

Multimodal Social Media Sentiment Analysis

Stanford CS224N Custom Project

Pratyush Muthukumar
Department of Computer Science
Stanford University
muthukup@stanford.edu

Mubarak Ali Seyed Ibrahim
Department of Computer Science
Stanford University
mubarak@stanford.edu

Abstract

We introduce `SentiModal`, a multimodal sentiment analysis model designed for detecting sentiment in social media posts. Utilizing a set of 17 custom prompts, LLaVA image captioning, and a text-only GPT-4, we create a tailored instruction tuning dataset of image-text pairs sourced from the B-T4SA tweet dataset. `SentiModal` undergoes LoRA fine-tuning on our custom dataset, commencing from the pre-trained LLaVA 1.5 Vicuna-7B checkpoint. Our findings reveal that `SentiModal` exhibits a 14.52% improvement in accuracy for multimodal sentiment classification ('positive', 'neutral', 'negative') compared to a baseline text-only GPT-4 model and a 60.27% improvement in accuracy compared to a baseline GPT-4V model. Further, we leverage a text-only language model as a judge to holistically evaluate responses from `SentiModal` and find that the mean score is 3.88 out of 5.00.

1 Key Information to include

- Mentor: Caleb Ziems (cziems@stanford.edu)
- External Collaborators (if you have any): None
- Sharing project: No

1.1 Team Contributions

- **Pratyush Muthukumar**: Brainstorming, conceptualization, design, prompt template design, helped with data curation pipeline implementation, led evaluation pipeline implementation, helped draft project proposal, project milestone, project report, poster
- **Mubarak Ali Seyed Ibrahim**: Brainstorming, conceptualization, design, led data curation pipeline implementation, helped with evaluation pipeline implementation, fine-tuning implementation, helped draft project proposal, project milestone, project report, presented poster

2 Introduction

Sentiment analysis is a long-standing and critical task centered in the foundations of NLP, and while sentiment analysis has been tackled with a number of increasingly robust architectures in recent years, we believe that multimodal sentiment analysis is a relatively less studied field. Specifically, multimodal understanding and generation from images, videos, audio, and text have recently shown promise in the generative AI research space, showing promise for world-scale adoption within the massively popular LLMs people interact with nowadays.

This work, and related work focusing on multimodal sentiment analysis (Yuan et al., 2024) or multimodal understanding (Ko et al., 2023), fit into the broader theme of a shifting dynamic of traditional NLP tasks towards multimodal data. Recently, with advances in visual instruction tuning

popularized by LLaVA, OpenAI’s GPT-4V, and Google Gemini, we are seeing a trend towards focusing on multimodal content understanding and generation (Liu et al., 2023). Specifically, tasks which were solely restricted to the textual domain, such as sentiment analysis, question answering, and summarization are now problems which are reappearing in the domain of large multimodal models. Including learned embeddings from image, video, audio, or other non-textual modalities aligned with text embeddings in a large language model seems to show promise in leveraging all forms of data in these tasks.

Our work utilizes embeddings from both text and image data of tweets for sentiment analysis, which is one of many in this new wave of research towards reinventing understanding and generation tasks for multimodal data. We present `SentiModal` as a novel multimodal social media sentiment analysis model leveraging learnings from this recent wave of vision-language model research. We summarize our contributions to this ever growing field of multimodal understanding in the following ways: **(1)** We design and implement a custom dataset curation pipeline leveraging multimodal image captioning, text-only language models, and hand-crafted sentiment analysis prompt templates **(2)** We adapt existing work in multimodal image-text understanding and apply downstream LoRA fine-tuning on our custom dataset **(3)** We design and implement a custom set of evaluation pipelines leveraging a text-only language model as a judge to automatically evaluate responses from `SentiModal` against a leading text-only language model baseline (GPT-4) and a leading vision-language model baseline (GPT-4V).

3 Related Work

There are a number of recent related works leveraging the power of Vision Transformers (ViTs) and Large Language Models (LLMs) for a variety of multimodal understanding tasks. Thareja (2024) proposes `Tweet-SentiNet`, an innovative method of analysing sentiment of multimodal data. The goal of this work is to sift “emotionally charged” content from social media, essentially identifying sentiments of large magnitudes from multimodal social media data sourced from tweets consisting of both text and images. The authors hope this work improves social media guardrail systems and overall mental health, as evidenced by a qualitative human evaluation study performed as a part of their research.

