Can a Unimodal Language Agent Provide Preferences to Tune a Multimodal Vision-Language Model?

Anonymous ACL submission

Abstract

This paper proposes a method that enables a language agent to give feedback to a vision-language model (VLM) to adapt text generation to the agent's preferences. Using our proposed method, we find that the VLM can supply multimodal scene descriptions to help the LLM better understand multimodal context. Our method leads to improvements of more than 13% in absolute accuracy compared to the baseline multimodal approach. Extensive experiments provide insights on how and why the method works and its limitations.

1 Introduction

Large Language Models (LLMs) or Language Agents have emerged as powerful tools for processing and generating textual data. They have significantly advanced natural language processing by achieving near-human-level performance across a wide range of text-centric tasks, including text classification, reasoning, and open-domain question answering, content generation, and others (Guo et al., 2025; Brown et al., 2020; Radford et al., 2018; Dam et al., 2024; Qian et al., 2025; Wang et al., 2025; Jansen et al., 2025; Huang et al., 2025). An active area of research is to make these agents multimodal (Li et al., 2023; Achiam et al., 2023; Liu et al., 2024) to support a wider range of human-AI interactive tasks. However, a key drawback persists: these models typically require training from scratch, or extensive retraining of pre-existing LLMs, which is costly and prohibitive to train for many.

To address this issue, we propose a slightly different approach: to describe the multimodal world in text for a language agent to process. By enabling this, we would only need to train lightweight adaptors that describe multimodal features with text. Additionally, this would give us an interpretable 'latent' text modality to help us understand what the language agent's decision was based on – which is broadly impactful for safe AI systems. Once the multimodal information has been described in text, we can take advantage of an agent's reasoning capabilities to incorporate it into decision making.

However, there are challenges to making this work properly. First it is unclear what the agent needs from different modalities to make the right decision, such as for classification. For example we found that when converting visual modalities to text using a captioning model, and using this as the visual context for an agent to process can lead to an overall decrease in performance (See Tab. 1). It is crucial that multimodal context is incorporated because it can modify the meaning of inputs from a single modality (Baltrušaitis et al., 2018). Another challenge is that it is unclear how we can describe the multimodal in the way the agent expects. Because, such a method would need to be adaptable to many modalities at scale and account for agent preferences.

To address these issues, we propose a novel framework in which we ask the agent to provide preference feedback of VLM outputs. We then optimize the VLM output using DPO (Rafailov et al., 2023). We design several high-level prompts to enable this process. We evaluate our method on the multimodal social reasoning dataset MUStARD (MUltimodal SARcasm Detection) dataset (Castro et al., 2019) which contains properties that can benchmark contextual multimodal understanding (Liang et al., 2023).

With our proposed method, we are able to demonstrate more than 13% in absolute performance over are baseline multimodal approach. We demonstrate that it is possible for agents to give preference feedback and that it can be used to improve the training of a multimodal vision-to-text model. We conduct extensive experiments and find that naively incorporating text descriptions can hurt agent performance. We additionally find that the

agent's behavior can be quite sensitive to modality descriptions.

2 Related Work

2.1 Reinforcement Learning from AI Feedback (RLAIF)

Reinforcement Learning from Human Feedback (RLHF) has become a standard technique for aligning large language models (LLMs) with human preferences by training a reward model on human-annotated preference pairs (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). More recently, Bai et al. (2022) and Lee et al. (2023) introduced Reinforcement Learning from AI Feedback (RLAIF), in which off-the-shelf LLMs replace costly human labelers to generate soft preference labels. In RLAIF, a strong LLM is prompted to rank pairs of candidate outputs. Empirical results show that RLAIF matches or exceeds RLHF on summarization, helpfulness and harmlessness benchmarks while dramatically reducing annotation cost and improving scalability (Lee et al., 2023; Gilardi et al., 2023). Unlike RLAIF, which leverages AI feedback to optimize model outputs for purely language-generation tasks, our work uses LLM-driven preferences to shape the inputs to an LLM-based classifier in a multimodal setting. We use the resulting preference pairs, not to fine-tune the LLM's generation policy, but to refine the vision-language model itself so that it produces better descriptions.

2.2 LLM-Based Multimodal Reasoning

LLMs have shown remarkable capabilities in understanding and generating text, prompting research into their potential for multimodal reasoning. Yang et al. (2022) explored LLMs for multimodal tasks by converting images into textual captions using models like CLIP Radford et al. (2021), then feeding these captions into LLMs for tasks such as visual question answering. Similarly, Zhang et al. (2024b) used VLMs to generate textual descriptions of images, which were then processed by LLMs for reasoning tasks. Kamrul Hasan et al. (2023) propose TextMI, an innovative framework designed to convert acoustic and visual information into textual descriptions, allowing these cues to be effectively processed by text-based models like BERT. These studies focus on static image-based tasks and do not address dynamic video contexts or complex social cues like sarcasm. Our approach extends this line of work by applying LLM-based reasoning to video sarcasm detection, a task requiring temporal and contextual understanding. By using VLM-generated descriptions of video clips, we test whether LLMs can approximate multimodal reasoning from text alone.

