What Do You MEME? Generating Explanations for Visual Semantic Role Labelling in Memes

Shivam Sharma [1,4], **Siddhant Agarwal** [1], **Tharun Suresh** [1], **Preslav Nakov** [2], **Md. Shad Akhtar** [1], **Tanmoy Chakraborty**[3]
[1]Indraprastha Institute of Information Technology Delhi, India
[2]Mohamed bin Zayed University of Artificial Intelligence, UAE
[3]Indian Institute of Technology Delhi, India
[4]Wipro AI Labs (Lab45), India
{shivams, siddhant20247, tharun20119, shad.akhtar}@iiitd.ac.in, preslav.nakov@mbzuai.ac.ae, tanchak@iitd.ac.in

Abstract

Memes are powerful means for effective communication on social media. Their effortless amalgamation of viral visuals and compelling messages can have far-reaching implications with proper marketing. Previous research on memes has primarily focused on characterizing their affective spectrum and detecting whether the meme's message insinuates any intended harm, such as *hate, offense, racism*, etc. However, memes often use abstraction, which can be elusive. Here, we introduce a novel task – EXCLAIM, generating explanations for visual semantic role labeling in memes. To this end, we curate ExHVV, a novel dataset that offers natural language explanations of connotative roles for three types of entities – *heroes*, *villains*, and *victims*, encompassing 4,680 entities present in 3K memes. We also benchmark ExHVV with several strong unimodal and multimodal baselines. Moreover, we posit LUMEN, a novel multimodal, multi-task learning framework that endeavors to address EXCLAIM optimally by jointly learning to predict the correct semantic roles and correspondingly to generate suitable natural language explanations. LUMEN distinctly outperforms the best baseline across 18 standard natural language generation evaluation metrics. Our systematic evaluation and analyses demonstrate that characteristic multimodal cues required for adjudicating semantic roles are also helpful for generating suitable explanations.

Introduction

In recent years, memes have become essential for communicating complex ideas and discussing societal issues. Besides being a developing phenomenon on the Internet, memetic visual recipes constantly mutate by adapting to the ever-evolving disparate social outlook, rendering them abstruse. Due to their design accessibility, memes are being increasingly disseminated to cause various kinds of harm (Pramanick et al. 2021a), including hate (Kiela et al. 2020a), offense (Shang et al. 2021), trolling (Suryawanshi and Chakravarthi 2021), etc. Recently, there have been developments towards automatically detecting such memes (Sharma et al. 2022b). However, the limitations start materializing while contemplating aspects like multimodality and contextual dependency.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

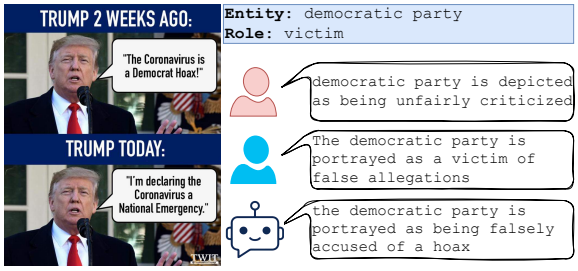

Figure 1: A novel task of EXCLAIM. Given a meme, a referred entity and its semantic role, generate a natural language explanation that emulates human-like reasoning.

For example, it is challenging to detect the multimodal narrative framing of social entities in memes (Sharma et al. 2022a,c), and to model the inherent contextual reasoning.

We attempt to leverage the features learned while detecting semantic role labels: *hero, villain*, and *victim* for memes towards generating plausible natural language explanations. In doing so, we aim to contextualize memes towards critical social-media based use-cases like *content moderation* and *digital marketing*. Generated explanations can help contextualize role labeling and assist in the retrospective deduction for strategic insights.

Memes often implicitly ply sarcasm, satire, humor, and irony while vilifying, victimizing, or glorifying target entities. Generating explanations for such aspects is challenging due to the need for abstract reasoning and multimodal contextualization, as role labels and corresponding explanations are *complementary* to each other. As an illustration, Figure 1 depicts a meme in which *Donald Trump* is portrayed as *falsely* implicating the *Democratic Party* for concocting the ruse of 'Coronavirus', without *explicitly* stating it. Therefore, *Donald Trump* needs to be adjudicated as a *villain*, based on the meme's semiotics involving his *ironic stance reversal* in the depicted scenario. Consequently, *Democratic Party's* role as a *victim* needs to be ascertained based upon the *sarcasm* implied. EXCLAIM asks for an explaining agent to factor in such intricacies while generating a suitable explanation for the aforementioned semantic roles.

The textual modality is well known to be instrumental to-

wards detection of *hateful memes* (Kiela et al. 2020a), *harmful memes, and their target types* (Pramanick et al. 2021a,b), *meme emotion* (Sharma et al. 2020), *intra-modal incongruity* (Pan et al. 2020), and *semantic role labels* for *multiple* affective connotations in memes (Sharma et al. 2022c). On the other hand, a rare task exhibiting visual influence is *detecting harmful entity* (Sharma et al. 2022a), which encompasses a *single* affective connotation. Moreover, the image modality has been observed to under-perform in detecting meme emotions (Singh, Bauwelinck, and Lefever 2020; Ruiz et al. 2020). Therefore, modality-influence typically varies across tasks. However, can something similar be generalized for the task complexity or modality-specific output configuration, like that solicited by EXCLAIM?

Bateman (2014) studied the text–image relation (Barthes and Heath 1978) graphically as a systematic network, and highlighted that text *amplifies* the image, while an image *inhibits* it. Memes were later deconstructed into three parts: (i) an initial verbal component influence *generic* users, (ii) the visual part for *experienced* users, and (iii) a conclusive second verbal part for the twists. This reinstates visual obscurity, emphasizing *abstract reasoning* for critical memetic analysis. We aim to investigate contextual augmentation for EXCLAIM via shared learning from multiple related objectives (Lee and Shen 2022).

