Teaching Sarcasm: Few-Shot Multimodal Sarcasm Detection via Distillation to a Parameter-Efficient Student

Soumyadeep Jana and **Sanasam Ranbir Singh**
Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
{sjana ,ranbir}@iitg.ac.in

Abstract

Multimodal sarcasm detection is challenging, especially in low-resource settings where subtle image-text contradictions are hard to learn due to scarce annotated data, which hinders the model's performance. Parameter-efficient finetuning (PEFT) methods like adapters, LoRA, and prompt tuning reduce overfitting but struggle to reach optimal performance due to limited supervision from few-shot data. We propose **PEKD**, a unified framework that enhances PEFT methods via distillation from an expert model trained on large-scale sarcasm data, which acts as the teacher. To mitigate unreliable signals from the teacher, we introduce an entropy-aware gating mechanism that dynamically adjusts the distillation strength based on teacher confidence. Experiments on two public datasets demonstrate that our PEKD framework enables PEFT methods to outperform both prior parameter-efficient approaches and large multimodal models, achieving strong results in the few-shot scenario. The framework is modular and adaptable to a wide range of multimodal models and tasks. The code is available at `https://github.com/mr-perplexed/kd_sarcasm`.

1 Introduction

The use of multimodal (image+text) sarcasm has grown rapidly on social media, as it allows the users to express their harsh opinions in a veiled manner. Creation of large sarcasm datasets for tackling this problem is both costly and challenging (Davidov et al., 2010; González-Ibáñez et al., 2011), as sarcasm is often context-dependent, culturally nuanced, and difficult to annotate consistently (Rockwell and Theriot, 2001; Dress et al., 2008; Oprea and Magdy, 2019).

In light of this, recent works in the multimodal domain have explored techniques like prompt-learning and adapter-learning with pretrained language models (PLMs) for few-shot sarcasm (Liang et al., 2022b; Jana et al., 2024) and sentiment analysis (Yu and Zhang, 2022; Yu et al., 2022; Yang et al., 2022; Wu et al., 2024). These techniques have shown promise in few-shot settings, allowing the model to adapt effectively by introducing only a small number of learnable parameters while keeping the large pretrained backbone intact. *The core motivation behind adopting these techniques is to achieve parameter-efficient learning, as large parameter size tend to overfit when training data is scarce.* However, while these methods mitigate overfitting, they often underperform because they have limited data to learn and generalize from—a challenge we refer to as **supervision scarcity**. From Table 4, we observe that across both datasets, vanilla PEFT methods (without distillation) fall short of their distillation-enhanced variants by approximately 1.7–3.5% in accuracy under the 1% data setting.

To mitigate this supervision scarcity problem, we take inspiration from the teacher-student framework for *knowledge distillation (**KD**)* (Hinton et al., 2015) and propose **PEKD** (Parameter-Efficient Knowledge Distillation), a plug-and-play framework designed for few-shot multimodal sarcasm detection. This framework allows any model built with parameter-efficient modules, such as LoRA (Hu et al., 2021), adapter (Houlsby et al., 2019), or prompt tuning (Lester et al., 2021), to be easily plugged in as the student, while an expert model trained on large-scale sarcasm data serves as the teacher. This setup enables the student to generalize from limited examples without overfitting, while learning inductive biases from the teacher through knowledge distillation. A major challenge in distillation is unreliable teacher predictions, which can mislead the student. We address this with an *entropy-aware gating mechanism* that weights the distillation loss based on teacher confidence, providing strong guidance when the teacher is confident and reducing its influence when the teacher is uncertain. This ensures effective knowl-

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1338–1349
December 20-24, 2025 ©2025 Association for Computational Linguistics

edge transfer while avoiding noisy signals.

Compared to SOTA multimodal baselines, our framework PEKD helps PEFT methods achieve superior performance in the 1% data regime (**Δ2.2% to Δ5.2%**). Additionally, when compared against SOTA LVLMs, PEKD consistently helps PEFT techniques outperform across 5/10/20 shot settings, with LoRA even outperforming the best LVLM LLaMA-3.2-11B using 18× fewer trainable parameters. In addition to strong quantitative results, we conduct qualitative evaluations, including embedding space visualization, student-teacher representation alignment, and error mitigation analysis to understand how distillation benefits different PEFT variants. **Our contributions are:**

1. We propose PEKD, the first teacher-student framework combining PEFT with entropy-aware knowledge distillation for few-shot multimodal sarcasm detection.

2. We systematically evaluate and compare three PEFT variants—LoRA, Adapters, and Prompt Tuning, under our KD framework, and show that distillation consistently enhances their performance.

3. We conduct a comprehensive empirical analysis to reveal how KD improves representation alignment, prediction confidence, and reduces mismatched errors between student and teacher.

2 Related Work

Image-Text Sarcasm Detection
The rise of visual content on social media has spurred interest in multimodal sarcasm detection. Early efforts by Schifanella et al. (2016) used hand-crafted image-text features, followed by hierarchical and contrastive fusion approaches (Cai et al., 2019; Xu et al., 2020; Pan et al., 2020) that model intra-modal and inter-modal incongruity. Graph-based techniques (Liang et al., 2021, 2022a) model token-level and global-level incongruity while knowledge-enriched (Liu et al., 2022) models enhanced cross-modal reasoning. More recent advances include dynamic routing on modalities for capturing dominant modalitiy of sarcasm. (Tian et al., 2023), Qin et al. (2023) improved MMSD dataset, and proposed CLIP-based modeling.Tang et al. (2024) used retrieval-augmented instruction tuning while Xie et al. (2024) introduced parameter-efficient learning using mixture of adapters. Jana et al. (2024) proposed the first dedicated few-shot multimodal sarcasm detection model using prompt-tuning approach on BERT.

In this study, we address the problem of few-shot multimodal saracsm detection with PEFT and distillation techniques.

Multimodal Few-Shot Detection
Early efforts for few-shot sentiment analysis used prompting and prompt tuning in PLM (Yu and Zhang, 2022). Yu et al. (2022) used a pre-training task to align image prompts before downstream sentiment analysis task. Yang et al. (2022) fused discrete prompts through bayesian fusion for improving sentiment detection. For object detection (Zhou et al., 2021) inserted prompts within CLIP text encoders while (Zhou et al., 2022) used image-conditioned prompts. Gao et al. (2021) introduced adapters in CLIP for object detection. Jana et al. (2024) used prompt tuning with attentive prompts for few-shot multimodal sarcasm detection.

Our approach differs orthogonally by leveraging knowledge distillation with PEFT methods under a teacher-student setup for robust few-shot multimodal sarcasm detection.