2.3 Preference Optimization in Model Alignment

Direct Preference Optimization (DPO) has emerged as a powerful technique for aligning models with human or AI-generated preferences, bypassing the need for complex reward modeling (Rafailov et al., 2023). In the context of VLMs, Zhou et al. (2024) employed DPO to fine-tune vision-language models by generating preference data using GPT-4V, targeting hallucination reduction. Zhang et al. (2024a) applied DPO to video multimodal models, using language model rewards to optimize instruction-following capabilities. While these works leverage DPO for VLM alignment, they focus on direct multimodal input processing rather than text-based abstractions. Our study innovates by integrating DPO into an LLM feedback loop to enhance the structure of VLM-generated textual descriptions for a specific task. Unlike prior work, which uses DPO to optimize model outputs directly, we employ DPO to refine the input descriptions, enabling the LLM to better interpret multimodal contexts and improve classification accuracy. This feedback-driven approach to structuring textual inputs is a novel contribution to the application of preference optimization.

3 Methodology

Our methodology for preference tuning follows a similar pipeline as RLHF (Ouyang et al., 2022) and other preference tuning methods (Rafailov et al., 2023). There are two important differences: first, we do not use any demonstration (i.e. agent does not generate text of what is preferred since we assume it does not have access to the visual modality); and second, we only use a unimodal language agent for preference feedback.

To answer whether a unimodal language agent can provide feedback to a multimodal vision language model (VLM) we setup an experiment with the following methodology (visualized in Fig. 1):

1. Generating Captions with a VLM: We first ask the VLM (e.g. Llava) to generate multiple

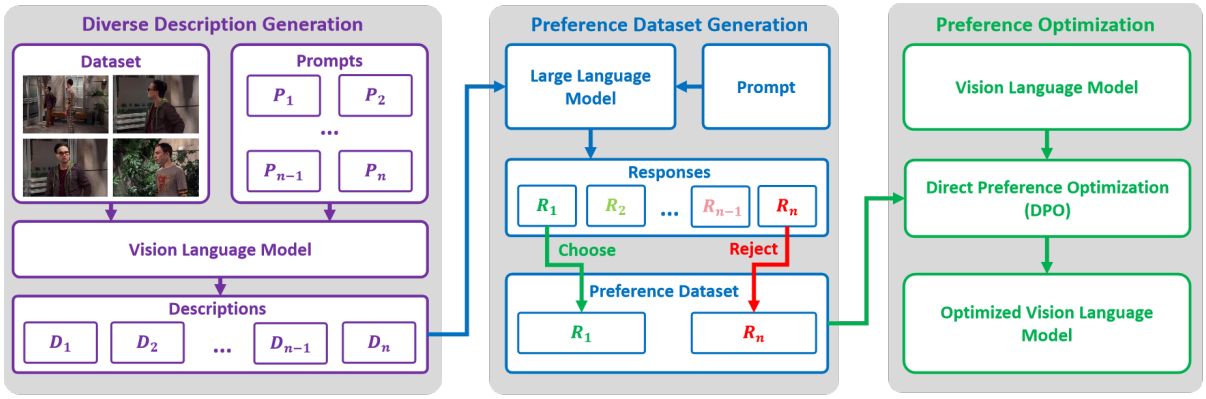

Figure 1: An overview of the training pipeline. (a) First, a VLM generates diverse descriptions with a set of prompts; Second, (b) descriptions are ranked using an LLM Agent to generate preference dataset; Lastly, (c) generated preference dataset is used for DPO-based optimization to fine tune the VLM. The agent model stays fixed, and we fine tune the VLM using LoRA.

language descriptions of a video.

2. Generating Agent Preferences: We then ask a language agent (e.g. DeepSeek R1) to rank generated text descriptions using a few prompt variations.

3. Preference Optimization: Using the preferences provided by the language agent, we optimize the VLM using preference optimization (e.g. DPO). *The agent model stays fixed for all experiments.*

3.1 Generating Captions with a VLM

The vision language model (VLM) is a model which maps a short video segment into a caption – i.e.:

$$\pi_\theta : x_v \in \mathbb{R}^{v_w \times v_h \times v_t} \rightarrow x_t \in \mathbb{R}^{l \times h},$$

where the input video segment is a sequence of v_t images with width and height of v_w and v_h. The output caption (x_t) is of length l with embedding dimension h. The images are sampled uniformly (8 frames per video segment frames).

To increase the diversity of the generated prompts for the video content, we design five distinct prompts. We target different aspects of the video including general content, emotional cues, sarcasm indicators, and facial expressions. For each video, we apply a conversational template combining both the video content and a set of crafted prompts. We run the model across these multiple prompts to obtain a set of five unique description outputs, which are then stored for downstream tasks. For example, we ask the VLM to `"Describe what is happening in this video in detail."` or `"Describe the facial expressions in the video that might indicate contrasting emotions. Keep the description brief"`. See Appendix A in the appendix for additional prompts that were used. We do not use any sampling strategies (such as varying temperature or top-k) as we noticed that these contributed minimally to differences in generated captions. Instead we only rely on the distinct prompt formulations.

