MemeInterpret: Towards an All-in-One Dataset for Meme Understanding

Jeongsik Park[1]*, **Khoi P. N. Nguyen**[2]*, **Jihyung Park**[3], **Minseok Kim**[2], **Jaeheon Lee**[2], **Jae Won Choi**[1], **Kalyani Ganta**[4], **Phalgun Ashrit Kasu**[2], **Rohan Sarakinti**[3], **Sanjana Vipperla**[2], **Sai Sathanapalli**[2], **Nishan Vaghani**[2], and **Vincent Ng**[2]
[1]University of Southern California [2]University of Texas at Dallas
[3]University of Texas at Austin [4]Virginia Tech
jeongsik@usc.edu khoi.nguyen6@utdallas.edu vince@hlt.utdallas.edu

Abstract

Meme captioning, the task of generating a sentence that describes the meaning of a meme, is both challenging and important in advancing Computational Meme Understanding (CMU). However, existing research has not explored its decomposition into subtasks or its connections to other CMU tasks. To address this gap, we introduce MemeInterpret, a meme corpus containing meme captions together with corresponding surface messages and relevant background knowledge. Strategically built upon the Facebook Hateful Memes dataset, MemeInterpret is the last piece in a set of corpora that unifies three major categories of CMU tasks for the first time. Extensive experiments on MemeInterpret and connected datasets suggest strong relationships between meme captioning, its two proposed subtasks, and the other two key categories of CMU tasks: classification and explanation. To stimulate further research on CMU, we make our dataset publicly available at https://github.com/npnkhoi/MemeInterpret.[1]

1 Introduction

Memes, which are user-created combinations of images overlaid with text, have become a prevalent means of online communication (Joshi et al., 2024). They are created with a variety of purposes: while some memes are simply used to express personal opinions, other memes can be malicious, such as inciting hatred or spreading manipulative propaganda. As such, recent years have seen increasing interest within NLP in the emerging area of Computational Meme Understanding (CMU), a term that we coined to refer to research on the automated comprehension of memes (Nguyen and Ng, 2024).

While numerous CMU tasks have been proposed in the past few years, researchers have essentially adopted the same methodology for tackling each newly-proposed task: (1) annotate a corpus of memes with task-specific labels and (2) fine-tune a Vision-Language Model (VLM) on the resulting corpus. While there is nothing inherently wrong with this methodology, it is not necessarily healthy for the long-term development of CMU research: as a field, existing CMU tasks are being tackled as if they have nothing to do with each other.

(a) (b)

Figure 1: Example memes from Kiela et al. (2020) (a) and Sharma et al. (2020) (b).

In light of this concern, we believe we should start thinking about developing task-agnostic rather than task-specific representations of memes. The question, then, is: what task-agnostic representation can be shared by and therefore benefit a range of CMU tasks? As many meme-based NLP tasks require an understanding of the *meaning* of a meme (i.e., what the meme author tries to convey) rather than its *form* (i.e., how the meaning is conveyed through the visuals and text), a good starting point would be to experiment with using the meaning of a meme as its task-agnostic representation.[2]

For this reason, we believe that it is important to examine an under-studied but important category of CMU tasks, *meme interpretation*, which is referred to as *meme captioning* by Hwang and

*These authors contributed equally to this work.

[1]For illustration purposes, we show in this paper memes from MemeInterpret, some of which could be offensive.

[2]We are by no means claiming that this representation would be useful for all CMU tasks. For instance, if the goal is to identify the persuasion strategies used in a meme, then the form of the meme would matter. Note also that this is an *unstructured* representation, as we express meaning in the form of natural language. We leave the development of a structured meaning representation to future work.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16073–16087
November 4-9, 2025

Shwartz (2023) and *intent description generation* by Park et al. (2024). Both tasks concern *generating* a description that captures the *meaning* of a meme. Given the meme in Figure 1a, the meme caption would be *"this meme maliciously makes fun of the trauma of the Vietnam War veterans"*.

Meme captioning could benefit a range of CMU tasks. For example, given the aforementioned meme caption, a Hateful Meme Detection (HMD) system can easily determine that this meme is hateful. We emphasize, however, that the significance of meme captions lies in the fact that memes that have different forms but convey the same meaning are being mapped to the same meme caption. This helps reduce data sparsity and remove information irrelevant to the meaning. Therefore, when training task-specific models for downstream CMU tasks, a model that takes meme captions (rather than the original memes) as input can potentially be fine-tuned on a smaller amount of task-specific data to achieve a given level of performance.

Meme captioning, however, is challenging for at least three reasons. First, for many memes, especially those that are malicious (e.g., the author's intent is to manipulate public opinion), the meaning of the meme is typically not explicitly stated and therefore can only be inferred.

Second, meme captioning relies on *background knowledge* (BK), which by definition is not explicitly stated in the meme either. Take Figure 1a as an example. To properly understand the terms "popcorn", "senior", and "Vietnam veteran" and their relationships in the meme, one needs to possess the BK that (1) *"the Vietnam war was a highly traumatic period for US soldiers who are now old veterans mostly living in senior centers"*, and (2) *"the sound of popcorn popping in the microwave can remind them of the gunfire sound during the war"*, but (3) *"generally, that sound is not to be scared of because there is no apparent danger"*. While large language models (LLMs) possess lots of knowledge, what is challenging is the generation of the BK *relevant* to a given meme.

Third, even if a model can identify relevant BK, a successful meme captioner needs to address another challenging task that we refer to as *surface message* (SM) generation, which involves generating a sentence to (1) capture the details of the meme that are crucial for its interpretation by (2) combining the meme's image and text in a coherent manner. Returning to Figure 1a, a good SM for this meme is *"It is an image of a 70-year-old Caucasian man sitting in front of a window with a stunned face, wearing a hat that says 'VIETNAM VETERAN'. The author describes the image as 'me: puts bag of popcorn into the microwave. every one else at the senior center'."* This SM captures the information on the image that enables the correct interpretation of the meme (e.g., the text "VIETNAM VETERAN" could have been easily ignored by a typical image captioner). Furthermore, it merges the information from the text and the image in a coherent manner by resolving "me" to the author.[3]

To advance research in meme captioning, we propose `MemeInterpret`, a meme corpus in which each meme is manually annotated with its meme caption and, given the aforementioned challenges associated with meme captioning, its surface message and the relevant background knowledge. In previous work, each meme in `MemeInterpret` is already labeled with whether they are hateful or not (a *categorization* task), and a subset of the hateful memes is additionally labeled with an explanation of why it is hateful (an *explanation* task). Therefore, `MemeInterpret` is the *first* meme corpus that bridges together three major *categories* of CMU tasks, namely categorization, explanation, and interpretation (see Section 2). Our empirical studies showed that surface message and background knowledge annotations significantly improved the performance on meme captioning. Moreover, stronger meme captioning systems can advance the performance in meme classification and explanation. Given these results, we envision that `MemeInterpret` can spark a new avenue of research in CMU that involves studying how different categories of tasks interact with and possibly benefit from each other.