3 Task Definition

The few-shot multimodal sarcasm detection task is a binary classification problem: given an image–text pair $\mathcal{X} = (I, T)$, the goal is to predict $y \in 0, 1$, where 1 denotes sarcasm. In the few-shot setting, we have a small support set $\mathcal{S} = \{(\mathcal{X}_i, y_i)\}_{i=1}^{N}$ with N equally split between sarcastic and non-sarcastic classes.

4 Proposed Approach

We choose CLIP (Radford et al., 2021) as the backbone for both student and teacher due to its strong multimodal grounding and proven effectiveness on sarcasm detection tasks (Qin et al., 2023; Xie et al., 2024), however, it can be extended to other vision-language models as well.

4.1 CLIP Preliminaries

Vision Encoder: The input image I is divided into patches and passed through L transformer blocks:

$$[\boldsymbol{z^i}, E^i] = \mathcal{E}_v^i([\boldsymbol{z^{i-1}}, E^{i-1}]), \quad i = 1, \dots, L \quad (1)$$

where $\mathcal{E}_v^i$ is the visual transformer at layer i, $E^i \in \mathbb{R}^{m \times d_v}$ denotes the patch embeddings from the i^{th} layer, and $\boldsymbol{z^i} \in \mathbb{R}^{1 \times d_v}$ is the embedding of the learnable class token. The final visual representation $\mathbf{h_{img}}$ is obtained by projecting the output class

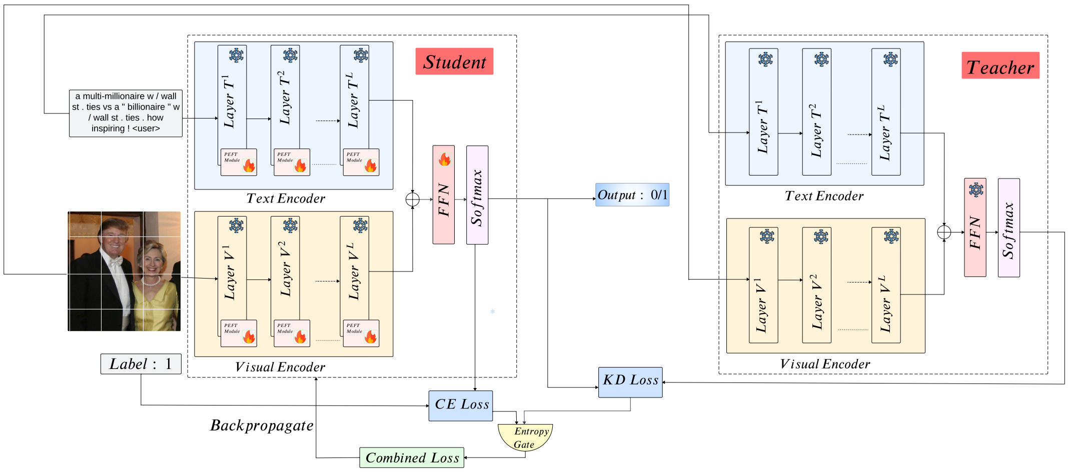


Figure 1: Architecture of PEKD framework.

token from the last layer L:

$$\mathbf{h_{img}} = \text{Proj}_{\mathcal{E}_v}(\boldsymbol{z^L}), \quad i \in \mathbb{R}^d \tag{2}$$

Text Encoder: A given input text sequence l is tokenized, embedded and passed through the L text transformer blocks $\mathcal{E}_t$:

$$W^i = \mathcal{E}_t^i(W^{i-1}), \quad i = 1, \ldots, L \tag{3}$$

where $W^i \in \mathbb{R}^{n \times d_t}$ represents the text embeddings from layer T^i. To derive the final textual representation, the embedding of the last token in the final layer L is projected into the shared embedding space:

$$\mathbf{h_{txt}} = \text{Proj}_{\mathcal{E}_t}(W^L[-1]), \quad t \in \mathbb{R}^d \tag{4}$$

4.2 Methodology

We propose PEKD, a parameter-efficient framework for few-shot multimodal sarcasm detection, illustrated in Fig. 1. It consists of two components: (a) A teacher model fully fine-tuned (updating all its parameters) on a large-scale sarcasm dataset, (b) a student augmented with PEFT modules for efficient adaptation. During training, both teacher and student process the same input, and their outputs are used to compute the KD loss along with the task loss. An entropy-aware gating mechanism combines these losses to regulate the teacher's influence based on confidence, and only the student's PEFT modules are updated. We elaborate on the details of the teacher, the student, and the fine-tuning of the student in the subsections below.

4.3 Teacher Model

Let $\mathcal{T}$ denotes the teacher, a pretrained CLIP model fine-tuned on a large sarcasm dataset. Due to its extensive parameterization and access to abundant training data, $\mathcal{T}$ captures intricate, cross-modal patterns of sarcasm. In few-shot settings, where direct fine-tuning of a large model can lead to overfitting, the teacher provides strong supervision and guides a parameter-efficient student, facilitating the transfer of its rich, sarcasm-sensitive knowledge. The teacher predicts the sarcasm label as:

$$\boldsymbol{y_{\mathcal{T}}} = \text{Softmax}(W_{\mathcal{T}} \cdot (\mathbf{h_{img}} \oplus \mathbf{h_{txt}})) \tag{5}$$

where $\mathbf{h_{img}}$ and $\mathbf{h_{txt}}$ are the image and text embeddings obtained from the respective CLIP encoders, $W_{\mathcal{T}}$ is the projection layer, $\oplus$ is the concatenation operator and $\boldsymbol{y_{\mathcal{T}}}$ is the soft logits from the teacher.

4.4 Parameter-Efficient Student Model

We design the student $\mathcal{S}$ as a parameter-efficient model that can be adapted to few-shot settings using a range of PEFT techniques. In this work, we focus on three such techniques, namely, *adapters*, *prompt-tuning*, and *LoRA* and apply them to CLIP. Our framework is flexible, allowing any PEFT method to be plugged in as the student model. Each technique adds only a small number of learnable parameters, keeping the rest of the CLIP backbone frozen, which reduces overfitting in few-shot settings. The trainable parameter sizes are specified in Table 1. We outline the adaptations of these techniques in CLIP in the following subsections.

