Reasoning Beyond Literal: Cross-style Multimodal Reasoning for Figurative Language Understanding

Seyyed Saeid Cheshmi Hahnemann Ortiz James Mooney Dongyeop Kang
University of Minnesota
{chesh014, orit0001, moone174, dongyeop}@umn.edu

Abstract

Vision–language models (VLMs) have demonstrated strong reasoning abilities in literal multimodal tasks such as visual mathematics and science question answering. However, figurative language—such as sarcasm, humor, and metaphor—remains a significant challenge, as it conveys intent and emotion through subtle incongruities between expressed and intended meanings. In multimodal settings, accompanying images can amplify or invert textual meaning, demanding models that reason across modalities and account for subjectivity. We propose a three-step framework for developing *efficient multimodal reasoning models* that can (i) interpret multimodal figurative language, (ii) provide transparent reasoning traces, and (iii) generalize across multiple figurative styles. Experiments across four styles show that (1) incorporating reasoning traces substantially improves multimodal figurative understanding, (2) reasoning learned in one style can transfer to others—especially between related styles like sarcasm and humor, and (3) training jointly across styles yields a generalized reasoning VLM that outperforms much larger open- and closed-source models. Our findings show that lightweight VLMs with verifiable reasoning achieve robust cross-style generalization while providing inspectable reasoning traces for multimodal tasks. The code and implementation are available at https://github.com/scheshmi/CrossStyle-MMR.

1 Introduction

Human communication often utilizes expressions that convey meanings beyond their literal interpretation. Figurative language, including sarcasm, humor, metaphor, and offense, is essential in expressing intent, emotion, and perspective. Understanding such expressions is vital for sentiment analysis, social media moderation, and human–AI interaction (Yang et al., 2022; Stappen et al., 2024). However, figurative communication often depends on subtle incongruities between what is said and what is meant (Camp, 2020), making it particularly difficult for language models (Jang and Frassinelli, 2025).

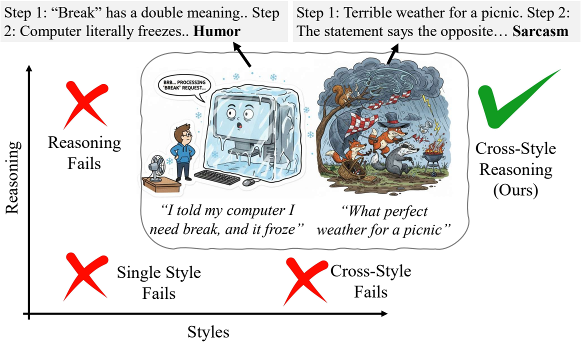

Step 1: "Break" has a double meaning.. Step 2: Computer literally freezes.. **Humor**
Step 1: Terrible weather for a picnic. Step 2: The statement says the opposite… **Sarcasm**
BRB... PROCESSING 'BREAK' REQUEST...
Reasoning Fails
"I told my computer I need break, and it froze"
"What perfect weather for a picnic"
Cross-Style Reasoning (Ours)
Single Style Fails
Cross-Style Fails
Reasoning
Styles

Figure 1: Previous work has focused on single styles and has not explored multimodal figurative language understanding. Our study examines whether incorporating reasoning can improve multimodal figurative understanding, and further, whether it can enable the development of a reasoning model capable of understanding multiple styles.

The challenge becomes even more complex in multimodal contexts, where text is paired with images, videos, or other signals. In such cases, visual cues can amplify or invert textual meaning (Wang et al., 2025), requiring models to detect semantic discrepancies across modalities and to reason about subjective interpretations (Zhou et al., 2025). Despite rapid progress in vision–language models (VLMs), current systems remain limited in handling these nuanced multimodal figurative phenomena (Bojić et al., 2025).

Recent advances in reasoning language models have shown that chain-of-thought (CoT) prompting can substantially improve interpretability and performance across analytical domains such as mathematics and science (Lu et al., 2023, 2022; Kembhavi et al., 2016). Yet, whether such reasoning mechanisms can enhance multimodal figurative lan-

Findings of the Association for Computational Linguistics: EACL 2026, pages 5942–5956
March 24-29, 2026 ©2026 Association for Computational Linguistics

guage understanding remains unexplored. Developing this capability requires datasets containing explicit reasoning traces—resources that are notably scarce for figurative language.

To bridge this gap, we introduce a three-step reasoning framework for multimodal figurative understanding. Building on Ho et al. (2022); Zhang et al. (2024), we first distill CoT reasoning paths from a large open-source teacher model (Grattafiori et al., 2024). We then enhance a lightweight student model through Supervised Fine-Tuning (SFT) on these distilled traces, followed by Reinforcement Learning with Verifiable Rewards (RLVR). This pipeline enables small VLMs to internalize structured reasoning and generate interpretable predictions for figurative multimodal tasks.

Extensive experiments across four figurative styles—sarcasm, humor, offense, and metaphor—demonstrate that: (1) Incorporating CoT reasoning markedly improves multimodal figurative understanding. (2) Reasoning knowledge learned in one style can transfer to related styles, particularly between sarcasm and humor. (3) Joint training across all styles produces a generalized reasoning VLM that outperforms much larger open- and closed-source models (e.g., Gemini 2.5 Flash, LLaMA-90B-Vision-Instruct) on most benchmarks.