2 Related Work

In this section, we review three categories of tasks in CMU. For a more comprehensive overview of CMU research, see Nguyen and Ng (2024).

Interpretation Interpretation, which is what we focus on, involves generating a description that

[3]A surface message is different from an image caption. The image caption for the meme in Figure 1a could be *"It is an image of a man with a long beard sitting in front of a window. He is wearing a dark-color hat with a jacket. The scene outside the window is blurry."* This description fails to identify crucial information on the image that enables the correct understanding of the meme, such as the text saying "VIETNAM VETERAN", which suggests that the man was a veteran of the Vietnam War, as well as the emotional expression of the man via his face and hands, which suggests fear and trauma. Also, the caption ignores the text in the meme.

Dataset	Size	Annotation Type	Annotation Quality	Topics
HatRed	3228	Hatefulness explanation	Two-stage Collect-And-Judge procedure	Hateful memes from the Facebook Hateful Memes dataset.
ExHVV	4680	Role of entities explanation	Two-stage Collect-And-Judge procedure	COVID-19 and US Politics. All memes are harmful.
MemeCap	6387	Meme caption	No explicit quality control, relied on prior performance of MTurkers	Only non-offensive memes from Reddit.
MemeIntent	950	Intent description	Two-stage Collect-and-Edit procedure	Politics, healthcare, and gender equality on Facebook
MemeInterpret	6810	Surface message, background knowledge, meme caption	Three-stage Collect-Edit-Judge procedure	Both hateful and non-hateful memes from Facebook Hateful Memes dataset

Table 1: Comparison between MemeInterpret and other free-text annotation datasets on memes.

captures the meaning of a meme. Research on meme interpretation is in its infancy: so far it has only been studied by Hwang and Shwartz (2023) and Park et al. (2024). As noted above, Hwang and Shwartz proposed the meme captioning task and created MemeCap, a dataset where each meme is annotated with its meme caption. Park et al. proposed the intent description generation task, which is essentially the meme captioning task, and created MemeIntent. Nevertheless, these authors failed to realize the potential of meme captions as a task-agnostic representation that can benefit a range of CMU tasks. While MemeIntent, like MemeInterpret, is composed of both hateful and non-hateful memes, MemeCap excludes hateful memes, which is a key weakness given the important role played by hateful memes in CMU (Kiela et al., 2020), and does not annotate surface messages and background knowledge.

Explanation Explanation tasks involve *generating* a description of *why* a meme should be assigned a particular *label* (e.g., sarcastic). For instance, if our task involves explaining why the meme in Figure 1b is *sarcastic*, the explanation would be *"languages used for people who fear long words are supposed to be short, but the medical name for the fear is actually long."* So far, only two meme corpora have been produced for explanation tasks, namely, HatRed (Hee et al., 2023) and ExHVV (Sharma et al., 2023).

Table 1 compares MemeInterpret with the above corpora, which are the *only* meme corpora with free-text annotations, along four dimensions. As can be seen, MemeInterpret (1) is the largest meme corpus containing free-text annotations; (2) supports multiple annotation types rather than just one type of annotations; (3) employs a stricter annotation protocol, namely a three-stage Collect-Edit-Judge procedure (see Section 3.2); and (4) contains both the hateful and non-hateful memes occurring in the wild, covering a wider variety of memes than existing corpora. As such, MemeInterpret contributes to the set of *much-needed* corpora regarding free-text annotations for CMU research.

Caregorization Categorization tasks involve *classifying* memes along various dimensions. These tasks can be broadly divided into two categories. The first group is composed of tasks that involve detecting malignity in memes, including offensiveness (Suryawanshi et al., 2020a), trolls (Suryawanshi et al., 2020b), hate (Kiela et al., 2020), antisemitism (Chandra et al., 2021), harm (Pramanick et al., 2021a,b), and misogyny (Fersini et al., 2022). The second group contains tasks that categorize memes along other dimensions such as types of persuasion techniques (Dimitrov et al., 2021), types of figurative language (e.g., irony, allusion, irony, sarcasm, and contrast) (Liu et al., 2022), entities' roles (e.g., hero, villain, or victim) (Sharma et al., 2022), emotions (e.g., humor) (Sharma et al., 2020), and attacked target groups (e.g., religion, race, sex, nationality, and disability) (Mathias et al., 2021).

3 The MemeInterpret Dataset

3.1 Meme Source

To assemble MemeInterpret, we sample the memes from the Facebook Hateful Meme dataset (FHM) (Kiela et al., 2020). Specifically, when assembling MemeInterpret, we inherit the entire training set and the test "seen" set from FHM, which has 8500 and 1000 memes, respectively. 42% of the memes in MemeInterpret are hateful.

We chose FHM for two reasons. First, and perhaps most importantly, each meme in FHM was already labeled as hateful (Kiela et al., 2020) and a subset of the hateful memes were annotated with

explanations of why the memes were hateful (Hee et al., 2023). Therefore, our additional annotations on the memes sampled from FHM will enable us to study how the three categories of CMU tasks — namely interpretation (meme captioning), explanation (hatefulness explanation), and categorization (hateful meme detection) — interact with and possibly benefit from each other. Second, the memes cover a variety of topics encountered in everyday life, including politics, race and ethnicity, sex and relationships, food and eating habits, animals and pets, historical events, social issues, religion and beliefs, lifestyle and daily life, stereotypes and prejudices, as well as hobbies and interests.

3.2 Annotation Procedure

Recall that in addition to the meme caption (MC), we annotate each meme in `MemeInterpret` with two kinds of supporting annotations, surface message (SM) and background knowledge (BK) (Figure 2). Below we describe our annotation procedure.

According to Wiegreffe and Marasovic (2021), the currently most advanced procedures for the form of our annotations — the *free-text* form — are Collect-And-Judge and Collect-And-Edit. In both approaches, "collect" means the initial round of free-text annotation. After that, one may "judge" the annotations by giving ratings or "edit" the annotations to make corrections and improve the annotations. Judging allows detecting poorly annotated instances and showing measurements of annotation quality while editing increases linguistic diversity by allowing multiple annotators per instance.

