors like CLIP (Radford et al., 2021) are not trained on humorous captions, which often contain unique expressions not found in standard image captioning tasks.

3.2 Data

The testing data comprised humorous image-text pairs from three different sources, including ImgFlip, Bokete, and The New Yorker Cartoon Caption Contest. ImgFlip and Bokete are meme-sharing websites where users can post, view, and vote on memes. ImgFlip primarily features English memes, while Bokete is a Japanese site dedicated to Japanese memes. The image-text pairs from Imgflip and Bokete were selected from the OxfordTVG-HIC dataset, a large-scale collection

Methods	SelfCIDEr(↑)	mBLEU(↓)	Div-1(↑)
Human	**0.868**	**0.014**	**0.361**
Simple GPT4-V	0.782	0.157	0.295
CoT GPT4-V	0.756	0.178	0.259

Table 1: Quantified content-specificity of captions generated by humans and GPT4-V. Human-generated captions exhibited higher content specificity compared to captions generated by GPT4-V.

of image-text pairs and humor ratings (Li et al., 2023). This dataset includes preprocessed English captions filtered to remove offensive content. We selected 131 images from each source to compile a testing set, choosing those with the highest-rated humorous captions.

The New Yorker Cartoon Caption Contest, held weekly by The New Yorker magazine, allows anyone to submit captions for provided cartoons, with three finalists chosen by the magazine's editors. We utilized the "Explanation test split" from previous work that evaluated GPT4's performance in evaluating humor, which had collected and preprocessed past contest results (Hessel et al., 2023).

In total, our dataset encompasses 393 unique images, each accompanied by a single human-created caption.

3.3 Experimental Setting

We conducted a comparative analysis of captions generated by humans and those produced by GPT4-V. We used a simple prompt and a CoT prompt. The former simply instructed the model to create a humorous caption from the image. In addition to the simple prompt, the CoT prompt instructed it to output the steps used to arrive at the final output. For detailed information on the prompts, versions, and parameters of GPT4-V used in our study, refer to Section B.1.

3.4 Result

The quantitative results are shown in Table 1. All metrics indicated that the humorous captions created by humans were more content-specific than those generated by GPT4-V.

Figure 2 presents two examples in which GPT4 generated captions that are similar, despite being associated with different images. Although these captions capture certain elements of each image, they fall short in some respects. For example, the caption for the image on the left accurately describes a person wearing a suit walking, but it fails to acknowledge the incongruity of the situation, namely that one of the businessmen has the face of a werewolf.

This result highlights the challenge of generating unique humorous captions that reflect the content of an image.

4 Content-Specific Humor Generation

In this section, we describe IRCoT, a novel prompting method that aims to improve the content-specificity of humorous captions generated by LMMs.

4.1 Incongruity Resolution Chain-of Thought (IRCoT)

As shown in Figure 1, IRCoT induces the model to reason in 5 consecutive steps, including image description, incongruity extraction, resolution, humor generation, and selection.

As shown in Figure 3, the LMM is first prompted to extract all fine details depicted in the image, including the incongruous element in the image description, and then performs further incongruity extraction steps.

Then, based on the descriptions of unique features depicted in the image, the LMM is instructed to generate 20 possible resolutions to the extracted incongruity. We generate multiple resolutions because it is known from previous research on the incongruity theory that humans can follow various paths to resolve an incongruity in a humorous way (Ritchie, 2009). This phenomenon is reflected in the fact that a standard dataset of humorous image captions contains over 10 times more captions per image compared to a standard image captioning dataset (Li et al., 2023), which highlights the variety of incongruity-resolution pairs that can be associated with a single image.

Finally, humorous captions are generated and selected based on the incongruity-resolution pairs generated in the previous steps.

4.2 Logit Bias

We hypothesize that even with IRCoT, the model may output generic resolutions that can explain any kind of incongruous element in the image. For example, using keywords that signify a fictitious or metaphorical setting, such as "dream" or "symbol" enables the model to output a simple resolution to any incongruous element. This would result in the model generating generic humor captions.

To prevent this from happening, we propose the use of logit bias to manipulate the logits of the

Human:
Oh, sure, they find one secretary in a pool of her own blood and everybody wants to blame the werewolf.
GPT4-V:
When you accidentally walk into the wrong conference and it's too late to leave.

Human:
We have to stop eating the seed money.
GPT4-V:
When you accidentally walk into the wrong room but play it cool.

Figure 2: Two examples in which GPT4-V generated similar captions. The human-generated captions are based on fine details of the image, whereas GPT4-V generated captions that focus only on broad elements of the image such as businessmen walking or the setting of a conference room.

model to suppress keywords that can resolve any kind of incongruous elements in the image. To determine which word to suppress, we created a resolution dataset by using GPT4-V to generate resolutions to incongruous elements in images. Then, for the generated resolutions, we used the following steps to calculate the document frequency.

1. Pre-processing to convert uppercase letters to lowercase, remove any punctuations included in the Python's "string.punctuation", and remove any stop words using NLTK [5] library.
2. Tokenize using the model's tokenizer.
3. For all the tokens used for resolution, calculate the percentage of images for which each token was used for resolution.

We performed an identical process to calculate the document frequency for the COCO Captions Dataset (Lin et al., 2014). Finally, we subtracted the document frequency of the COCO Captions Dataset from the resolution document frequency and extracted tokens with a positive subtracted value between 0 and 1 (penalty weight). Performing this process extracted keywords (penalty tokens) that appeared frequently only in generated resolutions and not in the COCO Captions Dataset. During generation, logits of penalty tokens are manipulated based on the penalty weight and logit bias weight β to suppress penalty tokens.

4.3 Negative Sampling

To reduce the generation of generic resolutions, we also employ negative sampling fine-tuning. Because the loss function of GPT4 cannot be changed by the end user, we fine-tuned a pretrained LLaVA 1.5 model using the resolution dataset introduced in Section 4.2 with the negative sampling loss. Given a previously generated sequence $(x_0, \cdots, x_{t-1})$, a set of penalty tokens $\mathcal{C}$, a penalty weight for each penalty token $pw(c)$, and a hyperparameter α, we define the negative sample loss for step t as

$$\mathcal{L}^t = -\log p(x_t|x_{<t}) - \alpha \sum_{c \in \mathcal{C}} pw(c) \log(1 - p(c|x_{<t})). \quad (1)$$

This step induces the model to focus on learning from examples that avoid using the identified penalty tokens. This method does not require human annotation because the training data are generated by GPT4.

5 Experimental setup

5.1 Data

To test the capability of large models to generate context-specific humorous caption using IRCoT, we used the testing set from the experiments described in Section 3.