To our knowledge, this is the first work to induce structured cross-style reasoning in VLMs using RLVR. Our approach establishes a lightweight but powerful framework for generalizable multimodal figurative language understanding with transparent decision paths.

2 Related Work

2.1 Figurative Language Understanding

Figurative language has been extensively studied in Natural Language Processing (NLP), with research spanning sarcasm, metaphor, idioms, humor, and hyperbole. Early approaches relied on textual cues and incongruity signals, while recent datasets have broadened coverage across multiple styles. Chakrabarty et al. (2022) released FLUTE, a dataset of 9,000 figurative instances (sarcasm, simile, metaphor, idioms), created via a model-in-the-loop approach with GPT-3 and human annotators, enabling progress in figurative language understanding. In Jang et al. (2023), authors probed the performance of several models on figurative language classification across sarcasm, similes, idioms, and metaphors, conducting experiments to highlight key differences in model behavior and to analyze the linguistic characteristics of these four figurative types. Moreover, Lai et al. (2023) investigated multilingual detection across several figurative styles, highlighting the promise of unified modeling.

Beyond text-only approaches, recent work has explored figurative language understanding in multimodal settings. Saakyan et al. (2023) introduced V-FLUTE, a dataset for visual figurative entailment with explanations across metaphors, idioms, sarcasm, and humor, showing that even strong language models struggle to generalize across styles, particularly in images. Building on V-FLUTE, the figurative language shared task consolidated multi-style datasets and emphasized identifying visual entailment relationships in multimodal instances (Kulkarni et al., 2024).

Additionally, the Multimodal Sarcasm Explanation (MuSE) task formalized the need for models not only to detect but also to justify figurative intent in image-text pairs (Desai et al., 2022). Although a few studies have investigated figurative language detection through reasoning (Tian et al., 2024; Yao et al., 2025) or explainable annotations or eye movements (Hayati et al., 2021; de Langis and Kang, 2023), they focus on a single style (sarcasm or metaphor) or text-only settings and cross-style setting (Kang and Hovy, 2019; Das et al., 2023) has not yet been explored. In this work, we study multimodal reasoning for figurative language and investigate cross-style knowledge transfer.

2.2 Multimodal Reasoning

The reasoning capabilities of VLMs have been studied primarily in domains such as mathematics (Wang et al., 2024) and science (Lu et al., 2022; Kembhavi et al., 2016). Approaches such as RePIC (Oh et al., 2025) and VLM-R1 (Shen et al., 2025) typically apply reinforcement learning to align a vision–language backbone with verifiable, task-level rewards but do not explicitly model multi-step chain-of-thought reasoning. These methods replace preference models with deterministic validators (e.g., exact-match or BLEU) and optimize the policy using Group Relative Policy Optimization (GRPO) or Proximal Policy Optimization (PPO) variants with a KL regularizer. Other approaches, such as SVQA-R1 (Wang and Ling, 2025) and STAR-R1 (Li et al., 2025), employ view-consistent or transformation-invariant objectives for Visual

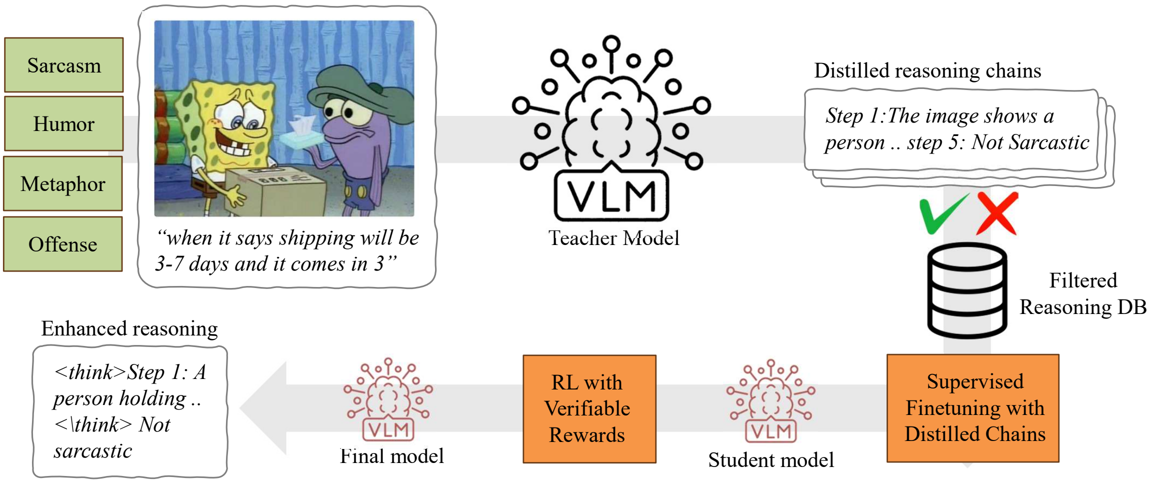

Sarcasm
Humor
Metaphor
Offense
"when it says shipping will be 3-7 days and it comes in 3"
VLM
Teacher Model
Distilled reasoning chains
Step 1:The image shows a person .. step 5: Not Sarcastic
✓ ✗
Filtered Reasoning DB
Supervised Finetuning with Distilled Chains
VLM
Student model
RL with Verifiable Rewards
VLM
Final model
Enhanced reasoning
<think>Step 1: A person holding .. <\think> Not sarcastic

Figure 2: Overall workflow of the proposed method.