To combine the strengths of both approaches, we employed a three-stage "Collect-Edit-Judge" procedure as follows:

Stage 1: Collect We asked four annotators who are native English speakers[4] to annotate the three fields (i.e., SM, BK, and MC) according to our annotation guidelines (see Appendix B), and indicate if they are confident with their understanding. Annotators have access to the entire Web for reference and were encouraged to provide multiple BK entries and MCs for a meme. If an annotator did not understand a meme or believed they had a bias (e.g., political or religious bias), they were instructed to skip the meme, and this instance would be reassigned to another annotator.

[4]Details on annotator recruitment and training can be found in Appdendix A.

Stage 2: Edit To increase linguistic diversity and reduce annotator bias, we implemented an "edit" stage. Given the annotation results from Stage 1, five new annotators conducted edits on the three fields and indicated whether they agreed with the Stage 1 annotations. These editors could select one of four options during their editing process: *Agree* (the editor semantically agrees with the annotation; only edits for tone or grammar errors are needed), *Disagree* (the editor semantically disagrees with the annotation; edits are necessary), *Add* (the annotation is insufficient; must add more information), or *Cut* (the annotation is too verbose; must shorten).

Editors were allowed to skip memes they did not understand. Additionally, if the Stage-1 annotator indicated a lack of confidence and the editor also chose to skip, the meme would be discarded due to a lack of clear meaning and intent.

Stage 3: Judge Four editors themselves served as judges, with each instance assigned to two judges. To avoid bias, no one evaluated their own annotation. A random 10%-subset of the training set and 100% of the test set were judged.

Consistent with related work in free-text annotation on memes (Hee et al., 2023; Sharma et al., 2023; Park et al., 2024), SMs and MCs were evaluated on two metrics: *Textual Completeness* (i.e., whether the text has complete English writing with good grammar) and *Correctness* (i.e., whether the text is semantically correct). Meanwhile, BK was evaluated on four metrics: *Textual Completeness*, *Factuality* (i.e., whether it is factually correct), *Relevance* (i.e., whether it is relevant to inferring the meme caption), and *Sufficiency* (i.e., taken together, whether it covers all the knowledge needed to properly infer the meme caption). All metrics are evaluated on the 5-point Likert scale (Likert, 1932).

3.3 Final Dataset

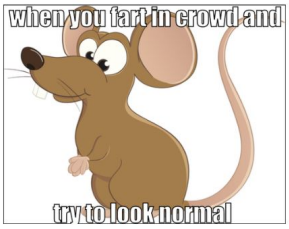
After filtering (to ensure high annotator confidence), we had 5,810 examples in the training set and 1,000 examples in the test set. In addition, 1,927, 509, and 447 memes in the training, development, and test sets contain more than two BKs, and 31, 7, and 11 contain more than two MCs, respectively.

As we used the Collect-Edit-Judge approach, inter-rater agreement is no longer applicable. However, to provide a rough idea of 'agreement', we measured the modification rates during the editing process (Table 2). SMs have the lowest modification rates of 24%, while BK and MCs have higher

Surface message: It is an image of the New York City skyline when the twin towers were still there. The author describes this image as: 'we said we would never forget why are you voting them into our governement?'

Background knowledge: The 9/11 terrorist attacks which destroyed the twin towers were perpetrated by Muslim terrorists. Since then, there has been distrust of Muslims by some Americans.

Meme caption(s): We should not let Muslims into the government because they are terrorists.



Surface message: It is an image of a cartoon mouse looking mischievous. The author describes this image as: 'when you fart in crowd and try to look normal'

Background knowledge: A fart is a natural human action to release gas but is seen as embarrassing.

Meme caption: Farting is embarrassing, and you should hide it.

Surface message: It is an image of a Caucasian woman in pink clothing looking at her smartphone. The author describes this image as: 'they might not be liking your pics, but i bet they're reading your comments and keeping track of who you respond to'

Background knowledge: Some people are obsessed with what others think of them on social media.

Meme caption: People are keeping track of your social media activities.

Figure 2: Annotation examples.

Field	Agree	Disagree	Add	Cut	Word count
SM	76%	6%	12%	6%	17.45
BK	62%	19%	11%	8%	21.61
MC	55%	31%	4%	10%	9.93

Table 2: Statistics of the fields in `MemeInterpret`. Types of edits include "Disagree" (serious disagreement), "Add" (annotation contains inadequate/missing information; changes involve adding information) and "Cut" (description is too verbose; changes involve deleting information). "Word Count" shows the average number of words.

rates of 38% and 45%, respectively. These numbers indicate the increasing annotation difficulty from SMs to BK to MCs and offer a rough idea of how often disagreements have occurred.[5]

The judging results indicate a high level of quality in all the annotation fields. On Textual Completeness, SM, BK, and MC got scores of 5, 4.97, and 4.98 respectively. On Correctness, SM, and MC got scores of 4.91 and 4.79, respectively. BK further got Relevance, Sufficiency, and Factuality scores of 4.91, 4.77, and 4.94, respectively.

4 Empirical Studies

We conducted a broad range of experiments on `MemeInterpret` to (1) gauge the usefulness of our annotations, and (2) empirically verify our hypotheses about meme captioning and its relationship with downstream CMU tasks.

4.1 Implementation Details

Model We conducted experiments with LLaVA 1.5-7b[6] (Liu et al., 2024), one of the top-performing open-source vision-language models. It uses CLIP-ViT-L-336px as the vision encoder and Vicuna v1.5 13B as the LLM. The two modalities are "bridged" using a multi-layer perceptron. The connector was pre-trained with a subset of the CC3M dataset (Sharma et al., 2018), and the whole model was fine-tuned in an end-to-end fashion using academic-task-oriented Visual Question Answering data.[7]

Training and hyperparameter tuning We used Parameter Efficient Fine Tuning (PEFT) by attaching and training a LoRA adapter (Hu et al., 2022) to all linear modules in the base model. We reserved 20% of our training set as development data. Across all PEFT runs, we set `lr=1e-5`, `lora_alpha=8`, `lora_dropout=0.1`, `num_epochs=3`, `batch_size=2`, and `r=8`, and selected the best checkpoint based on the performance on the development set. For evaluation on

[5]The major sources of disagreement are discussed in Appendix C.

