Hope 'The Paragraph Guy' explains the rest : Introducing MeSum, the Meme Summarizer

Anas Anwarul Haq Khan[1]*, **Tanik Saikh**[2], **Arpan Phukan**[3], **Asif Ekbal**[4]
[1]Department of CSE, IIT Bombay [2]School of Computer Engineering, KIIT, Bhubaneswar
[3]Department of CSE, IIT Patna [4]School of Artificial Intelligence, IIT Jodhpur, Jodhpur
anas290816007@gmail.com, tanik.saikhfcs@kiit.ac.in
arpan_2121cs33@iitp.ac.in, asif@iitj.ac.in

Abstract

Over the years, memes have evolved into multi-faceted narratives on platforms like Instagram, TikTok, and Reddit, blending text, images and audio to amplify humor and engagement. The objective of the task described in this article is to bridge the gap for individuals who may struggle to understand memes due to cultural, geographical, ancillary insights, or relevant exposure constraints, aiming to enhance meme comprehension across diverse audiences. The lack of large datasets for supervised learning and alternatives to resource-intensive vision language models have historically hindered the development of such technology. In this work, we have made strides to overcome these challenges. We introduce "MMD" a **M**ultimodal **M**eme **D**ataset comprising **13,494** instances, including 3,134 with audio, rendering it the largest of its kind, with **2.1** times as many samples and **9.5** times as many words in the human annotated meme summary compared to the largest available meme captioning dataset, MemeCap. Our framework, MeSum (**Me**me **Sum**mariser), employs a fusion of Vision Transformer and Large Language Model technologies, providing an efficient alternative to resource-intensive Vision Language Models pioneering the integration of all three modalities, we attain a ROUGE-L score of 0.439, outperforming existing approaches such as zero-shot Gemini, GPT4 Vision, LLaVA and QwenVL which yield scores of 0.259, 0.213, 0.177 and 0.198. We have made our codes and datasets publicly available.[1]

1 Introduction

The term "*meme*" originates from the Greek word "*mimoúmai*," meaning "*to imitate*." Within the domain of social media, memes serve as humorous narratives, crafted to mimic everyday situations, with the purpose of entertaining and amusing audiences. Over the years, human being have been witnessed a significant transformation in humor, shifting from traditional sources. Initially, jokes were mainly shared as written text in books and newspapers. However, with the advent of the internet, the way people shared humor changed dramatically. Social media platforms like Facebook, Instagram and Twitter have been playing a crucial role in this shift, giving rise to a new form of humor: memes.

Memes are essentially funny images paired with witty captions or text overlays, and they quickly gained popularity online. As social media continued to evolve, so did the content shared on these platforms. Users began to incorporate various multimedia elements into their memes, including audios. Platforms like Reddit, TikTok and Instagram Reels further expanded the possibilities by allowing users to add background audio to their memes, making them more dynamic and engaging. This evolution demonstrates how humor has adapted creatively to the digital age, offering users new ways to express themselves and entertain others. Here's a rundown of our task, we start the process by extracting the textual content embedded within meme images. This extraction has facilitated through the utilization of advanced Optical Character Recognition (OCR)[2] techniques alongside pre-trained image caption generators (Li et al., 2022). Moreover, we integrate the image and audio components with the extracted text to create a cohesive synthesis of multimedia elements. Subsequently, our model is based on supervised learning to predict concise summaries of the meme content. Additionally, we have crafted the gold standard summaries that is appealing to viewers, adding an element of fun to the explanation. This approach enhances the overall engagement and enjoyment, making it both

*Work done during internship at IIT Patna.
[1]https://github.com/anas2908/MeSum
[2]https://github.com/JaidedAI/EasyOCR

Findings of the Association for Computational Linguistics: EACL 2024, pages 6654–6668
November 12-16, 2024

informative and entertaining for the audience.

Motivation : In today's digital era, memes have become a ubiquitous form of communication, transcending cultural boundaries and spreading rapidly across social media platforms. However, the interpretation of memes can often vary from person to person, influenced by individual experiences, intellectual reasoning and cultural references. To bridge this gap in understanding, our approach focuses on providing concise, and engaging explanations that enhance the viewers' comprehension of the meme's context and humor. By weaving together relatable anecdotes and witty observations, we aim to make the explanation not only informative but also enjoyable for the readers. Our goal is to ensure that everyone, regardless of their background or familiarity with meme culture, can appreciate and enjoy the humor behind each meme.

The main contributions of our proposed research are as follows: **(1)** To best of our knowledge, **MeSum** is the first work that utilizes all three modality i.e Text, Image and Audio for understanding and summarization of memes. **(2)** We have curated the largest multi-modal dataset to date, comprising 13,494 instances, with more than 23% of them featuring background audio. Each instance in the dataset is accompanied by a meticulously crafted Gold Standard summary that encapsulates the essence of the meme. **(3)** In lieu of resource-intensive Vision-Language Models (VLMs), we propose MeSum, an efficient approach that involves fusing vision transformers and Large Language Models (LLMs). Specifically, we modify the encoder of BART to seamlessly integrate embeddings from all the three modalities.

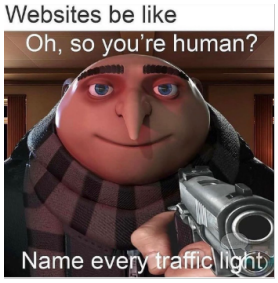
2 Related work

Memes are copied and spread rapidly by internet users, often with slight variations. Various shared tasks have been organized recently, with a recent one on detecting the hero, the villain and the victims entities in memes (Sharma et al., 2022b). There are tasks such as troll meme classification, defined in (Suryawanshi and Chakravarthi, 2021) and meme-emotion analysis through their sentiment, types and intensity prediction (Sharma et al., 2020). The task of hateful meme detection was introduced by Kiela et al. (2020), further Zhou and Chen (2020) carry forwarded this task, proposing various solutions to the problem that showed lots of interest to the community. Sharma et al. (2022a) provided a very concise and good survey on Harmful Memes including different types of harmful memes like *hate, racist, Misogynistic/Sexist, offensive memes, Propaganda, Harassment/Cyberbullying, Violence and Self-Inflicted Harm*. Kiela et al. (2021); Qu et al. (2022); Sharma et al. (2023b) worked on detection of harmful or hateful content in the memes. The task defined in (Sharma et al., 2023a) presented a novel task - EXCLAIM that generates explanations for visual semantic role labeling in memes. The dataset utilized for humor-related images is the New Yorker Cartoon Caption Contest (Hessel et al., 2023), addressing tasks such as caption-to-cartoon matching, caption quality assessment, and joke explanation. However, (Hessel et al., 2023) primarily focuses on cartoons rather than real-life objects or images encompassing ancillary knowledge or everyday experiences. This limitation restricts its comparison to meme explanations, where cartoons are just a subset of the broader spectrum. Our research, on the other hand, emphasizes meme summarization, ensuring that every aspect of humor is meticulously structured and connected to provide comprehensive yet detailed explanations. The most relevant dataset to our MMD is MemeCap (Hwang and Shwartz, 2023). It comprises 6.3K memes with human-annotated meme captions and textual metadata, focusing on metaphorical elements. However, Hwang and Shwartz (2023) overlooks the significance of the audio modality and offers brief captions that do not qualify as explanations or summaries. Our proposed MMD dataset goes beyond by incorporating audio cues, ensuring comprehensive connections across all three modalities during human annotation of summaries. Additionally, each meme instance summary in MMD is roughly 4.5 times larger than MemeCap (Hwang and Shwartz, 2023) captions. A detail comparison is shown in Table 3.