Question Answering (VQA). More recent work such as Visionary-R1 (Xia et al., 2025), enforces image interpretation prior to reasoning to mitigate shortcut reliance and improve generalization. However, these reasoning frameworks have primarily been validated on objective tasks like mathematics and science, where ground truth is objectively verifiable. It remains an open question whether such verifiable reasoning pipelines can be effectively adapted to the subjective and nuanced domain of figurative language understanding.

3 Method

Our goal is to equip compact VLMs with explicit multimodal reasoning capabilities that enable them to interpret and generalize across different figurative styles.

Building on recent findings that large models can serve as reasoning teachers (Ho et al., 2022) and that RLVR can enhance self-evolved reasoning in language models (Wen et al., 2025; Guo et al., 2025), we propose a three-stage reasoning framework illustrated in Figure 2. The approach consists of: (1) Chain-of-Thought Distillation from a large teacher VLM to extract structured reasoning traces; (2) Supervised Fine-Tuning on a smaller model to internalize these reasoning patterns; and (3) RLVR-based optimization using verifiable rewards from ground-truth labels to further refine reasoning consistency and factual correctness. Importantly, this framework can be applied both per style and in a combined training regime, the latter producing our best-performing, generalized reasoning VLM.

3.1 Step 1: Chain-of-Thought Extraction from the Teacher

We first employ a large VLM as a reasoning teacher to generate structured CoT explanations for multiple figurative styles (sarcasm, humor, offense, and metaphor) under zero-shot prompting. Each CoT follows a standardized five-step schema: image description, caption interpretation, mismatch detection, intent inference, and final label prediction. Table 7 provides some examples of generated CoTs for different styles.

Generated samples are automatically filtered to remove malformed or incorrect reasoning paths, retaining only valid CoTs for downstream use. This step yields a high-quality reasoning corpus that captures multimodal cues, pragmatic intent, and contextual inconsistencies—knowledge rarely available in existing datasets. Prompts and filtering rules are provided in Appendix A.1.

3.2 Step 2: Chain-of-Thought Distillation via Supervised Fine Tuning

Next, we fine-tune a smaller student model on the distilled reasoning corpus. This SFT stage acts as a reasoning warm-up, teaching the model to emulate the teacher's explicit CoT style and to produce interpretable intermediate steps. Through this distillation, the smaller VLM acquires a foundation for step-wise multimodal reasoning while preserving computational efficiency. The SFT model serves as the reference policy for subsequent RLVR training, ensuring stable optimization and alignment between reasoning structure and prediction accuracy.

3.3 Step 3: Reinforcement Learning with Verifiable Rewards

Finally, we refine the SFT model using RLVR implemented via GRPO algorithm (Guo et al., 2025). Unlike Reinforcement Learning from Human Feedback (RLHF), RLVR replaces human preference models with deterministic verification functions, enabling scalable, objective evaluation of prediction accuracy and structural adherence.

Given an input instance q, the model's output o—comprising both a reasoning trace and a final answer—is evaluated using a binary verifiable reward function $R(q, o)$ that measures factual correctness and adherence to the required reasoning format:

$$R = R_{\text{acc}} + R_{\text{format}}$$

where

$$R_{\text{acc}} = \begin{cases} 1, & \text{if prediction matches ground truth,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The RL objective is to maximize the expected verifiable reward while maintaining proximity to the SFT reference policy through KL-regularized optimization:

$$\max_{\pi_\theta} \mathbb{E}_{o\sim\pi_\theta(q)}[R(q, o)] - \beta \, \mathrm{KL}[\pi_\theta(o|q) \,||\, \pi_{\text{ref}}(o|q)] \,. \tag{2}$$

This process reinforces accurate and well-structured reasoning traces without relying on subjective human labels.

The model's output is expected to include two parts: a reasoning process, enclosed within `<think></think>` tags, which explains how the model integrates visual and textual cues to arrive at its prediction, and a final prediction, enclosed within `<answer></answer>` tags. If the output satisfies these formatting constraints, the format reward is assigned a value of 1; otherwise, it is set to 0. The prompt used here can be found in the Appendix A.2.

3.4 Unified Training for Generalized Figurative Reasoning

While the three stages can be executed independently for each figurative style, our **combined training strategy**, performing SFT + RLVR jointly on all styles, achieves the best overall performance. This unified optimization enables **cross-style knowledge transfer**, allowing the model to leverage shared reasoning patterns across different styles.

4 Experimental Setup

4.1 Style Selection and Datasets

We focus on four figurative language styles—*sarcasm*, *humor*, *offense*, and *metaphor*. We leverage three multimodal datasets covering diverse figurative styles. Each dataset provides paired textual and visual information, enabling analysis of cross-modal incongruities central to figurative expression.

MMSD2.0 (Qin et al., 2023). The Multi-Modal Sarcasm Detection 2.0 benchmark contains roughly 25,000 image–text pairs annotated for sarcasm. It offers rich context for studying multimodal incongruity and intent recognition.

Memotion (Sharma et al., 2020). This SemEval-2020 Task 8 dataset comprises about 9,800 memes annotated for multiple emotional dimensions, including *humor* and *offense*. Its varied labels and human-annotated judgments make it ideal for evaluating subjectivity and overlapping stylistic cues.