We also created two types of training datasets, including an image-caption training dataset and a resolution dataset. The image-caption training set contains 361K image-caption pairs with 65K unique images that are not included in the testing set. The resolution dataset contains 10K pairs of images and results generated by GPT4-V from IRCoT step 3. The images were randomly sampled from the image-caption training set.

[5] https://www.nltk.org/

5.2 Methods Used for Comparison

Trained Baselines: To compare the capability of LMMs prompted with IRCoT with that of LMMs trained using large humorous image-caption datasets, we trained two types of LLaVA 1.5 7b models using the image-caption training dataset. The first optimized a cross-entropy loss. For the second model, we implemented the position-conditioned loss that was proposed in a previous study on increasing the diversity of generated captions (Li et al., 2023).

w/o IRCoT: We also compared the results with humorous captions generated using a simple prompt. We used the same prompt as in Section 3.

IRCoT: For experiments with IRCoT, we used 5 different settings that differed in terms of how the resolution (step 3) was performed. Note that we used the same GPT4-V model for steps 1, 2, 4, and 5. First, the "GPT4-V" setting used GPT4-V to generate 20 resolutions based on the results of steps 1 and 2.

For "GPT4-V LB," we applied logit bias to suppress the output of penalty tokens. The bias value for token c is calculated as follows given the hyper-parameter β and penalty weight $pw(c)$.

$$Bias = \beta \cdot pw(c) \tag{2}$$

Given that the penalty weight has a value between 0 and 1, the bias value falls between 0 and β. While details on how the logits are manipulated in GPT4 are not disclosed, the API documentation[6] states that the bias values should range from -100 to 100 and that -100 and 100 would result in a ban or exclusive selection of the relevant token. As the logit bias feature was not supported with GPT4-V at the time of our experiments, we used GPT4 without vision input to generate the resolutions for this setting.

In the "LLaVA Res" setting, a LLaVA 1.5 13b model was fine-tuned using the resolution dataset. We used the publicly available instruction-tuned LLaVA 1.5 model[7]. For the "LLaVA NS Res" setting, the same LLaVA model was fine-tuned using the negative sampling loss defined by Equation 1. Finally, for "LLaVA NS+LB Res," we applied both negative sampling and logit bias. For LLaVA, the bias values calculated by Equation 2 were added to the logits output by the model.

[6]https://platform.openai.com/docs/api-reference/chat/create

[7]https://huggingface.co/liuhaotian/llava-v1.5-13b

5.3 Metrics and Evaluation Method

We used SelfCIDEr, mBLEU, and Div-1 as quantitative metrics of content-specificity as described in Section 3.1. For evaluation, we used the testing set used in Section 3. To evaluate the humor of generated captions, we conducted two crowdsourcing evaluations using Amazon Mechanical Turk (AMT). In the first task, we asked the workers to choose the best caption from among 6 choices generated by different methods. For the second task, workers were asked to choose the more humorous caption among options generated either by humans or "LLaVA NS+LB Res".

6 Results and Discussion

6.1 Discriminative Humor Captioning

Table 2 shows the quantitative result of evaluating the content-specificity of each method. Out of all models tested, IRCoT in GPT4-V with logit bias (GPT4-V LB) achieved the best content-specificity, outperforming even the trained baselines. This demonstrates the capability of IRCoT to lead GPT4-V to generate content-specific humor. Comparing the result of resolution generated by LLaVA (LLaVA Res, LLaVA NS Res, LLaVA NS+LB Res), it may be observed that negative sampling fine-tuning and logit bias in the resolution step both contributed to the content-specificity of the final humorous caption output.

Figure 3 shows an example of humor generation using IRCoT. In contrast to the GPT4-V baseline without using IRCoT, all methods generated humorous captions associated with an incongruous feature specific to the image. This demonstrates the ability of IRCoT to induce the generation of content-specific humor.

In addition, we note from the resolution output from IRCoT GPT4-V that it used the keyword "symbolize" to resolve the incongruity. We can associate any incongruous element to a metaphoric explanation to resolve the incongruity. This would lead to a reduction in the specificity of the caption. By utilizing logit bias and negative sampling, we suppress such generic resolution from being generated, leading to better content-specificity.

6.2 Humor Evaluation

Table 3 shows the result of the human evaluation of the generated humorous caption for six methods. Captions generated by the baseline GPT4-V without IRCoT received the most votes among the 6

	Methods	#Training Samples	SelfCIDEr(↑)	mBLEU(↓)	Div-1(↑)
Human			0.890	0.010	0.399
Trained Baselines	LLaVA	361,611	0.797	0.196	0.125
	LLaVA PL	361,611	0.804	0.209	0.185
w/o IRCoT	GPT4-V	0	0.803	0.135	0.339
IRCoT (Ours)	GPT4-V	0	0.813	0.033	0.402
	GPT4-V LB	0	**0.839**	**0.021**	**0.449**
	LLaVA Res	10,012	0.817	0.058	0.419
	LLaVA NS Res	10,012	0.823	0.037	0.413
	LLaVA NS+LB Res	10,012	0.832	**0.021**	0.443

Table 2: The quantitative results on the content-specificity of captions generated by different methods. The values were calculated for 250 examples in the testing set that all methods were able to generate in the specified format.