To summarize, we benchmark our proposed dataset, ExHVV using several unimodal and multimodal baselines, typically involving uni/multi-modal encoder-decoder architectures. We further propose LUMEN, a multimodal encoder-decoder framework that incorporates the entity-specific role-label information and explanation generation capability via multi-task learning. We compare the performances of these systems using multiple standard natural language generation (NLG) evaluation measures to assess their generation quality. We finally divulge LUMEN's limitations while highlighting the challenges posed towards addressing EXCLAIM. Our contributions are summarized as follows:[1]

1. EXCLAIM: A novel task formulation, soliciting explanation generation for semantic role labelling in memes.
2. ExHVV: A multimodal dataset comprising natural language explanations accompanying the sets of memes, entities, and semantic roles.
3. LUMEN: A novel multimodal, multi-task learning framework that facilitates shared feature learning, from semantically related tasks.
4. An extensive study of the explanation generation quality, across 18 standard NLG evaluation measures.

Related Work

This section briefly discusses relevant studies on meme analyses that primarily attempt to capture a meme's affective aspects, such as *hostility* and *emotions*, along with other noteworthy research on memes.

[1]The source code for this work can be found at https://github.com/LCS2-IIITD/LUMEN-Explaining-Memes.

Meme Analysis. Several shared tasks have been organized lately, with a recent one on detecting the hero, the villain, and the victim entities in memes (Sharma et al. 2022c). Others include troll meme classification (Suryawanshi and Chakravarthi 2021) and meme-emotion analysis via their sentiments, types and intensity prediction (Sharma et al. 2020). Notably, hateful meme detection, introduced by Kiela et al. (2020b) and later followed up by Zhou, Chen, and Yang (2021), garnered significant interest, with various solutions being proposed. A few of these efforts included fine-tuning Visual BERT (Li et al. 2019) and UNITER (Chen et al. 2020), along with using Detectron-based representations (Velioglu and Rose 2020; Lippe et al. 2020) for hateful meme detection. On the other hand, there were systematic efforts involving unified and dual-stream encoders using Transformers (Muennighoff 2020; Vaswani et al. 2017), ViLBERT, VLP, UNITER (Sandulescu 2020; Lu et al. 2019; Zhou et al. 2020; Chen et al. 2020), and LXMERT (Tan and Bansal 2019) for dual-stream ensembling. Besides these, other tasks addressed anti-semitism (Chandra et al. 2021), propaganda techniques (Dimitrov et al. 2021), harmfulness (Pramanick et al. 2021b), and harmful targeting in (Sharma et al. 2022a) and detection of memes (Sharma. and Pulabaigari. 2020; Sharma, Pulabaigari, and Das 2020).

Visual Question Answering (VQA). Early prominent work on VQA with a framework encouraging *open-ended* questions and candidate answers was done by Antol et al. (2015). Since then, there have been multiple variations observed. Antol et al. (2015) classified the answers by jointly representing images and questions. Others followed by examining cross-modal interactions via attention types not restricted to co/soft/hard-attention mechanisms (Lu et al. 2016; Anderson et al. 2018; Malinowski et al. 2018), effectively learning the explicit correlations between question tokens and localized image regions. Notably, there was a series of attempts toward incorporating common-sense reasoning in (Zellers et al. 2019; Wu et al. 2016, 2017; Marino et al. 2019). Many of these studies also leveraged information from external knowledge bases for addressing VQA tasks. General models like UpDn (Anderson et al. 2018) and LXMERT (Tan and Bansal 2019) explicitly leverage non-linear transformations and Transformers for the VQA task, while others like LMH (Clark, Yatskar, and Zettlemoyer 2019) and SSL (Zhu et al. 2021) addressed the critical language priors constraining the VQA performances, albeit with marginal enhancements.

Cross-Modal Association. Due to an increased influx of multimodal data, the cross-modal association has received significant attention lately. For cross-modal retrieval and vision-language pre-training, accurate measurement of cross-modal similarity is imperative. Traditional techniques primarily used concatenation of modalities, followed by self-attention to learn cross-modal alignments (Wang et al. 2016). Following the object-centric approaches, Zeng, Zhang, and Li (2021) and Li et al. (2020) proposed a multi-grained alignment approach, which captures the relation between visual concepts of multiple objects while simultaneously aligning them with text and additional meta-data.

Split	U.S. Politics		Covid-19		Test	Total
	Train	Val	Train	Val		
Villain	1708	217	654	78	347	**3004**
Victim	531	71	357	47	104	**1110**
Hero	276	35	185	20	50	**566**
Total	**2515**	**323**	**1196**	**145**	**501**	**4680**

Table 1: Dataset summary, tabulating the entity-specific sample counts w.r.t. different *domains* and *splits*.

On the other hand, several methods also learned alignments between coarse-grained features of images and texts while disregarding object detection in their approaches (Huang et al. 2020; Kim, Son, and Kim 2021). Later approaches attempted diverse methodologies, including cross-modal semantic learning from visuals and contrastive loss formulations (Yuan et al. 2021; Jia et al. 2021; Radford et al. 2021).

Despite a wide coverage of cross-modal and meme-related applications in general, there are still several fine-grained aspects of memes that are yet to be studied. Here, we attempt to address one such novel task: EXCLAIM.