MultiMET (Zhang et al., 2021). MultiMET includes 10,437 image–text pairs annotated for *metaphor* detection. It distinguishes text-dominant, image-dominant, and complementary metaphors, supporting analysis of abstract, cross-modal reasoning.

For all datasets, we adopt standard splits: the default training and test sets for MMSD2.0 and Memotion, and an 80/20 split for MultiMET. CoT reasoning traces are distilled exclusively from the training data of each dataset.

4.2 Models

We employ LLaMA3.2-90B-Vision-Instruct (Grattafiori et al., 2024) as the reasoning teacher model for the first step of our approach, and use Qwen2.5-VL-3B-Instruct (Bai et al., 2025) as the student model in the second and third steps. All experiments are formulated as binary classification tasks, where the goal is to determine whether an image–text pair expresses a specific figurative meaning. Our primary model backbone is `Qwen2.5-VL-3B-Instruct`, a compact yet capable VLM chosen to evaluate reasoning scalability under constrained model size. All model evaluations and training are performed with **zero-shot prompting**, ensuring fair and consistent comparisons across methods.

Table 1: Training model configurations.

Setup	Description
Zero-shot CoT	Prompt teacher/student to produce CoT without gradient updates
SFT-Binary	Supervised fine-tuning on gold binary labels only
SFT-CoT	Supervised fine-tuning on distilled teacher CoTs
GRPO-only	RLVR via GRPO starting from base model (no SFT)
SFT→GRPO	SFT-CoT warm-up followed by RLVR via GRPO (full pipeline)

We evaluate five experimental configurations to assess the contribution of each training stage: (1) `Zero-shot CoT Prompting`, (2) `SFT on Binary Labels`, (3) `SFT on Distilled CoTs`, (4) `GRPO-only (RLVR without SFT)`, and (5) `SFT → GRPO (Full Pipeline)`. Table 1 includes detailed configurations.

4.3 Implementation Details

SFT is performed for 5 epochs with a learning rate of 2e–4 using cosine annealing. RLVR via GRPO is conducted for 2 epochs with a learning rate of 1e–5 and 8 rollouts per sample.

All experiments are executed on two NVIDIA H100 GPUs. The SFT stage equips the model with baseline reasoning capabilities, while RLVR refines its verifiable reasoning and alignment between reasoning traces and final predictions. Together, these stages enable the development of a compact, interpretable, and generalized multimodal reasoning VLM.

5 Results

We evaluate our approach on three questions: *(1) Does CoT reasoning improve multimodal figurative understanding? (2) Can reasoning learned for one figurative style transfer to others? (3) Can a single generalized reasoning VLM match or surpass much larger models?*

5.1 Does CoT reasoning enhance multimodal figurative understanding?

Observation. Table 2 presents the accuracy and F1 score results for Qwen2.5-VL-3B-Instruct across four different styles under the following setups: zero-shot CoT prompting, SFT on binary labels, SFT on distilled CoT, GRPO-only, and SFT + GRPO on the same style. As shown, fine-tuning the model on binary labels outperforms zero-shot CoT prompting across all styles, which aligns with previous findings (Bojić et al., 2025) highlighting the challenges of figurative language understanding for LLMs and VLMs. Furthermore, performing SFT on distilled CoTs to enable reasoning capabilities yields notable gains: $\sim 1.9\%$ accuracy improvement on sarcasm detection, $\sim 4.5\%$ on humor detection, $\sim 2.4\%$ on offense detection, and $\sim 2.6\%$ on metaphor detection. Building on this, applying RLVR with GRPO on top of the SFT models leads to even greater improvements $\sim 3\%$ for sarcasm detection, $\sim 8\%$ for humor detection, $\sim 7\%$ for offense detection, and $\sim 7\%$ for metaphor detection. These results demonstrate that RLVR effectively enhances CoT reasoning when the model is first equipped with initial reasoning capabilities through SFT on distilled CoTs. Additionally, results from the GRPO-only setup show that while some improvement over zero-shot CoT prompting is possible, the absence of initial CoT reasoning and the inherent complexity of figurative language highlight the critical role of SFT on CoTs as a warm-up step.

5.2 Can CoT learned in one style transfer to other styles?

We test cross-style transfer by performing SFT-CoT on a source style and RLVR on a distinct target style. Here, we conduct cross-style experiments where we first perform distillation SFT on one style and then apply GRPO on a different style. We repeat this process for all style combinations.

Observation. Table 3 presents the results of these experiments. Transfer is strongest between semantically related styles (sarcasm↔humor). For example, performing SFT on distilled sarcasm CoTs followed by RLVR on humor data yields a $\sim 10\%$ accuracy improvement compared to performing GRPO on the base model. Figure 3 visualizes absolute gains over GRPO-only; the diagonal (same-style SFT→GRPO) is highest, while off-diagonal improvements confirm cross-style knowledge sharing. Metaphor shows weaker transfer into other styles, consistent with its greater semantic distinctiveness (Skalicky and Crossley, 2018). These findings suggest that RLVR on top of an SFT model trained on a related style can achieve improved performance without requiring the first two steps of

Table 2: Pipeline ablation: accuracy (Acc) and F1 across four figurative styles. Best score in each subcolumn is **bold**. All runs use zero-shot prompting.