3 Dataset

To support our research into modern meme formats, we compiled a dataset with 13,494 meme instances, including 3,134 associated audio, along with images and texts. These memes were sourced from Reddit utilizing the publicly accessible API[3]. We employ annotators to annotate the memes (Refer Section 3.2). The motivation for creating this dataset stemmed from the lack of similar tasks previously undertaken. The closest related work to

[3] https://www.reddit.com/dev/api/



[A] *Explanation :* So, when we enter a website or sign up on a site, in order to sign up or enter, it wants to be sure if we are not a bot. In order to do this, the website presents us with a number of pictures, and usually asks us to select every picture with a traffic light on it. If we're wrong we can try again, but if we suceed we can enter that website.

[B] *Explanation :* The joke here starts with text that says "bus drivers when they see each other." We then see two images that are meant to show that the bus drivers might look excited to see another person who's working in the same profession as them. Below this there's more text that says "pilots when they see each other" with two images that show people freaking out and absolutely horrified. What this is saying is that because pilots fly planes way up in the sky, they'd be super scared if they saw another one while flying because that would mean that they're super close to each other which could result in them crashing into each other and that would be a valid reason to be scared.

[C] *Explanation :* One astronaut asking another "is there anywhere to get a beer around here?" With all the words smushed into one long one because there's no spaces. The other astronaut then responds by saying "there is no space bar" with the words again being a smashed into one long word. There are two meanings to this short conversation. The first one is that when the guy says "there's no space bar," he's saying that there's no bars in space so the guy is unable to get a beer which is what he originally asks. And because there is no space bar (meaning a bar in space) there's also no space bar to type spaces in between each word which is why the words are all crushed together.

[D] *Explanation :* Kintsugi ("golden repair" or "golden joinery") is the Japanese art of repairing broken pottery with lacquer dusted or mixed with powdered gold or similar material, highlighting the cracks instead of disguising them. This means that in Japan, broken objects are often repaired with gold. The person in this meme now goes to Japan, thinking they would cover his whole body in gold because he is so broken (maybe because he is depressed).

Figure 1: Few examples of human-annotated summaries corresponding to memes from the MMD dataset.

Meme	Observational	Slapstick	Wordplay	Depreciating	Teamplate	Ancillary Insight	Text Dominant
[A]	✓	×	×	×	✓	✓	✓
[B]	✓	✓	×	×	✓	×	×
[C]	×	×	✓	×	×	×	×
[D]	×	×	×	✓	×	✓	×

Table 1: With reference to Figure 1, meme's categories (Observational, Slapstick, Wordplay, Depreciating) are provided, along with properties(Template, Ancillary Insight, Text Dominant) they retain.

our task is MemeCap (Hwang and Shwartz, 2023), which captioned memes using texts and images. However, it has several drawbacks. Firstly, its dataset is relatively small, comprising only 6.3K instances. Additionally, (Hwang and Shwartz, 2023) lacks an audio modality, and its captions are often short, dull, and lack engagement. In comparison, our dataset is **2.1** times larger in terms of the number of instances and **9.5** times larger in terms of the number of words in the gold standard annotations.

Diversity: We collect memes from diverse, publicly available Reddit accounts known for their content, capturing memes from over 36 unique cultural backgrounds for 1517 instances and 72 unique geographic locations for 4743 instances, enriching our dataset's diversity further elaborated in Section A. These memes are broadly classified into four categories: *"Observational," "Slapstick," "Wordplay," and "Depreciating"*. An example is illustrated in the Table 1.

- **Observational:** Memes that humorously comment on or observe everyday situations and behaviors often highlight common experiences or quirks that many people can relate to, making them widely appealing. Everyday situations can be universal truths or may reference specific cultural or societal norms, adding layers of humor and relatability to observational memes.

- **Slapstick:** Form of humor that frequently incorporates physical comedy, exaggerated gestures, and facial expressions to generate amusement. These memes commonly depict characters engaging in actions like falling down, getting hit, or clumsily navigating situations. Additionally, slapstick memes may reference memorable moments from movies or television shows.

- **Wordplay:** Utilizing clever or humorous manipulation of language, often incorporating puns, double meanings, or creative spelling to evoke humor. An example of wordplay in a meme context could be transforming the word "*brownie*" into "*brow knee*", playing on the similarity in sound between "*brown*" and "*brow*" and humorously suggesting a knee-related association.

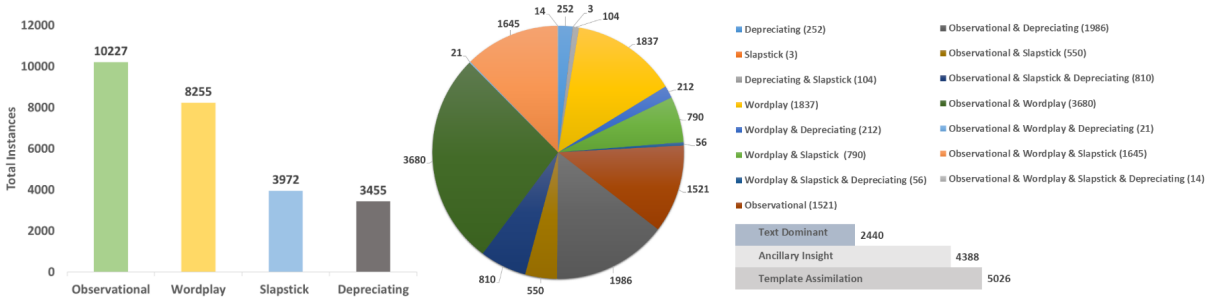


Figure 2: Statistical insights into MMD (**M**ulti-modal **M**eme **D**ataset) categorization and the properties they encompass. Each meme within the dataset must belong to at least one of the categories, while it is not mandatory for a meme to possess any specific properties. The categorization of memes is related to what makes them humorous, while properties depend on how the humor is presented.