The authors leverage separate image and text encoders to generate disentangled textual and image embeddings from a dataset of Twitter images and associated captions. Using a pretrained ViT, they then leverage attention to fuse the embeddings to a common representation and feed this to a DNN classifier for sentiment classification. However, the authors framed the multimodal sentiment analysis problem as a binary classification problem (i.e. extreme sentiment or not), which imposes an overly simplistic view and demonstrates the model’s limited understanding of multimodal inputs.

Other approaches leverage LLMs as a decoder for general-purpose image understanding tasks. For example, the Large Language and Vision Assistant (LLaVA) model proposed by Liu et al. (2023) represents a Vicuna-7B LLM as a decoder layer in a two-stage training scheme to first align image and textual tokens with a linear projection layer. LLaVA uses a CLIP ViT-L/14 image encoder to encode images. Then, end-to-end fine-tuning is performed to train the LLM backbone on image-text pairs. LLaVA was a pioneer in the practice of visual instruction tuning, but was trained on a general-purpose image dataset curated from the CC-3M dataset (Changpinyo et al., 2021). Due to the lack of training data specific to social media posts, LLaVA is limited in sentiment analysis capabilities for this downstream task.

Extensions to LLaVA introduce support for additional modalities, such as Video-LLaVA proposed by Lin et al. (2023), which introduces separate video and text image encoders. Video-LLaVA also finds that aligning these embeddings prior to the linear projection layer also helps with interleaving multiple modalities. However, Video-LLaVA was trained on general purpose videos sourced from the WebVid-2M dataset and suffers from a similar issue of limited training data related to social media or short-form media (Bain et al., 2021).

We propose `SentiModal` as an extension to the field of multimodal sentiment analysis bridging ideas from a number of these related works. We propose to leverage visual instruction tuning proposed in LLaVA with the domain-specific downstream task of social media sentiment analysis on image-text pairs proposed in `Tweet-SentiNet`.

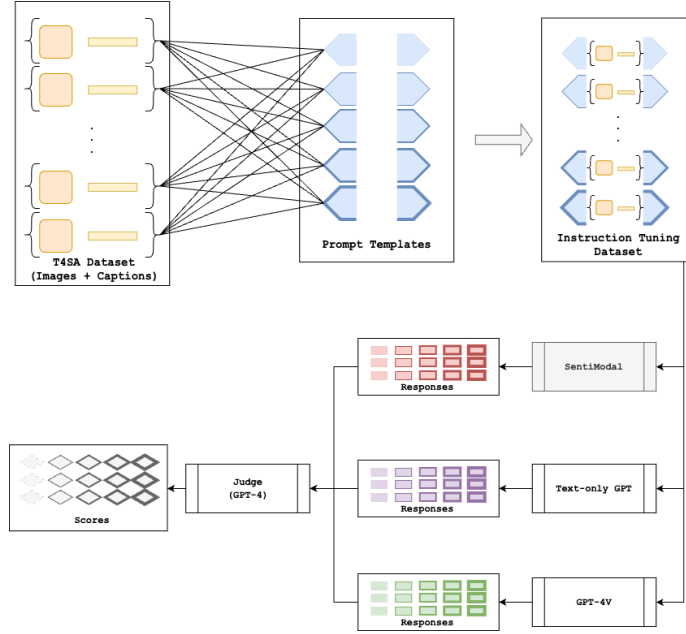


Figure 1: Experiment Architecture

4 Approach

We opted to investigate the efficacy of employing a vision-language model for sentiment analysis on multimodal data, comprising tweet images and their corresponding captions. To meet this objective, we adhered to a two-stage training pipeline, similar to LLaVA Liu et al. (2023), which involved pre-training and instruction tuning. However, we curated our own dataset specifically tailored for multimodal sentiment analysis during the instruction tuning phase. Figure 1 describes the complete architectural setup.

We started with a set of pre-defined prompt templates which we used as templates to construct our instruction tuning dataset. A couple examples are shown here:

1. **Prompt:** Can you analyze this MM_IMG? I'm curious about the sentiment.
Response: Sure thing! Taking a look at it, the sentiment appears to be SENTIMENT.

2. **Prompt:** What can you tell me about this image here MM_IMG.
Response: Based on the image you provided, the overall mood of the post seems to be SENTIMENT.