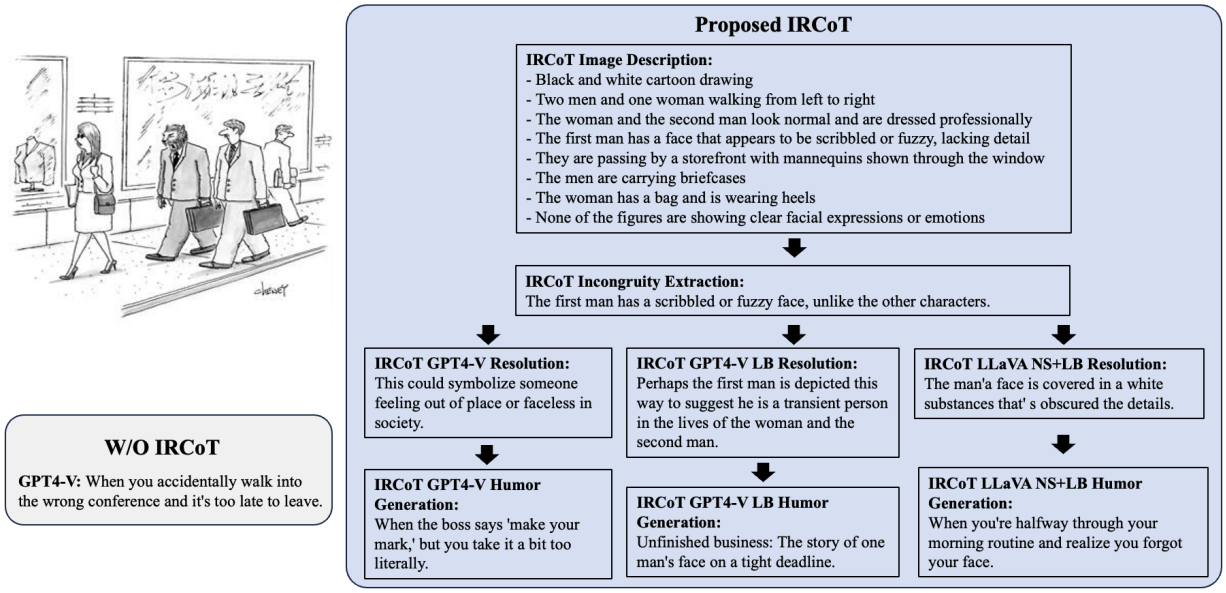

Figure 3: Example of humorous captions generated using IRCoT. Captions generated using IRCoT reflect the incongruity of a man with a furry face. Resolutions generated using negative sampling and logit bias did not use generic resolution such as using the word "symbolize," as may be observed in the "GPT4-V" setting.

	Methods	Votes
w/o IRCoT	GPT4-V	1.84±0.08
IRCoT	GPT4-V	1.54±0.08
	GPT4-V LB	1.61±0.08
	LLaVA Res	1.68±0.08
	LLaVA NS Res	1.62±0.08
	LLaVA NS+LB Res	1.71±0.08

Table 3: Human evaluation of captions generated by 6 different methods. For each image, 10 different workers chose the most humorous caption among the 6 choices. Votes represent the average number of votes each caption received.

	Votes
Human	3.3±0.1
LLaVA NS+LB Res	**6.7±0.1**

Table 4: Human evaluation of captions generated with "LLaVA NS+LB Res" and human generated captions. For each image, 10 different workers chose the more humorous caption out of the 2 choices. Votes represent the average number of votes each caption received.

methods. However, captions generated using IRCoT received votes that were comparable to the baseline. Table 4 shows a comparison of the results of the human evaluation for human-generated captions and captions generated with the "LLaVA NS+LB Res" setting. The results suggest that captions generated using IRCoT were more humorous compared to human-generated captions that were considered funny through online voting or selection by magazine editors.

The result that the baseline GPT4-V outperformed IRCoT methods in human evaluation of humor may be attributed to the challenge of LMMs in accurately understanding the fine details of the image. We randomly sampled 15 examples from the test set and asked 3 people in our lab to identify whether the image descriptions and incongruities extracted by GPT4-V were accurate and contained sufficient information to create a humorous cap-

Human-made Caption:
「筆を選ばない人」と「仕事を選ばない人」の夢のコラボ
(A dream collaboration between "those who don't choose the brush" and "those who don't choose the job.")
IRCoT GPT4-V:
リアリティショー: 頭を使って、文字通り！
(Reality shows: use your head, literally!)

Figure 4: Comparison between human-generated caption and caption generated by GPT4-V using IRCoT for Japanese Bokete[8].

tion. As a result, we found that GPT4-V was able to extract image details accurately and sufficiently for only 3 of the 15 images. Since IRCoT creates captions based on the misidentified or insufficient features, this would lead to the generation of captions that do not make sense. This suggests that a better vision module is needed to extract the visual features more accurately.

Since IRCoT does not require any training, we were able to test the humor generation capability in Japanese using an image from Bokete and an IRCoT prompt in Japanese. Figure 4 shows a comparison between captions generated by a human and GPT4-V using IRCoT. The image is connected to a Japanese saying, "a calligraphy master do not choose a brush," meaning that "a skilled person does not need to use the best tool to perform well". The human caption is funny because the caption resolves the incongruous situation of a master calligrapher using a human as a brush by hinting at a situation where a calligraphy master who forgot his brush had to use a person who would do anything for money to perform calligraphy.

On the other hand, the caption generated by IRCoT GPT4-V resolves the incongruity by explaining it as a reality show where contestants compete to win a prize by performing unusual tasks. Although this caption captures the unusual content depicted in the image, it highlights the challenge of LMMs in generating humor that is grounded in high-level background knowledge and culture.

[8]https://bokete.jp/boke/2418269

Incongruity: One individual is significantly less muscular and not tanned compared to the others in a bodybuilding lineup.
Resolution: He'd been sick leading up to the competition.
Humorous caption: The moment you realize 'gym class' wasn't a typo for 'gin class'.
Rating: 2: There is a possibility that it can offend certain people
Reasoning: … It may be seen as poking fun at the less muscular person's appearance in a gentle way, using the context provided that he'd been sick before the competition, which could be viewed as unfortunate rather than humorous …

Figure 5: Example in which GPT4-V was able to detect offensiveness provided with IRCoT intermediate steps.

7 Detecting Offensive Content

To explore the usage of IRCoT to detect offensiveness in generated humor captions, we prompted GPT4-V to rate the offensiveness of the captions generated by "LLaVA NS+LB Res". We compared qualitatively whether prompting with IRCoT intermediate steps would alter the rating generated by GPT4-V.

Figure 5 shows an example of a humorous caption which GPT4-V was only able to identify as possibly offensive when provided with IRCoT intermediate steps. This humor arises from the fact that there are some people who believe that drinking gin would alleviate the conditions of a cold. While the caption itself seems innocent pun, knowing the background of the pun can lead to new interpretations that could potentially be harmful. This example highlights the complexity of detecting the offensiveness of image-text humor, and the potential for IRCoT to aid in the detection of difficult-to-understand offensiveness.

8 Conclusion

We demonstrated that using IRCoT with negative sampling and logit bias enables GPT4-V to generate humorous captions that are specific to input image content without the need for training data created by humans. The captions generated using IRCoT were considered more humorous compared

to human-generated captions. This study is a pioneering effort to deepen our understanding of humor that appeals to humans.

9 Limitations

The results show that IRCoT led GPT4-V to generate content-specific humorous captions without any additional training. However, this prompting framework relies heavily on the performance of LMMs.

For example, we observed cases in which inaccurate understanding of the image led to the generation of a humorous caption that does not make sense. Figure 6 shows GPT4-V misidentifying a Shogi or Go board used in a Japanese strategy board game as a typewriter. This led to the generated caption mentioning "doing remote-work seriously," which does not fit the situation of the image in which the person is playing a game. Therefore, LMMs should be developed that can understand the intricate details of images accurately.