ExHVV: Dataset Curation and Annotation

Dataset. Since EXCLAIM requires a multimodal dataset with memes, semantic role labels for entities referred, and associated natural language explanations, we leverage HVVMemes. This dataset was recently released as part of a shared task at CONSTRAINT-2022 (Sharma et al. 2022c). The original dataset, HVVMemes, was annotated considering the connotative labels for meme's entities by associating them with roles: *hero, villain, victim* or *others* across the domains of COVID-19 and US Politics. We augment HVVMemes with natural language explanations for the connotative labels provided for entities present in memes by re-annotating them and occasionally rectifying the role labels wherever obvious. Table 1 shows the summary of ExHVV. We consider only *hero, villain* and *victim*, as *other* cases would be either inherently ambiguous or would solicit explanations that could be out-of-scope w.r.t. this work. This resulted in the increment of entity-specific sample count in ExHVV for category *hero* by 78, *villain* by 616, and *victim* by 159. Annotators are requested to explain why an entity could be a hero, villain, or victim. This resulted in a total of 4680 labeled samples (c.f. Table 1). The explanations reasonably varied across annotators, in terms of the *vocabulary* and length owing to their different explanation styles. The subjectivity of the task is also highlighted in the annotations, as memes often lead to several valid interpretations.

Annotation Process. Two annotators were trained especially for obtaining explanations w.r.t. EXCLAIM. They were prescribed standard annotation guidelines (c.f. Table 2), which were drafted to assist them w.r.t. the task requirements and ambiguity resolution. The annotation guidelines explicitly emphasized the consideration of the meme author's viewpoint as a standardized reference. The annotators were also issued a list of verbs with certain use cases

S. No.	Annotation Guidelines
1	Explanations should consider meme author's perspective.
2	Explanations should emphasize meme's content only.
3	Annotations should be narrated/written- • Using reported speech for opinions. • In the simple present form when stating facts.
4	Explanations should be lexically diverse.
5	Cases labeled as *others* should be skipped.
6	Entity should constitute as explanation's primary subject.
7	Obvious erroneous labels should be rectified.
8	Explanation may follow a standardized format: `[Entity] [Connotative Word] [Description].`
9	Each explanation should refer to a single entity.
10	Explanations may borrow facts from the OCR text.

Table 2: Prescribed guidelines for ExHVV annotation.

to maintain consistency across the annotations. Both the annotators were briefed w.r.t. annotation guidelines by a consolidator and were tasked with annotating an initial common set of 20 random memes with 125 entities for assessing and streamlining the annotation patterns. The test set was then annotated, considering the feedback from the consolidator. The remaining memes were then equally distributed amongst the annotators, with the consolidator adjudicating ambiguous cases. These aspects were accommodated in the final dataset to ensure consistent annotations.

The annotation quality was assessed via human evaluation by explicitly capturing the semantic validity and diversity of the explanations proposed. The questions posed in this assessment, along with the possible answer choices (and corresponding scores assigned) were:

- Does the annotator provide a valid explanation? No/Partially/Yes (0/1/2), no label (-1)?
- Are the given pair of annotations diverse or not? No/Yes (0/1), invalid (no) annotation (-1)?

The average normalized *validity* and *diversity* scores captured from this assessment were 0.81 and 0.84, respectively, for both the annotators.

Methodology

This section presents the details of LUMEN, our proposed multimodal, multi-task learning framework to address EXCLAIM (c.f. Fig. 2). As exemplified while motivating EXCLAIM in Section , the role labels and the corresponding explanations are likely complementary; we hypothesize that predicting one should help infer the other. LUMEN comprises *four* key components: (a) *Visual Recognition*, (b) *Entity Semantics*, (c) *Explanation Generation*, and (d) *Role-Label Prediction*. The latter three also contribute toward modeling the sub-tasks (ST) – sequence classification (ST-1), explanation generation (ST-2), and role-label prediction (ST-3), as part of the *multi-tasking* framework in LUMEN.

The *visual recognition* module contributes via OCR-based meme's text extraction, caption generation, and visual

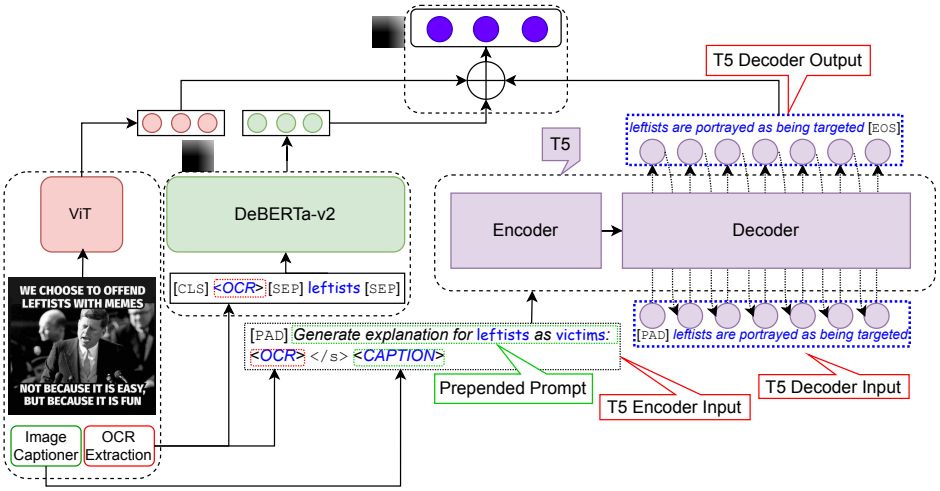

Figure 2: An illustration of LUMEN's architecture, constituting modules: (1) Visual Recognition, (2) Entity Semantics, (3) Explanation Generation, and (4) Role-Label Prediction.

feature extraction. The *entity semantics* module jointly encodes the meme's embedded text and the candidate entity while optimizing for *sequence classification* task. On the other hand, the *explanation generation* module firstly encodes a specially configured *prompt*, along with the meme text and its captions, followed by optimizing over the autoregressive decoding of the encoded hidden representations. Finally, the outputs from the first three modules are exploited towards *role-label prediction* task via their multimodal fusion. Effectively, LUMEN jointly trains for sequence classification, role prediction, and explanation generation by optimizing the joint-loss formulation from the three corresponding objectives. We explicate each of these components in detail in the following subsections.