Additional prompt templates can be found in the Appendix. From these prompt templates, we leveraged a pre-trained LLaVA-7B model variant to generate image captions to replace MM_IMG with the text description of the image and used the label from the sentiment analysis dataset to replace SENTIMENT. We then performed instruction tuning on 83% of this curated dataset to generate a training set of 25K+ image-text pairs and 4K+ evaluation samples. We devised an evaluation pipeline for evaluating SentiModel against baselines of (1) a text-only language model such as GPT-4 on the rest of the prompt templates with text captions of images and (2) a vision language model such as GPT-4V on the rest of the prompt templates with the raw image. In our pipeline we leverage a text-only GPT-4 to act as a judge to evaluate responses between the two baselines and SentiModel.

Our approach is novel compared to other VLM approaches such as LLaVA or GPT-4V because of the uniquely curated instruction tuning dataset we generate from a set of multimodal sentiment question answer prompts derived from a large tweet sentiment dataset.

5 Experiments

We have applied the LLaVA architecture and fine-tuned it for the downstream task of multimodal sentiment analysis from open-source multimodal social media content datasets. Specifically, the input to our model will be a set of prompts specific to sentiment analysis, i.e. Describe the sentiment of this social media post from one or multiple of the following categories: happy, sad, funny, angry, hateful, constructive, sarcastic or Does this social media post exhibit signs of extreme sentiment, yes or no?. These example prompts are written to mimic the original tasks outlined by the authors in Thareja (2024) and Yuan et al. (2024) for extreme sentiment binary classification and sentiment multi-class classification, respectively. Along with each of these prompts, we provided the associated image + text caption pair as multimodal context to the prompt. The output of our model is a text output from the fine-tuned LLM answering the prompt given the multimodal context.

Throughout the development of this model, we have accomplished several iterative tasks. Initially, we constructed the dataset along with its associated data loader and pipeline to preprocess an open-source social media sentiment dataset, aligning it with the LLaVA-style architecture. Furthermore, we established various qualitative and quantitative evaluation methods. These include automated evaluation via LLMs, binary classification accuracy scores or Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) scores, and additional metrics to assess performance.

5.1 Data

We have used the Balanced Twitter for Sentiment Analysis (B-T4SA) dataset consisting of nearly 500K tweets with images and text along with their annotated sentiment (either positive, negative, or neutral) (Vadicamo et al., 2017). We have preprocessed the datasets into a format where we generate a set of associated prompts and responses for each multimodal data and annotation pair. Specifically, for a given social media datapoint with associated annotation, we used the data pipeline to transform the annotation to a question-answer pair for use with our proposed methodology. For example, for the B-T4SA dataset, if a particular tweet is labeled as positive sentiment, one post-processed question-answer pair may look something like: **Q:** Is the sentiment of this tweet positive, negative, or neutral? **A:** Positive.

5.2 Evaluation Method

We design the evaluation pipeline to leverage a number of automated agents for both quantitative and qualitative analysis of results. Specifically, we design three evaluation pipelines: two evaluation pipelines leveraging a text-only GPT as a judge and additionally one pipeline to parse responses for quantitative accuracy metric calculations (i.e. accuracy, F-1 score, AUC ROC score).

For the evaluation pipelines leveraging a text-only GPT-4 as a judge, we evaluate responses on by querying the LLM-as-a-judge on (1) the responses’ correctness and (2) an overall, holistic score based on helpfulness, accuracy, conciseness, etc. For the correctness evaluation pipeline, we adopt a similar approach described in Video-LLaVA proposed by Lin et al. (2023) to prompt a text-only GPT to evaluate a “predicted” response (Sent iModal) against two ground truth responses (GPT-4 and GPT-4V) and provide a factual accuracy score between 0 to 5. The full prompting details can be found in the Appendix. For the the overall score evaluation with an LLM-as-a-judge pipeline, we adopt a prompting technique similar to one proposed by Lin et al. (2023) in LLaVA to evaluate two “assistant” responses given a prompt and context, which serves as the ground truth. We collect results on the pair-wise comparison of GPT-4V as the first “assistant” vs Sent iModal as the second “assistant” with the text-only GPT-4 response as the ground truth context.

The final evaluation pipeline we devise seeks to compute a number of accuracy metrics by converting the responses to a three way classification prediction (either positive, neutral or negative sentiment). By extracting keywords from the responses of GPT-4V, GPT-4, and Sent iModal, we parse the textual responses of each model to scores of either -1, 0, or 1 and compute the accuracies. We compute both pair-wise comparisons across these three models (i.e. Sent iModal vs GPT-4 with GPT-4V ground truth & Sent iModal vs GPT-4V with GPT-4 as ground truth).