4.4.1 Adapter-CLIP

In this setup, we insert *adapters* into every layer of CLIP's text and visual encoders. These adapters are lightweight bottleneck layers, consisting of a down-projection, a non-linearity, and an up-projection layer. Adapter Ad is realized as:

$$Ad(\boldsymbol{h}) = M_{up} \cdot \theta(M_{down} \cdot \boldsymbol{h}) \tag{6}$$

where M_{up} and M_{down} are upsample and downsample projection layers respectively and θ is ReLU non-linearity. The visual and text encoders in equations (1) and (3) can be modified as:

$$\begin{aligned}[\boldsymbol{z^i}, E^i] &= \mathcal{E}_v^i([\boldsymbol{z^{i-1}}, E^{i-1}]) \\ &\quad + \text{Ad}^i_{\mathcal{E}_v}([\boldsymbol{z^{i-1}}, E^{i-1}]) \\ &\qquad i = 1, .., L \end{aligned} \tag{7}$$

$$\begin{aligned}[W^i] &= \mathcal{E}_t^i([W^{i-1}]) + Ad^i_{\mathcal{E}_t}([W^{i-1}]) \\ &\qquad i = 1, .., L \end{aligned} \tag{8}$$

where $Ad^i_{\mathcal{E}_v}$ and $Ad^i_{\mathcal{E}_t}$ are the adapter layers in the i^{th} transformer block of the visual and text encoders, respectively.

The insertion of adapters reduces the model's trainable parameter size to around $3 \sim 4\%$ of the original CLIP model. During fine-tuning, only the adapters get trained while the CLIP backbone is frozen.

4.4.2 Prompt-CLIP

In this configuration, we introduce a small set of learnable embeddings, called *prompts*, that are added to the input of the text and visual encoders at every layer. After applying prompts, the visual and text encoders in equations (1) and (3) can be reformulated as:

$$\begin{aligned}[\boldsymbol{z^i}, E^i, P^i_{\mathcal{E}_v}] &= \mathcal{E}_v^i([\boldsymbol{z^{i-1}}, E^{i-1}, P^{i-1}_{\mathcal{E}_v}]) \\ &\qquad i = 1, .., L \end{aligned} \tag{9}$$

$$\begin{aligned}[P^i_{\mathcal{E}_t}, W^i] &= \mathcal{E}_t^i([P^{i-1}_{\mathcal{E}_t}, W^{i-1}]) \\ &\qquad i = 1, .., L \end{aligned} \tag{10}$$

where $P^{i-1}_{\mathcal{E}_v}$ and $P^{i-1}_{\mathcal{E}_t}$ are learnable prompt embeddings added to the input of the visual and text encoders i. This approach reduces the number of trainable parameters to roughly $0.02 \sim 0.03\%$ of the original CLIP model. During fine-tuning, only the prompts are trained, while the CLIP backbone remains frozen.

4.4.3 LoRA-CLIP

LoRA (Low-Rank Adaptation) updates a pre-trained weight matrix $W \in \mathbb{R}^{d_1 \times d_2}$ using a low-rank decomposition $\Delta W = BA$, where $A \in \mathbb{R}^{r \times d_2}$, $B \in \mathbb{R}^{d_1 \times r}$, and $r \ll \min(d_1, d_2)$. For input X, the adapted projection becomes:

$$\boldsymbol{h} = WX + \gamma(BA)X, \tag{11}$$

Model	Trainable Parameters
CLIP ViT-B/16 (Teacher)	149 M
LoRA-CLIP (Student)	2.9 M
Prompt-CLIP (Student)	0.03 M
Adapter-CLIP (Student)	4.1 M

Table 1: Trainable parameter sizes of teacher and student models.

Model	Train	Valid	Test
		MMSD	
Teacher (99%)	8543 / 11075 / 19618	860 / 1351 / 2211	959 / 1450 / 2409
Student (1%)	99 / 99 / 198	99 / 99 / 198	959 / 1450 / 2409
		MMSD2.0	
Teacher (99%)	9477 / 10141 / 19618	943 / 1269 / 2212	1037 / 1072 / 2409
Student (1%)	99 / 99 / 198	99 / 99 / 198	1037 / 1072 / 2409

Table 2: Train, valid, and test splits (Pos/Neg/Total) for teacher and student. Student sees only 1% of the training data, while the teacher sees the remaining 99%.

with γ as a scaling factor. We apply LoRA to the query, key, and value matrices of each attention layer in both text and vision encoders:

$$Q_{Lo} = XW_q + \gamma(B_q A_q)X \tag{12}$$

$$K_{Lo} = XW_k + \gamma(B_k A_k)X \tag{13}$$

$$V_{Lo} = XW_v + \gamma(B_v A_v)X \tag{14}$$

The attention operation now becomes:

$$\text{Attn} = \text{Softmax}\left(\frac{Q_{Lo} K_{Lo}^T}{\sqrt{d_k}}\right) V_{Lo} \tag{15}$$

LoRA reduces trainable parameters by around 1.9% of the original CLIP. Only the low-rank matrices A and B are optimized during training.

4.5 Fine-tuning the Student Model

To fine-tune the student, we use knowledge distillation (KD) in addition to the task-specific loss to enable the student to learn from both ground truth labels and the teacher's rich output distribution. This allows the student to mimick the teacher's rich sarcasm-specific cross-modal representations, which is difficult for the student to generalize from limited training examples.

We follow the same operations as the teacher (similar to equation 5) to get the final soft logits $\boldsymbol{y_{\mathcal{S}}}$ from the student. We fine-tune $\mathcal{S}$ on the few-shot data split, updating only the PEFT modules and the projection layer $W_{\mathcal{S}}$. The combined objective for fine-tuning the student is:

$$\mathcal{L} = g\mathcal{L}_{CE} + (1 - g)\mathcal{L}_{KD} \tag{16}$$

where, $\mathcal{L}_{CE}$ is the task-specific cross-entropy loss while $\mathcal{L}_{KD}$ is the knowledge-distillation loss from the teacher to the student.

The KD loss is computed as the KL divergence between teacher and student predictions:

$$\mathcal{L}_{KD} = T^2 \sum_{i=1}^{C} y_{\mathcal{T}}(i) \log \frac{y_{\mathcal{T}}(i)}{y_{\mathcal{S}}(i)}, \qquad (17)$$

where C is the number of classes and T^2 compensates for the temperature scaling effect. There might be cases when the teacher's predictions are less confident. In such scenarios, we want the student to rely on learning from the data rather than rely on the teacher. To achieve this, we introduce an entropy-aware gating parameter g based on normalized entropy of the teacher:

$$g = \frac{-\sum_{i=1}^{C} y_{\mathcal{T}}(i) \log y_{\mathcal{T}}(i)}{\log C}, \qquad (18)$$

where $\log C$ is the maximum possible entropy, ensuring $g \in [0, 1]$.

Lemma. Weighting $\mathcal{L}_{KD}$ by $(1 - g)$ ensures it approaches zero as teacher uncertainty increases ($g \to 1$) and remains intact when the teacher is confident ($g \to 0$). (Proof in Appendix A.4.)