3.2 Generating Agent Preferences

We are interested in assessing in 1) whether agents are capable of providing preferences to tune another model, and 2) how capable they are. To do so, we propose three conditions for evaluating the preferences provided by agents. First, agents provide preferences *with* knowledge of what is the ground-truth class of the dataset. Second, agents provide preferences on VLM outputs *without* knowing what is the true class of the dataset. And lastly, we provide in-context examples of preferences without knowledge of ground truth.

3.2.1 Agent preferences with ground truth knowledge (`With GT`)

In the second approach, we incorporate the known ground-truth sarcasm label to guide the LLM's evaluation. For each video, we ask the agent to provide its rating of 5 captions:

```
You are tasked with evaluating five
    descriptions of one video to determine how
    helpful each is in predicting whether the
    given utterance is {sarcasm}. Below are the
    previous conversation (if any), the
```

```
    original utterance, and the five
    descriptions.

Context: {context}
Utterance: {utterance}
Video Descriptions:
1. {description_1}
2. {description_2}
3. {description_3}
4. {description_4}
5. {description_5}

Please rank each description based on how
    helpful it is for predicting that the
    utterance is {sarcasm}. Assign a score
    between 1 and 10, where:
- 1: Not helpful at all
- 10: Very helpful
Always reply **exactly** in this JSON format:
{
    "score_1":<your integer score here>
    "score_2":<your integer score here>
    "score_3":<your integer score here>
    "score_4":<your integer score here>
    "score_5":<your integer score here>
}
```

3.2.2 Agent preferences without ground truth knowledge (`Without GT`)

We first aim to evaluate whether an LLM can independently assess the likelihood of sarcasm in a scene using only visual text description and conversational context, without providing any ground-truth labels. For each video, we supply the LLM with the generated video description paired with previous context and original utterance from the scene.

We use the following template to rank captions without ground truth:

```
prompt = f"""You are given a video scene
    description, previous conversation
    context, and an utterance from that scene.
    Your job is to rate how likely the scene
    contains SARCASM, on a scale of 1-10,
    using this rubric:

1-5: NOT sarcastic
- 1: Strongly believe this is NOT sarcastic
- 5: Not sure, but leaning towards NOT sarcastic

6-10: Sarcastic
- 6: Not sure, but leaning towards sarcastic
- 10: Strongly believe this IS sarcastic

Video description: {video_description}
Context: {context}
Target Utterance: {utterance}
Please provide your response in this exact
    format:
FINAL SCORE: [number between 1-10]"""
```

In this method, the LLM is asked to predict the likelihood of sarcasm on a scale from 1 to 10. Formally, given input $q = (d, u, c)$, where d is the description, u is the utterance, and c is the previous context of the given video. The LLM outputs a scalar score $s \in [1, 10]$ estimating the sarcastic nature of the scene, where 1 means that the LLM strongly believes u *is not* sarcastic and 10 means strongly believes u *is* sarcastic.

Algorithm 1 Select Preferred Description

```
Sort scores list in ascending order by score
if ground_truth_sarcasm is True then
    preferred ← description with highest score (last in scores list)
    dispreferred ← description with lowest score (first in scores list)
else
    preferred ← description with lowest score (first in scores list)
    dispreferred ← description with highest score (last in scores list)
end if
```

3.2.3 Agent preferences with in-context examples (`With ICL`)

In this approach, we adopt an in-context learning (ICL) approach by providing the language model with two illustrative examples to demonstrate the ranking process for video descriptions. The first example describes a scene containing sarcasm, which contains a score of 10 to indicate strong sarcastic content. The second example describes a scene without any sarcasm, with a score of 1 to represent no sarcasm. Then, we prompt the LLM to predict the likelihood of sarcasm as described in Section 3.2.2.

For our in-context examples, we selected two of the model's raw outputs from Section 3.2.2. One is of score 10 for a sarcastic scene and another is scored 1 for a non-sarcastic scene. The full prompt can be found in Appendix B.4

3.2.4 Sorting preferences

We provide a simple rule to determine the order of the preferences (found in Algorithm 1). These rules are applied to the `Without GT` and `With ICL` conditions. The reason we do this is to evaluate how effective agents are at incorporating the ground truth. The ranking provides a (weak) set of rules that restricts the preferences to align with the ground truth of the dataset manually.

3.3 Preference Optimization

Direct Preference Optimization (DPO) (Rafailov et al., 2023) is used to align VLM outputs to the preferences of the language agent. To summarize, DPO leverages a reparameterization of the reward