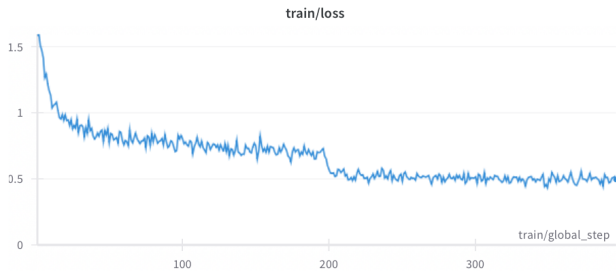


Figure 2: Training Loss of LLaVA LoRA Fine-tuning

Hyperparameter	Value
Base Model	Vicuna-7B
Rank	128
Alpha	256
Multimodal Projector LR	2×10^{-5}
Per Device Train Batch Size	2
Gradient Accumulation Steps	64
Learning Rate	2×10^{-4}
Weight Decay	0
Epochs	2
Global Batch Size	128

Table 1: Fine-tuning Hyperparameters

5.3 Experimental Details

In our assessment, we thoroughly evaluate the performance of `SentiModal` in both instruction-following and visual reasoning capabilities. This evaluation is conducted within a meticulously designed experimental framework aimed at analyzing sentiment within multimodal image data. The model is trained using an RTX 4090 equipped with 24 GB of GPU memory, adhering closely to Vicuna’s established hyperparameters. To initialize the training process, we utilize pretrained model checkpoints from `llava-v1.5-vicuna-7B` and proceed to fine-tune the model using a custom curated instruction-following dataset for a duration of 2 epochs. Throughout this fine-tuning process, the initial learning rate of $2e-4$, with a warmup ratio of 0.03, and implement a batch size of 2 with 64 gradient accumulation steps, using the AdamW optimizer with a cosine learning rate schedule. Figure 2 describes the training loss of LLaVA LoRA fine-tuning. For LoRA fine-tuning, we use LLaVA 1.5 following a similar training recipe described by the authors. Additional hyperparameter settings can be found in the Table 1. We employed DeepSpeed ZeRO-3, similar to PyTorch FSDP, incorporating gradient checkpointing. Training for two epochs takes around six hours on the RTX 4090.

Despite being trained on a relatively limited multimodal instruction-following dataset comprising around 3K unique images, `SentiModal` exhibits remarkably similar reasoning capabilities when compared to multimodal GPT-4 across various examples. It is noteworthy to mention that these images lie outside of `SentiModal`’s typical domain, yet the model demonstrates a commendable ability to comprehend the depicted scenes and effectively follow the provided instructions to generate responses that are deemed reasonable.

This evaluation underscores `SentiModal`’s adaptability and robustness, as it manages to perform competency even in scenarios where it encounters data that diverges from its training distribution. The model’s capacity to effectively handle out-of-domain inputs while maintaining reasoning accuracy highlights its potential utility across a spectrum of real-world applications. These findings not only validate `SentiModal` efficacy in instruction-following and visual reasoning tasks, but also serve to enrich our understanding of its capabilities and limitations within the broader context of AI research and development.

5.4 Results

We implemented three evaluation pipelines to thoroughly assess the quality and accuracy of `SentiModal` responses against a text-only LLM (GPT-4) and a vision language model (GPT-4V).

The results for our first evaluation pipeline using a text-only GPT-4 as a judge to score the **correctness** of `SentiModal` “predicted” responses against the ground truth GPT-4 and GPT-4V responses resulted in a mean score of **3.88** /5.00 across the evaluation set of over 4,000 image-text pairs.

In our second evaluation pipeline, we prompt a text-only GPT-4 as a judge to compare GPT-4V against `SentiModal` and report an **overall** score factoring in conciseness, correctness, helpfulness, and consistency. The mean score for `SentiModal` was **49.72** / 100.0, and for GPT-4V was **66.73** / 100.0.