Alternative Variant. We also experimented with a variant that sets the KD loss to zero when the teacher's prediction is incorrect; however, this approach underperformed compared to entropy-based gating. Detailed discussion is provided in Appendix A.5.

5 Experiments

5.1 Datasets

We assess our approach, PEKD, on the few-shot splits of MMSD (Cai et al., 2019) and MMSD2.0 (Qin et al., 2023), proposed in the study (Jana et al., 2024). They extract two 1% few-shot splits for each dataset, with an equal number of samples per class. The size of the test set remains unchanged from the original dataset. **We utilize the 1% split to train the student while the remaining 99% to train the teacher**. Detailed statistics are provided in Table 2.

5.2 Experimental Settings

We employ `ViT-B/16 CLIP` backbone for both the teacher and student. Few-shot training can exhibit significant performance variability. To capture this, we train the student 6 times (3 trials × 2 splits) for each dataset and report the average Accuracy (Acc), average Macro-F1 (F1), and the standard deviation across all six runs. Additional hyperparameter details for training the teacher and the student are reported in Appendix A.1.

5.3 Baselines

We evaluate our approach against two groups of baselines: PEFT and non-PEFT-based. For the non-PEFT based multimodal baselines, we consider the following SOTA models for sarcasm detection; **HFM** (Cai et al., 2019), **Att-BERT** (Pan et al., 2020), **HKE** (Liu et al., 2022), **DIP** (Wen et al., 2023), **MV-CLIP** (Qin et al., 2023), **DynRT** (Tian et al., 2023) and **RAG-LLaVA** (Tang et al., 2024). For the PEFT-based multimodal group, we consider **PVLM** (Yu and Zhang, 2022), **UP-MPF** (Yu et al., 2022), and **CAMP** (Jana et al., 2024) which are based on prompt-tuning in PLMs. **CoOp** (Zhou et al., 2021) and **CoCoOp** (Zhou et al., 2022) are prompt-tuning adaptations for CLIP. **MoBA** (Xie et al., 2024) is an adapter-based adaptation of CLIP. The details of these baselines are provided in the Appendix A.2. We evaluate all baselines on the few-shot data splits and report the performances.

6 Main Results

Following Jana et al. (2024), we report the few-shot performance of the models in Table 3. Our observations are: **(1)** PEFT methods often outperform or match their non-PEFT counterparts by leveraging task-specific parameters while keeping the pretrained backbone frozen, thus reducing overfitting. **(2)** When we train the teacher, `Teacher (CLIP)` (row 8 in Non-PEFT), on 1% few-shot split, we observe that it overfits and performs poorly due to its large parameter size. In contrast, the student models (`LoRA-CLIP (w/ KD)`, `Adapter-CLIP (w/ KD)`, and `Prompt-CLIP (w/ KD)`) benefit from KD by acquiring inductive biases from the teacher (trained on the remaining 99% split). As a result, they outperform the best baseline on MMSD, `MoBA` by gains ranging from **Δ2.2% to Δ4.6%** in Acc and `CoOp` on MMSD2.0 by **Δ2.2% to Δ5.2%**. **(3)** Among the students, `LoRA-CLIP (w KD)` demonstrates superior performance in the KD setup due to LoRA's ability to directly modify the attention mechanism of the pretrained model. By injecting low-rank updates into the Q, K, V matrices of the attention layers, LoRA enables the student to more precisely align its internal attention patterns with those of the teacher, unlike prompts (which affect only inputs)

	MMSD		MMSD2.0	
Method	Acc	F1	Acc	F1
Multimodal (Non PEFT)				
HFM	0.612 (1.3)	0.598 (1.1)	0.561 (0.2)	0.361 (0.3)
Att-BERT	0.707 (1.7)	0.696 (1.3)	0.659 (1.6)	0.683 (1.8)
HKE	0.503 (2.3)	0.667 (2.8)	0.408 (1.5)	0.579 (1.3)
DIP	0.704 (2.7)	0.698 (2.3)	0.685 (2.8)	0.658 (2.6)
DynRT	0.583 (0.1)	0.487 (0.6)	0.518 (2.9)	0.513 (3.2)
MV-CLIP	0.780 (0.2)	0.770 (0.2)	0.742 (0.4)	0.740 (0.3)
RAG-LLaVA	0.483 (9.1)	0.406 (4.6)	0.569 (0.1)	0.446 (8.3)
Teacher (CLIP)	0.777 (0.5)	0.768 (0.6)	0.740 (1.2)	0.739 (1.1)
Multimodal (PEFT)				
PVLM	0.712 (0.6)	0.699 (0.2)	0.665 (2.2)	0.658 (2.1)
UP-MPF	0.707 (2.4)	0.701 (2.6)	0.669 (0.4)	0.663 (0.1)
CAMP	0.729 (0.9)	0.717 (1.0)	0.692 (2.8)	0.681 (2.3)
CoOp	0.772 (1.6)	0.769 (1.5)	0.759 (0.9)	0.759 (0.8)
CoCoOp	0.782 (1.0)	0.779 (1.0)	0.746 (1.6)	0.745 (1.6)
MoBA	0.799 (1.2)	0.790 (1.1)	0.758 (0.7)	0.753 (0.9)
PEKD (Ours)				
Adapter-CLIP (w/ KD)	0.824 (0.2)	0.819 (0.3)	0.799 (0.1)	0.792 (0.3)
Prompt-CLIP (w/ KD)	0.821 (0.3)	0.811 (0.4)	0.781 (0.5)	0.798 (0.4)
LoRA-CLIP (w/ KD)	0.845 (0.3)	0.843 (0.1)	0.811 (0.2)	0.811 (0.2)

Table 3: Performance comparison of multimodal methods across few-shot MMSD and MMSD2.0 datasets. Our PEKD variants achieves SOTA performance. Numbers in brackets indicate standard deviation. The best , second-best , and third-best results are highlighted respectively. Our method outperforms baselines significantly with p < 0.05.

or adapters (which act more peripherally).

7 Comparison with LVLMs

We compare our PEKD-based models with SOTA LVLMs, LLaVA-1.6-7B (Liu et al., 2023), LLaMA-3.2-11B (et al., 2024), and Qwen-2.5-VL-7B (Bai et al., 2025) across extremely low-resource settings: 5/10/20-shots, as well as 1% supervision, as shown in Fig 4 for MMSD2.0 dataset. For fine-tuning of these LVLMs, we resort to LoRA with rank 8, due to resource constraints. Despite having drastically fewer trainable parameters (e.g., `Prompt-CLIP`: 0.03M, `Adapter-CLIP`: 4.1M, `LoRA-CLIP`: 2.9M vs. 40–52M for LVLMs), our models consistently outperform LVLMs in the 5/10/20-shot settings. Even at 1% supervision, PEKD-based methods reach near-comparable accuracy to LVLMs while `LoRA-CLIP` outperforms the LVLMs in this setting.