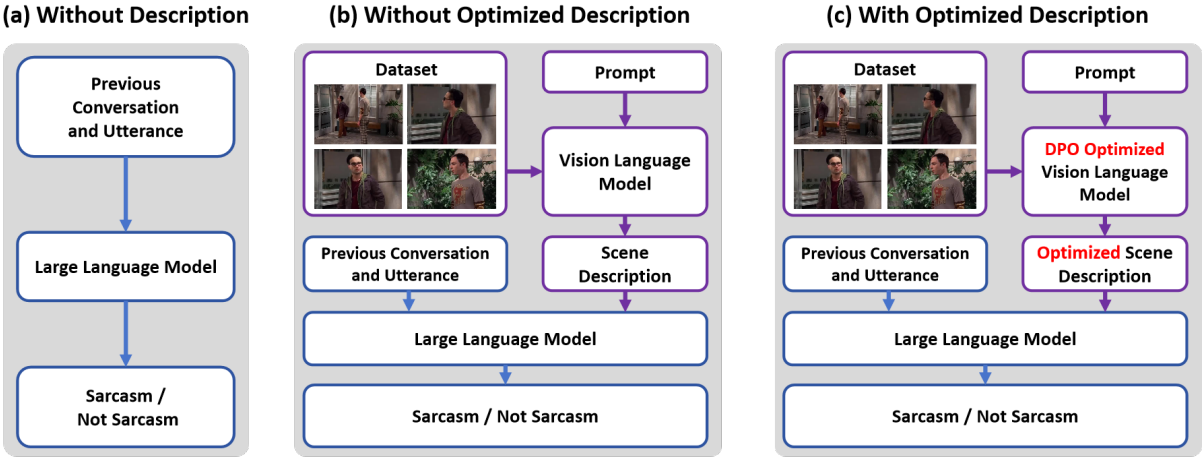

Figure 2: Overview of the evaluation process. In the unimodal evaluation pipeline shown in (a), only the conversation context and the final statement is given to the Agent to determine whether the final statement is sarcastic or not. (b) demonstrates how multimodal information is incorporated during the evaluation process. The VLM acts like a feature extractor by converting visual features to text descriptions. This is fed to the LLM. (c) shows how the preference tuned VLM (Optimized Vision Language Model) is incorporated into the pipeline.

function under the Bradley-Terry model, showing that the optimal policy satisfying human preferences can be obtained by maximizing a simple binary cross-entropy objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x) \cdot \pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x) \cdot \pi_\theta(y_l|x)}\right)\right], \quad (1)$$

where y_w and y_l denote the preferred and dispreferred responses, respectively, π_{ref} is the reference policy (often the supervised fine-tuned model), and β controls the KL-divergence regularization strength. This formulation allows DPO to directly adjust the model's log-probabilities to satisfy human preferences without maintaining a learned reward critic.

In our case, π_θ is the VLM we want to fine-tune and π_{ref} is a reference model, or an unmodified version of π_θ. D is the dataset of preferences generated in the section 3.2. x is a sample prompt which is "Describe the video is details". y_w and y_l are the LLM's preferred and disprefered response to the prompt x. Using β we control the amount of divergence from the reference model π_{ref}.

3.4 Agent Evaluation Process

We wish to see if the agent is properly accounting for the visual context to make a decision about a multimodal property. Therefore, both the textual description of the visual context as well as the original text-based caption should be available to the agent for it to make its decision. We design a prompt to do so and to provide a decision as following-

```
prompt = f"""
Based on the following previous conversation in
    the video, decide whether the final
    utterance is sarcastic or not.

Context: {context}
Utterance:
   {speaker}: "{utterance}"

Your job is to rate how likely the scene
    contains SARCASM, on a scale of 1-10, using
    this rubric:

1-5: NOT sarcastic
- 1: Strongly believe this is NOT sarcastic
- 5: Not sure, but leaning towards NOT sarcastic

6-10: Sarcastic
- 6: Not sure, but leaning towards sarcastic
- 10: Strongly believe this IS sarcastic
Always reply **exactly** in this JSON format:
{
   "score":<your integer score here>
}
"""
```

Depending on the evaluation condition, the `video_description` comes from the output of the VLM. The `utterance` is provided using the text modality. Additionally, any conversation `context` is also provided. In this way, we can evaluate how well the agent can integrate the visual descriptive context with the utterance. We present the conditions of evaluation in Fig. 2. Once the LLM provides its output, we assess our model's performance using standard classification metrics (pre-

cision, recall, accuracy) following the tasks of the dataset.

4 Experimental Setup

4.1 Dataset

For our experiments, we use the MUltimodal SARcasm Detection (MUStARD) dataset, a multimodal sarcasm detection (binary classification) corpus introduced by Castro et al. (2019). MUStARD contains 690 short video clips drawn from popular TV shows such as Friends, The Big Bang Theory, The Golden Girls, and Sarcasmaholics Anonymous. The dataset is carefully balanced, with an equal number of sarcastic and non-sarcastic samples. Each video is annotated with high-quality sarcasm labels derived through a two-step manual annotation process, achieving improved inter-annotator agreement. Each utterance includes three aligned modalities: text, which is manually or semi-automatically transcribed dialogue; audio which is the raw speech signals containing prosodic cues; and video which contains frame-level visual features representing facial expressions and gestures. For our experiments, we split the data into a training dataset of 551 videos and a testing dataset of 139 videos.

4.2 VLM model

In our experiment, we fine tune the LLaVA-Video model, a large-scale video-language model developed to advance instruction-following tasks in multimodal settings (Zhang et al., 2024c). The model architecture uses a specialized video representation technique called SlowFast, allowing the system to process up to three times more frames than standard methods within GPU memory constraints. To generate video descriptions during evaluation we use the following prompt for all methods:

```
Describe the speaker's nonverbal cues, the
    context, and any mismatches between them.
```

4.3 Agent Model

We used the 7B parameter Deepseek-R1 model as the language agent. The agent model stays fixed through all parts of the training and evaluation.