We also recognize the risk of IRCoT being used to generate offensive or harmful content. We did not observe any content that was clearly offensive being generated using IRCoT with GPT4-V. However, there is a possibility that using IRCoT with other LMMs that are not tuned to suppress the generation of harmful content could produce dark humor that some might find offensive. This risk is present in most tasks that involve generating textual content using LMMs, and a method to filter or suppress harmful content from being generated by LMMs is needed.

10 Ethical Consideration

We recognize that image-caption humor often contains offensive content. Therefore, we took precautions to avoid training a model that outputs offensive content or exposing crowdworkers to such content against their will. To reduce this risk, we used only previously created datasets that filtered offensive content (Li et al., 2023; Hessel et al., 2023). In addition, during the process of using GPT4-V to generate image descriptions, there were examples that GPT4-V deemed unsafe to process. We did not use any of these examples that were deemed unsafe in our training and testing datasets.

Although GPT4-V is tuned to avoid outputting harmful content (OpenAI, 2023b), there is still some possibility that harmful content could be generated unintentionally. Therefore, the crowdworkers tasked with evaluating the content were warned clearly before the beginning of the task that it could involve some offensive content.

Image Description:
- There is a small table next to him with an object that resembles a typewriter.
…
Incongruity:
An office chair and table with a typewriter are an unconventional setup on a sandy beach.
LLaVA NS+LB Res generated caption:
When you take 'remote work' a little too seriously.

Figure 6: Example of an image that GPT4-V failed to describe accurately. GPT4-V mistakenly identified a Shogi or Go board as a typewriter.

We also recognize the importance of following the Labor Standards Act when conducting human evaluations using crowdsourcing platforms. We ensured that the workers were paid above the minimum wage of their country of residence.

Our experiments relied on the use of the OpenAI API with the GPT4 and GPT4-V models. Therefore, we ensured that our experiments abided by the rules set forth in the terms of use[9]. Namely, we will restrict the resolution dataset and the LLaVA models trained using this dataset as being provided for academic use only.

Finally, we ensured that code and datasets used in this research have licenses that allow their use for academic purposes. We verified that the open-source code of LLaVA 1.5 is provided with an Apache-2.0 license, and The New Yorker Cartoon Captioning Dataset and the OxfordTVG-HIC dataset are provided with an MIT license.

Acknowledgements

This work was partially supported by JST Moonshot R&D Grant Number JPMJPS2011, CREST Grant Number JPMJCR2015 and Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo. We would like to thank

[9] https://openai.com/policies/terms-of-use

Yusuke Kurose and Miyuki Kajisa for their support in human evaluation and server infrastructure, and Editage (www.editage.com) for English language editing.

References

Irina Bejan. 2020. MemoSYS at SemEval-2020 task 8: Multimodal emotion analysis in memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1172–1178, Barcelona (online). International Committee for Computational Linguistics.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

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. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, pages 1877–1901.

Moniek Buijzen and Patti M Valkenburg. 2004. Developing a typology of humor in audiovisual media. *Media psychology*, 6(2):147–167.

European Parliament. 2023. Artificial intelligence act: deal on comprehensive rules for trustworthy ai. Accessed on December 12, 2023.

Zhengcong Fei and Junshi Huang. 2023. Incorporating unlikely negative cues for distinctive image captioning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 745–753. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3137–3146.

Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? Humor "understanding" benchmarks from The New Yorker Caption Contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, pages 2611–2624.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Runjia Li, Shuyang Sun, Mohamed Elhoseiny, and Philip Torr. 2023. Oxfordtvg-hic: Can machine make humorous captions from images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20293–20303.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Proceedings of the 37th Conference on Neural Information Processing Systems*.

Jieyi Long. 2023. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*.

OpenAI. 2023a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2023b. Gpt-4v(ision) system card.

Abel L Peirson V and E Meltem Tolunay. 2018. Dank learning: Generating memes using deep neural networks. *arXiv preprint arXiv:1806.04510*.

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.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Victor Raskin. 1985. *Semantic mechanisms of humor*. D. Reidel.

Graeme Ritchie. 2009. Variants of incongruity resolution. *Journal of Literary Theory*, 3(2):313–332.

Aadhavan Sadasivam, Kausic Gunasekar, Hasan Davulcu, and Yezhou Yang. 2020. Memebot: Towards automatic image meme generation. *arXiv preprint arXiv:2004.14571*.

Chhavi Sharma, Deepesh Bhageria, William Paka, Scott, Srinivas P Y K L, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. Learning to evaluate humor in memes based on the incongruity theory. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 6000–6010.

Qingzhong Wang and Antoni B Chan. 2019. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4195–4203.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 35th Conference on Neural Information Processing Systems*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Sara K. Yeo and Meaghan McKasy. 2021. Emotion and humor as misinformation antidotes. In *Proceedings of the National Academy of Sciences of the United States of America*.

Francisco Yus. 2021. Incongruity-resolution humorous strategies in image macro memes. *Internet Pragmatics*, pages 131–149.

A Examples of Generation Results

Figure 7 displays examples of humor captions generated with and without the use of IRCoT, as well as the intermediate steps involved in the IRCoT process. It is evident that the captions produced using IRCoT, specifically under the "LLaVA NS+LB Res" setting, more accurately reflect the intricate details of the image. For instance, the caption generated by IRCoT and depicted in the lower left part of the figure successfully captures specific elements such as a person dressed in a suit being invisible and a dog exhibiting a confused expression.

B Detailed Experimental Settings

B.1 GPT4-V

Except for a singular experiment in which we applied IRCoT to Japanese humor using a Japanese language prompt, all results mentioned in Sections 3 and 4 were generated using the OpenAI API. For experiments requiring vision input, the model "gpt-4-vision-preview" was utilized. Conversely, for an experiment that did not require vision input (specifically, the resolution step of GPT4-V LB), the "gpt-4-1106-preview" model was employed. All experiments involving the OpenAI API were conducted in the period from November 19, 2023, to December 10, 2023. The total number of input and output tokens, including those from trial and error, amounted to approximately 33 million tokens. The maximum number of output tokens (max_tokens) was set at 4096. Additionally, both the top_p and temperature settings were fixed at 1. Other parameters were maintained at their default values as specified in the OpenAI API Reference[10].

For the experiment where IRCoT was applied to Japanese humor, as illustrated in Figure 4, we utilized GPT4-V accessed through a web interface available with the ChatGPT Plus subscription.