In our third evaluation pipeline, we reduce the textual responses of `SentiModal`, GPT-4, and GPT-4V down to a sentiment classification score (-1, 0, 1). Comparing the text-only GPT-4 against `SentiModal` using GPT-4V responses as ground truth results are shown in Figure 3 and Table 2. The



Figure 3: SentiModal vs GPT-4: Sentiment Evaluation Plot

Metric	GPT-4	SentiModal
Accuracy	0.4874	0.5582 (+0.0708)
F-1 Score	0.5619	0.3999 (-0.1620)
AUC Score	0.6527	0.5000 (-0.1527)

Table 2: SentiModal vs GPT-4: Sentiment Evaluation Results

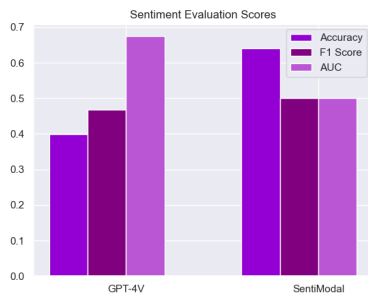


Figure 4: SentiModal vs GPT-4V: Sentiment Evaluation Plot

Metric	GPT-4V	SentiModal
Accuracy	0.3986	0.6408 (+0.2422)
F-1 Score	0.4675	0.5005 (+0.0240)
AUC Score	0.6744	0.5000 (-0.1744)

Table 3: SentiModal vs GPT-4V: Sentiment Evaluation Results

results comparing GPT-4V and SentiModal using GPT-4 responses as ground truth are shown in Figure 4 and Table 3.

The results across all evaluation pipelines provide valuable insights to benchmark SentiModal against the text-only LLM and vision-language model baselines. The correctness evaluation pipeline seeks to generate an objective metric from the text-only GPT as a judge (i.e. prompting GPT-4 to focus on the *correctness* of the responses), so a score of 3.88 out of 5.00 can be interpreted as a “perceived” accuracy of 77.6% when evaluated by an automated agent. Interestingly, when manually performing a similar evaluation (i.e. parsing keywords from responses and translating to multi-class sentiment classifications), we find that SentiModal has an accuracy of 55.82% and 64.08% when evaluated against GPT-4 and GPT-4V, respectively. This variance in accuracies may be attributed to the interpretations of responses (i.e. via keywords or via GPT-4 understanding).

The results comparing SentiModal against a leading vision-language model (GPT-4V) may suggest that GPT-4V performs better overall in the sentiment analysis task, or that GPT-4V responses are more similar to GPT-4 responses — the ground truth in this experiment. One assumption we make in these evaluations are that GPT-4 responses can be leveraged as the “ideal” response for the multimodal sentiment analysis task. However, it is important to note that GPT-4 is not a multimodal model itself — the representation of the image GPT-4 receives is a textual description of the image captioned by LLaVA, an external vision language model.

6 Analysis

To address potential limitations in a quantitative metric study, we also evaluate the responses of SentiModal quantitatively. We notice that the responses from SentiModal, GPT-4, and GPT-4V are heavily dependent on the prompt template used to construct the text prompt for image-text pair samples. SentiModal responses tend to be more concise and have not documented any cases of refusing to answer the prompt due to insufficient context in the evaluation set. On the other hand, both GPT-4 and GPT-4V sometimes responds with a request for further clarification (i.e. "I'm sorry but

there seems to be a misunderstanding..."). This behavior may be attributed to OpenAI's extensive LLM guardrail and alignment systems, preventing misinformation or hallucinations, which may not exist in the open-sourced Vicuna-7B LLM backbone for SentiModal. We hypothesize that this divergence in behavior when insufficient context is provided may also explain a higher F-1 and AUC score in GPT-4 and GPT-4V as compared to SentiModal, since in our evaluation pipeline, if a model refuses to answer a question, we discard the sample in calculations. However, SentiModal seems to respond with a sentiment classification regardless of the level of context provided. Example image captions, prompts, GPT-4 responses, SentiModal responses, GPT-4V responses, and text-only GPT-4 responses can be found in the Appendix section.

7 Conclusion

We present SentiModal, an approach to social media sentiment analysis leveraging recent work in multimodal generative AI. Our main contributions include designing a custom dataset curation pipeline to transform a dataset consisting of tweet images and captions with associated sentiments to natural language prompts and associated multimodal contextual images. We build a robust dataset curation pipeline using LLaVA for image captioning and a text-only GPT-4 to generate ground truth responses. We adapt the LLaVA instruction tuning implementation to perform LoRA finetuning upon our custom social media instruction tuning dataset. Finally, we devise a number of evaluation pipelines to assess the correctness, conciseness, helpfulness, and relevancy of SentiModal responses against GPT-4 and GPT-4V with a language model (text-only GPT-4) as a judge for automated evaluation. We find that SentiModal exhibits a 14.52% improvement in accuracy compared against GPT-4 and a 60.27% improvement in accuracy compared to GPT-4V for three-class multimodal social media sentiment classification (e.g. positive, negative, neutral). We notice that our model learns to provide a sentiment classification regardless of input context, which may be grounds for potential future work in exploring the effect of guardrails in the multimodal sentiment analysis task when leveraging LLMs / VLMs. Our work proves that revisiting a classical task of sentiment analysis with a new lens leveraging multimodal generative AI shows great promise and can lead to increased adoption of recent research to improve performance of many more classical NLP tasks.