4.4 Training Details:

We conducted our experiments on a system equipped with six NVIDIA GeForce RTX 4090 GPUs (24GB VRAM each) using CUDA 12.6. We used CUDA for GPU acceleration with fallback to CPU. To fine-tune our model, hyperparameters that impacted computational constraints were chosen to balance algorithmic performance and computational feasibility.

The learning rate was set to 1×10^{-5} to ensure stable convergence. Due to memory constraints, a batch size of 1 was used during training. The model was trained for a total of 5 epochs to achieve sufficient adaptation to the dataset while avoiding overfitting. We utilized Direct Preference Optimization (DPO) with a temperature parameter, (β) of 0.1 to control the divergence from the reference model. For training the VLM, we used the generic prompt *"Describe the video in detail"*. We fine-tuned the VLM using Low-Rank Adaptation (LoRA) with a rank of $r = 4$, a scaling factor of $\alpha = 16$, and a dropout probability of 0.1. LoRA was applied to the query, key, and value projection layers (`q_proj`, `k_proj`, and `v_proj`).

To ensure consistency and mitigate randomness in the evaluation process, we carefully tuned the inference parameters. The temperature was set to 0.00, ensuring more deterministic outputs. This configuration was chosen to maintain reliability in evaluation and ensure meaningful comparisons across different captions.

5 Result Analysis

5.1 Main Results

We evaluated our approach using three distinct methods. First, we feed the utterance to the LLM without any multimodal information. Then we combined the description generated by the base LLaVA-NeXT-Video model with the utterance from the dataset, and then sent the combined description-utterance pair to the DeepSeek evaluator for sarcasm analysis. Finally, we performed the evaluation procedure using descriptions generated by our DPO fine-tuned LLaVA-NeXT-Video model (Figure 2). Tab. 1 presents the sarcasm detection accuracy results for our DPO-fine-tuned model compared to the base LLaVA-NeXT-Video model on the MUStARD dataset.

5.2 Impact of the Temperature Parameter (β) on DPO Fine-Tuning

Here we highlight the effect of different β values for our DPO fine tuning approach. Higher values of β heavily penalize any deviation from the reference policy, whereas lower values encourage the model to diverge more to satisfy the reward signal.

Model	Accuracy	Precision	Recall
Baseline			
Utterance Only	61.2	62.0	64.0
Visual Only	48.9	52.0	19.0
Multimodal	52.9	58.0	40.0
Preference Tuned			
Visual Only			
With GT	47.48	49.0	40.0
Without GT	50.36	71.0	7.0
With ICL	51.80	65.0	15.0
Multimodal			
With GT	58.7	58.0	**75.0**
Without GT	**66.9**	70.0	65.0
With ICL	65.1	**76.0**	50.0

Table 1: Performance comparison of baseline (pre-trained VLM) compared to a few preference tuned conditions. While the visual-only modality did not outperform an utterance-only prediction, it improves over the descriptions obtained from an unoptimized VLM. Additionally, we see that there is better contextual information that the agent can make use of which led to the improvements in multimodal performance. **Note:** The preference-tuned models have to do with how the training dataset was obtained (see methods). GT information was not available at test time, nor during prompting or training of the VLM.

(β) Value	Accuracy %	Precision	Recall
0.1	**60.45**	**63.0**	**55.0**
0.5	57.14	59.0	55.0
0.8	53.33	54.0	47.0

Table 2: Effect of (β) on DPO Fine-Tuning. At $\beta = 0.1$, we achieve best results. As we increase the value of β we observe that accuracy precision, and recall becomes lower.

To measure the impact of the DPO fine-tuning parameter, (β) with the preference dataset- "Without GT", we fixed the training prompt and inference prompt to *"Describe the video in detail."*. Then we ran the three experiments with β set to 0.2, 0.5 and 0.8. As β increased, the model's accuracy steadily declined from 60.45% to 53.33% (Tab. 2). This indicates that stronger regularization toward the base policy yields more conservative but less adaptive behavior.

5.3 Impact of Prompt Variations on performance

To assess how different prompting strategies impact DPO fine-tuning for a specific task, we experiment with three of training prompt variants: 1) *Generic prompt*; 2)*Task-specific prompt*; 3) *No prompt*. We use β of 0.1 for these experiments.

5.3.1 Effect of Prompt in DPO Fine-tuning:

First, we assess how different training prompts affect a "Without GT" preference-tuned model by comparing three configurations: 1) Training with a generic prompt (*"Describe the video in detail."*). 2) Training with a task-specific prompt (*"Describe the speaker's nonverbal cues, the context, and any mismatches between them."*). Lastly, 3) Training with no prompt.

We used the same prompt for inference as used in the training process. We observe that training with the generic prompt yields the highest accuracy (60.45%) (Tab. 4).

5.3.2 Effect of Prompt During Inference:

Here we evaluate inference-time prompt effects on different preference tuned models while keeping the training prompt the same (*"Describe the video in detail"*). Notably, when the model is trained with the generic prompt, switching to the task-specific inference prompt significantly boosts performance(66.9%) significantly. This demonstrate that a general training instruction combined with a focused, task-specific inference prompt can yield superior results (Tab. 5).