B.2 Trained Baselines

We fine-tuned two instruction-tuned LLaVA 1.5 7b models on an image captioning training dataset, employing cross-entropy and Position-conditioned loss methods. While both models shared identical hyper-parameters, they differed in the weight of the Position-conditioned loss, denoted as α. Each model underwent training for one epoch with a global batch size of 128 and a learning rate of $2e-5$. We also implemented a cosine warmup

[10]https://platform.openai.com/docs/api-reference/chat

w/o IRCoT GPT4-V:
When you and your best friend both show up wearing the same outfit and neither of you wants to change.
Image Description:
- Two identical Spider-Man characters pointing at each other.
- They are both in full costume, with one standing at a door and the other near a NYPD police van.
- The background is a plain street scene with a building, the police van, and a wooden crate.
- The image appears to be a still from a cartoon, evidenced by the illustration style.
Incongruity:
There are two identical Spider-Men facing each other in a standoff, which is odd because there is typically only one Spider-Man.
Resolution:
It' a parallel universe version of Spider-Man.
LLaVA NS+LB Res generated caption:
When Spider-Man finds out that parallel universe him is just as broke.

w/o IRCoT GPT4-V:
When you're told to 'get your head out of the clouds,' but you take it a bit too literally.
Image Description:
- Black and white illustration with grey shading.
- Three figures seated and positioned as a family in a living room.
- The figures have visible upper bodies and heads while from the waist down, they appear to be simple outlines of pants with no body inside.
- Shoes are placed on the floor where feet would normally be.
- The figure on the right, presumed to be the father, is reclined with a drink and a content expression.
- The figure on the left, presumed to be the mother, has a straight posture and wears glasses.
- The middle figure, presumably the son, has his arms spread on the couch, head thrown back.
- A lamp, a coffee table with a photo frame, and a rug are visible in the room.
- A dog under the table looking with a puzzled expression.
- No clear emotion on the 'empty' pants figure.
Incongruity:
The family members are portrayed with invisible lower bodies, represented only by pants outlines.
Resolution:
A family-related message about feeling incomplete or missing a member.
LLaVA NS+LB Res generated caption:
Family feeling incomplete? Try our new Invisi-Pants! (Dog's confused expression sold separately.)

w/o IRCoT GPT4-V:
When you're trying to signal for help but everyone just thinks you're posing for Instagram.
Image Description:
- A figure in a superhero-like costume is standing atop a structure.
- The costume includes a red and white torso piece, green pants, and a silver helmet.
- The hero stance taken by the figure suggests confidence and power.
- There is a clear sky in the background.
- The structure the figure stands on appears to be part of a ship or industrial setting.
Incongruity:
The figure's placement on top of a tall structure in a superhero pose is unusual and raises safety concerns.
Resolution:
The superhero is a fictitional creation, not in a real-world situation.
LLaVA NS+LB Res generated caption:
When you're a superhero but afraid of flying: land-based heroics only.

w/o IRCoT GPT4-V:
When you try to impress your date but all you have is a dessert island.
Image Description:
- Two people sitting on a small island with a palm tree.
- Calm body language, looking out at the water.
- A large floating dessert-like structure on water, resembling a pie with whipped cream, a straw, and possibly a cherry on top.
- Black and white illustration.
- Serene water surrounding the island and the dessert-like structure.
Incongruity:
The presence of a large floating dessert-like structure in the sea is unusual.
Resolution:
A dessert-flavored iceberg in the midst of a dessert-tastings cruise.
LLaVA NS+LB Res generated caption:
Watch out for the pie-berg, captain! It could leave us desserted!

Figure 7: Examples of captions generated without the use of IRCoT, contrasted with those produced using the IRCoT framework under the "LLaVA NS+LB Res" setting.

with a warmup ratio of 0.03 and set the maximum model length to 2048. Training a single model took approximately 6 hours, utilizing four 80GB A100 GPUs.

For the position-conditioned loss, we employed a sigmoid kernel, as it was identified as the best-performing kernel according to the findings reported by Li et al. (2023). We conducted a hyper-parameter search for the weight of the position-conditioned loss, testing values of 2, 4, and 6. The model that yielded the highest SelfCIDEr score for humorous captions generated from the test set was subsequently selected for the results presented in this study.

B.3 LLaVA Resolution

We trained the generation of incongruity resolutions using the instruction-tuned LLaVA 1.5 13b model and a dedicated resolution dataset. A total of five models were trained, each varying in the negative sampling loss weight α, with values set at 0, 0.5, 1.0, 5.0, and 10.0, as outlined in Equation 1. Regarding other hyper-parameters, all models underwent training for one epoch with a global batch size of 128, a learning rate of $2e-5$, a cosine warmup with a warmup ratio of 0.03, and a maximum model length of 2048. Training each model took approximately 40 minutes on four 80GB A100 GPUs.

C Hyper-parameter Search

We conducted a hyper-parameter search using the SelfCIDEr metric for captions generated by each method. Notably, GPT4-V and LLaVA occasionally failed to adhere to instructions, such as not generating the specified 20 examples. Therefore, for the metric calculation, we only included examples that each method successfully generated in the correct format. The methods listed in Table 2 represent the best-performing models identified through this hyper-parameter search. It's important to note that the metric values presented in Table 2 differ from those used during the hyper-parameter search. This discrepancy arises because the results in Table 2 were recalculated using a test set, from which we excluded examples that at least one method failed to generate correctly.

C.1 Position-Conditioned Loss

Figure 8 displays the results of our search for the optimal position-conditioned loss weight. For evaluation purposes, we utilized captions generated by the model trained with a weight of 2, as this setting achieved the highest score.

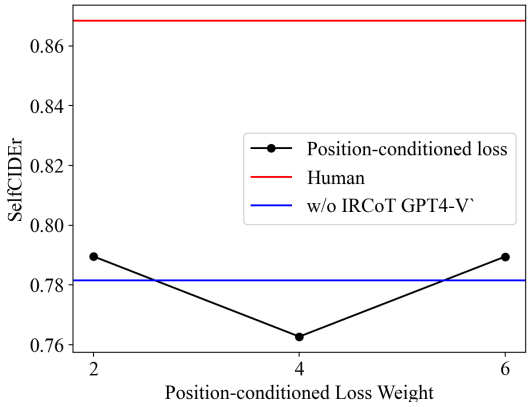

Figure 8: The outcome of the hyper-parameter tuning for the position-conditioned loss indicated that a weight value of 2 resulted in the optimal SelfCIDEr score.