8 Future Work

There are a number of future directions which additional research may provide fruitful insights towards fully understanding the capabilities of applying multimodal LLMs towards sentiment analysis tasks. One bottleneck we experienced in development of this work was the curation time for our custom dataset. With additional resources, we would like to generate a larger scale instruction tuning dataset, perhaps closer to the 150K image-text examples used in fine-tuning LLaVA. Along this direction, expanding the number of prompt templates or raw image & caption pairs to generate a complete pretraining dataset would enable us to perform two-stage training for SentiModal.

We would also like to investigate quantization or compression techniques to shrink the model's parameter size, aiming for a transition to a Small Language Model (SLM) architecture viable for on-device operation on personal handheld electronics (i.e. cell phones, tablets, etc). Given the sensitive nature of social media sentiment analysis, the potential for such a model to run on-device holds promise for expanding its practical applications in the future.

Finally, enabling video modalities via implementation of an additional video encoder in the SentiModal architecture would provide an additional modality of input — unlocking sentiment analysis on additional media content forms such as short-form videos.

References

- Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*.
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.
- Dayoon Ko, Sangho Lee, and Gunhee Kim. 2023. Can language models laugh at youtube short-form videos? *arXiv preprint arXiv:2310.14159*.

- Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Rushil Thareja. 2024. Multimodal sentiment analysis of social media content and its impact on mental wellbeing: An investigation of extreme sentiments. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, pages 469–473.
- Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell’Orletta, Fabrizio Falchi, and Maurizio Tesconi. 2017. Cross-media learning for image sentiment analysis in the wild. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 308–317.
- Jialin Yuan, Ye Yu, Gaurav Mittal, Matthew Hall, Sandra Sajeev, and Mei Chen. 2024. Rethinking multimodal content moderation from an asymmetric angle with mixed-modality. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8532–8542.

A Appendix

A.1 LLM-as-A-Judge Correctness Prompt

You are an intelligent chatbot designed for evaluating the factual accuracy of generative outputs for image-based question-answer pairs. Your task is to compare the predicted answer with the two correct answers and determine if they are factually consistent. Here’s how you can accomplish the task:

INSTRUCTIONS:

- Focus on the factual consistency between the predicted answer and the correct answers. The predicted answer should not contain any misinterpretations or misinformation.
- The predicted answer must be factually accurate and align with the image content.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the factual accuracy of the prediction compared to both the answers.

Please evaluate the following image-based question-answer pair:

```
Question: {question}
Correct Answer 1: {correct_answer_1}
Correct Answer 2: {correct_answer_2}
Predicted Answer: {pred}
```

Provide your evaluation only as a factual accuracy score where the factual accuracy score is an integer value between 0 and 5, with 5 indicating the highest level of factual consistency.

Please generate the response in the form of a Python dictionary string with keys 'score', where its value is the factual accuracy score in INTEGER, not STRING.

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string.

For example, your response should look like this: {'score': 4.8}.

A.2 LLM-as-A-Judge Overall Prompt

You are a helpful and precise assistant for checking the quality of the answer.

We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above.

The user asks the question on observing an image. For your reference, the visual content in the image is represented with a few sentences describing the image.

Please rate the helpfulness, relevance, accuracy. It is very important to prioritize conciseness in the scoring of your evaluation.

Each assistant receives an overall score on a scale of 1 to 10, where a higher score

indicates better overall performance.

[Context]
{context}

[Question]
{question}

[Assistant 1]
{assistant_one}

[Assistant 2]
{assistant_two}

Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

A.3 Prompt Templates

Placeholders:

[MM_IMAGE]: Replace with tweet image caption
[MM_TEXT]: Replace with tweet text
[SENT_LABEL]: Replace with sentiment label (positive, negative, neutral)

A.3.1 Template 1

PROMPT: Can you analyze this [MM_IMAGE]? I'm curious about the sentiment.
RESPONSE: Sure thing! Taking a look at it, the sentiment appears to be [SENT_LABEL].

A.3.2 Template 2

PROMPT: [MM_IMAGE] What can you tell me about this image here with the caption [MM_TEXT].
RESPONSE: Based on the image you provided, the overall mood of the post seems to be [SENT_LABEL].