8 Ablation

8.1 KD Boosts PEFT Performance

Model	MMSD			MMSD2.0		
	Acc	F1	Δ	Acc	F1	Δ
Prompt-CLIP (w/o KD)	0.802	0.799	+1.9%	0.762	0.762	+1.9%
Prompt-CLIP (w/ KD)	0.821	0.811		0.781	0.798	
Adapter-CLIP (w/o KD)	0.807	0.803	+1.7%	0.771	0.778	+2.8%
Adapter-CLIP (w/ KD)	0.824	0.819		0.799	0.792	
LoRA-CLIP (w/o KD)	0.810	0.803	+3.5%	0.783	0.782	+2.8%
LoRA-CLIP (w/ KD)	0.845	0.843		0.811	0.811	

Table 4: Ablation on KD. Δ shows % improvement in Acc with KD. Green highlights indicate positive gains.

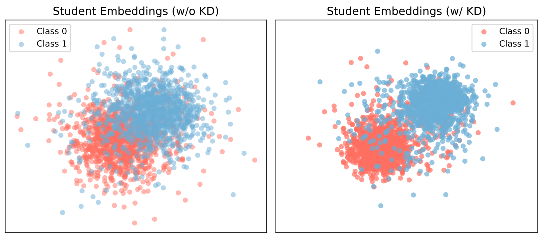


Figure 2: Comparison of student embeddings with and without KD.

Table 4 compares each PEFT variant with and without KD. Across all methods, KD consistently improves the performance of the student in the 1% few-shot setting. In contrast, `LoRA-CLIP (w/o KD)`, `Adapter-CLIP (w/o KD)`, and `Prompt-CLIP (w/o KD)`, without KD, underperform due to insufficient supervision. The students trained with KD outperforms as we mitigate the supervision scarcity problem by inducting knowledge from the teacher.

8.2 Effect of Entropy-Aware Gating

To assess entropy-aware gating, we analyze its effect on the weighting of KD loss during training. Figure 5 shows that higher teacher confidence corresponds to lower entropy (g), resulting in larger weights $(1 - g)$ for KD loss, thus amplifying reliable teacher signals. When confidence is low, g increases, reducing KD influence. Table 5 confirms that entropy-based gating consistently improves performance across all student variants.

9 Analysis

9.1 Impact of KD on Student

We study the impact of distilling knowledge from the teacher to the student from two perspectives: **(a)** *Effect of KD on the student's embedding space:*We visualize the logit representations of the student model trained with and without KD in Fig 2. The KD-trained student exhibits better class-wise separation and structured embeddings, suggesting improved discriminative capacity and alignment with the teacher's feature space. **(b)** *Improved predic-*

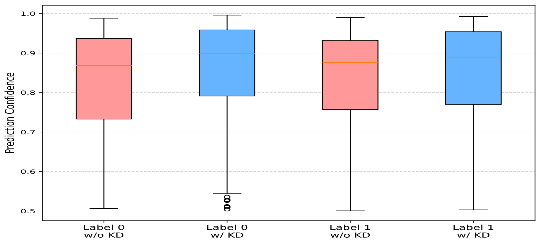


Figure 3: Class-wise prediction confidence comparison of student with and without KD.

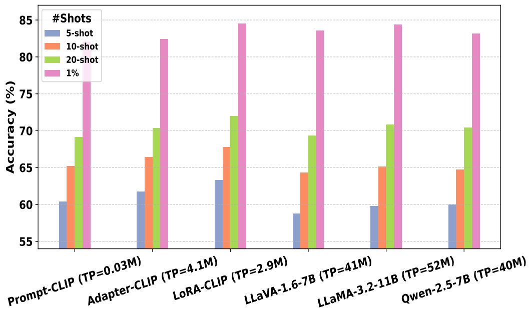


Figure 4: Comparison of PEKD-based models with LVLMs in 5/10/20 shots and 1% data setting for MMSD2.0 dataset. TP denotes trainable parameters.

tion confidence of the student: As shown in Fig 3, the student model trained with KD exhibits consistently higher prediction confidence compared to its non-KD counterpart, across both class labels. This is evident from the upward shift in the median values and a tighter interquartile range.

9.2 Strength of LoRA-CLIP

To evaluate the strength of `LoRA-CLIP` over `Adapter-CLIP` and `Prompt-CLIP`, we compute the the mean Canonical Correlation (CCA) between the hidden states of the teacher model and each student model across corresponding transformer layers, for both the text and vision branches. This metric quantifies how closely the student preserves the internal representational structure of the teacher. As shown in Fig 6, LoRA-CLIP consistently exhibits the highest CCA scores across all layers. This suggests that LoRA is more effective at preserving and transferring internal structural representations from the teacher model. Notably, the alignment gap between LoRA and other methods widens in deeper layers, indicating LoRA's superior ability to absorb hierarchical knowledge during distillation.

Model	MMSD			MMSD2.0		
	w/o Gating	w/ Gating	Δ	w/o Gating	w/ Gating	Δ
Prompt-CLIP	0.812	**0.821**	**+0.9%**	0.771	**0.781**	**+1.0%**
Adapter-CLIP	0.811	**0.824**	**+1.3%**	0.777	**0.799**	**+2.2%**
LoRA-CLIP	0.835	**0.845**	**+1.0%**	0.803	**0.811**	**+0.8%**

Table 5: Ablation on Entropy-Aware Gating: Accuracy improvements when applying entropy-based gating for KD. Δ shows percentage gain.

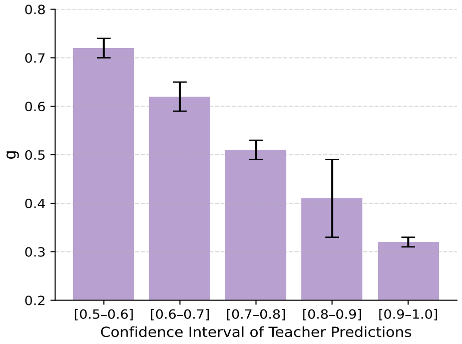


Figure 5: Effect of teacher confidence on KD gating: Lower confidence (higher uncertainty) results in reduced KD weight (1-g), demonstrating the adaptive behavior of entropy-aware gating.