Dataset	Modalities	Total Meme Instance	Total words in Explanation	Avg words per Explanation
MemeCap	T + I	6,387	134,328	21.03
MMD	**T + I + A**	**13,494**	**1,275,540**	**94.52**

Table 2: A concise comparison between MMD and MemeCap, the largest and most relevant dataset in the field.

Categories	**Instances**	**Total Words**	**Average Words per Instance**	**Instances having Audio**	**Visual Inclusion**
Observational	10,227	944,736	92.38	2,505	1 image per meme
Slapstick	3,972	378,510	95.29	772	1 image per meme
Wordplay	8,255	804,850	97.50	1,639	1 image per meme
Depreciating	3,455	308,551	89.31	919	1 image per meme

Table 3: Detailed breakdown analyses for text, audio, and image components in the dataset, along with categorical distributions.

- **Depreciating:** Memes that humorously highlight perceived flaws, weaknesses, or embarrassing moments, whether in oneself or others, often in a lighthearted or exaggerated manner. *These memes may poke fun at non-existing TV series characters or existing individuals, such as celebrities or public figures, without intending harm or causing offense.*

Data Analysis: MMD is the largest annotated collection for the meme summarization task, accompanying 13,494 meme instances. Notably, over 23% of these memes are enriched with background audio, aimed at enhancing the overall meme experience. Delving deeper, the cumulative word count across all meme summaries reaches **1,275,540** words, resulting in an average explanation length of **94.52** words per meme. A few of the meme summaries are shown in Figure 1. During the annotation process, memes were categorized into four to fifteen categories, including *"Observational", "Wordplay", "Slapstick", and "Depreciating"* as well as various combinations thereof, enriching its complexity and humor, the detail analysis layout is given in the Figure 2. Expanding our analysis, we conducted an examination of three distinct meme properties (c.f Table 1). First, we explored the property of "*Text-Dominant*," representing memes where text holds more significance compared to the accompanying image. This property constituted approximately 18.1% of the dataset. Second, *"Ancillary Insight"* refers to memes where a deeper understanding requires knowledge of external facts or widely known cultural references, encompassing around 32.5% of the memes. Lastly, *"Template Assimilation"*, where creators incorporate recognizable scenes from popular media to evoke specific emotions or reactions from viewers. This property accounted for nearly 37.2% of the dataset. We noted that one template can be used for many memes, highlighting the potential benefits of supervised learning in better understanding of meme creation and usage patterns (c.f Figure 3).

Categorical and Modal Analysis The Table 3 provides insights into the categorization and distribution of memes based on their textual content, image inclusion, and audio presence. The dataset includes 13,494 meme instances distributed across four overlapping categories: Observational, Slap-

Figure 3: An illustration of a single meme template employed to convey various jokes.

stick, Wordplay, and Depreciating. Each meme is associated with a single image, and the audio length ranges between 5 to 10 seconds. Observational memes have the highest number of instances with audio (2,505), while Depreciating has the least (919). On average, Wordplay memes contain the highest number of words per instance (97.50), and Depreciating memes the least (89.31).

3.1 Importance of All three modalities

Our dataset encompasses a combination of three modalities: text, audio, and image. Each modality plays a distinct role in conveying the essence of a meme. Text serves as the backbone, offering context and references essential for understanding the meme's message. It can manifest as a witty caption, a punchline, or keywords strategically placed within the image. Images are the visual elements that capture attention and evoke emotions ranging from simple depictions of facial expressions to intricate compositions laden with symbolism. Some memes follow recognizable templates derived from popular iconic movie scenes or recurring motifs in internet culture. As shown in Figure 3, these templates serve as shorthand references. Moreover, certain images require a degree of intellectual engagement or logical inference to decode fully. These memes often employ visual puns, optical illusions, or complex visual metaphors. Audio complements the visual elements to the meme-watching experience. It can heighten emotions, set the tone, or provide additional context to the visual content. For instance, sarcastic memes may feature accompanying sounds that reinforce the irony, while funny memes may incorporate laughter or comedic sound effects. Similarly, serious memes may integrate somber music to convey gravitas and depth.

3.2 Dataset Creation

We scraped memes from Reddit using the publicly available API. We employ two in-house annotators having doctorate degree in linguistics and three hourly-paid employees, compensating them at an average rate of $11.12 per hour. We measure the inter-annotator agreement (IAA) by calculating the METEOR score between the sets of annotations.The observed score of a substantial 0.793 indicates a strong consensus among annotators regarding the answer's relevance and accuracy. The annotation process involves several steps as follows: **(1)** During the scraping of memes, we extracted metadata such as captions and comments from the posts. Annotators were asked to identify and filter relevant data using keywords like "Explanations," "The Paragraph guy," and "meaning." **(2)** If an explanation was available in the filtered data, it was utilized to further enhance the meme content. **(3)** The explanations were then manually refined by eliminating unnecessary words or promotional content. **(4)** The annotators carefully structured the explanations into well-formed sentences and made modifications to ensure that they were appealing and engaging to read. **(5)** Annotators also classified memes based on their type and determined whether ancillary insights were required to understand them. **(6)** We made efforts to eliminate duplicate memes from our dataset and conducted a thorough review of all instances after curation. *Any memes containing vulgar, racist, or discriminatory content were immediately discarded to maintain the ethical standards of our model.* (c.f Section A.1). **(7)** Once the dataset was finalized, we reviewed the explanations again to ensure they met our standards of ethics, clarity and engagement.

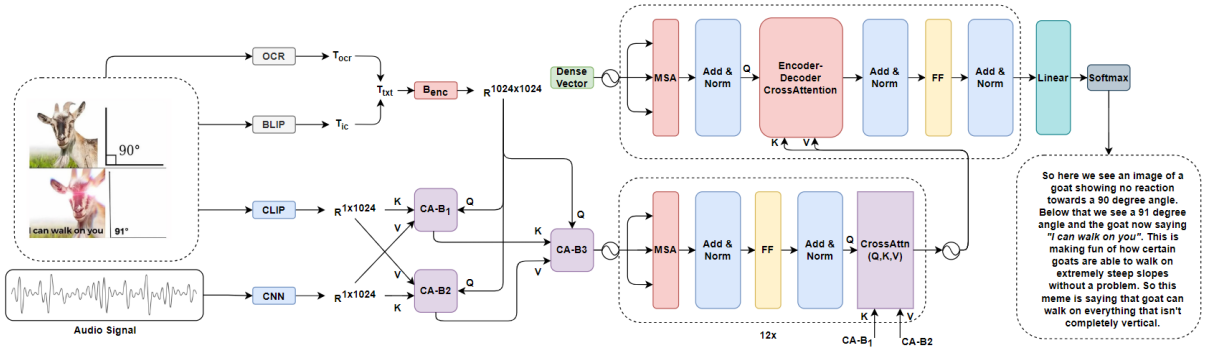


Figure 4: A framework that enriches the fused information from text, audio, and visual modalities through various cross-attention blocks before or within the BART encoder. Ensuring integration of multi-modal cues, facilitating the generation of a supervised summary of the meme.