[6]https://huggingface.co/llava-hf/llava-1.5-7b-hf

[7]In our experiments, we always fed both the text extracted from a meme and the image to LLaVA, leveraging its visual understanding capability.

generation tasks, we set `max_new_tokens=100` and used greedy generation.

Prompt templates We provide the prompt templates used in the experiments in Appendix D.

Evaluation metrics We conducted both human and automatic evaluations. For human evaluation, we had annotators manually review the generated outputs of all models on 250 random test samples, which is 25% of the test set. The outputs were scored on the 5-point Likert scale w.r.t. *Correctness* (i.e., how close the generated caption semantically is from the ground truth) and *Fluency* (i.e., whether the output is in good English language). For BK, owing to its nature as a listing, we added two more metrics, *Relevance* (i.e., whether the BK is relevant to the meme) and *Sufficiency* (i.e., whether the BK is sufficient for someone from a different culture to make sense of the meme). Each model output was rated by two independent annotators and the model's names were not shown to them.

For automatic evaluation, we evaluated each model variant on the entire test set using popular metrics: BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020), and the entailment score from an NLI model[8] (Manakul et al., 2023). While the first two metrics are based on n-gram overlaps, which tend to measure textual fluency, the last two measure the semantic similarity of the generated and ground-truth texts.

4.2 Sanity Checks

The goals of our first experiment are two-fold. First, we check whether our annotations are useful by comparing results of models fine-tuned on our annotations with the corresponding zero-shot results.[9] Second, we gauge the difficulty of generating good MCs, BK, and SMs.

Setup LLaVA was evaluated on three tasks — SM generation, BK generation and meme captioning — under *two* settings: the *zero-shot* setting, where no data from `MemeInterpret` was used to train LLaVA, and the *fine-tuned* setting, where we fine-tuned LLaVA on the training split of `MemeInterpret` with the hyperparameters tuned on the development set.

[8] https://huggingface.co/potsawee/deberta-v3-large-mnli

[9] This check is motivated by the experiments conducted on MemeCap by Hwang and Shwartz (2023), who showed that their fine-tuned results on meme captioning were worse than their zero-shot results but did not provide any explanations for their unexpected results.

Task	Setup	Human				Automatic			
		Cor.	Flu.	Rel.	Suf.	BLE	ROU	BER	NLI
MC	Z	**2.38**	3.87	-	-	.001	.051	.804	.162
	FT	2.37	**4.60**	-	-	**.015**	**.189**	**.889**	**.265**
BK	Z	3.10	3.83	2.13	2.08	.006	**.152**	**.845**	.234
	FT	**3.88**	**3.88**	**3.57**	**3.46**	**.017**	.142	.830	**.336**
SM	Z	2.45	3.66	-	-	.018	.230	**.897**	.281
	FT	**3.38**	**3.81**	-	-	**.172**	**.370**	.877	**.385**

Table 3: **LLaVA's results on meme captioning (MC), background knowledge generation (BK), and surface message generation (SM)**. "Z" and "FT" refer to the zero-shot and fine-tuned settings, respectively. The best result for each task and each setting is **boldfaced**.

Results As can be seen in Table 3, fine-tuning the model on **SM, BK, and MC annotations yielded better results** across all metrics. There are a few cases where the fine-tuned models scored lower than the zero-shot models, but the difference is negligible compared to the significant outperformance on most of the metrics. While the fine-tuned results are better than the zero-shot results, we can see that they are still far from perfect.

4.3 Using BK and SM for Meme Captioning

Since BK and SM are meant to support meme captioning, our next experiment involves determining the usefulness of these two types of annotations.

4.3.1 Using Gold BK and SM Annotations

We first evaluate the usefulness of *gold* BK and SM annotations for meme captioning, with the goal of obtaining upper-bound performance on MC generation given perfect SM and BK information.

Setup We trained three models. First, to determine whether BK and SM are useful for meme captioning when applied in *combination*, we fine-tuned LLaVA using both SM and BK in the prompt. Next, to determine whether BK and SM are useful when applied in *isolation*, we conducted ablation experiments in which we fine-tuned LLaVA using exactly one of the two knowledge sources. In these experiments, *gold* BK and SM annotations were used.

Results Rows 1-2 of Table 4 show the results. The fine-tuned LLaVA with both inputs achieves better performance across all metrics except Fluency. This suggests that BK and SM, when applied in combination, can improve MC generation except that the outputs are slightly less fluent.

An interesting question is: is it more challenging to generate MCs for hateful memes or non-hateful

#	Inputs	Human		Automatic			
		Cor.	Flu.	BLE	ROU	BER	NLI
MC prediction without BK and SM							
1	Meme only	2.37	4.60	.015	.189	.889	.265
Pipeline MC prediction with gold annotations							
2	Meme+BK+SM	**3.54**	4.49	**.041**	**.280**	**.898**	**.507**
3	Meme+SM	2.68	3.61	.004	.090	.824	.382
4	Meme+BK	2.92	3.41	.007	.102	.824	.462
Pipeline MC prediction with predicted inputs							
5	Meme+AutoBK&SM	1.95	2.37	.002	.054	.778	.213
Joint MC, SM, and BK prediction							
6	Meme only	2.90	**4.65**	.013	.186	.877	.364

Table 4: **Fine-tuned LLaVA's results on meme captioning with varying inputs**. For comparison purposes, row 1 shows the fine-tuned results for meme captioning that were taken verbatim from Table 3. The best result for each metric is **boldfaced**.

memes? To answer this question, we examined the MC generation performance of the best-performing model (row 2) separately on the hateful and non-hateful memes. While the model scored slightly lower on the hateful memes across all metrics, the differences are statistically indistinguishable.[10]

Ablation results are shown in rows 3 and 4. Comparing them with the results in row 2, we see that removing either knowledge source causes considerable precipitation in MC generation performance. These results suggest that both BK and SM contribute positively to MC generation performance.

4.3.2 Using Automatically Generated BK and SM Annotations

When applying meme captioning models, it is not practical to assume that gold BK and SM annotations exist. Hence, we examined the impact of automatically generated BK and SM annotations on MC generation. Note that our goal is *not* to design new BK and SM generation models. Rather, we seek to understand whether the BK and SM annotations generated by *existing* models can benefit MC generation, thus establishing lower-bound performance on MC generation.

Setup We experimented with two MC generation models that exploit BK and SM, a *pipeline* model and a *joint* model. The *pipeline* model operates as follows. We first used the fine-tuned BK and SM generation models in Table 3 to produce automatic BK and SM annotations. Then the MC generation model shown in row 2 of Table 4 was used to generate meme captions by replacing the gold BK and SM annotations with the automatic BK and SM annotations. For the *joint* model, we fine-tuned it to generate the concatenation of SM, BK, and MC.