9.3 Influence of KD on Student Errors

We study the impact of KD on student performance from two perspectives: **(a)** *Absolute reduction in student errors:* Fig 7 shows a class-wise comparison of the number of misclassified samples by the student with and without KD. Across both labels 0 & 1, KD consistently reduces the total number of prediction errors, demonstrating its effectiveness in improving student generalization. **(b)** *Mitigation of errors where the teacher was correct:* To assess whether the student learns from the teacher's correct predictions, we isolate those samples where the teacher was correct but the student was not. Fig 8 presents these student-teacher mismatches, which effectively represent cases where distillation has the potential to guide the student. After KD, the number of such mismatches drops substantially across both classes, showing that the student better aligns with the teacher's correct decisions.

9.4 Cross-Dataset Generalization

To evaluate the generalizability of our proposed framework, we conduct cross-dataset experiments by training all models on just 1% of the MMSD2.0 dataset and testing them on two unseen datasets: MCMD (Maity et al., 2022) and RedEval (Tang et al., 2024). The details of these datasets are

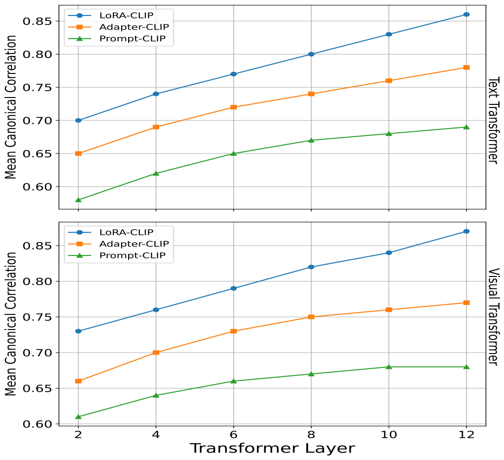


Figure 6: Layer-wise CCA similarity between teacher and student models across text and visual transformers.

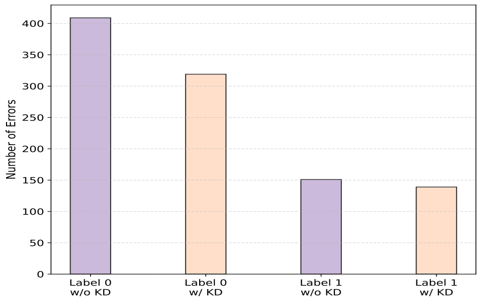


Figure 7: Class-wise error comparison of the student with and without KD.

provided in Appendix A.3. As shown in Table 6, PEKD-based models consistently outperform both PEFT and non-PEFT baselines, and even LVLMs across both datasets. Notably, the best-performing PEKD variant (`LoRA-CLIP`) achieves a substantial improvement over the strongest baseline (`Qwen2.5-VL-7B`). These results prove the generalizability of PEKD framework.

10 Error Analysis

To better understand the failure cases of our PEKD framework, we qualitatively analyze prediction errors and identify two broad patterns, shown in Table 7. First, both teacher and student models fail on (i) **OCR-heavy cases**, where sarcasm cues are embedded as image text (e.g., memes, screenshots), and (ii) **context-dependent cases**, where interpretation relies on external knowledge such as political or cultural context, inaccessible to the models (row 1). Second, we observe **student-specific failures** where the teacher predicts correctly but the student does not—often due to limited data exposure or weaker multimodal reasoning capacity (row 2).

Method	MCMD		RedEval	
	Acc	F1	Acc	F1
Multimodal (Non-PEFT)				
Att-BERT	0.477 (0.3)	0.474 (0.1)	0.461 (1.1)	0.457 (0.8)
DIP	0.545 (1.2)	0.545 (0.8)	0.532 (0.2)	0.513 (0.7)
DynRT	0.519 (1.6)	0.518 (1.4)	0.541 (0.6)	0.537 (0.9)
MV-CLIP	0.653 (0.2)	0.641 (0.2)	0.623 (1.1)	0.620 (1.3)
RAG-LLaVA	0.624 (1.7)	0.623 (1.5)	0.617 (1.9)	0.611 (1.4)
Multimodal (PEFT)				
PVLM	0.564 (1.8)	0.541 (1.3)	0.553 (1.2)	0.552 (1.1)
UP-MPF	0.582 (2.1)	0.577 (1.9)	0.569 (0.1)	0.561 (0.3)
CoOp	0.663 (0.4)	0.662 (0.2)	0.629 (0.9)	0.618 (0.7)
CoCoOp	0.658 (1.0)	0.649 (0.5)	0.637 (1.2)	0.631 (1.1)
CAMP	0.601 (1.3)	0.591 (1.6)	0.631 (0.7)	0.628 (1.2)
Large Vision-Language Models (LVLMs)				
LLaMA3.2-11B	0.693 (0.4)	0.682 (0.3)	0.641 (0.1)	0.653 (0.4)
LLaVA1.6-7B	0.674 (0.5)	0.669 (0.3)	0.642 (0.4)	0.641 (0.6)
Qwen2.5-VL-7B	0.691 (0.3)	0.685 (0.6)	0.656 (0.7)	0.659 (0.5)
PEKD (Ours)				
Prompt-CLIP (w/ KD)	0.713 (0.1)	0.698 (0.2)	0.686 (0.4)	0.683 (0.1)
Adapter-CLIP (w/ KD)	0.724 (0.2)	0.719 (0.3)	0.698 (0.2)	0.691 (0.4)
LoRA-CLIP (w/ KD)	**0.742 (0.1)**	**0.739 (0.2)**	**0.704 (0.2)**	**0.698 (0.4)**

Table 6: Cross-dataset evaluation. All models are trained on 1% of MMSD2.0 and tested on two unseen datasets: MCMD and RedEval. PEKD consistently boosts all PEFT variants. Numbers in brackets denote standard deviation. Best results are in **bold**.

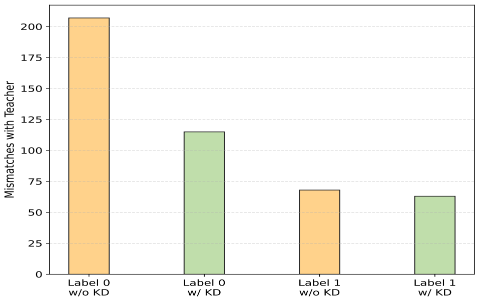
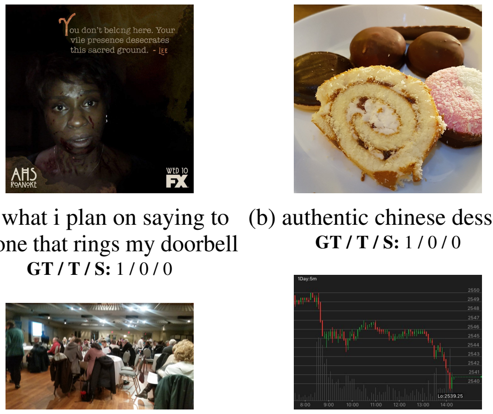


Figure 8: Unique prediction errors made by the student (i.e., errors made by the student but not by the teacher) across KD and non-KD settings.