Visual Recognition

Since memes contain everything primarily as visuals, every input characteristic cue needs to be mined utilizing some visual processing mechanism first. To this end, we first extract the meme's embedded text and perform OCR-based extraction of embedded text from a given meme image using Google's OCR Vision API We consider the meme's text entirely and discard any inherent line breaks while preprocessing. The meme's visual background in the form of imagery might not holistically capture the meme's intended message. However, it does provide a harmonizing comprehension for gaining a complete perspective on the meme's message. Evidence by Blaier, Malkiel, and Wolf (2021) suggests that utilizing meme captions improves hateful meme identification results. Furthermore, additional cues such as the person, the location, and the entities present in the meme are helpful for downstream tasks. Thus, we use such ancillary information along with the OCR text. For image captioning, we make use of the recently-released OFA model (Wang et al. 2022). Finally, we encode the meme visuals using a ViT-based model and extract the pooled output from its last hidden layer, to obtain visual embedding $\mathbf{V}_{h_i} \in \mathbb{R}^{768}$.

Entity Semantics

As discussed previously, verbal content within the memes prominently constitutes associated semantic undertones. Towards capturing the semantics from the textual content embedded within memes, we empirically designate the output from the last hidden layer of DeBERTa (He et al. 2021), as our preferred choice. Besides demonstrating its superiority for a wide range of tasks, namely MNLI, SQuAD, RACE, etc., DeBERTa has demonstrated superior performance in detecting semantic role labels for entities in memes (Kun, Bankoti, and Kiskovski 2022). We employ its `deberta-v2-xlarge` pre-trained variant (# layers: 24, H: 1536, and # parameters: 900M) for an objective similar to the latter, except for considering the 'other' category as a target semantic role within EXCLAIM formulation. While using the last hidden layer output, we also fine-tune this setup toward the sequence classification objective. The inputs to this model are configured as a pair of sequences: `[CLS] A [SEP] B [SEP]`, wherein `A` corresponds to OCR-extracted text and `B` is the candidate entity. We fine-tune this model jointly within LUMEN w.r.t. three target labels, *hero, villain* and *victim*, and obtain a mean-pooled representation for embedding *entity-semantics*: $\mathbf{T}_{h_i} \in \mathbb{R}^{1536}$, along with a multi-class classification loss ($\mathcal{L}^{\text{SEQ}}$) for the given sequence,

$$\mathcal{L}^{\text{SEQ}} = -\sum_{c=1}^{3} y_{o,c} \log(p_{o,c}) \quad (1)$$

Explanation Generation

Generating natural language explanations conditioned on complex multimodal cues requires optimal modeling of entity-specific semantic role w.r.t. a given meme. To fully leverage the required information and compensate for the missing modality (Ma et al. 2021) induced due to obscure memetic visuals, we formulate the required encoder's input as the combination of only text-based inputs, which inherently factors in the visual modality as well. To this end, we make use of T5, a transformer-based

encoder-decoder model, designed specifically to cater to tasks that can be reframed in *text-to-text* format (Raffel et al. 2020a). For encoder inputs, we consider meme's *OCR-extracted text* (source A) and *caption* (source B) and configure them as `A [SEP] B [SEP]`. We also prepend every input with a *task-specific prompt* as follows:

> Generate explanation for ENTITY as ROLE: OCR-TEXT [SEP] CAPTION

where X items above are replaced by the corresponding values for each sample. This is also depicted in Figure 2.

We use `t5-large` variant of T5, a 770 million parameter model checkpoint, originally pre-trained on a multi-task mixture of unsupervised and supervised tasks and was evaluated on a set of 24 tasks. We fine-tune it for the conditional generation objective along with the teacher-forcing strategy for training, and thereby obtain the T5-decoder's mean-pooled last hidden layer representation ($\mathbf{E}_{h_i} \in \mathbb{R}^{1024}$) and a language modeling loss ($\mathcal{L}^{\text{EXP}}$),

$$\mathcal{L}^{\text{EXP}} = -\log(p_{y_t}) = -\log(p(y_t|y_{<t})) \tag{2}$$

Role-label Prediction

Predicting semantic role labels for a given meme is another key sub-task we aim to incorporate as part of LUMEN design. Towards this, we fuse the representations gleaned via ViT-based visual features ($\mathbf{V}_{h_i} \in \mathbb{R}^{768}$), DeBERTa-based entity semantics ($\mathbf{T}_{h_i} \in \mathbb{R}^{1536}$) and T5-based representations obtained whilst decoding explanations ($\mathbf{E}_{h_i} \in \mathbb{R}^{1024}$). This not only takes into account modality-specific information for image and textual modalities via $\mathbf{V}_{h_i}$ and $\mathbf{T}_{h_i}$ respectively but also from the joint multimodal representation ($\mathbf{E}_{h_i}$), representing the interaction of visual description, textual context from meme and explanation-specific decoded hidden states. We first non-linearly project these representations individually into 512 dimensional space before concatenating them. Finally, the fused representation is condensed into a 512-sized embedding that represents the fused multimodal feature before eventually being linearly projected to $\mathbf{C}^{\text{OUT}} \in \mathbb{R}^3$ and utilized towards a 3-way multi-class classification. The corresponding cross-entropy loss ($\mathcal{L}^{\text{RP}}$) for semantic role prediction, culminated from this realization, is further incorporated into the overall multi-task learning objective.

$$\mathcal{L}^{\text{RP}} = -\sum_{c=1}^{3} y_{o,c} \log(p_{o,c}) \tag{3}$$

Finally, we combine the loss terms obtained from entity-semantics ($\mathcal{L}^{\text{SEQ}}$), explanation generation ($\mathcal{L}^{\text{EXP}}$), and role prediction ($\mathcal{L}^{\text{RP}}$) modeling objectives. Optimizing using this joint-loss formulation facilitates leveraging the similarities of these three related tasks, leading to enriched feature learning in LUMEN. The joint loss is configured as follows:

$$\mathcal{L}^{\text{LUMEN}} = \beta_1 \mathcal{L}^{\text{SEQ}} + \beta_2 \mathcal{L}^{\text{EXP}} + \beta_3 \mathcal{L}^{\text{RP}} \tag{4}$$

where βs are set as $\beta_1 = 0.2$, $\beta_2 = 0.5$, and $\beta_3 = 0.3$.