4 Methodology

We have extracted features from the different modalities separately.

Image Processing: We extracted the dense vector representation of images from a Vision Transformer, Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021) and passed through a linear function reshaping the vector to $(1, 1024)$, denoted as $\mathbf{R}_{img}^{1\times1024}$, which was pretrained on a large-scale dataset comprising 400 million image-text pairs collected from the internet. CLIP embeddings facilitated understanding facial expressions, emotions, object recognition, etc. Moreover, due to supervised learning, it was able to create a correlations between similar templates used in multiple memes as shown in Figure 3, resulting in robust results.

Audio Processing: The audio segments accompanying memes typically ranged from 5 to 10 seconds in duration and exclusively featured music, devoid of any dialogues or lyrics, contributing to the overall humor atmosphere. To effectively process this audio data, we employed Mel-Frequency Cepstral Coefficients (MFCCs), which encapsulate information regarding rate changes across various spectrum bands. This approach enabled our model to discern the tonal nuances of the meme content and seamlessly integrate audio cues into the summarization process. The vector representation of the audio after passing through linear function is denoted as $\mathbf{R}_{aud}^{1\times1024}$.

We utilize BLIP (Bootstrapping Language-Image Pre-training) (Li et al., 2022), which produces attribute-centric captions T_{ic}, and Optical Character Recognition (OCR) for text identification T_{ocr}. These outputs are then passed through a BART encoder (Lewis et al., 2019) (B_e) to generate vector representations for the text modality.

$$\mathbf{R}_{txt}^{1024\times1024} = B_e\left(T_{ocr} \cdot T_{ic}\right) \quad (1)$$

Simultaneously, we process the image and audio to obtain $\mathbf{R}_{img}^{1\times1024}$, $\mathbf{R}_{aud}^{1\times1024}$. To fuse all vector representations of all the three modalities, we employ the Cross-Attention mechanism three times to ensure proper fusion of the embeddings. In the first configuration, textual embeddings serve as the Query (Q_{txt}), image embeddings as the Key (K_{img}), and the audio vector as the Value (V_{aud}), feeding into Cross Attention Block 1 (CA-B1), the resultant vector is represented by α. In the second configuration, textual embeddings serve as the Query (Q_{txt}), audio as the Key (K_{aud}), and image vector as the Value (V_{img}), feeding into Cross Attention Block 2 (CA-B2). The resultant vector is represented by β. The output from CA-B1 serves as the Key (K_α), and the output from CA-B2 as the Value (V_β), in the final Cross Attention Block 3 (CA-B3). The main difference between CA-B1 and CA-B2 lies in which modality provides the contextual information (Key) and which modality is influenced by it (Value). In the first configuration, the model attends to extract features from the image based on the text query and integrates information from the audio input.

$$\alpha_{1024,1024} = Attention\left(Q_{txt}, K_{img}, V_{aud}\right) \quad (2)$$

In the second configuration, the model attends to extract features from the audio input based on the text query and integrates information from the image input.

$$\beta_{1024,1024} = Attention\left(Q_{txt}, K_{aud}, V_{img}\right) \quad (3)$$

After fusion, the outputs from both the blocks are combined with the Text Query, CA-B1 key (K_α),

and CA-B2 value (V_β). This ensures proper binding of all three modalities thrice in all the cross attention blocks, helping the model to learn from a richer binded representation, with more preference given to text data because BART is well-equipped with textual data, while the importance of audio and image features is learned during supervised training of the model.

$$\gamma_{1024,1024} = Attention\left(Q_{txt}, K_\alpha, V_\beta\right) \quad (4)$$

The fused embeddings are then applied with positional encoding as described in (Vaswani et al., 2017) before passing to the BART encoder. ($\gamma'_{1024,1024}$)

$$r_{(k,2i)} = \sin\left(\frac{k}{10000^{2i/100}}\right)$$

$$r_{(k,2i+1)} = \cos\left(\frac{k}{10000^{2i/100}}\right)$$

$$\gamma'_k = \gamma_k + r_{k/i}, \quad (5)$$

Additionally, cross-modal attention is introduced at the end of every encoder, where the dense vector from the last Layer is passed as the Query (Q_σ), the representation from CA-B1 as the Key (K_α), and from CA-B2 as the Value (V_β).

$$Y_{1024,1024} = Attention\left(Q_\sigma, K_\alpha, V_\beta\right) \quad (6)$$

The inclusion of cross-attention at the end of every encoder has been found to enhance the performance compared to scenarios without cross-attention or only including it in the last encoder.

5 Evaluation Metrics

We evaluated our model using both Automatic metrics and Human Evaluation schemes. For Automatic Evaluation, we selected BLEU (Papineni et al., 2002) for its effectiveness in measuring translation quality, CIDEr (Vedantam et al., 2015) for its focus on capturing consensus between human judgments and model predictions, METEOR (Banerjee and Lavie, 2005) for its robustness to lexical variations, BERTScore (Zhang et al., 2019) for its ability to assess fluency and coherence, and Distinct-N (Li et al., 2015) for quantifying diversity in generated outputs. Additionally, we employ ROUGE (Lin, 2004) for their capability to evaluate text summarization quality. To capture semantic similarities between words we utilize Embedding based metrics (Rus and Lintean, 2012; Landauer and Dumais, 1997; Forgues et al., 2014). Automatic evaluation metrics are reliable and applicable to meme summarization, but they do not capture all necessary information. To address this, we manually evaluate the clarity, engagingness, and faithfulness of the generated summaries. **Clarity** measures how clear and concise the summary is and whether it effectively communicates the meme's message. **Engagingness** assesses how well the summary maintains the user's interest and evokes reactions similar to the original meme. **Faithfulness** Assesses if the summary reflects only the meme's content, without any fabricated elements, measured as the percentage of samples without errors. Scores for clarity and engagingness range from 1 (poor) to 5 (excellent). Further elaborated in section A