11 Conclusion

In this work, we introduce **PEKD**, a framework for few-shot multimodal sarcasm detection that combines parameter-efficient tuning with knowledge distillation from a strong CLIP-based teacher. By guiding lightweight CLIP student variants through KD and incorporating an entropy-aware gating mechanism to prioritize reliable teacher signals, PEKD enhances robustness and performance in few-shot settings. Experiments on two benchmarks show consistent gains over parameter-efficient and large multimodal baselines, highlighting its potential for scalable vision-language understanding under data scarcity.


You don't belong here. Your vile presence desecrates this sacred ground.
AHS ROANOKE
WED 10 FX

(a) what i plan on saying to anyone that rings my doorbell
GT / T / S: 1 / 0 / 0

(b) authentic chinese desserts
GT / T / S: 1 / 0 / 0

(c) capacity room at basic income consultation in kingston
GT / T / S: 0 / 0 / 1

(d) major pullback ! ! 5pts
GT / T / S: 1 / 1 / 0

Table 7: Qualitative error examples with image, caption, and prediction (GT: ground truth, T: teacher, S: student). Refer to appendix A.6 for an explanation of the errors.

12 Limitations

While PEKD delivers notable improvements in few-shot multimodal sarcasm detection, it has some limitations. Both teacher and student primarily leverage visual-textual content and can struggle in OCR-heavy scenarios or cases requiring external context. Lastly, the entropy-based weighting of the KD signal is a simple confidence proxy and may not fully capture nuanced uncertainties of sarcasm. Future work could explore richer reliability estimation and integration of external knowledge to improve robustness and reasoning.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *ArXiv*.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, Uppsala, Sweden. Association for Computational Linguistics.

Megan L. Dress, Roger J. Kreuz, Kristen E. Link, and Gina M. Caucci. 2008. Regional variation in the use of sarcasm. *Journal of Language and Social Psychology*, 27:71 – 85.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *ArXiv*.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Jiao Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *ArXiv*.

Roberto I. González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Annual Meeting of the Association for Computational Linguistics*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *ArXiv*.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*.

Soumyadeep Jana, Animesh Dey, and Ranbir Singh Sanasam. 2024. Continuous attentive multimodal prompt tuning for few-shot multimodal sarcasm detection. *Proceedings of the 28th Conference on Computational Natural Language Learning*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. *Proceedings of the 29th ACM International Conference on Multimedia*.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022a. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Annual Meeting of the Association for Computational Linguistics*.

Sheng Liang, Mengjie Zhao, and Hinrich Schütze. 2022b. Modular and parameter-efficient multimodal fusion with prompting. *ArXiv*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hui Liu, Wenya Wang, and Haoliang Li. 2022. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Silviu Oprea and Walid Magdy. 2019. Exploring author context for detecting intended vs perceived sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. MMSD2.0: Towards a reliable multi-modal sarcasm detection system. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Patricia Rockwell and Evelyn M. Theriot. 2001. Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis. *Communication Research Reports*, 18:44 – 52.

Rossano Schifanella, Paloma de Juan, Joel R. Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. *Proceedings of the 24th ACM international conference on Multimedia*.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. 2024. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In *North American Chapter of the Association for Computational Linguistics*.

Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023. Dynamic routing transformer network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Chan Shao Wen, Guoli Jia, and Jufeng Yang. 2023. Dip: Dual incongruity perceiving network for sarcasm detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zichen Wu, Hsiu-Yuan Huang, Fanyi Qu, and Yunfang Wu. 2024. Mixture-of-prompt-experts for multi-modal semantic understanding. In *International Conference on Language Resources and Evaluation*.

Yifeng Xie, Zhihong Zhu, Xin Chen, Zhanpeng Chen, and Zhiqi Huang. 2024. Moba: Mixture of bi-directional adapter for multi-modal sarcasm detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*. Association for Computing Machinery.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Xiaocui Yang, Shi Feng, Daling Wang, Pengfei Hong, and Soujanya Poria. 2022. Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts. *Proceedings of the 31st ACM International Conference on Multimedia*.

Yang Yu and Dong Zhang. 2022. Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling. *2022 IEEE International Conference on Multimedia and Expo (ICME)*.

Yang Yu, Dong Zhang, and Shoushan Li. 2022. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. *Proceedings of the 30th ACM International Conference on Multimedia*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2021. Learning to prompt for vision-language models. *International Journal of Computer Vision*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

A Appendix

A.1 Hyperparameter Details

The teacher was fine-tuned till the best val accuracy was reached, with learning rates of 1e-5 for the backbone and 1e-4 for the head and a batch size of 32. The student was fine-tuned with a learning rate of 1e-4, batch size of 32, with a LoRA rank of 32. The best model on the validation set was used for testing. The value of temperature T for the KD loss was set to 2. All experiments were conducted on an Nvidia RTX A5000 GPU with 24GB of memory.

A.2 Baseline Details

To establish a comprehensive comparison, we evaluate our method against two categories of multimodal sarcasm detection baselines: (i) **Non-PEFT-based methods**, which primarily focus on modeling image-text interactions through hierarchical fusion, graph networks, co-attention, or dynamic routing, and (ii) **PEFT-based methods**, which leverage prompts and adapters to address few-shot multimodal sentiment, sarcasm and image recognition tasks.

Non-PEFT Baselines:

- **HFM** (Cai et al., 2019): Introduces hierarchical early and late fusion to integrate image, text, and image attributes.
- **D&R Net** (Xu et al., 2020): Uses decomposition and relational reasoning to capture semantic associations.
- **Att-BERT** (Pan et al., 2020): Employs co-attention to detect intra- and inter-modal incongruity.
- **HKE** (Liu et al., 2022): Leverages a hierarchical network to model coarse- and fine-grained incongruities.
- **MV-CLIP** (Qin et al., 2023): Extends CLIP with an interaction layer for enhanced image-text incongruity modeling.
- **DIP** (Wen et al., 2023): Captures sarcasm via semantic reweighting, uncertainty modeling, and contrastive learning at factual and affective levels.
- **DynRT** (Tian et al., 2023): Employs dynamic routing to identify sarcasm-relevant tokens in multimodal data.
- **RAG-LLaVA** (Tang et al., 2024): Incorporates retrieval-based demonstrations to assist LLaVA for sarcasm detection.