Baselines

We benchmark ExHVV, using several multimodal and unimodal baselines. The baselines use transfer learning and are encoder-decoder models with unimodal and multimodal pre-trained encoders and a pre-trained text-based decoder fine-tuned on ExHVV.

Unimodal Text-only. T5 (Raffel et al. 2020b), BERT (Devlin et al. 2018) and GPT2 (Radford et al. 2019).

Unimodal Image-only. ViT (Dosovitskiy et al. 2021) and BEiT (Bao, Dong, and Wei 2021).

Multimodal. For the multimodal (text+image) systems, we employ several combinations of these vision encoders, text encoders, and text decoders as baseline models.

Experimental Details

As part of our experimental setup, we first conduct multiple benchmarking experiments leveraging state-of-the-art systems as baselines and compare their performances with LUMEN's. Since EXCLAIM emulates the natural language generation task family, we adopt an exhaustive set of standard evaluation metrics accordingly. Metrics used for comparing baseline performances: BLEU (B-[1,4]), METEOR (M), ROUGE-L (R-L), and CIDEr (C). Further, we discuss and compare explanations generated by unimodal (image/text-only), multimodal, and LUMEN systems. Next, we divulge ablation study-based performances towards establishing the component-wise relevance. Additionally, we evaluate best-performing models across modality configurations, using additional *closeness-measuring* metrics: BERTScore (BERTs), BP, chrF, GLEU, LASER (LAS), and RIBES (RIB), and *error rates* based metrics: TER, WER, WER-D, WER-I, and WER-S. [2] The experimental observations also capture the performance change induced by the proposed system, LUMEN, w.r.t. the second best comparative baseline. Despite the metric-specific variations, the comparison is primarily made at the system level, not at the individual metric level.

Benchmarking ExHVV

Since our primary objective is to generate natural language explanations, we observe patterns contrary to what other related tasks conventionally exhibit, such as multi-class/label multimodal classification, wherein multimodal systems typically outperform unimodal ones. This performance behavior stems from the output requirement of EXCLAIM, which is text generation. Also, this performance shift has varying implications regarding the explanation generation quality, which we will elaborate on later.

Unimodal text-only systems for BERT, GPT2, and T5-based decoders open with good median scores of 0.498, 0.396, 0.313, and 0.230 for BLEU-1/2/3/4, respectively, and METEOR, ROUGE-L and CIDEr values 0.247, 0.468 and 0.840 scores, respectively. This suggests not only better linguistic grounding but also objective completeness. Semantic alignment can be better adjudged by gleaning the generated explanations. In terms of *uni-gram* overlap, the text-only

[2] We refer to the evaluation metrics for *baseline* as `gen-eval`; closeness-based comparison of top-performing systems set as $\texttt{gen-eval}_{\varsigma}$ and error-rate-based metrics as $\texttt{gen-eval}_{\vartheta}$.

Modality	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
UM	TXT-T5†	**0.509**	0.408	0.315	0.235	0.250	0.468	**1.022**
	TXT-BERT.BERT	0.498	0.396	0.313	0.230	**0.254**	0.468	0.878
	TXT-BERT.GPT2	0.316	0.205	0.136	0.086	0.140	0.296	0.319
	IMG-BEiT.GPT2	0.504	**0.410**	0.332	**0.250**	0.247	0.470	0.840
	IMG-ViT.GPT2	0.489	0.389	0.309	0.230	0.238	0.456	0.816
MM	ViT.BERT.BERT	0.498	0.404	**0.404**	0.239	0.253	**0.473**	0.890
	ViT.BERT.GPT2	0.454	0.348	0.264	0.178	0.223	0.421	0.671
	ViT.DeBERTa.GPT2	0.435	0.338	0.263	0.187	0.214	0.413	0.627
	BEiT.BERT.GPT2	0.447	0.350	0.271	0.193	0.227	0.422	0.679
	BEiT.DeBERTa.GPT2	0.445	0.350	0.274	0.198	0.222	0.427	0.706
	LUMEN	**0.578**	**0.485**	**0.399**	**0.313**	**0.294**	**0.530**	**1.380**
	$\Delta_{\text{LUMEN}-\dagger}$	↑ 6.94%	↑ 7.66%	↑ 8.38%	↑ 7.73%	↑ 4.40%	↑ 6.12%	↑ 0.36

Table 3: Performance comparison for EXCLAIM, using gen-eval metrics across unimodal, multimodal and LUMEN. Top two scores across the metrics are presented in bold; *higher* scores are better; † represents the second best model.