6 Experiments and Results

We experimented with our model using various combinations of modalities and architectural modifications. Additionally, we compare the results with those obtained from OpenAI's GPT-4 Vision(Achiam et al., 2023), Google's Gemini(Team et al., 2023), LLaVA(Liu et al., 2023) and QwenVL(Bai et al., 2023) zero-shot approach. The prompt used was: *"The bot is provided with an image that is a meme. Bot has to provide a summary of the meme, capturing its humor and what makes it humorous. The summary should be clear and engaging."* In our exploration of model architectures, One of our approaches involves employing a single cross-attention mechanism, where the query was the text input (Q_{txt}), the key was the image input (K_{img}), and the value was the audio input (V_{aud}), passing the resulting dense vector to the BART encoder, along with positional encoding. We observed **20.7%** decrease in performance of ROUGE-L for MeSum illustrated in Table 4. Additionally, we conducted experiments where we removed the cross-attention mechanism from the end of the BART encoder. This modification resulted in drop of **64%** ROUGE-L score illustrated in Table 4. The traditional architecture where the query is always the text representation ($\mathbf{R}^{1024\times1024}txt$) features CA-B1 with key-value pairs from the image ($\mathbf{R}^{1\times1024}img$), CA-B2 with key-value pairs from the audio ($\mathbf{R}^{1\times1024}_{aud}$), and CA-B3 as the only attention block binding all three representations together. This configuration results in a **6.55%**

Model	bleu	cider	meteor	dist_1	dist_2	bert_F1	rouge_L	E_avg	E_Grdy	E_extrm
GPT-4 V	0.030	0.079	0.251	0.050	0.295	0.551	0.213	0.974	0.669	0.976
Gemini	0.036	0.167	0.217	0.089	**0.396**	0.562	0.259	0.975	0.679	0.976
LLaVA	0.017	0.052	0.184	0.089	0.386	0.504	0.177	0.964	0.653	0.972
QwenVL	0.028	0.076	0.196	0.111	0.327	0.532	0.198	0.967	0.661	0.972
Text	0.127	1.052	0.302	0.112	0.386	0.575	0.351	0.957	0.696	0.968
Image	0.076	0.853	0.243	0.131	0.336	0.521	0.288	0.948	0.682	0.955
Text + Image	0.154	1.290	0.322	0.109	0.387	0.621	0.407	0.966	0.725	0.972
Text + Audio	0.133	1.096	0.298	0.107	0.384	0.590	0.359	0.961	0.703	0.971
Image + Audio	0.081	0.861	0.251	0.128	0.339	0.539	0.292	0.952	0.697	0.953
MeSum(Txt+Img+Aud)	**0.161**	**1.411**	**0.381**	0.113	0.382	**0.652**	**0.439**	**0.978**	**0.764**	**0.985**
MeSum[only 1 CA]	0.131	0.355	0.286	0.112	0.382	0.602	0.348	0.984	0.732	0.979
MeSum[no CA in B_e]	0.054	0.113	0.218	0.109	0.384	0.561	0.281	0.970	0.706	0.968
MeSum[Traditional CA]	0.107	0.718	0.215	**0.135**	0.376	0.598	0.412	0.977	0.721	0.983

Table 4: ROUGE-L scores for MeSum are statistically significantly higher than all baselines. Two-tailed t-tests between MeSum and (Gemini, GPT-4 V, LLaVA, QwenVL) yield p-values < 0.05, confirming MeSum's superiority for meme summarization over state-of-the-art vision models.

lower ROUGE-L score, as illustrated in Table 4. We observed addition of modalities consistently resulted in the improvement of results. Trimodal (Text + Image + Audio) combination performing the best among all. *Our work, therefore, demonstrates the importance of all three modalities and that for binding more than two modalities, utilizing multiple cross-attention blocks before passing through the encoder and at the end of every encoder with key-value pairs from different representations of different modalities results in a richer fused representation, yielding comparatively better results.* Experimental setup can be found in Section A.2.

Impact of Audio Modality : The integration of text with images resulted in a performance gain of **15.95%**, while the subsequent addition of audio contributed an additional **7.86%** improvement in ROUGE-L scores, also improved other metrics, highlighting the importance of the audio signal, illustrated in Table 5. The extracted audio features help set the atmosphere of the generated summary. Memes that were previously misclassified when relying solely on text and image were better understood with the inclusion of audio. Often, depreciating memes that include plenty of sarcasm were confused with other categories. By correlating the audio signals with words like "sarcastic," "humorous," "nostalgic," etc., from the human-annotated meme summary during training, the model showed improvement in interpreting and categorising memes during testing, leading to better summaries.

Category	ROUGE-L		Performance gain
	(T + I)	(T + I + A)	
obsv	0.398	0.436	**9.54%**
wordplay	0.419	0.440	**5.01%**
slapstick	0.413	0.441	**6.77%**
dep	0.389	0.449	**15.42%**

Table 5: Observational (obsv), Depreciating (dep). Modalities: Text (T), Image (I), Audio (A). Audio signals significantly improved the summarization of depreciating memes, followed by observational memes.

6.1 Human Evaluation

The evaluators diligently assessed the test instances generated by GPT-4 V, Gemini, LLaVA, QwenVL and MeSum, adhering to the guidelines outlined in Sections 5,A. MeSum demonstrated notable superiority over its counterparts, as depicted in Table 6. Gemini slightly exceeded GPT-4 V in clarity but lagged in engagingness. QwenVL performed similarly to GPT-4 V in automatic metrics but scored lower in engagingness during human evaluation. LLaVA received the lowest scores in both automatic and human-defined metrics. MeSum surpassed LLaVA and QwenVL in faithfulness but fell short compared to Gemini and GPT-4 V.

Models	Clarity	Engagingness	Faithfulness
GPT-4 V	2.909	3.225	91.5%
Gemini	3.495	2.621	**94.5%**
LLaVA	2.257	1.982	71%
QwenVL	2.865	2.349	84%
MeSum	**3.879**	**4.157**	85.5%

Table 6: Comparison of human evaluation results with MeSum, Gemini, GPT-4V, LLaVA and QwenVL on defined metrics.