Results Results of MC generation using these two models are shown in rows 5 and 6 of Table 4. A few points deserve mention. First, the joint model (row 6) outperforms the pipeline model w.r.t. all of the metrics used in both human and automatic evaluations. This is perhaps not surprising, as pipeline models are known to suffer from error propagation. Second, the joint model outperforms the basic model (row 1) on both human evaluation metrics, with a wide margin on Correctness (0.53). This shows that even noisily computed SMs and BKs can benefit MC prediction. Perhaps impressively, the joint model achieved a higher Correctness score than one of the models that has access to gold SM during test time (row 3) while achieving the highest Fluency score overall. This result shows that the idea of training MC generators with SM and BK is very promising.

4.4 Evaluation on Hateful Meme Detection

We hypothesize that meme captions, being the core task in CMU, together with the surface messages and background knowledge, could be profitably exploited for downstream CMU tasks. In our next experiment, we tested this hypothesis on HMD, the task of classifying whether a meme is hateful or not, using the hatefulness labels provided by the FHM dataset (See Section 3.1 for details).

State of the art The top half of Table 5 summarizes the state-of-the-art (SOTA) results on HMD, which are expressed in terms of AUROC and accuracy.[11] The current best systems are PaLI-X-VPD (Hu et al., 2024), which has an AUROC of 0.892, and RGCL HateCLIPper (Mei et al., 2024), which has an accuracy of 0.788. They are followed by Flamingo 80B (Alayrac et al., 2022) and HateCLIPer (Kumar and Nandakumar, 2022). Among these top-performing HMD systems, only HateCLIPer has publicly available source code.

Our systems To determine the usefulness of our annotations for HMD, we used them to construct five HMD systems. The first four are fine-tuned LLaVA models, which differ in terms of what is

[10] Further details are shown in Appendix E.

[11] https://paperswithcode.com/sota/meme-classification-on-hateful-memes (retrieved in October 2024)

	Model	Performances	
		AUROC	Acc.
1	PaLI-X-VPD (55B)	**0.892**	-
2	RGCL HateCLIPper	0.870	0.788
3	Flamingo 80B	0.866	-
4	Hate-CLIPper	0.858	0.740
5	LLaVA-AutoMC	0.724	0.668
6	LLaVA-AutoAll	0.739	0.662
7	LLaVA-GoldMC	0.856	0.771
8	LLaVA-GoldAll	0.876	0.786
9	LLaVA-GoldAll × Hate-CLIPper	0.890	**0.800**

Table 5: **Results of the state-of-the-art models (top) and our systems (bottom) on Hateful Meme Detection**. The × symbol indicates a joint inference system. For each metric, the best result is **boldfaced** and the second-best result is underlined.

fed to the model beside the meme — "MC" only or "All" (SM+BK+MC), as well as whether those inputs are gold ("Gold") or predicted ("Auto") annotations. Next, to determine if our annotations can be used to improve a SOTA HMD system, we employ a model that performs joint inference over LLaVA-GoldAll and HateCLIPper (the only open-sourced SOTA HMD model shown above). In doing so, we first replicated the fine-tuning procedure of HateCLIPper on our training set. During inference, let p_1 and p_2 be the probabilities that HateCLIPper and LLaVA-GoldAll predict as the hatefulness probability of the input meme respectively. The *combined* prediction from both models is then $p = tp_1 + (1-t)p_2$, where the weight $t \in [0; 1]$ is chosen using the development set.[12]

Results Table 5 shows the results of four SOTA models (rows 1–4) and our five HMD systems (rows 5–9). As can be seen, LLaVA-GoldAll × Hate-CLIPper, our strongest model, *provisionally* achieved SOTA performance in terms of accuracy and was competitive with the strongest model in AUROC (only 0.0025 point less). Note that this system has only 7.5 billion parameters, while PaLI-X-VPD is roughly seven times larger in size. Furthermore, LLaVA-GoldAll alone has higher performance than all but the best model on the leaderboard in terms of AUROC. These results showcase the **effectiveness of our annotations in HMD**, predicting that improvements in modeling SM, BK, and MC can enhance HMD systems. As of now, the results of the models using predicted inputs, LLaVA-AutoMC and LLaVA-AutoAll (rows 5 and 6), are about 11% away from their counterparts. This performance drop should not be surprising since (1) the models we used to generate these predicted outputs are very simplistic and (2) errors from the MC/SM/BK predictions propagate to HMD as these HMD models are effectively pipeline models.

[12]The final mixing weight was 0.2452.

Model	Human		Automatic			
	Cor.	Flu.	BLE	ROU	BER	NLI
T5 Large	2.44	3.80	.033	.135	.420	.401
LLaVA-FT-GoldAll	**2.53**	3.75	.024	.161	.800	**.435**
LLaVA-FT-AutoAll	2.30	**3.90**	**.036**	**.200**	**.827**	.364

Table 6: **System performances on hateful meme explanation**. The best performances are **boldfaced**.

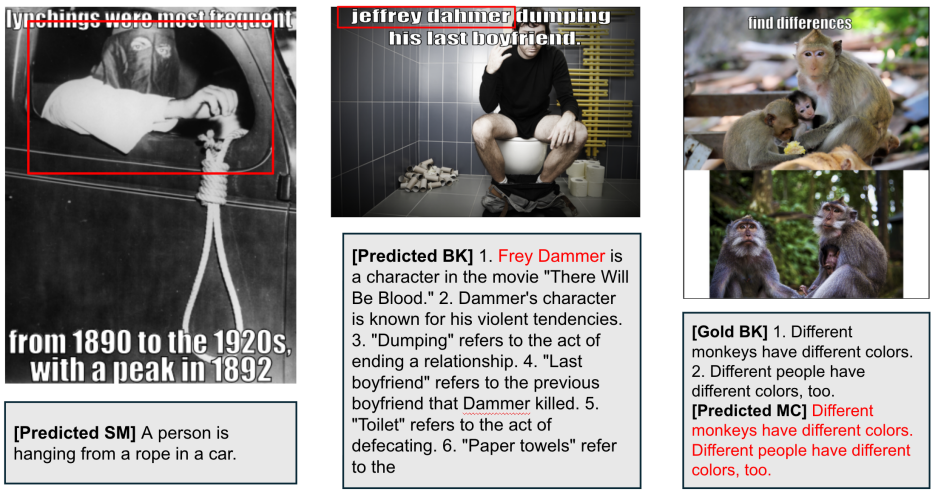
4.5 Evaluation on Explaining Hateful Memes

Finally, to complete the study of CMU task interaction, we also demonstrated the usefulness of `MemeInterpret`'s annotations in hateful meme *explanation*. Proposed by Hee et al. (2023), this task asks systems to generate a textual explanation for why a given meme is hateful. They also released HatReD, a set of hateful meme explanation annotations for the images in the FHM dataset. This allows us to again examine the effects of our annotations on this downstream task.