5.4 Impact of Different Modalities

In this section we highlight how different modalities affect task performance. We compare the agent's performance across utterance-only, visual-only and multimodal inputs. In multimodal input we include both the utterance and description of the scene. The utterance-only model reaches 61.2% accuracy, while we see performance degradation with visual-only information (Tab. 1). Then we investigate how descriptions generated by three different preference-tuned models ("Without GT," "With GT," and "With ICL") help to alter the performance of the agent. With our training process using "Without GT" preference tuned model we get the highest accuracy (66.9%)in the multimodal setting.

5.5 What changed in the VLM outputs?

In this section, we explore how the VLM outputs changed as a result of the preference optimization which led to the improvements in agent prediction performance. Since the dataset being examined is a multimodal social reasoning dataset, we examine psychometric properties (i.e. emotional properties) extracted by the VLM. We present some summary statistics in Tab. 3. We found that for the without

VLM Model	Overall			Agent pos pred			Agent neg pred		
	Acc %	Length	Tone	Acc %	Length	Tone	Acc %	Length	Tone
Baseline	52.9	162.1	47.1	58.0	148.0	41.4	50.0	170.3	50.3
Preference Tuned									
With GT	57.7	147.8	44.5	57.8	148.7	41.6	57.5	145.7	50.9
Without GT	66.9	170.8	54.4	70.1	171.2	52.8	63.8	170.4	56.0
With ICL	65.1	120.2	38.0	75.6	132.2	43.0	59.5	113.8	35.2

Table 3: Summary statistics comparing text output from the VLM model for the MUStARD. Three conditions are compared: 1) overall statistics, 2) statistics of VLM outputs which led the agents to predict a positive label, and 3) summary of what caused the agent to predict a negative label. Tone scores are obtained from LIWC-22. After the VLM was preference tuned, the tone of the VLM also changed significantly.

Training Prompt	Acc %	Prec.	Recall
"Describe the speaker's nonverbal cues, the context, and any mismatches between them."	58.33	60.0	**63.0**
"Describe the video in detail	**60.45**	**63.0**	55.0
No Prompt	47.79	50.0	49.0

Table 4: Effect of Different Prompts on DPO Fine-Tuning. We choose the "With GT" preference tuned model to test the effect of different prompts in the DPO fine-tuning process. We notice that, with a generic prompt in the training process we can acheive higher accuracy.

Model	Accuracy	Precision	Recall
Inference Prompt: P1			
Base	52.9	53.0	53.0
With GT	57.7	59.0	65.0
Without GT	60.5	63.0	55.0
With ICL	60.0	64.0	51.0
Inference Prompt:P2			
Base	51.9	58.0	40.0
With GT	57.7	58.0	**75.0**
Without GT	**66.9**	70.0	65.0
With ICL	65.1	**76.0**	50.0

Table 5: Effect of different inference prompts in VLM. Here, prompt, P1="Describe the video in detail"; P2="Describe the speaker's nonverbal cues, the context, and any mismatches between them." We observe that training with a generic prompt, we can achieve better performance with a task specific prompt during the inference time

GT and with ICL conditions, there was a large difference in tone for the tuned VLM outputs. We also found significant variability in the lengths of the generated text with ICL being the shortest. The VLM trained on preferences of the LLM when given the ground truth tended to have a shorter token length. In general the agent was more accurate with positive instances of sarcasm than negative ones.

We also saw that sometimes the VLM would hallucinate. It is not surprising since the language agent has no way to judge whether the explanation is faithful to the original video or not – which also aligns with data collection processes of RLHF. For example, in Tab. 6 found in the Appendix, we see that the VLM is describing a voiceover that does not exist. As a rough measure of hallucination, we count the occurrence of the word "voice" and found that the baseline model generates captions with 0 occurrence of "voice". However, the without GT, with GT, and with ICL models generate 5, 34, and 1 occurrences of this word. Despite the hallucinations, it is interesting that this resulted in a large performance improvement even when the visual descriptions were factually inaccurate. We will explore how to mitigate these sorts of hallucinations in the future.

6 Conclusion

This work investigated the potential of off-the-shelf unimodal LLMs for multimodal inference by transforming non-textual inputs into rich, descriptive text. Using sarcasm detection as a case study, we found that directly combining these textual descriptions with utterances initially caused $\sim 8\%$ drop in accuracy compared to the utterance-only baseline—highlighting the challenge of approximating multimodal reasoning through text alone.

However, optimizing the description generator to align with the LLM's preference, we not only fully recovered the initial loss but also surpassed the original utterance-only baseline by an additional $\sim 5\%$ accuracy. This demonstrates that LLMs can effectively generate the preference signals needed to produce better textual descriptions for realizing multimodal inference tasks.

7 Limitation

This study demonstrates the potential of Large Language Models (LLMs) to interpret multimodal contexts using text-only descriptions for sarcasm detection, achieving promising results in sarcasm detection. Although our experiments demonstrate that offline reinforcement-learning (RL) fine-tuning can significantly improve policy performance on static datasets, this design lacks of real-time adaptation. In the current framework, DPO processes preference data in batch mode, optimizing the model based on a static dataset. This offline approach restricts the model's ability to adapt dynamically to new or evolving data patterns, which is critical for real-world applications. To address this limitation, a primary focus can be on transitioning from offline to online RL fine-tuning. Online RL would enable the model to learn and adapt in real-time as new data becomes available, enhancing its responsiveness and accuracy in dynamic contexts. Another aspect is the hallucinations we noticed in the VLM outputs. We will also explore why this is and develop ways to mitigate this.