Experiment	Sarcasm		Humor		Offense		Metaphor	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Zero-shot CoT Prompting	63.88	0.41	49.82	0.68	46.72	0.44	53.89	0.63
SFT on Binary Labels	82.02	0.79	66.29	0.74	50.57	0.62	59.67	0.69
SFT on Distilled CoTs	83.91	0.77	70.73	0.78	52.15	0.59	62.19	0.73
GRPO-only	72.19	0.43	65.85	0.74	48.33	0.49	59.21	0.66
SFT → GRPO	**86.82**	**0.81**	**78.91**	**0.89**	**59.51**	**0.75**	**69.24**	**0.80**

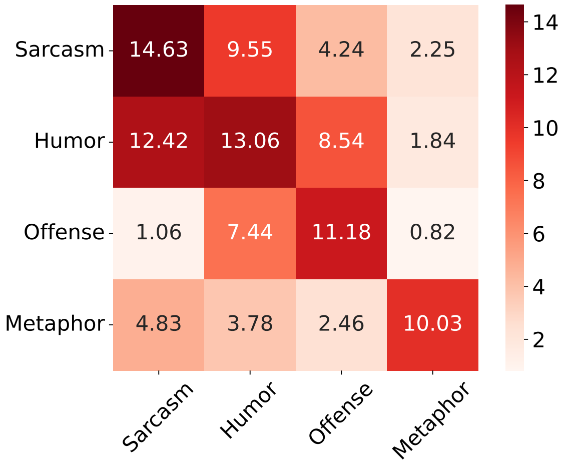


Figure 3: Cross-style gains over GRPO-only. Rows: SFT-CoT source style; columns: RLVR target style. Diagonal cells (same-style) are highest; sarcasm↔humor shows the strongest transfer

our proposed approach for the target style.

5.3 Can we build a single generalized model across styles?

We investigate whether our proposed approach can produce a generalized reasoning small VLM capable of understanding all figurative styles (in this case, four styles). We train a unified model by distilling CoTs from the union of all training data and then running SFT→GRPO on either (i) each target style or (ii) the combined dataset. We compare against much larger open- and closed-source systems under zero-shot prompting.

Observation. Joint SFT→GRPO on the combined dataset yields the best overall small-model performance (Table 4), surpassing much larger models on sarcasm, humor, and offense, while trailing Gemini 2.5 Flash only on metaphor. This validates that *a single lightweight reasoning VLM* can generalize across styles and remain competitive with far larger systems.

5.4 Effect of Data Scale vs. Cross-Style Diversity

To isolate the impact of cross-style diversity from training data scale, we conduct a controlled experiment with a fixed training budget of 5k samples. In the *style-specific* setting, separate SFT→GRPO models are trained using 5k samples from a single style. In the *combined* setting, a unified model is trained on the same total number of samples, uniformly sampled across all four styles (1.25k per style). Evaluation is performed on the original test sets.

As shown in Table 5, the combined model achieves clear improvements on sarcasm and humor, despite having fewer per-style training examples. This indicates that the gains observed in Section 5.3 are not solely attributable to increased training data size, but also to shared reasoning learned across related figurative styles. In contrast, offense and metaphor exhibit slight performance degradation under unified training, suggesting weaker cross-style transfer for these categories when the training budget is constrained. Overall, these results show that cross-style diversity does not benefit all categories equally, with the most consistent improvements observed for sarcasm and humor.

5.5 Disagreement Analysis

To better understand the practical advantage of our approach over general-purpose models, we conducted a systematic disagreement analysis between our best model (SFT → GRPO on combined dataset) and Gemini 2.5 Flash. We isolated instances in the test sets where the two models produced conflicting predictions and calculated which

Table 3: Cross-style transfer results: accuracy (Acc) and F1 for each target style. Best score per subcolumn is **bold**.

Experiment	**Sarcasm**		**Humor**		**Offense**		**Metaphor**	
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
GRPO-only	72.19	0.43	65.85	0.74	48.33	0.49	59.21	0.66
SFT Humor → GRPO	84.61	0.78	**78.91**	**0.89**	56.87	0.67	61.05	0.72
SFT Sarcasm → GRPO	**86.82**	**0.81**	75.40	0.82	52.57	0.58	61.46	0.72
SFT Metaphor → GRPO	77.02	0.63	69.63	0.74	50.79	0.53	**69.24**	**0.80**
SFT Offense → GRPO	73.25	0.57	73.29	0.80	**59.51**	**0.75**	60.03	0.64

Table 4: Generalized model vs. large VLMs: accuracy (Acc) and F1 per style. Best in each subcolumn is **bold**. All results use zero-shot prompting.

Model / Setup	**Sarcasm**		**Humor**		**Offense**		**Metaphor**	
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
Gemini 2.5 Flash	89.34	0.84	79.13	0.89	60.95	0.77	**73.19**	0.76
Qwen2.5-VL 32B Instruct	76.78	0.64	77.14	0.87	49.37	0.52	61.31	0.69
LLaMA 90B Vision Instruct	88.85	0.83	78.71	0.87	55.47	0.75	67.56	0.76
Gemma 3 27B	54.63	0.61	62.23	0.68	45.33	0.57	58.44	0.64
Phi-4 Multimodal	68.28	0.56	70.33	0.73	48.74	0.64	55.71	0.55
LLaVA 1.5 7B	59.09	0.48	61.54	0.61	43.22	0.52	53.68	0.56
SFT on Combined → GRPO (Style specific)	87.76	0.82	78.59	0.91	60.27	0.77	68.19	0.79
SFT on Combined → GRPO (Combined)	**90.23**	**0.85**	**80.71**	**0.93**	**61.49**	**0.79**	69.14	**0.80**

Table 5: Comparison of Style-specific vs. Combined models with a fixed training budget (5k samples).