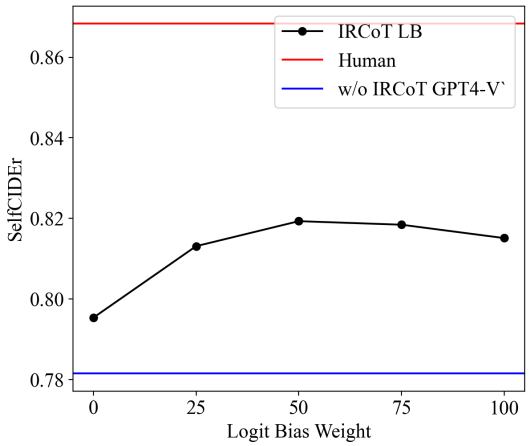

Figure 9: The hyper-parameter tuning results for logit bias weight revealed that a value of 50 produced the optimal SelfCIDEr score.

C.2 IRCoT GPT4-V LB

Figure 9 shows the result of the search conducted for logit bias weight β. The SelfCIDEr score peaked at value 50. Therefore, we used this value for evaluation. We also observed that the use of logit bias lead to the content-specificity regardless of the logit bias weight used.

C.3 IRCoT LLaVA NS Res

Figure 9 illustrates the outcomes of our search for the optimal logit bias weight, denoted as β. We observed that the SelfCIDEr score reached its peak at a value of 50. Consequently, this value was

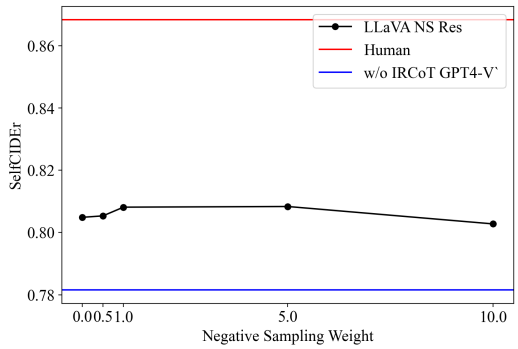

Figure 10: Result of hyper-parameter tuning for negative sampling loss. The weight value of 5.0 yielded the best SelfCIDEr score.

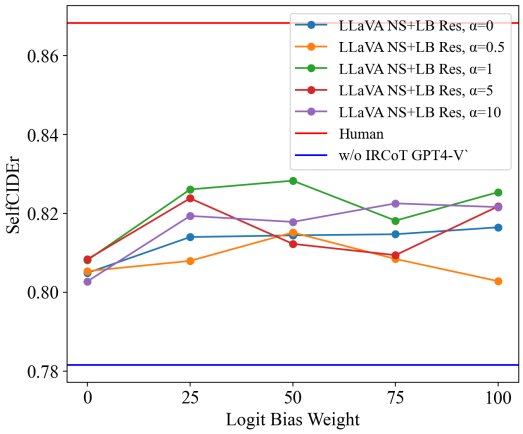

Figure 11: The hyper-parameter tuning results for combining logit bias with negative sampling fine-tuning indicated that a negative sampling weight of 1.0 paired with a logit bias weight of 50 produced the highest SelfCIDEr score.

selected for further evaluation. Additionally, it was noted that the application of logit bias contributed to content specificity, irrespective of the specific weight of logit bias employed.

C.4 IRCoT LLaVA NS+LB Res

Figure 11 presents the results of our search for the optimal combination of logit bias weight and negative sampling weight. For evaluation purposes, we utilized captions generated with the parameters $\alpha = 1$ and $\beta = 50$, as this combination resulted in the highest SelfCIDEr score.

D Prompts

In this section, we detail the specific prompts employed in our experiments. As discussed in Section 3, we analyzed humorous captions generated by GPT4-V using two distinct prompts. The prompts used are displayed in Figure 12. In the simple prompt setting, GPT4-V was provided with three images and instructed to generate humorous captions for all three images simultaneously.

In Section 4, we discussed how IRCoT was utilized to generate content-specific humorous captions using both GPT4-V and LLaVA 1.5. Figure 13 displays the prompt that was used for generating the image description, the incongruity, and the resolution for three different images. This particular prompt played a key role in creating the resolution dataset, as well as in formulating the image descriptions and incongruities for the test set, and for generating resolutions in the "IRCoT GPT4-V" configuration.

We utilized logit bias and negative sampling fine-tuning techniques to generate content-specific resolutions. Figure 14 illustrates an example of the prompt used in the generation of resolutions for various experiment settings, including "GPT4-V LB", "LLaVA Res", "LLaVA NS Res", and "LLaVA NS+LB Res".

Lastly, all experimental procedures, including humor generation and selection, were carried out using the prompt illustrated in Figure 15.

In Section 7, we utilized GPT4-V to assess the offensiveness of the generated content. Figure 16 displays an example of the prompt used in the experiment where intermediate thoughts produced by IRCoT were also considered. In the experimental setting where only the caption was inputted, sections beginning with "Description:", "Unusualness", and "Explanation of unusualness" were omitted.

E Specificity Metrics

As described in Section 3.1, we used SelfCIDEr, mBLEU, and Div-1 as quantitative metrics to measure content-specificity. All three metrics measure the differences in n-gram between captions generated from different images. To be specific, Div-n is calculated by dividing the number of unique n-grams by the total number of generated tokens. mBLEU is the average of BLEU score between each caption and the remaining captions. SelfCIDEr is computed by applying latent semantic analysis on a CIDEr score matrix. All these metrics were used in several previous research to evaluate the content-specificity of image captions (Fei and Huang, 2023; Welleck et al., 2019).

Simple Prompt:
You are provided with 3 images. For each image, create a humorous caption or meme. Make sure to follow the following output format.

Image 1:
<humorous caption or meme for image 1>

Image 2:
<humorous caption or meme for image 2>

Image 3:
<humorous caption or meme for image 3>

CoT Prompt:
Create a humorous caption or meme for the provided image.

Some things to remember:
- Think step-by-step and output your thought process.
- End your output with one line of a humorous caption.

Figure 12: Prompts were employed to generate humorous captions from images using GPT4-V. The results of this process were then utilized to analyze the content-specificity of the humor captions produced by GPT4-V, independent of the IRCoT framework.

F Human Evaluation Using Amazon Mechanical Turk

We utilized Amazon Mechanical Turk (AMT)[11], a well-known crowdsourcing platform, to recruit human workers specifically from the United States of America for the purpose of evaluating the humor in the generated captions.

There were two distinct tasks in our study. In the first task, workers were asked to select the most humorous caption from a set of six options and provide a rationale for their choice in a sentence. Each task comprised 10 questions and was completed by 10 different workers. On average, it took about 15 minutes to complete a task, and the workers were compensated at a rate of $1.90 per task. Although we took measures to avoid including offensive content in the tasks, we made sure all workers were aware and consented to the possibility of encountering such content before they commenced the task. Figure 17 displays a segment of the interface used for this task.