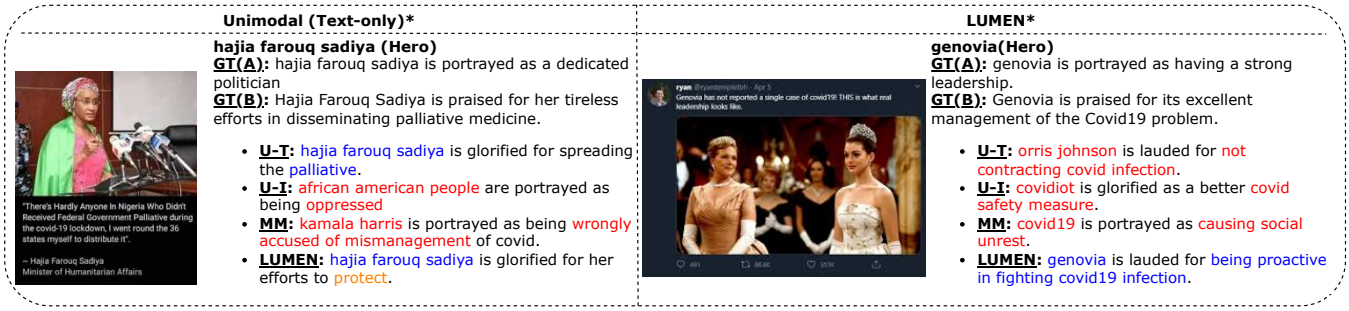

Figure 3: Illustration of explanation generation quality, for unimodal text (U-T), unimodal image (U-I), multimodal (MM), and LUMEN. Left: U-T explains better; Right: LUMEN explains better. Prediction scheme: Correct, Incorrect and Partially-correct.

Model	B-1	B-2	B-3	B-4	M	R-L	C
+ self-attend	0.554	0.453	0.365	0.277	0.282	0.506	1.240
LUMEN	**0.578**	**0.485**	**0.399**	**0.313**	**0.294**	**0.530**	**1.380**
- adafactor	0.561	0.462	0.375	0.290	0.287	0.517	1.280
- wtd loss	0.559	0.461	0.372	0.283	0.285	0.504	1.313
- captions	0.531	0.433	0.347	0.265	0.276	0.486	1.198
- T5 + GPT2	0.459	0.361	0.282	0.205	0.234	0.418	0.727
- MTL	0.435	0.338	0.263	0.187	0.214	0.413	0.627
- deBERTa-v2	0.489	0.389	0.309	0.230	0.238	0.456	0.816
- ViT	0.509	0.408	0.315	0.235	0.250	0.468	1.022

Table 4: Ablation results w.r.t. LUMEN and its components, evaluated using gen-eval. Top scores across the metrics are presented in bold; *higher* scores are better.

T5-based system subtly outperforms the BEiT-based vision-only one by 0.5%, whereas, in terms of fluency (BLEU-2/3/4), the latter is better. The T5-based unimodal model yields a sizable gain of 18% in terms of consensus-based CIDEr score, which propounds the dominance of unimodal text-only systems as being *closer* to the ground-truth explanations (c.f. Table. 3). On the other hand, multimodal baselines (except ViT.BERT.BERT based model) exhibit a significant drop of 4% median scores across the gen-eval scores. This suggests a significant downgrade in the explanation generation quality that, to a reasonable extent, can also be characterized by the gen-eval suite. Finally, LUMEN, owing to its systematic, multimodal, and multi-task regime, shows exceptional gen-eval performance.

As part of exploring MTL towards explanation generation for semantic role labeling of meme entities, our primary focus in this work is on explanation generation; therefore, we exhaustively evaluate it. As for the sequence generation task, we observe a steady decrease in validation loss from 0.1345 to 0.1322 over 15 epochs, with an average decrement rate of $1.5e-4$/epoch. For role-label prediction, we observe validation accuracies of up to 98%.

t

EXCLAIM: Analyzing Explanations

The explanations by unimodal systems are sufficiently *adequate* and *fluent* while being brief and frequently accurate; decent median BLEU-1, BLEU-4, and ROUGE-L scores of 0.498, 0.230, and 0.468, respectively, indicate this. Their integrity is demonstrated by optimal median METEOR and CIDERr scores of 0.247 and 0.840, respectively. Text-based systems are often observed to get their short yet coherent explanations correct; for example, in Figure 3 (left), the T5-based generation correctly predicts the keyword - *palliative* as the primary subject of the explanation, which other models struggle with. LUMEN, on the other hand, shows its contribution to the gen-eval suite by not only frequently generating semantically aligned explanations but also complete

Model	BERTs	BP	chrF	GLEU	LAS	RIB
UM-TXT-T5†	**0.892**	0.963	**0.322**	0.255	0.730	**0.590**
UM-IMG-BEiT.T5	0.890	**0.979**	0.312	0.266	0.725	0.563
MM-ViT.BERT.BERT	0.890	0.938	0.307	**0.267**	**0.733**	0.588
LUMEN	**0.902**	**1.000**	**0.368**	**0.280**	**0.762**	**0.610**
$\Delta_{\texttt{LUMEN}-\dagger}(\%)$	↑ 1.0	↑ 3.7	↑ 4.6	↑ 2.5	↑ 3.2	↑ 2.0

Table 5: Evaluation using gen-eval$_\varsigma$, that measures generated text's closeness with ground-truth. Top two scores across the metrics are presented in bold; *higher* scores are better; † represents the second best model.

and factually relevant ones: see Figure 3 (right).

Ablation Study

This section presents the performance details (c.f. Table 4), induced by sequentially removing critical components from LUMEN. We begin by replacing the simple concatenation-based fusion in LUMEN with popular multi-headed self-attention-based fusion and observe an average decline of 2% across gen-eval suite. This potentially indicates a self-aligning tendency of LUMEN that operates using a simple concatenation of jointly projected visual, textual, and multimodal signals. Adafactor optimizer reflects better convergence and low memory footprint (1% drop on removal). Whereas *wtd. loss* implies performing a weighted combination of the constituent loss terms (0.5% drop on removal) in multi-task learning (MTL) setup (c.f. Table 4) in LUMEN. Further, the textual captions from the meme visuals contribute crucial 2% to the gen-eval scores. Replacing the primary decoder, T5, with GPT2 brings down the average gen-eval score massively from 0.390 to 0.327, reinstating the T5's robustness for NLU tasks against that of GPT2. Multi-task learning setup demonstrates its utility with the performance reduction of 2% when the system is evaluated with only text generation objective. Finally, removing deBERTa and ViT-based components prompts a reduction of 4% and 3%, respectively, in the non-MTL configuration compared to the system under the MTL regime.