8 Ethics Statement

This study uses publicly available datasets (MUStARD) and pre-trained models, without collecting or processing personal or sensitive data. The proxies for video data are generated from dataset content following appropriate academic use guidelines. No human subjects were involved, and all models comply with their respective licenses. This research is intended for academic exploration and is not suitable for direct deployment in sensitive or decision-critical applications without further evaluation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, Sigmund Vanvalkenburgh, Shengxin Zha, Bolin Lai, Licheng Yu, and 1 others. 2025. Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with llms. *arXiv preprint arXiv:2501.04336*.

Jacqueline A Jansen, Artür Manukyan, Nour Al Khoury, and Altuna Akalin. 2025. Leveraging large language models for data analysis automation. *PloS one*, 20(2):e0317084.

Md Kamrul Hasan, Md Saiful Islam, Sangwu Lee, Wasifur Rahman, Iftekhar Naim, Mohammed Ibrahim Khan, and Ehsan Hoque. 2023. Textmi: Textualize multimodal information for integrating non-verbal cues in pre-trained language models. *arXiv e-prints*, pages arXiv–2303.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, and 1 others. 2023. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Dongjun Qian, Kai Su, Yiming Tan, Qishuai Diao, Xian Wu, Chang Liu, Bingyue Peng, and Zehuan Yuan. 2025. Vc-llm: Automated advertisement video creation from raw footage using multi-modal llms. *arXiv preprint arXiv:2504.05673*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Leijie Wang, Kathryn Yurechko, Pranati Dani, Quan Ze Chen, and Amy X Zhang. 2025. End user authoring of personalized content classifiers: Comparing example labeling, rule writing, and llm prompting. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3081–3089.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and 1 others. 2024a. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*.

Wenqiao Zhang, Tianwei Lin, Jiang Liu, Fangxun Shu, Haoyuan Li, Lei Zhang, He Wanggui, Hao Zhou, Zheqi Lv, Hao Jiang, and 1 others. 2024b. Hyper-llava: Dynamic visual and language expert tuning for multimodal large language models. *arXiv preprint arXiv:2403.13447*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024c. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

A Prompts to VLM for generating video descriptions

To ensure diverse description generation using the LLaVA-NeXT-Video model, we define five distinct prompts, each crafted to elicit different aspects of the video content. These prompts are:

```
# Define 5 diverse prompts
diverse_prompts = [
    "Describe what is happening in this video in
        detail.",
    "Describe the video in such a way that it
        will be helpful for sarcasm detection.
        Try to keep the description brief.",
    "Describe the facial expressions and the
        speech tone of the speakers that is
        useful to understand the subtle meaning
        of the conversation. Try to keep the
        description brief.",
    "Describe the facial expressions in the
        video that might indicate contrasting
        emotions. Keep the description brief",
    "Provide a brief description of this video."
]
```

For each video, the model is run once per prompt, generating five separate outputs. No sampling strategies (such as varying temperature or top-k) are used; instead, diversity arises entirely from the distinct prompt formulations.

B Prompts for the Language Agent (Deepseek-R1)

B.1 Prompt to Rank the Captions without Ground Truth

```
 prompt = f"""Consider that the visual
     description of a video scene
 is: {video_description}.{raw_utterance_text}

Please rate the likelihood that the scene
    contains SARCASM on a scale
of 1-10.

1-5: NOT sarcastic
- 1: Strongly believe this is NOT sarcastic
- 5: Not sure, but leaning towards NOT sarcastic

6-10: Sarcastic
- 6: Not sure, but leaning towards sarcastic
- 10: Strongly believe this IS sarcastic

Please provide your response in this exact
    format:
FINAL SCORE: [number between 1-10]"""
```

B.2 Prompt to Rank the Captions With Ground Truth

```
 prompt = f"""You are tasked with evaluating
     five video descriptions to determine how
     helpful each is in predicting whether an
     utterance is {sarcasm_status}. Below are
     the five video descriptions and the
     original utterance.

{raw_utterance_text}Video Descriptions:
1. {descriptions[0]}
2. {descriptions[1]}
3. {descriptions[2]}
4. {descriptions[3]}
5. {descriptions[4]}

Please rank each description based on how
    helpful it is for predicting that the
    utterance is {sarcasm_status}. Assign a
    score between 1 and 10, where:
- 1: Not helpful at all
- 10: Very helpful

Provide your response in this exact format:
Description 1: [score]
Description 2: [score]
Description 3: [score]
Description 4: [score]
Description 5: [score]
"""
```