Style	**Style specific**		**Combined**	
	Acc	**F1**	**Acc**	**F1**
Sarcasm	80.25	0.78	**83.91**	**0.81**
Humor	78.13	0.88	**79.95**	**0.89**
Offense	**56.36**	**0.73**	55.82	0.71
Metaphor	**66.44**	**0.77**	63.15	0.72

model matched the ground truth in these contested cases.

As shown in Table 6, the models disagreed on a significant portion of the test set. In the sarcasm style, the models disagreed on 529 out of 2,390 samples. Within this disagreement subset, our model correctly predicted 309 samples (58.4%), whereas Gemini was correct in only 220 cases. This indicates that our specialized reasoning framework aligns better with human labels in ambiguous sarcasm instances.

We observed similar trends for humor and offense. In the humor style, our model demonstrated its strongest relative performance, winning 70.0% of the disagreement cases (21 vs. 9). For offense, our model maintained a lead with a 55.7% win rate. However, consistent with the main accuracy results, Gemini 2.5 Flash retained an advantage in the metaphor style, correctly classifying 350 of the 561 contested samples. These results highlight that our lightweight model effectively corrects errors made by larger VLMs in affect-heavy styles like sarcasm and humor, though larger models may still hold an edge in abstract semantic associations required for metaphor.

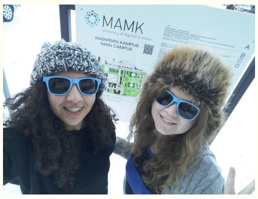
6 Qualitative Examples Analysis

In this section, we analyze a few examples where our best model (SFT → GRPO on the combined dataset) makes correct predictions, while Gemini 2.5 Flash fails to predict the final label correctly. We focus only on samples for sarcasm style from the MMSD2.0 test set.

Table 6: Comparison of correctly classified samples when our model and Gemini 2.5 Flash differ.

Style	**Disagreed**	**Ours Correct**	**Gemini Correct**	**Win Rate (%)**
Sarcasm	529	**309**	220	58.4%
Humor	30	**21**	9	70.0%
Offense	244	**136**	108	55.7%
Metaphor	561	211	**350**	37.6%

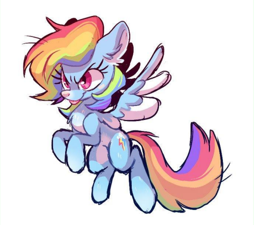
Example 1:



Text: mikkeli university of applied sciences . it 's too sunny . # mamk # finland # finnips # uas # suomi
Ground Truth: Sarcastic
Our model CoT: <think>Step 1:The image shows a person wearing sunglasses and a hat, standing in front .... Step 2: The caption reads: mikkeli university of applied sciences.... Step 3: There is a mismatch between the image and the caption...... Step 4: it can be inferred that the intent behind the post is sarcastic. The person ....it.</think><answer>sarcastic</answer>
Gemini CoT: Step 1: The image shows two young women posing for a .... .Step 2: The caption states: mikkeli ... Step 3: The image depicts a snowy, winter environment. The women are dressed in warm winter attire, ... the bright conditions and the use of sunglasses shown in the image. There is no clear mismatch. Step 4: Given the bright, snowy conditions and the fact that both individuals are wearing sunglasses ..... caption's claim; instead, they are congruent. Step 5: not sarcastic.
Our Prediction: Sarcastic
Gemini Prediction: Not sarcastic

Explanation for example 1: Our model recognizes sarcasm by noting the contrast between the cheerful photo and the caption complaining it's "too sunny," interpreting it as a lighthearted exaggeration rather than a real complaint. However, Gemini takes the post literally, reasoning that bright snow could indeed make it "too sunny," and therefore sees no contradiction. In other words, our model captures the playful irony, while Gemini overanalyzes environmental realism and misses the humorous intent.

Example 2:



Text: not sure why i drew rainbuw dersh but i did
Ground Truth: Not sarcastic
Our model CoT: <think>Step 1: The image depicts a cartoon-style ..... Step 2:The caption reads, not sure .... Step 3: . The caption's tone is casual and conversational, suggesting that the .... .Step 4: Based on the congruence between the image and caption, it can be inferred that the intent behind the image is genuine and not sarcastic. .... </think><answer>not sarcastic</answer>
Gemini CoT: Step 1: The image shows a vibrant, stylized illustration of a light blue pegasus .... Step 2: The caption says: not sure why ..... Step 3: The image clearly depicts Rainbow Dash, ... The mismatch lies in the phrase not sure why i drew... Drawing an elaborate piece of fan art typically implies some level of intention, desire, or motivation from the artist...... This is a common form of self-deprecating or ironic humor .... Step 4: The artist created a clear, recognizable, and detailed drawing of Rainbow Dash. The caption's statement not sure why i drew rainbuw dersh is likely not meant to be taken literally. Instead, it serves as a humorous, ironic, or self-aware comment. .... This playful downplaying of the cre-

ative act, juxtaposed with the evident effort and clear subject, indicates a sarcastic or ironic intent, rather than genuine confusion.
Step 5: Sarcastic
Our Prediction: Not sarcastic
Gemini Prediction: Sarcastic