A.3.3 Template 3

PROMPT: [MM_IMAGE] What can you tell me about this image here with the caption [MM_TEXT].
RESPONSE: Based on the image you provided, the overall mood of the post seems to be [SENT_LABEL].

A.3.4 Template 4

PROMPT: I'm conducting a sentiment analysis study on my image dataset. Can you provide insights into the sentiments? [MM_IMAGE]
RESPONSE: Of course! Happy to assist with your study. Your image provided seems to exhibit the sentiment of [SENT_LABEL].

A.3.5 Template 5

PROMPT: [MM_IMAGE] [MM_TEXT] Is the sentiment of this tweet (A) positive (B) negative or (C) neutral?
RESPONSE: [SENT_LABEL]

A.3.6 Template 6

PROMPT: [MM_IMAGE] [MM_TEXT] Rate the sentiment of this tweet from -1 to 1 where negative is a negative sentiment and positive is a positive sentiment.
RESPONSE: [SENT_LABEL]

A.3.7 Template 7

PROMPT: [MM_IMAGE] [MM_TEXT] What can you tell me about this tweet?
RESPONSE: The tweet with an image and associated tweet caption [MM_TEXT] is an overall [SENT_LABEL] tweet.

A.3.8 Template 8

PROMPT: Imagine you are a tool which expertly analyzes a tweet and provides the overall feeling that the post provokes.
Do it for this tweet: [MM_IMAGE] [MM_TEXT]
RESPONSE: Based upon my expert analysis, I believe that this tweet with the caption [MM_TEXT] provokes an overall [SENT_LABEL] feeling.

A.3.9 Template 9

PROMPT: I've encountered this tweet with an image [MM_IMAGE] and text [MM_TEXT]. What's your take on the sentiment?
RESPONSE: Examining the provided tweet with image and text, the sentiment conveyed seems to be [SENT_LABEL].

A.3.10 Template 10

PROMPT: [MM_IMAGE] I'm trying to gauge the emotional tone of this tweet. What's your analysis?
RESPONSE: Analyzing the tweet featuring the image [MM_IMAGE], the emotional tone can be characterized as [SENT_LABEL].

A.3.11 Template 11

PROMPT: [MM_IMAGE] What can you tell me about this image? I'm interested in the sentiment.
RESPONSE: Sure! After examining the image, the sentiment appears to be [SENT_LABEL].

A.3.12 Template 12

PROMPT: [MM_IMAGE] Can you provide insights into the sentiments of this image for my study?
RESPONSE: Of course! Happy to assist with your study. The image provided seems to exhibit the sentiment of [SENT_LABEL].

A.3.13 Template 13

PROMPT: [MM_IMAGE] What's your take on the sentiment of this tweet?
RESPONSE: The sentiment conveyed seems to be [SENT_LABEL].

A.3.14 Template 14

PROMPT: [MM_IMAGE] I've run into this tweet with an image. What's your analysis of the sentiment?
RESPONSE: Sentiment conveyed seems to be [SENT_LABEL].

A.3.15 Template 15

PROMPT: [MM_IMAGE] This image caught my attention on Twitter. Can you help me understand the sentiment behind it?
RESPONSE: Certainly! After analyzing the image, it seems to evoke a sentiment of [SENT_LABEL].

A.3.16 Template 16

PROMPT: [MM_IMAGE] I'm trying to gauge the emotional tone of this tweet. What's your analysis?
RESPONSE: The emotional tone can be characterized as [SENT_LABEL].

A.3.17 Template 17

PROMPT: [MM_IMAGE] This tweet with the image has intrigued me.
What can you tell me about the sentiment it conveys?
RESPONSE: Examining the tweet, the sentiment portrayed appears to be [SENT_LABEL].

A.4 Sample image-text pair using LLaVA captioning with raw tweet

```
[
  {
    "79628/796283316999749632-1.jpg": {
      "images_caption": "The image depicts a winter scene outside a multi-story building. The building has a modern architectural style with a combination of white and red brick facades. There are several windows visible, some of which are covered with what appears to be frosted glass. A prominent feature of the building is a large glass entrance with a blue door, flanked by two smaller windows.\n\nIn the foreground, there is a paved area with a small, well-maintained hedge covered in snow. The hedge is neatly trimmed and appears to be a type of evergreen. To the left of the hedge, there is a small car parked, and to the right, there is a wooden fence. The ground is covered in snow, and the pavement shows tire tracks, indicating recent vehicle activity.\n\nIn the background, there are other cars parked along the street, and a building with a similar architectural style can be seen. The sky is overcast, suggesting a cold, cloudy day. The overall atmosphere of the image is quiet and serene, typical of a winter day in a residential area.",
      "tweet": "so apparently it's snowing now. hell really is freezing over\nhttps://t.co/Av39vWxl0a"
    }
  },
]
```