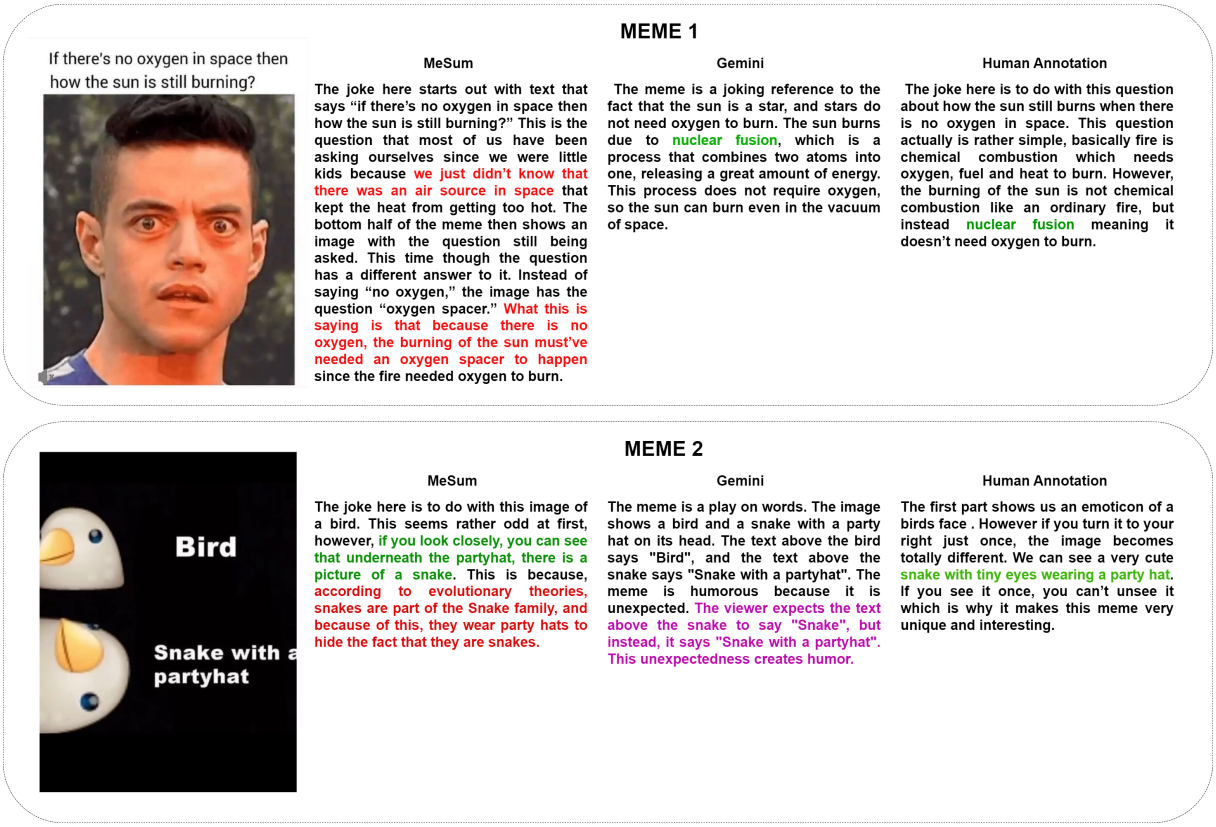


Figure 5: A succinct overview of the erroneous outputs generated by the model is provided, along with a comparison of Gemini and GPT Vision's performance on these instances.

7 Conclusion :

In this research, we summarize memes using MeSum, capable of processing all three modalities (text, image, and audio) to provide comprehensive explanations. Our contribution includes the creation of the largest **M**ultimodal **M**eme **D**ataset (MMD), featuring **2.1** times more instances and **9.5** times more words in explanations compared to the existing largest dataset *MemeCap*. MMD is manually annotated with summaries and includes classification and categorization of memes based on their properties. We deliberately opted not to utilize a vision-language model due to its high computational demands and resource consumption. Instead, we devised a method that effectively harnesses a large language model with fewer parameters. This approach involves fine-tuning the model while appropriately integrating various modalities to achieve efficient performance.Our experiments achieve state-of-the-art results, outperforming zero-shot vision-language models such as Google's Gemini, OpenAI's GPT Vision, QwenVL, and LLaVA by **72.83%**, **106.10%**, **121.71%**, and **148%** on ROUGE-L, and by **347.22%** (BLEU) and **16.01%** (BERT-F1) when comparing MeSum to the best baseline, Gemini. The success of our method underscores the importance of considering multiple modalities. In our future research endeavors, we plan to explore the domain of meme creation. By analyzing existing meme templates and categories, we aim to identify patterns and trends that can inform the generation of new memes. To address the issue of hallucination, we plan to experiment with retrieval-augmented generation techniques, which we believe will significantly enhance the accuracy and reliability of the generated content.

8 Limitation

We conducted a detailed analysis to identify Limitations in MeSum. We identified three key weaknesses: **Limited Ancillary insight:** In Figure 5 (Meme 1), the model should have connected the concept of the burning sun with knowledge of nuclear fusion. However, due to being pretrained on a smaller dataset compared to Gemini, our model failed to provide an accurate explanation, while Gemini succeeded. **Hallucination:** In Figure 5 (Meme 1) and Figure 5 (Meme 2), our model produced incorrect theories when unable to accurately

predict the reason. These hallucinations (marked in red) were mistaken as explanations for the memes. Although MeSum performed better than Gemini in inferring Figure 5 (Meme 2) (marked in green), it still struggled to fully comprehend it. Conversely, Gemini provided incorrect explanations (marked in purple). **Faithfulness:** The faithfulness of a summary is critical to ensuring that generated content remains true to the original material. In Figure 5 (Meme 1), MeSum introduced hallucinated or fabricated content, highlighted in red, resulting in unfaithful summary. Conversely, Gemini provided more faithful summary by accurately reflecting the meme's content. Faithfulness is measured by the percentage of samples free from hallucinated content, with a higher percentage indicating better alignment with the original material (c.f Table 6).

9 Dataset Collection and Ethical Considerations:

The dataset utilized in this study has been collected and annotated with meticulous attention to ethical norms and considerations.To filter any inappropriate content, we followed a two-step approach, starting with an automatic process followed by a manual review. In the first step, inspired by the filtering process of MemeCap (Hwang and Shwartz, 2023), we passed the OCR and image captions through Google's banned word list[4] and filtered out images with sexual content if the NudeNet Classifier[5] returned an unsafe score higher than 0.9. In the second step, we instructed the annotators to discard any memes that, while annotating the summaries, exhibited even the slightest hint of racism, sexism, vulgarity, or discriminatory content. Finally, all the memes were reviewed by the authors to ensure no compromise was made with ethical standards.

References

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2, page 168.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

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

EunJeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes. *arXiv preprint arXiv:2305.13703*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, et al. 2021. The hateful memes challenge: Competition report. In *NeurIPS 2020 Competition and Demonstration Track*, pages 344–360. PMLR.

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. *CoRR*, abs/2005.04790.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

[4] https://github.com/coffee-and-fun/google-profanity-words

[5] https://github.com/notAI-tech/NudeNet

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

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

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

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

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.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022. Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation. *arXiv preprint arXiv:2205.12617*.