Compared to `MemeInterpret`, HatReD covers a *different* subset of the FHM dataset. As a result, `MemeInterpret` and HatReD overlap on only 2,359 images. Thus, in this experiment, we split their intersection into 60% for training, 20% for development, and 20% for testing. In that way, all images involved have SM, BK, MC, and explanation annotations. To make the test set challenging, all memes with multiple social targets (based on HatReD's annotations) were put into that split.

Setup Similar to previous generative tasks, we fine-tuned two LLaVA-based models to generate the explanation given all three annotation types as input. The first one used gold inputs and the second one used predicted inputs. The prompt template for this task can be found in Appendix D.5.

Results Results are shown in Table 6. For comparison purposes, we replicated the best system in Hee et al. (2023) on our dataset split, which is T5 Large. We see that the system using the gold inputs (row 2) outperformed the SOTA model (row 1), while the simple pipeline system (row 3) established a strong lower bound. This further underlines that the annotations in `MemeInterpret` are a good representation of the memes' meaning.



(a) SM error (b) BK error (c) MC error

Figure 3: Errors made by the top model variants.

5 Error Analysis

To gain further insights into the challenges in modeling meme interpretations, we analyzed the errors made by our best performing models, which are the fine-tuned BK generation model (row 4, Table 3), the fine-tuned SM generation model (row 6, Table 3), and the fine-tuned MC generation model with gold SM and BK inputs (row 5, Table 4). Figure 3 shows examples of their failure modes.

SM We observed that the model frequently omits specific details in its descriptions, such as the overlaid text, group names, race, celebrities, age, or disabilities. For example, in Figure 3a, the model failed to identify the entity "KKK group" and the overlaid text. In this case, the model failed to pick up the hints from the text, which refers to the activities of the KKK group. Subsequently, it produced a generic image caption that excludes meaningful details of the meme. This *selective vision* issue remains an open question for image understanding in general (Chung et al., 2024).

BK The fine-tuned BK generation model sometimes produced knowledge that is irrelevant to the meme. In Figure 3b, the model incorrectly included Frey Dammer, a character unrelated to the meme, and failed to capture the significance of "Jeffrey Dahmer". This error perhaps stemmed from the limited text recognition capability of LLaVA 1.5.

MC A common error in MC generation is the mere repetition of BK or SM from the input. For instance, in Figure 3c, the MC merely regurgitates the BK without attempting to unpack the implication behind the meme. This underscores the need for the model to exhibit more advanced reasoning capabilities.

6 Conclusion and Future Work

In light of the fact that existing CMU tasks are largely tackled independently of each other, we took the first step towards suggesting that this current research practice can possibly be reshaped by proposing a task-agnostic representation of memes based on meaning rather than form. Subsequently, we (1) advocated the importance of advancing research on the under-studied task of meme captioning, (2) identified two challenging subtasks of meme captioning, background knowledge generation and surface message generation, and (3) proposed `MemeInterpret`, which could spark research on studying the interactions of different categories of CMU tasks. Extensive experiments showed that (1) the `MemeInterpret` annotations are useful, as the fine-tuned results are better than the zero-shot results; (2) meme captioning could benefit other categories of CMU tasks, including classification and explanation tasks; and (3) `MemeInterpret` could facilitate the development of a unified framework that can simultaneously address all three categories of CMU tasks.

As the fine-tuned models still have a lot of room for improvement, we believe future work should include (1) the development of novel models that can better exploit our annotations, especially joint models that allow interaction of multiple CMU tasks and (2) an exploration of whether our annotations can benefit additional downstream CMU tasks.

Limitations

Given that `MemeInterpret` covers topics on US social media, it is not expected to generalize to other cultures. Outside of the American context, there has been work in Bengali (Ahsan et al., 2024) and Tamil (Suryawanshi et al., 2020b). Future work should consider constructing multicultural meme dataset by combining existing datasets and collecting new memes for underrepresented cultures.

Besides, our findings are based on an open-sourced model. Future work should extend the investigation to close-sourced models to deepen our understanding of CMU task interactions in these types of systems.

Ethics Statement

Broader implications As mentioned before, the solution to the meme captioning task is of practical significance. From a practical perspective, knowledge of the meaning being conveyed in a meme (and its caption) could be useful for other meme-related processing tasks. For instance, knowing what the meaning is could facilitate the determination of whether a meme contains harmful content. Theoretically speaking, being able to generate messages like humans requires that a machine read between the lines and achieve a deeper level of understanding of perceptual input, enabling machine perception to get one step closer to human perception.

Ethical considerations Having said that, we are all aware that some memes contain harmful content, so when our models are applied to these harmful memes, they will generate harmful captions that could have a negative psychological impact on the users, especially if they are the target of the harmful content. Therefore, as with many other AI/NLP technologies, our models should be used with care. We should emphasize that our intent is to build models for interpreting memes, hoping that readers of memes will less likely be manipulated after understanding the intention of the meme authors.

Steps taken to protect annotators from harmful content All annotators were provided with a thorough instructional training session in which they were instructed on how to annotate the data and how to go about the whole task. During training, annotators were shown the types of memes that they will work with so that they have an idea of the dataset's nature. The annotators have full autonomy to withdraw from the project at their own judgement. They also gave consent for the collected data to be used for research purposes.

Terms of use This dataset is consistent with the terms of use and the intellectual property and privacy right of people with the Facebook Hateful Meme dataset. The dataset was licensed from ©Getty Images. There is nothing about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses.

Data distribution We open-sourced the data produced from this work at `https://github.com/npnkhoi/MemeInterpret`.

References

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M. Dewan. 2024. A multimodal framework to detect target aware aggression in memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2487–2500, St. Julian's, Malta. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736, New Orlanes, LA, USA.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "Subverting the Jewtocracy": Online antisemitism detection using multimodal deep learning. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, pages 148–157, New York, NY, USA. Association for Computing Machinery.