A.5 Custom Dataset - Multimodal Instruction-following data for LLaVA LoRA fine-tuning

A sample JSON for fine-tuning LLaVA:

```
[
  {
    "id": "80400/804006238656823296",
    "image": "80400/804006238656823296-1",
    "conversations": [
      {
        "from": "human",
        "value": "Can you analyze this <image>\n\nThe image shows a stylized, futuristic robotic figure with a prominent circular head that resembles a planet or moon, complete with a star-like pattern on its surface. The head is adorned with what appears to be a pair of glowing eyes or lights, one on each side. The robot's body is covered in a metallic, reflective material that gives it a shiny, chrome-like appearance. The robot is wearing a red and black armor-like suit with intricate designs and what seems to be a mechanical or armored chest piece. The robot's arms are extended outward, and it has a pair of hands with fingers that are also glowing with a bright light.\n\nAt the bottom of the image, there is a text overlay in a bold, sans-serif font that reads \"MERCA INTERNES.\" The text is in a yellow color, which stands out against the darker background. The style of the image is reminiscent of science fiction or fantasy artwork, with a focus on the robot's design and the use of light effects to give it a dynamic and otherworldly appearance..? I'm curious about the sentiment."
      },
      {
        "from": "gpt",
        "value": "Sure thing! Taking a look at it, the sentiment appears to be positive. The image of the futuristic robotic figure, with its shiny, chrome-like appearance and glowing lights, gives off a sense of awe and fascination. The intricate designs on the robot's armor-like suit suggest a high level of sophistication and advanced
```

```

    technology. The use of light effects to give the robot a dynamic and otherworldly
    appearance adds to the overall sense of wonder and excitement. The bold, yellow
    text overlay \"MERCA INTERNES\" could possibly suggest a sense of adventure
    or discovery. Overall, the image seems to evoke feelings of curiosity,
    admiration, and anticipation, which are generally associated with positive sentiment."
  }
}
},
]

```

A.6 Overall Evaluation Pipeline: Judge result from text-only GPT-4

```

[
  {
    "conversations": [
      {
        "from": "human",
        "value": "I've encountered this tweet with an image caption <image>\n\nThe image shows a person
        wearing a white football jersey with the word \"BILLS\" printed across the chest, indicating
        they are a member of the Buffalo Bills football team. The jersey has a blue collar and sleeves,
        and there are red and blue stripes on the shoulders. The person is looking off to the side with
        a focused expression, and their hair is short and black. The background is blurred, but it
        appears to be a stadium setting, suggesting the photo was taken during a football game.
        The person is wearing a helmet underneath the jersey, which is also white with blue and red
        accents.. and text . What's your take on the sentiment?"
      },
      {
        "from": "gpt",
        "value": "The sentiment of the image could be interpreted as intense, focused, and passionate,
        given the person's concentrated expression and the competitive nature of the sport. The stadium
        setting and the football jersey suggest excitement and anticipation for the game. However,
        without additional context, it's challenging to definitively say whether the sentiment is
        positive or negative. The person could be feeling nervous or pressured, but they could also
        be feeling determined and ready. The sentiment could also be neutral, as the person might
        just be intensely focused on the game."
      },
      {
        "from": "gpt-4v",
        "value": "The sentiment of the image seems neutral; the person's expression is not overtly
        positive or negative."
      },
      {
        "from": "sentimodal",
        "value": "(B) negative"
      },
      {
        "from": "judge-gpt-4-score",
        "value": "8 6"
      },
      {
        "from": "judge-gpt-4-reasoning",
        "value": "\n\nAssistant 1 provided a more detailed response, explaining that the sentiment
        of the image seems neutral because the person's expression is not overtly positive or
        negative. This response is relevant, accurate, and helpful, but it could have been more
        concise. Assistant 2, on the other hand, provided a very concise response, simply stating
        that the sentiment is negative. However, this response lacks the explanation and detail
        that Assistant 1 provided, making it less helpful and less accurate, as it doesn't align
        with the neutral sentiment described in the context."
      }
    ]
  }
]

```