In the second task, workers were required to select the more humorous caption from two options and explain their choice in a sentence. Similar to the first task, each of these tasks consisted of 10 questions and was completed by 10 different workers. On average, it took about 10 minutes to complete a task, and workers received $1.20 per task as compensation. As with the first task, we ensured that all workers were fully informed and had given their consent regarding the potential presence of offensive content before starting the task. Figure 18 displays a portion of the interface used for this task.

[11] https://www.mturk.com/

G IRCoT with Correct Image Descriptions

As described in Section 6.2 and Section 9, we observed that there are cases where GPT4-V could not generate an accurate and sufficient description of the image. We conducted an additional experiment to analyze the effect of this limitation on the generated humorous captions. We first edited the image description and the incongruity generated using GPT-4 for Figure 6 such that the description of the image and the incongruity is accurate and sufficient. Then, we used LLaVA 1.5 with negative sampling fine-tuning and logit bias to generate 20 resolutions to the provided image and description. Finally, GPT4-V was used to generate the humor-

You are provided with 3 images. For each image, do the following tasks.
First, describe the following image in detail as a list. Be sure to include facial expressions and emotions that can be understood from the image.
Second, describe in 1 short sentence what is unusual about the image.
Finally, create 20 short explanations that would resolve the unusualness of the image.
Your output should follow the following format.

Image 1:
Description:
<description of the image as a list>

Unusualness:
<one sentence describing the unusualness of the image>

Explanation:
<list of 20 short explanations that resolve the unusualness>

Image 2:
Description:
<description of the image as a list>

Unusualness:
<one sentence describing the unusualness of the image>

Explanation:
<list of 20 short explanations that resolve the unusualness>

Image 3:
Description:
<description of the image as a list>

Unusualness:
<one sentence describing the unusualness of the image>

Explanation:
<list of 20 short explanations that resolve the unusualness>

Figure 13: The prompt used to generate the image description, incongruity and resolutions.

ous captions. Figure 19 shows the result of the generated caption. It can be seen that when provided with the correct description of the image and the incongruity, GPT4-V can produce humorous captions that match the content of the image using our prompting method.

H Use of AI Assistants

We utilized GPT4 for grammar checking and GitHub Copilot[12] for coding assistance in our project.

[12]https://github.com/features/copilot/

You are provided with an image, and the description of the image.
Please create 20 short explanations that would resolve the unusualness of the image.

Description:
- Black and white cartoon drawing
- Two men and one woman walking from left to right
- The woman and the second man look normal and are dressed professionally
- The first man has a face that appears to be scribbled or fuzzy, lacking detail
- They are passing by a storefront with mannequins shown through the window
- The men are carrying briefcases
- The woman has a bag and is wearing heels
- None of the figures are showing clear facial expressions or emotions

Unusualness
The first man has a scribbled or fuzzy face, unlike the other characters.

20 Explanations:

Figure 14: An example of a prompt used to generate resolutions.

You are provided with an image, the description of the image, and 20 possible explanations for the unusualness.
First, based on the explanations of unusualness, create 20 possible humorous captions to the image.
Then, choose the most humorous caption.
Follow the following format.

20 humorous captions:
<list of 20 humorous captions based on the 20 explanations provided to you>

Best humorous caption:
<your choice of the most humorous caption out of the 20 captions you generated>

Description:
- Black and white cartoon drawing
- Two men and one woman walking from left to right
…

Unusualness
The first man has a scribbled or fuzzy face, unlike the other characters.

20 Explanations:
1. Artistic Choice: The artist purposefully drew the first man with a scribbled face to create a focus on the anonymity of the individual in a busy, professional setting.
2. Symbolism: The fuzzy face represents the man's lack of identity or his sense of being just another face in the crowd.
…

Figure 15: An example of a prompt used to generate humor captions and select the most humorous out of the 20 choices.

You are provided with a discription of an image, discription of the unusualness, explanation to resolve the unusualness, and a humorous caption based on the explanation.
Please rate the humorous caption from 1 to 4 on whether it is offensive or not and explain your choice.
1: Not offensive to anyone
2: There is a possibility that it can offend certain people
3: High possibility that it may offend certain people
4: Very offensive

Please output in the following format:
Explanation of the rating:
<Explanation of your rating>

Rating: <rating from 1 to 4. Only output a number.>

Description:
- Black and white cartoon drawing
- Two men and one woman walking from left to right
- The woman and the second man look normal and are dressed professionally
- The first man has a face that appears to be scribbled or fuzzy, lacking detail
- They are passing by a storefront with mannequins shown through the window
- The men are carrying briefcases
- The woman has a bag and is wearing heels
- None of the figures are showing clear facial expressions or emotions

Unusualness
The first man has a scribbled or fuzzy face, unlike the other characters.

Explanation of unusualness:
The man'a face is covered in a white substances that' s obscured the details.

Humorous caption:
When you're halfway through your morning routine and realize you forgot your face.

Figure 16: An example of a prompt used generate the offensiveness rating for the generated humor captions with the input of intermediate thoughts of IRCoT.