Jiwan Chung, Sungjae Lee, Minseo Kim, Seungju Han, Ashkan Yousefpour, Jack Hessel, and Youngjae Yu. 2024. Selective vision is the challenge for visual reasoning: A benchmark for visual argument understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2423–2451, Miami, Florida, USA. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021.

SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5995–6003, Macau, SAR China. International Joint Conferences on Artificial Intelligence Organization.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the Tenth International Conference on Learning Representations*, Virtual.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2024. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9590–9601, Seattle, WA, USA.

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.

Saurav Joshi, Filip Ilievski, and Luca Luceri. 2024. Contextualizing internet memes across social media platforms. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1831–1840, Singapore.

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 *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624, Virtual. Curran Associates, Inc.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26286–26296, Seattle, WA, USA.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 2021. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 201–206, Online. Association for Computational Linguistics.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2024. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5333–5347, Bangkok, Thailand. Association for Computational Linguistics.

Khoi P. N. Nguyen and Vincent Ng. 2024. Computational meme understanding: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21251–21267, Miami, Florida, USA. Association for Computational Linguistics.

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

Jeongsik Park, Khoi P. N. Nguyen, Terrence Li, Suyesh Shrestha, Megan Kim Vu, Jerry Yining Wang, and

Vincent Ng. 2024. MemeIntent: Benchmarking intent description generation for memes. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 631–643, Kyoto, Japan. Association for Computational Linguistics.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

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

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. What do you MEME? Generating explanations for visual semantic role labelling in memes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9763–9771.

Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020a. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020b. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, Virtual.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia.

A Annotator Recruitment and Training

We started with 15 annotator candidates who are all undergraduate students in computer science[13] at our institution and provided them with one hour of training before assigning them a sample task involving 30 annotations. The nine candidates who met the required standards (regarding Textual Completeness and Correctness) were recruited. They participated in this project as part of the "Undergraduate Research in Computer Science" course they signed up for, during which they acquired experience and skills involved in a research project involving data annotation and model training. No additional compensation was thus provided to these students. Among the nine participants, there were four Americans, four Koreans, and one Indian. As for the gender distribution, two participants were females and seven were males. Four of them served as annotators (during the Collect stage) while five acted as editors and judges. These participants provided full consent to the annotation process.

When the participants first started annotating, the two first authors discussed the results of the first 100 instances with both the annotators and the editors. Throughout the process, we also conducted bi-weekly one-hour meetings where all participants annotated the same five instances and reviewed them together to better understand the guidelines.

[13]While we agree that it would be ideal for social science students to conduct this annotation task, we noted that memes are commonly encountered by users in everyday social media, and their interpretations are performed by people who do not necessarily have training in social sciences.

(a) (b)

Figure 4: Two example memes.

B Annotation Guidelines

The annotation guidelines that we provided to the annotators are shown in Table 7.

C Sources of Annotator Disagreement

To gain insights into the sources of disagreement, we further inspected the typical edits our editors made. For SM, edits mostly involved adding or correcting information about celebrities and the context of the image (e.g., the first annotator did not know or misrecognized the character in the image, which was later corrected by the editor). For BK, edits mostly involved adding new pieces of information that the second annotator deemed necessary to interpret the meme, with occasional disagreements about the factual accuracy of the previous BK. For MC, edits varied and could simply be about the specificity of the caption, or as serious as disagreements about the *target* of the meme and the *sentiment* of the meme toward the target. Figure 4 and Table 8 illustrate two example memes and their annotations before and after editing.

We observed that the disagreements in the MCs were typically accompanied by a lack of knowledge or an overlook by the annotators. In such cases, SM and BK annotations came in handy as they helped the editors understand the viewpoint of the annotator and conduct fact-checking on the BK when appropriate. On occasions when the editors were unsure about how to edit the annotations, they raised the meme instance in the annotator group for discussion. When a consensus was reached in the group, the editor went ahead and made the edit. Otherwise, if the group as a whole did not have a good idea of the meme's meaning, the meme would be discarded. Finally, if there are multiple strong opinions on the meme's meaning (e.g., due to the differences in cultural backgrounds and personal beliefs), all annotations were considered reasonable and added to the dataset. This final case is very rare — out of nearly 7K memes, only 49 memes have more than one meme caption (less than 1%).

D Prompt Templates and Implementation Details

D.1 Prompts for the Sanity Check Experiments

The prompt for meme captioning is:

```
You will be provided with a meme. Your task is to infer the message that the author is trying to convey through the meme. The message must be in one single short sentence. The final message of this meme is:
```

The prompt for surface message generation is:

```
You will be provided with a meme. Your task is to identify the explicit or surface-level message conveyed by the meme. The surface message is what the meme is saying directly, including any text, images, or symbols present. Describe this surface-level message as simply and clearly as possible without interpretation of deeper meaning. Surface message must be in one single short sentence.
### Surface message:
```

The prompt for background knowledge generation is:

```
You will be provided with a meme. Your task is to infer the background knowledge that a reader of the meme needs to possess before they can understand the ultimate intent behind the creation or sharing of a meme, as perceived by its audience. Background knowledge is the minimum amount of knowledge that is missing from the meme. It is the knowledge that needs to be combined with visual and textual cues from the meme in order to understand its meaning. Give me background knowledge in the form of a list. For example: '1. Soccer is the sports that children likes a lot. 2. There are two main political parties in the US: Democratic and Republican.' Each background knowledge must be in one single short sentence.
### Background knowledge:
```

D.2 Prompt for Meme Captioning with varying inputs

Below is the prompt format for meme captioning with SM and BK in the inputs. Note that `[SM]` and `[BK]` were substituted with the actual SM and BK ground truth annotations. Whenever an input is absent, the corresponding line (starting with ###) was removed.

Guideline
The meme originally has two components: the **Image** and the **Text**.
A **surface message** is a complete and standalone representation (1-3 sentences) of the whole meme, describing both the Image and the Text. It includes any identifiable information about races, religions, genders, sexual orientations, specific celebrities, and any other characteristics of people in the meme. It must start with `"It is an image about ..."`, `"In the image, ..."` or a similar structure. If a meme has multiple images, you can write `"It is two images. The top image is ..., and the bottom is ..."` Finally, depending on how the text fits into the image, it can be written as (1) `"{Description of the image}. The author describes the image as {Text}."`, (2) `"{Description of the image}. The character in the meme says that {Text}."`, or (3) some other structure that appropriately combines the Text with the image caption. For Figure 1a: *It is an image of a 70-year-old Caucasian man sitting in front of a window with a stunned face, wearing a hat that says "VIETNAM VETERAN". The author describes the image as "me: puts bag of popcorn into the microwave. every one else at the senior center:"*
Background knowledge is additional knowledge you needed to use to derive the meme caption beyond the surface message. You can write multiple sentences, representing multiple pieces of knowledge. For Figure 1a: *(1) The Vietnam War, also known as the Second Indochina War, was a protracted and highly controversial conflict that took place in Vietnam, Laos, and Cambodia from November 1, 1955, to April 30, 1975. (2) The sound of popcorn popping in the microwave can remind veterans of the gunfire sound during the war. (3) The popcorn popping sound is a satisfying sound that is not to be scared of.*
A **meme caption** is a message the author intends to convey through the meme. There can be multiple meme captions.