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

Vasile Rus and Mihai Lintean. 2012. An optimal assessment of natural language student input using word-to-word similarity metrics. In *Intelligent Tutoring Systems: 11th International Conference, ITS 2012, Chania, Crete, Greece, June 14-18, 2012. Proceedings 11*, pages 675–676. Springer.

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.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023a. 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, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022a. Detecting and understanding harmful memes: A survey. *Preprint*, arXiv:2205.04274.

Shivam Sharma, Atharva Kulkarni, Tharun Suresh, Himanshi Mathur, Preslav Nakov, Md Shad Akhtar, and Tanmoy Chakraborty. 2023b. Characterizing the entities in harmful memes: Who is the hero, the villain, the victim? *arXiv preprint arXiv:2301.11219*.

Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022b. 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 and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132, Kyiv. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yi Zhou and Zhenhao Chen. 2020. Multimodal learning for hateful memes detection. *CoRR*, abs/2011.12870.

A Appendix

One of the objectives of our task is to bridge the gap in meme understanding due to cultural and geographical constraints. This is particularly addressed by creating a diverse dataset that captures memes from over 36 unique cultural backgrounds, comprising 11.2% of the dataset, and from 72 unique geographical locations, comprising 35.2% of the dataset, as indicated in Table 8. Furthermore, we have observed consistent results across different categories of memes in all the automatically defined metrics, as shown in Table 7, indicating that

MeSum does not encounter any difficulties in summarization with respect to any specific category.

	MeSum			
	observational	wordplay	slapstick	depreciating
bleu	0.163	0.160	0.179	0.166
cider	1.447	1.371	1.527	1.589
meteor	0.383	0.383	0.408	0.396
dist_1	0.088	0.087	0.083	0.091
dist_2	0.385	0.373	0.375	0.398
bert_F1	0.653	0.654	0.671	0.662
rouge_L	0.436	0.440	0.441	0.449
E_avg	0.979	0.977	0.980	0.980
E_Grdy	0.769	0.757	0.772	0.776
E_extrm	0.986	0.985	0.987	0.986

Table 7: MeSum, demonstrates consistent scores across all automatic metrics, indicating its unbiased nature towards specific categories. Additionally, audio significantly impacts performance, as evidenced in Table 5, where its exclusion leads to inconsistent and low ROUGE-L scores for certain meme types.

Human Evaluation Rules : We chose 200 samples (50 randomly chosen from each category) for human evaluation. Clarity is assessed based on understandability, grammar, completeness, structure, and coherence. Engagingness is assessed based on interest, reaction, humor, creativity, and relevance. Faithfulness is measured as the percentage of samples without errors. Scores for clarity and engagingness range from 1 (poor) to 5 (excellent):
1 (Poor): The aspect performs poorly.
2 (Fair): The aspect demonstrates some strengths but significant improvements are needed.
3 (Average): The aspect meets basic expectations but lacks notable strengths.
4 (Good): The aspect performs well, showing clear strengths.
5 (Excellent): The aspect excels, demonstrating outstanding performance and noteworthy strengths. For better understanding, refer to Table 9.

Validity of evaluation results with automatic metrics : The concern regarding the validity of automatic evaluation results arises from the fact that recent large multimodal models, such as Gemini and GPT-4V, generate high-quality outputs with paraphrased words. Given that MeSum is fine-tuned to generate words found in the dataset, it may produce captions that perform better on automatic evaluation metrics like BLEU and ROUGE. This brings into question the reliability of these scores. To address this issue, we incorporated the BERTScore (BERTF1) as part of our evaluation framework. BERTScore leverages pre-trained BERT models to compute similarity scores between candidate and reference texts based on contextual embeddings, rather than relying solely on exact word matches. This method captures semantic meaning and can effectively handle paraphrasing, thus providing a more robust evaluation of text quality. BERTScore operates by first encoding both the reference and candidate texts into contextual embeddings using BERT. It then computes precision, recall, and F1 scores based on the cosine similarity between these embeddings. The key advantage of BERTScore is its ability to recognize semantically similar sentences, even if they use different words or structures, thereby mitigating the issues caused by paraphrasing. Notably, MeSum performed better in the BERTF1 score as well. Moreover, we conducted human evaluations focusing on the "engagingness" and "clarity" of the generated captions. These evaluations revealed that MeSum still performed better, demonstrating its superior capability in generating meaningful and engaging content. This dual approach of using both BERTScore and human evaluation ensures a comprehensive assessment of our model's performance, addressing the limitations of traditional automatic metrics.

A.1 Annotator Guidelines and Potential Bias

The annotators' backgrounds include two PhDs in Linguistics, two PhD scholars in Computer Science, and one Master's student in Computer Science, along with two experienced annotators with a bachelor's in computer science engineering and a background in social media content creation and management. The detailed written instructions given to annotators are as follows:

A.1.1 Guidelines for Content Categorization

To maintain ethical standards, annotators followed these guidelines. Each category—vulgar, racist, and discriminatory content—was defined with specific examples to ensure accurate content assessment. Any meme falling into the following categories was to be immediately discarded.

Vulgar Content Definition: Vulgar content includes any form of language, imagery, or behavior that is explicit, offensive, or designed to shock, insult, or degrade individuals or groups.
Examples:

- *Profanity:* Use of obscene language or slang recognized as inappropriate or offensive. **Example:** Foul language or expletives.

- *Explicit Sexual Content:* Visual or textual content depicting sexual acts, nudity, or explicit sexual behavior. **Example:** Images of nudity or sexually suggestive behavior.
- *Derogatory Insults:* Use of language intended to demean individuals. **Example:** Derogatory remarks aimed at any individual or group.

Racist Content Definition: Racist content involves derogatory expressions targeting individuals or groups based on race, ethnicity, or national origin, which could incite hatred or discrimination.

Examples:

- *Racial Stereotyping:* Content that perpetuates stereotypes about racial or ethnic groups. **Example:** Memes that depict negative racial stereotypes.
- *Hate Speech:* Any text or imagery that encourages violence or hostility towards individuals based on their race or ethnicity. **Example:** Inciting violence against a racial group.
- *Racial Slurs:* Use of offensive language or slurs aimed at racial or ethnic groups. **Example:** Derogatory terms used against specific races.

Discriminatory Content Definition: Discriminatory content targets individuals based on characteristics such as gender, sexual orientation, religion, or disability, often reinforcing negative stereotypes or biases.