Extended Evaluation

The n-gram-based metrics constituting gen-eval overlook aspects like distributional semantics-based similarity, sentence length consideration, sub-word level overlap, multi-linguality, etc. Therefore, we further compared the best systems across modalities, including LUMEN, using gen-eval$_\varsigma$, a suite of 6 additional evaluation metrics that represent explanation generation quality w.r.t. the ground-truth explanations (c.f. Table 5). The trend observed earlier for gen-eval is reflected again in gen-eval$_\varsigma$, with the models being in the decreasing order of LUMEN, UM-TXT-T5, followed by UM-IMG-BEiT.GPT2 and MM-ViT.BERT.BERT, in terms of performance. The lead of 1.6% and 1.1% by the image-only model over the text-only model suggests the subtle closeness that image-only models exhibit.

Models	TER	WER	WER-D	WER-I	WER-S
UM-TXT-T5†	**0.874**	0.658	0.952	0.513	**0.335**
UM-IMG-BEiT.T5	0.900	0.652	**0.828**	**0.497**	0.341
MM-ViT.BERT.BERT	**0.861**	**0.634**	0.952	**0.331**	**0.332**
LUMEN	0.907	**0.634**	**0.519**	0.633	0.361
$\Delta_{\texttt{LUMEN}-\dagger}(\%)$	↑ 3.3	↓ 2.4	↓ 43.3	↑ 12.0	↑ 2.6

Table 6: Error Evaluation using gen-eval$_\vartheta$. Top two scores across metrics are presented in bold; *lower* scores are better; † represents the second best model.

Error Analysis

Despite LUMEN's stellar performance and consistent explanation generation quality, it is observed to fall short in exhibiting the required inductive biases acquired from ExHVV. This is likely due to the inherent complexity posed by multimodal interactions, which either resist accurate prediction or induce additional noise in the pipeline. Another prominent limitation discerned from LUMEN's explanations is the strong intra/cross-modal grounding, which in a few cases results in LUMEN picking up some exact verbal or visual cues from the input memes and missing out on the correspondingly required semantic coherence. The impact of these challenges can be observed occasionally within the generated explanations. As for the *error-rates* evaluated using gen-eval$_\vartheta$, interestingly, LUMEN yields increments of 3.3%, 12%, and 2.6% on TER, WER-I, and WER-S, respectively. This not only implies the complexity involved in bridging the gap between the reference sentences and the candidate explanations but could also indicate the novelty, sufficiency, and creativity observed in LUMEN's explanations. This is corroborated by 2.4% and 43% deductions, LUMEN induces in WER and WER-D scores, respectively. Also, LUMEN-generated explanations exhibit greater diversity, with 45 more unique non-stop words as against that for a unimodal text-only system.

Conclusion

We introduced a novel EXCLAIM, for generating natural language explanations for semantic role labels within memes. We first presented ExHVV, a new multimodal dataset that curates natural language explanations suited for EXCLAIM, and a benchmarking setup encompassing multiple unimodal and multimodal baselines. Next, we proposed LUMEN, a robust multimodal, multi-task learning framework to address EXCLAIM. The empirical observations reflect within the quality of generated explanations for models from different modality configurations. Besides showcasing LUMEN's adequacy and fluency, we also highlighted its inherent limitations, which constitute strong intra/cross-modal grounding w.r.t. the generated explanations and induced multimodal noise. We hope the insights from this work and the resources we release will nourish thought-provoking ideas to be explored in future work.

Acknowledgements

The work was supported by Wipro research grant.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR'18*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV'15*.

Bao, H.; Dong, L.; and Wei, F. 2021. BEiT: BERT Pre-Training of Image Transformers. *ArXiv*, abs/2106.08254.

Barthes, R.; and Heath, S. 1978. Image, Music, Text. *Journal of Aesthetics and Art Criticism*, 37(2): 235–236.

Bateman, J. A. 2014. *Text and image: A critical introduction to the visual-verbal divide*. Routledge, 1 edition.

Blaier, E.; Malkiel, I.; and Wolf, L. 2021. Caption Enriched Samples for Improving Hateful Memes Detection. In *EMNLP'21*, 9350–9358.

Chandra, M.; Pailla, D.; Bhatia, H.; Sanchawala, A.; Gupta, M.; Shrivastava, M.; and Kumaraguru, P. 2021. "Subverting the Jewtocracy": Online Antisemitism Detection Using Multimodal Deep Learning. WebSci '21, 148–157.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *ECCV'20*, 104–120.

Clark, C.; Yatskar, M.; and Zettlemoyer, L. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *EMNLP-IJCNLP'19*, 4069–4082.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Dimitrov, D.; Bin Ali, B.; Shaar, S.; Alam, F.; Silvestri, F.; Firooz, H.; Nakov, P.; and Da San Martino, G. 2021. Detecting Propaganda Techniques in Memes. In *ACL-IJCNLP'21*, 6603–6617. Online.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. In *ICLR'21*.

Huang, Z.; Zeng, Z.; Liu, B.; Fu, D.; and Fu, J. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML'21*, 4904–4916.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020a. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *NeurIPS'20*, volume 33, 2611–2624.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020b. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *NeurIPS'20*, 33.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML'21*, 5583–5594.

Kun, L.; Bankoti, J.; and Kiskovski, D. 2022. Logically at the Constraint 2022: Multimodal role labelling. In *CONSTRAINT, ACL'22*, 24–34.