Table 7: Annotation guidelines.

Field	Agree.	Before	After	Reason
Figure 4a				
SM	Cut	The text on top is a list spoken by the author and the bottom text is spoken by the character in the image.	It is an image of a Caucasian man with a beard and mustache. The text on top is a list narrated by the author, reading, "Brett Kavanaugh, 17. Christina Ford, 15." The bottom text is spoken by the character in the image, reading, "I'll take shit that never happened for 1000."	changed the format and made it more detailed.
BK	Agree	This requires understanding the Christine Ford and Brett Kavanaugh case, which involved sexual assault allegations against Kavanaugh during his Supreme Court nomination process and an understanding that Jeopardy is a game show where people choose categories and point amounts.	1. This requires understanding the Christine Ford and Brett Kavanaugh case, which involved sexual assault allegations against Kavanaugh during his Supreme Court nomination process and an understanding that Jeopardy is a game show where people choose categories and point amounts.	changed the format by indexing BKs.
MC	Cut	Kavanaugh and Ford had no sexual relations.	The author implies that the Christine Blasey Ford and Brett Kavanaugh case was a farce, meant only to attack Kavanaugh's character during his Supreme Court nomination. The message brushes off Ford's allegations of sexual assault against Kavanaugh, and insults her ability to provide truthful testimony.	made it more detailed.
Figure 4b				
SM	Agree	2 pictures side by side of Pope Francis and a Muslim man with a beard. The character in the meme says that 'i'll have 72 virgins when i die i'll have 72 virgins before i die.'	2 pictures side by side of Pope Francis and a Muslim man with a beard.. The character in the meme says that 'i'll have 72 virgins when i die i'll have 72 virgins before i die.'	NaN
BK	Cut	1. Knowledge of Pope Francis and his role as the head of the Catholic Church. 2. Knowledge of the Islamic belief in 72 virgins in paradise. 3. Knowledge of the misconduct of the Catholic Church with molestation.	1. Knowledge of Pope Francis and his role as the head of the Catholic Church. 2. There is an Islamic belief that there are 72 virgins in paradise. 3. The Catholic Church has been associated with molesting children. 4. The same statement means completely different things depending on who is saying it.	made it more detailed.
MC	Agree	The pope will molest virgins before he dies.	The Pope will molest virgins before he dies.	NaN

Table 8: The annotations of the two memes shown in Figure 4 before and after editing.

```
You will be provided with a meme, a description of its text and image, and the background knowledge that a reader of the meme needs to possess before they can understand the message. Your task is to infer the message that the author is trying to convey through the meme. The message must be in one single sentence.
### Description of its text and image: [SM].
### Background knowledge: [BK].
### Message:
```

D.3 Prompt for Joint Modeling of SM, BK, and MC

We used the following prompt to fine-tune LLaVA to generate the concatenation of SM, BK, and MC:

```
You will be provided with a meme. Generate a surface message (SM), background knowledge (BK), and a meme caption (MC) for the meme. A surface message is defined as "what the meme is saying directly, including any text, images, or symbols present, and excluding interpretation of deeper meaning." A background knowledge list is defined as "the minimum list of factual statements that is missing from the meme. It is the knowledge that needs to be combined with visual and textual cues from the meme in order to understand the meme's meaning." A meme caption is defined as "the message that the author is trying to convey through the meme, written in one single short sentence." Format your answers as "<SM_start> Your SM <SM_end> <BK_start> Your BK <BK_end> <MC_start> Your MC <MC_end>"
```

D.4 Details of Hateful Meme Detection Experiments

Below is the prompt used to instruct LLaVA to perform HMD detection given SM, BK, and MC as inputs. During evaluation, only one new token was generated and evaluated against a "0" or "1" label.

```
You will be provided with a meme, a description of its text and image, and the background knowledge that a reader of the meme needs to possess before they can understand the message. Your task is to determine whether the meme is hateful or not. Format your answer as a binary classification, where 1 indicates that the meme is hateful and 0 indicates that the meme is not hateful. Only return the number.
### Description of its text and image: [SM].
### Background knowledge: [BK].
### Message: [MC]
Is this meme hateful? (0/1)
```

	Cor.	Flu.	BLE	ROU	BER	NLI
Non-Hate	3.51	4.47	.035	.274	.897	.496
Hate	3.58	4.51	.046	.286	.898	.518
p-value	0.65	0.69	0.27	0.44	0.76	0.41

Table 9: Comparison of non-hateful and hateful memes in performance of the fine-tuned meme captioning model. Unpaired t-tests were conducted to determine whether the performance differences are statistically significant, with the p-values shown in the last row.

D.5 Prompt for Hateful Meme Explanation

Below is the prompt used to instruct LLaVA to perform the hateful meme explanation task.

```
You will be provided with a meme, a description of its text and image, the background knowledge that a reader of the meme needs to possess before they can understand the message, and the message conveyed by the meme. Your task is to explain why the meme is hateful. The explanation must be in one of the form (i) '<verb> <target> <predicate>' or (ii) 'use of derogatory terms against <target> <predicate>', where <target> represents the attacked social target and <predicate> highlights the hateful implication.
### Description of its text and image: [SM].
### Background knowledge: [BK].
### Message: [MC]
### Explanation:
```

E Performance Differences on Hateful vs. Non-Hateful Memes

We investigated the difference in effects of hateful and non-hateful memes on meme captioning performance. Among the 250 manually evaluated examples, we selected the output from the best setup based on human and automatic evaluation — the fine-tuned MC generation model with gold SM and BK inputs (row 5, Table 4). Then, we divided them into two groups — 120 hateful memes and 130 non-hateful memes, and calculated the metrics for each (Table 9). We can see that non-hateful memes have slightly lower scores than their hateful counterparts, though the differences are not statistically significant (unpaired t-tests with $p < 0.05$).