Examples:

- *Gender Discrimination:* Content that reinforces harmful gender stereotypes or biases. **Example:** Memes that portray women or men in a demeaning or stereotypical manner.
- *Sexual Orientation Bias:* Content that mocks or excludes individuals based on sexual orientation. **Example:** Jokes or images that discriminate against LGBTQ+ individuals.
- *Religious Intolerance:* Content that ridicules or denigrates religious beliefs or practices. **Example:** Memes that mock religious symbols or practices.
- *Disability Discrimination:* Content that belittles or excludes individuals with disabilities. **Example:** Derogatory references or imagery related to physical or mental disabilities.

A.1.2 Potential Bias

While Pre-trained Language Models (PLMs) like BART are advantageous for various natural language processing tasks, they can introduce biases present in their training corpora (Gallegos et al., 2023; Navigli et al., 2023). Despite efforts to mitigate bias, it is challenging to completely eliminate biased or discriminatory content in the model's representations.

A.2 Parameters and Computational Resources

In our experiments, we used the GELU activation function and the Adam optimizer, with a batch size of 8 and training over 70 epochs. The dataset was split into 80% for training, 10% for validation, and 10% for testing. We ensured the model parameters were trainable by setting them to unfrozen. The learning rate was set at 3×10^{-6} , and a weight decay of 0.001 was applied for regularisation. We employed a grid search to determine the optimal parameters. The experiments were conducted on a 40GB A100 GPU, taking approximately 12-14 hours per session.

Location	Instances	Location	Instances	Culture	Instances
United States	1108	Texas	16	Western	307
Australia	294	Alabama	16	American	180
United Kingdom	280	Rome	16	Consumerism	131
America	263	Arctic	16	Satirical	110
Europe	256	North Pole	16	British	75
Japan	240	Ireland	15	Resilience	67
Italy	180	California	14	Materialistic	64
Africa	126	Atlantis	14	English	58
Western Europe	111	Singapore	13	Capitalism	57
France	103	Finland	13	Superstition	49
India	103	Vatican City	12	Italian	46
England	103	Korea	12	Environmentalism	34
China	102	Norway	11	Influencer	29
Asia	100	Brazil	11	Traditional	28
Russia	93	London	11	Ancient Egyptian	25
Germany	76	South Korea	10	Generational	23
Canada	72	Ukraine	9	Mexican	23
Greece	70	Belgium	9	Japanese	22
Egypt	70	South America	9	Spanish	22
Antarctica	60	Paris	8	Southern	17
Middle Earth	57	Britain	8	Philosophical	16
Mexico	54	Nigeria	7	Astrology	16
Florida	47	Himalayas	7	Masculinity	14
Caribbean	43	West Virginia	6	Feminism	13
Netherlands	42	Las Vegas	6	Jewish	12
Spain	42	Oregon	6	Tribal	12
New York	40	New Zealand	6	British Monarchy	10
Sweden	40	Greenland	6	Domestication	10
Denmark	38	Colorado	6	Mughal	9
Hawaii	38	Argentina	6	Japanese art	8
Silicon Valley	36	Pakistan	5	Biblical	6
Switzerland	26	Zimbabwe	5	Brazilian	6
Austria	21	Tibet	3	Dutch	6
Mount Everest	20	Turkey	3	Greek	5
Scotland	20	Malaysia	3	Arabian	5
North America	18	Malaysia	3	Polish	2

Table 8: Columns 1 and 2 illustrate geographical diversity, indicating locations directly or indirectly referenced or hinted at, or the corresponding target audience. Column 3 displays various cultural instances directly or indirectly referenced or hinted at, or the corresponding target audience.

Generated Summary	Score
Kintsugi ("golden repair" or "golden joinery") is the Japanese art of repairing broken pottery with lacquer dusted or mixed with powdered gold or similar material, highlighting the cracks instead of disguising them. This means that in Japan, broken objects are often repaired with gold. The person in this meme now goes to Japan, thinking they would cover his whole body in gold because he is so broken (maybe because he is depressed).	5 (Excellent Clarity) Comments : Clearly understandable.
Kintsugi is the Japanese art of repairing broken pottery with lacquer and gold, highlighting the cracks. In Japan, broken objects are often fixed with gold. The person in this meme goes to Japan, thinking they would cover his body in gold because he is broken and maybe depressed.	4 (Good Clarity) Comments : Captures key points with slightly less details.
Kintsugi is a Japanese art of fixing broken pottery with gold, highlighting the cracks. In Japan, broken objects are fixed with gold. The person in the meme goes to Japan, thinking they will cover his body in gold because he is broken and maybe sad.	3 (Average Clarity) Comments : Some unclear phrases, Captures key points with slight less detail.
Kintsugi is a Japanese art of fixing things with gold. Broken things are fixed in Japan. The person in the meme thinks they will cover him in gold because he is broken and sad.	2 (Fair Clarity) Comments : Somewhat understandable, requires effort, awkward phrasing, Weak structure.
Kintsugi is fixing with gold. Japan fixes broken things. The meme person goes to Japan, thinks he will be covered in gold because he is broken and sad.	1 (poor Clarity) Comments : Confusing, unclear phrasing, Weak structure, logical gaps.
Kintsugi ("golden repair" or "golden joinery") is the Japanese art of repairing broken pottery with lacquer dusted or mixed with powdered gold or similar material, highlighting the cracks instead of disguising them. This means that in Japan, broken objects are often repaired with gold. The person in this meme now goes to Japan, thinking they would cover his whole body in gold because he is so broken (maybe because he is depressed).	5 (Excellent Engagingness) Comments : Maintains Interest throughout the summary.
In Japan, broken objects are often fixed with gold, a practice known as Kintsugi, which highlights the cracks with lacquer and gold. This meme depicts a person who goes to Japan, believing they will cover their body in gold due to feeling broken and perhaps depressed.	4 (Good Engagingness) Maintains interest, though less vivid.
Japanese art of fixing broken pottery with gold, highlighting the cracks. In Japan, broken objects are fixed with gold. The person in the meme goes to Japan, thinking they will cover his body in gold because he is broken and maybe sad.	3 (Average Engagingness) Comments : Lacks vividness, Lacks Relevant terminologies.
Kintsugi is a Japanese art of fixing things with gold. The person in the meme thinks they will cover him in gold because he is broken and sad.	2 (Fair Engagingness) Comments : Very basic explanation, Repetitive.
Japan fixes broken things. The meme person goes to Japan, thinks he will be covered in gold because he is broken and sad.	1 (poor Engagingness) Comments : Dull and basic explanation, Fails to maintain interest, Repetitive.

Table 9: Further elaboration on the clarity and engagingness metrics concerning MEME-D in Figure 1, along with some useful comments that were the basis for the score it received.