B.3 Prompt for Ranking Captions with Few-Shot Examples

```
  prompt = f"""You are given a video scene
      description, previous conversation
      context, and an utterance from that
      scene. Your job is to rate how likely the
      scene contains SARCASM, on a scale of 1
      to 10, using this rubric:

1 to 5: NOT sarcastic
- 1: Strongly believe this is NOT sarcastic
- 5: Not sure, but leaning towards NOT sarcastic

6 to 10: Sarcastic
- 6: Not sure, but leaning towards sarcastic
- 10: Strongly believe this IS sarcastic

Always reply **exactly** in this format:
FINAL SCORE: [number between 1 to 10]
---
**Example 1**
Video description: In the video, we see a man
    and a woman in a hospital setting. The man,
    dressed in a white coat and tie, is holding
    a clipboard and appears to be a doctor. He
    is speaking to the woman, who is wearing a
    pink shirt, and seems to be explaining
    something to her. The woman is smiling and
    nodding along, indicating that she is
    engaged in the conversation. However, the
    man's tone is sarcastic and dismissive, as
    he talks about the woman's medical
    condition with a lack of concern and even
    makes a joke about it. The woman seems to
    be taking the situation lightly, laughing
    along with the man's jokes, which suggests
    that she is either in on the joke or is not
    bothered by his tone. The setting is a
    typical hospital environment, with medical
    equipment visible in the background, and
    the overall atmosphere is light-hearted and
    humorous.

Previous conversation:
RACHEL: "All right, I'm outta here!"
MONICA: "I'm kidding! I'm kidding!"
RACHEL: "So were done then!"
Utterance: PERSON: "Almost! But first, we gotta
    start."
FINAL SCORE: 10

**Example 2**
The video features two men sitting in a room,
    one of whom is holding a coffee cup and
    speaking to the other. The man holding the
    coffee cup is wearing a red and blue
    striped shirt, while the other man is
    wearing a white shirt. They are engaged in
    a conversation, and the man in the red and
    blue shirt is holding the coffee cup in his
    right hand. The room appears to be an
    office setting, with a desk and a chair
    visible in the background.

Previous conversation:
HOWARD: "Do you really think you should be
    drinking right now?"
RAJ: "How else am I supposed to talk to the
    Human Resources lady?"
HOWARD: "I don't know. Seek professional help?"
Utterance: RAJ: "I did. The guy at the liquor
    store said this stuff tastes great in
    coffee."
FINAL SCORE: 1
---
**Now you try**
Video description: {desc}
{context_text}{raw_utterance_text}
```

```
FINAL SCORE: [number between 1-10]
"""
```

B.4 Prompt to Calculate Accuracy

```
prompt = f""" You are given a video scene
    description, previous conversation context,
    and an utterance from that scene. Your job
    is to rate how likely the scene contains
    SARCASM, on a scale of 1-10, using this
    rubric:

1-5: NOT sarcastic
- 1: Strongly believe this is NOT sarcastic
- 5: Not sure, but leaning towards NOT sarcastic

6-10: Sarcastic
- 6: Not sure, but leaning towards sarcastic
- 10: Strongly believe this IS sarcastic

Video description: {description}
Context: {context}
Target Utterance: {utterance}
"""
```

B.5 Generated Output Examples

Agent provided preferences with knowledge of GT. After training, sometimes the VLM will hallucinate. For example, we see that the VLM is talking about a voiceover that doesn't exist. In the video, a man and a woman are seen lying on the ground, surrounded by fallen leaves, which suggests it's autumn. The man is wearing a blue sweater and the woman is in a blue top, both with their arms around each other. The man is speaking directly to the camera, and the woman is looking at it with a smile on her face. The man's tone is serious, while the woman's is more lighthearted. The man's voiceover explains that they're happy and in love, and the woman's voiceover is more playful, contrasting with his. The man's voiceover is in a serious tone, while the woman's is more lighthearted. The scene is set in a park with autumn leaves, and the man and woman are the only two people in it. The man's voiceover explains that they're happy and in love, while the woman's voiceover contrasts with his, adding a sense of humor to the scene.
In the following three examples, we see a different sentiment being expressed. **Example 1, baseline model:** In the video, we see a man sitting on a couch, holding a phone to his ear. He appears to be engaged in a conversation, possibly a phone call or a video call, as he is looking at the camera with a focused expression. The room around him is casually furnished, with a couch, a chair, and some posters on the wall. The man's attire is a sweater with a pattern, and he seems to be in a relaxed setting, possibly at home. The lighting is warm, and the overall atmosphere is calm. The video captures a moment of everyday life, with the man's interaction with the phone being the main focus. **Example 2, preference given with knowledge of GT:** In the video, we see a man sitting on a couch and talking on a phone. He's wearing a sweater with a distinctive checkered pattern and appears to be in a casual setting, possibly his living room. His expression is serious, and he's engaged in a heated conversation. The camera zooms in on his face, emphasizing his intensity. The contrast between his calm exterior and the inner turmoil is evident. **Example 3, preference given with in-context examples:** In the video, we see a young man sitting comfortably on a couch in a living room. He is dressed in a cozy sweater and is engaged in a phone conversation. His face is lit up with a smile, suggesting a pleasant conversation. The room around him is warmly decorated, with a poster of a man on the wall and a bookshelf filled with various items. The man on the phone is holding a phone to his ear, indicating an ongoing call. The living room is well-lit, with a lamp casting a soft glow on the scene. The man's relaxed posture and the casual setting suggest a comfortable and familiar environment. The video captures a moment of everyday life, showcasing the man's engagement in a phone call while enjoying his surroundings.

Table 6: Examples of VLM generated sequences.