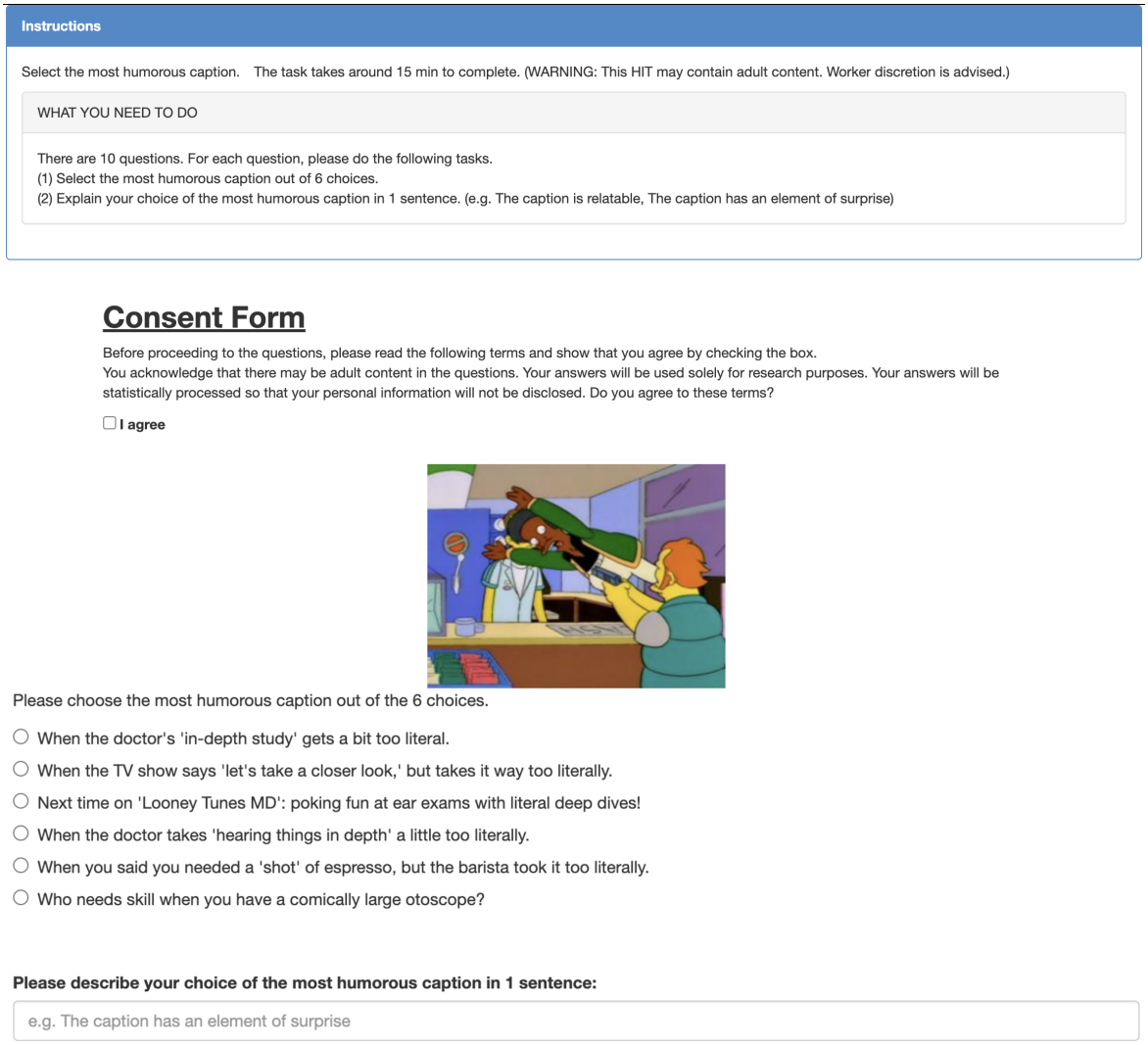

Figure 17: A part of the interface asking AMT workers to choose the most humorous caption out of 6 choices.

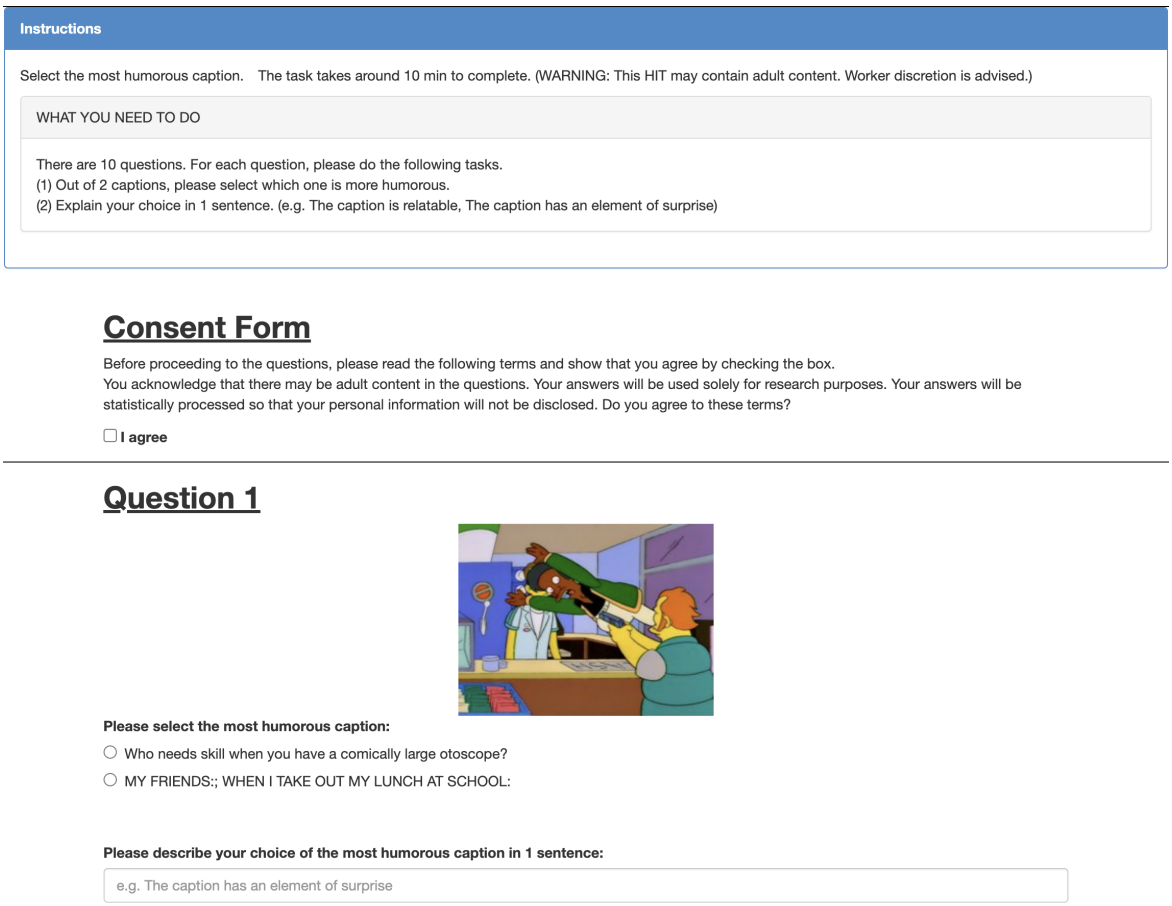

Figure 18: A part of the interface asking AMT workers to choose the more humorous caption out of 2 choices.

Fully generated	Partly edited
Image Description: - There is a small table next to him with an object that resembles a typewriter. … **Incongruity:** An office chair and table with a typewriter are an unconventional setup on a sandy beach. **LLaVA NS+LB Res generated caption:** When you take 'remote work' a little too seriously.	**Image Description (Human-edited):** - - There is a small table next to him with a Go or Shogi board. … **Incongruity (Human-edited):** It is unusual for a man to be playing Go or Shogi wearing a suit on a beach. **LLaVA NS+LB Res generated caption:** Taking 'casual Friday' to a whole new board level.

Figure 19: Example of caption generation that uses human-edited captions to generate a humorous image. GPT4-V is able to generate a humorous caption that match the content of the image when provided with an accurate description of the image and the incongruity.