PEFT baselines:

- **PVLM** (Yu and Zhang, 2022): Adopts prompt-based fine-tuning with integrated image tokens for few-shot multimodal sentiment analysis.
- **UP-MPF** (Yu et al., 2022): Pretrains on image tasks to bridge the gap between textual and visual prompts before applying to few-shot sentiment tasks.
- **CoOp** (Zhou et al., 2021): Learns continuous text prompts for CLIP for image recognition.
- **CoCoOp** (Zhou et al., 2022): Introduces image-conditioned text prompts for improved adaptation for CLIP for image recognition.
- **CAMP** (Jana et al., 2024): Proposes continuous attentive prompt tokens in BERT for few-shot multimodal sarcasm detection.
- **MoBA** (Xie et al., 2024): Proposed mixture of adapters-based technique to combine visual and textual information for multimodal sarcasm detection, achieving parameter efficiency.

Dataset	Sarcastic	Non-Sarcastic	Total
MCMD	183	123	306
RedEval	395	609	1004

Table 8: Statistics of datasets used for cross-dataset evaluation.

A.3 Cross-Dataset Details

To evaluate the generalization ability of multimodal sarcasm detection models, we introduce two out-of-distribution datasets: **MCMD** and **RedEval**.

- **MCMD (Multi-modal Code-Mixed Memes Dataset)** (Maity et al., 2022): This dataset consists of code-mixed memes originally meant for hateful meme detection but also has annotations for sarcasm. We remove samples that are code-mixed in nature or has no sarcasm annotations.
- **RedEval** (Tang et al., 2024): Constructed to assess domain shift, RedEval includes image-text pairs from *Reddit*, since existing MMSD datasets are Twitter-centric. Sarcastic examples are sourced from the *sarcasm* subreddit, while non-sarcastic examples come from subreddits like *aww*, *funny*, and *pics*.

  The detailed statistics for both these datasets are in Table 8

Model	MMSD2.0			MMSD		
	Hard-Gate	Entropy	Δ	Hard-Gate	Entropy	Δ
Prompt-CLIP	0.775	**0.781**	+0.6%	0.815	**0.821**	+0.6%
Adapter-CLIP	0.791	**0.799**	+0.8%	0.817	**0.824**	+0.7%
LoRA-CLIP	0.803	**0.811**	+0.8%	0.834	**0.845**	+1.1%

Table 9: Comparison of KD Variants. Entropy-gated KD outperforms Hard-Gate KD across models and datasets. Δ shows accuracy gain.

A.4 Proof for Lemma 1

Proof. The entropy-aware gating factor is defined as:

$$g = \frac{H(y_{\mathcal{T}})}{\log C}, \quad H(y_{\mathcal{T}}) = -\sum_{i=1}^{C} y_{\mathcal{T}}(i) \log y_{\mathcal{T}}(i), \tag{19}$$

where $\log C$ is the maximum possible entropy for C classes, normalizing $g \in [0, 1]$. The gated KD loss:

$$\mathcal{L}_{KD}^{gated} = (1 - g)\, \mathcal{L}_{KD}. \tag{20}$$

Case 1: Teacher is highly uncertain. If $y_{\mathcal{T}}(i) \approx \frac{1}{C}$ for all i, then:

$$H(y_{\mathcal{T}}) = -\sum_{i=1}^{C} \tfrac{1}{C} \log \tfrac{1}{C} = \log C, \tag{21}$$

$$g = \tfrac{\log C}{\log C} = 1, \quad \mathcal{L}_{KD}^{gated} = 0. \tag{22}$$

Case 2: Teacher is highly confident. If teacher predicts one class with high probability:

$$H(y_{\mathcal{T}}) \approx 0, \quad g = 0, \quad \mathcal{L}_{KD}^{gated} = \mathcal{L}_{KD}. \tag{23}$$

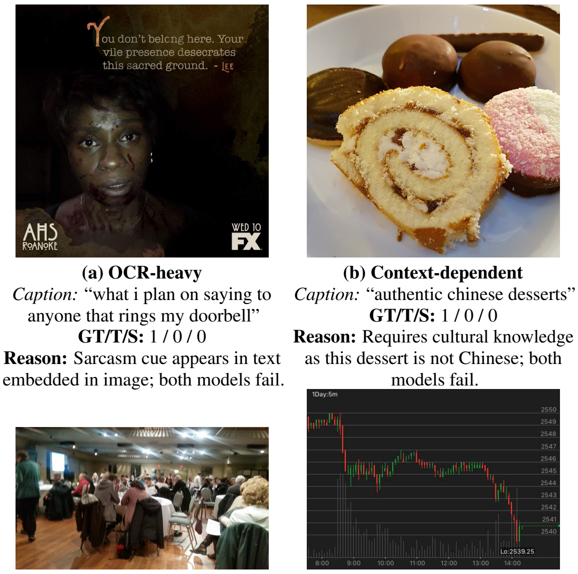
Conclusion: Weighting by $(1 - g)$ makes KD vanish when $g \to 1$ (uncertain teacher) and fully retain it when $g \to 0$ (confident teacher). This prevents noisy guidance while exploiting strong teacher signals.

□

A.5 Effect of Hard-Gate on KD Loss

We experimented with an alternative KD strategy, **Hard-Gate KD**, which discards the KD loss whenever the teacher prediction is incorrect. Table 9 compares this approach against our **Entropy-Gated KD**. The entropy-based approach consistently outperforms the hard-zero variant by **0.6–1.1% in accuracy** across both datasets and all PEFT configurations. *Reason:* Hard-Gate KD eliminates valuable soft-label information even in cases where the teacher prediction is slightly off but provides useful class-probability distribution. Entropy gating, on the other hand, dynamically reduces KD influence without discarding it entirely, leading to better knowledge transfer.

A.6 Detailed Error Analysis


You don't belong here. Your vile presence desecrates this sacred ground. - LEE
AHS ROANOKE
WED 10 FX
1Day:5m
Lo:2539.25

(a) OCR-heavy
Caption: "what i plan on saying to anyone that rings my doorbell"
GT/T/S: 1 / 0 / 0
Reason: Sarcasm cue appears in text embedded in image; both models fail.

(b) Context-dependent
Caption: "authentic chinese desserts"
GT/T/S: 1 / 0 / 0
Reason: Requires cultural knowledge as this dessert is not Chinese; both models fail.

(c) Student-only
Caption: "capacity room at basic income consultation"
GT/T/S: 0 / 0 / 1
Reason: Student overfits sarcastic patterns, might think that sparse room contradicts caption.

(d) Student-only
Caption: "major pullback!! 5pts"
GT/T/S: 1 / 1 / 0
Reason: Sarcasm via exaggeration; student misses numeric reasoning.

Figure 9: **Qualitative Error Analysis:** Failure cases with GT (Ground Truth), T (Teacher), and S (Student) predictions.