Lee, G. G.; and Shen, M. 2022. Multi-modal, multi-task learning for Memotion 2.0 challenge. In *AAAI 2022 DEFACTIFY Workshop: Multi-Modal Fake News and Hate-Speech Detection*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV'20*, 121–137.

Lippe, P.; Holla, N.; Chandra, S.; Rajamanickam, S.; Antoniou, G.; Shutova, E.; and Yannakoudakis, H. 2020. A Multimodal Framework for the Detection of Hateful Memes. *arXiv preprint arXiv:2012.12871*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *NeuIPS'19*, volume 32, 13–23.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *NeurIPS'16*, 289—-97.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021. SMIL: Multimodal Learning with Severely Missing Modality. In *AAAI'21*, 2302–2310.

Malinowski, M.; Doersch, C.; Santoro, A.; and Battaglia, P. 2018. Learning Visual Question Answering By Bootstrapping Hard Attention. In *ECCV'18*, 3–20.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR'19*, 3195–3204.

Muennighoff, N. 2020. Vilio: State-of-the-art Visio-Linguistic Models applied to Hateful Memes. *arXiv preprint arXiv:2012.07788*.

Pan, H.; Lin, Z.; Fu, P.; Qi, Y.; and Wang, W. 2020. Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection. In *EMNLP'20 (Findings)*, 1383–1392.

Pramanick, S.; Dimitrov, D.; Mukherjee, R.; Sharma, S.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021a. Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2783–2796.

Pramanick, S.; Sharma, S.; Dimitrov, D.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021b. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *EMNLP'21 (Findings)*, 4439–4455.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML'21*, volume 139, 8748–8763.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020a. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020b. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Ruiz, G.; Tellez, E. S.; Moctezuma, D.; Miranda-Jiménez, S.; Ramírez-delReal, T.; and Graff, M. 2020. Infotec + CentroGEO at SemEval-2020 Task 8: Deep Learning and Text Categorization approach for Memes classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1141–1147.

Sandulescu, V. 2020. Detecting Hateful Memes Using a Multimodal Deep Ensemble. *arXiv preprint arXiv:2012.13235*.

Shang, L.; Zhang, Y.; Zha, Y.; Chen, Y.; Youn, C.; and Wang, D. 2021. AOMD: An Analogy-aware Approach to Offensive Meme Detection on Social Media. arXiv:2106.11229.

Sharma, C.; Bhageria, D.; Scott, W.; PYKL, S.; Das, A.; Chakraborty, T.; Pulabaigari, V.; and Gambäck, B. 2020. SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, 759–773.

Sharma., C.; and Pulabaigari., V. 2020. A Curious Case of Meme Detection: An Investigative Study. In *WEBIST*, 327–338. ISBN 978-989-758-478-7.

Sharma, C.; Pulabaigari, V.; and Das, A. 2020. Meme vs. Non-meme Classification using Visuo-linguistic Association. In Marchiori, M.; Mayo, F. J. D.; and Filipe, J., eds., *Proceedings of the 16th International Conference on Web Information Systems and Technologies, WEBIST 2020, Budapest, Hungary, November 3-5, 2020*, 353–360. SCITEPRESS.

Sharma, S.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2022a. DISARM: Detecting the Victims Targeted by Harmful Memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 1572–1588.

Sharma, S.; Alam, F.; Akhtar, M. S.; Dimitrov, D.; Da San Martino, G.; Firooz, H.; Halevy, A.; Silvestri, F.; Nakov, P.; and Chakraborty, T. 2022b. Detecting and Understanding Harmful Memes: A Survey. In *IJCAI-ECAI '22*, IJCAI-ECAI '22.

Sharma, S.; Suresh, T.; Kulkarni, A.; Mathur, H.; Nakov, P.; Akhtar, M. S.; and Chakraborty, T. 2022c. Findings of the CONSTRAINT 2022 Shared Task on Detecting the Hero, the Villain, and the Victim in Memes. In *CONSTRAINT, ACL'22*, 1–11.

Singh, P.; Bauwelinck, N.; and Lefever, E. 2020. LT3 at SemEval-2020 Task 8: Multi-Modal Multi-Task Learning for Memotion Analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1155–1162.

Suryawanshi, S.; and Chakravarthi, B. R. 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*, 126–132.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS'17*, 5998–6008.

Velioglu, R.; and Rose, J. 2020. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge. *arXiv preprint arXiv:2012.12975*.

Wang, K.; Yin, Q.; Wang, W.; Wu, S.; and Wang, L. 2016. A Comprehensive Survey on Cross-modal Retrieval. *CoRR*, abs/1607.06215.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. *CoRR*, abs/2202.03052.

Wu, Q.; Shen, C.; Wang, P.; Dick, A.; and Van Den Hengel, A. 2017. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1367–1381.

Wu, Q.; Wang, P.; Shen, C.; Dick, A.; and Van Den Hengel, A. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR'16*, 4622–4630.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. 2021. Florence: A New Foundation Model for Computer Vision. *arXiv preprint arXiv:2111.11432*.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR'19*, 6720–6731.

Zeng, Y.; Zhang, X.; and Li, H. 2021. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. *arXiv preprint arXiv:2111.08276*.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; and Gao, J. 2020. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, volume 34, 13041–13049.

Zhou, Y.; Chen, Z.; and Yang, H. 2021. Multimodal Learning For Hateful Memes Detection. In *ICMEW*, 1–6.

Zhu, X.; Mao, Z.; Liu, C.; Zhang, P.; Wang, B.; and Zhang, Y. 2021. Overcoming Language Priors with Self-Supervised Learning for Visual Question Answering. In *IJCAI*, IJCAI'20. ISBN 9780999241165.