Explanation for example 2: Our model correctly interprets this pair as not sarcastic, by focusing on the literal meaning of the caption and the cheerful cartoon image. It sees the creator as simply sharing artwork without any hidden irony, recognizing the direct congruence between image and text. Gemini, however, overanalyzes the situation. It interprets the phrase "not sure why I drew rainbuw dersh" as ironic or self-deprecating humor, reasoning that the effort put into the detailed drawing contradicts the claimed uncertainty. This leads Gemini to label the post as sarcastic, even though the creator is genuinely sharing the artwork in a casual, conversational tone.

Observation. Gemini tends to take things too literally or analyze them too much, while our model can pick up on social and figurative signs of sarcasm. In the first example, Gemini focuses on realistic details and logic, so it misses the playful exaggeration that our model correctly sees as sarcastic. In the second example, which isn't sarcastic, Gemini overthinks a casual phrase like "not sure why I drew," interpreting it as ironic or self-deprecating. Our model, however, correctly sees that it's just a simple statement and understands it literally. Therefore, Gemini overanalyzes and misses the subtle, often illogical nature of sarcasm, while our model successfully captures the intended meaning and context.

7 Conclusion

In this work, we introduced a three-stage framework that effectively induces reasoning capabilities to compact VLMs for the challenging task of figurative language understanding. Our results demonstrate that explicit reasoning not only enhances figurative comprehension but can also be successfully transferred across related styles. Most notably, our unified model, trained jointly on multiple styles, sets a new benchmark for efficiency, outperforming larger open- and closed-source VLMs while providing transparent reasoning traces that facilitate model inspection. Qualitative analysis further shows that our model captures subtle, nonliteral intent that larger models often miss due to overly literal interpretations or excessive contextual over-analysis. The success of our framework has implications beyond figurative language. The core methodology—distilling structured reasoning from a powerful teacher model and refining a smaller student model with verifiable, task-aligned rewards—offers a scalable and computationally efficient paradigm for other complex multimodal tasks. Nevertheless, it is important to recognize the inherent limitations of this approach. The generalization of the student model is fundamentally constrained by the quality and depth of the reasoning traces provided by the teacher. Benefits plateau because the student cannot acquire reasoning patterns that the teacher itself cannot express. Any subtle biases or overlooked nuances in the teacher are inevitably transferred, and potentially amplified, in the student. Future work should investigate how teacher model architectures, sizes, and pre-training domains influence the quality of distilled reasoning to optimize knowledge transfer. While we observed strong cross-style transfer between semantically related styles such as sarcasm and humor, this effect may weaken as the conceptual distance between tasks increases. Moreover, the effectiveness of the RLVR stage relies on the availability of clear, objective ground-truth labels. Exploring more sophisticated reward functions—such as nuanced metrics for reasoning quality or penalties for overconfident incorrect predictions—could further enhance the framework, particularly in sensitive or high-stakes applications.

8 Limitations

Although our proposed framework achieves strong performance and outperforms larger models in figurative language understanding, our study has some limitations. First, the experiments should be extended to different models in both the CoT distillation and subsequent training steps to examine the behavior of larger models under various settings. Second, our study lacks out-of-distribution experiments to evaluate whether the CoT reasoning capability learned by VLMs can generalize across different datasets. Third, incorporating additional figurative styles and new datasets would be valuable. Fourth, a comparison with both open-source and closed-source models in a few-shot setting would be helpful. Finally, conducting an ablation study to isolate the effects of CoT filtering and reward formatting is essential.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaite, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. 2025. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific reports*, 15(1):11477.

Elisabeth Camp. 2020. Sarcasm, irony, and satire. In Raymond W. Gibbs, editor, *The Oxford Handbook of Figurative Language*, pages 321–339. Oxford University Press.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding through textual explanations. *arXiv preprint arXiv:2205.12404*.

Debarati Das, David Ma, and Dongyeop Kang. 2023. Balancing effect of training dataset distribution of multiple styles for multi-style text transfer. *Preprint*, arXiv:2305.15582.

Karin de Langis and Dongyeop Kang. 2023. A comparative study on textual saliency of styles from eye tracking, annotations, and language models. *Preprint*, arXiv:2212.09873.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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*.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does bert learn as humans perceive? understanding linguistic styles through lexica. *Preprint*, arXiv:2109.02738.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Hyewon Jang and Diego Frassinelli. 2025. The difficult case of intended and perceived sarcasm: a challenge for humans and large language models. In *16th International Conference on Computational Semantics*, page 279.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative language processing: A linguistically informed feature analysis of the behavior of language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832, Toronto, Canada. Association for Computational Linguistics.

Dongyeop Kang and Eduard Hovy. 2019. Style is not a single variable: Case studies for cross-style language understanding. *arXiv preprint arXiv:1911.03663*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Shreyas Kulkarni, Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2024. A report on the figlang 2024 shared task on multimodal figurative language. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 115–119.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. *arXiv preprint arXiv:2306.00121*.

Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. 2025. Star-r1: Spatial transformation reasoning by reinforcing multimodal llms. *arXiv preprint arXiv:2505.15804*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Yeongtak Oh, Jisoo Mok, Dohyun Chung, Juhyeon Shin, Sangha Park, Johan Barthelemy, and Sungroh Yoon. 2025. Repic: Reinforced post-training for personalizing multi-modal language models. *arXiv preprint arXiv:2506.18369*.

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*, pages 10834–10845, Toronto, Canada. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, Yudi Zhang, Artemis Panagopoulou, and Smaranda Muresan. 2023. V-FLUTE: Visual Figurative Language Understanding with Textual Explanations. *arXiv preprint arXiv:2303.15445*. Accepted at ACL 2023.

Chhavi Sharma, William Paka, Scott, Deepesh Bhageria, Amitava Das, Soujanya Poria, Tanmoy Chakraborty, and Björn Gambäck. 2020. Task Report: Memotion Analysis 1.0 @SemEval 2020: The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain. Association for Computational Linguistics.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Stephen Skalicky and Scott Crossley. 2018. Linguistic features of sarcasm and metaphor production quality. In *Proceedings of the Workshop on Figurative Language Processing*, pages 7–16, New Orleans, Louisiana. Association for Computational Linguistics.

Lukas Stappen and 1 others. 2024. The muse 2024 multimodal sentiment analysis challenge: Social perception and humor recognition. *arXiv preprint arXiv:2406.07753*.

Yuan Tian, Nan Xu, and Wenji Mao. 2024. A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7738–7755, Mexico City, Mexico. Association for Computational Linguistics.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Peiyao Wang and Haibin Ling. 2025. Svqa-r1: Reinforcing spatial reasoning in mllms via view-consistent reward optimization. *arXiv preprint arXiv:2506.01371*.

Xinyu Wang, Yue Zhang, and Liqiang Jing. 2025. Can large vision-language models understand multimodal sarcasm? *arXiv preprint arXiv:2508.03654*.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, and 1 others. 2025. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*.

Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, and Kaiyang Zhou. 2025. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*.

Zheng Yang and 1 others. 2022. Multimodal sarcasm detection with image-text fusion. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ben Yao, Yazhou Zhang, Qiuchi Li, and Jing Qin. 2025. Is sarcasm detection a step-by-step reasoning process in large language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25651–25659.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. MultiMET: A multi-modal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225, Online. Association for Computational Linguistics.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Jingxuan Zhou, Yuehao Wu, Yibo Zhang, Yeyubei Zhang, Yunchong Liu, Bolin Huang, and Chunhong Yuan. 2025. Semirnet: A semantic irony recognition network for multimodal sarcasm detection. In *2025 10th International Conference on Information and Network Technologies (ICINT)*, pages 158–162. IEEE.

A Appendix

A.1 CoT Prompt

We use a decomposed CoT prompting strategy, inspired by Khot et al. (2022) and Zhou et al. (2022) to break the problem into four steps. At each step, we instruct the model to focus on a specific aspect of the input data. Prompt 1 detects sarcasm, Prompt 2 humor, Prompt 3 offense, and Prompt 4 metaphor. After generating the CoT, we check, using regex, whether it contains all five steps and whether the final prediction matches the ground truth, discarding any samples that do not meet these criteria. Table 7 presents examples of generated CoTs.

> You are an expert at detecting sarcasm in images and text. Analyze the provided image and caption to determine if the pair is sarcastic or not sarcastic. Provide your reasoning in the following format:
>
> **Step 1:** What the image shows: [Detailed description of the image content]
> **Step 2:** What the caption says: [Quote or paraphrase the caption]
> **Step 3:** Detecting mismatch: [Explain if there is a mismatch or congruence between the image and caption, and why]
> **Step 4:** Inference of intent: [Conclude whether the intent is sarcastic or not based on the mismatch/congruence]
> **Step 5:** [Provide your final answer in the form of 'sarcastic' or 'not sarcastic']

Prompt 1: CoT prompt to detect sarcasm.

A.2 GRPO Prompt

We use the prompt structure described by Guo et al. (2025) for GRPO training as illustrated in Prompt 5 to detect sarcasm. We apply the same GRPO structure to the other styles, which are omitted here since their prompts were introduced in the previous section.

> You are an expert at detecting humor in images and text. When given an image and text, analyze whether the content is humorous or not. Provide your reasoning process in the following format:
>
> **Step 1:** What the image shows: [Detailed description of the image content]
> **Step 2:** What the caption says: [Quote or paraphrase the caption]
> **Step 3:** Humor cues: [Explain if there are elements such as exaggeration, wordplay, absurdity, or incongruity between the image and caption that make the content humorous]
> **Step 4:** Inference of intent: [Conclude whether the intent is humorous or not based on the cues]
> **Step 5:** [Provide your final answer in the form of "humorous" or "not humorous"]

Prompt 2: CoT prompt to detect humor.

> You are an expert at detecting offensive content in images and text. When given an image and text, analyze whether the content is offensive or not. Provide your reasoning process in the following format:
>
> **Step 1:** What the image shows: [Detailed description of the image content]
> **Step 2:** What the caption says: [Quote or paraphrase the caption]
> **Step 3:** Offense cues: [Explain if there are elements such as hate speech, slurs, derogatory language, demeaning stereotypes, harassment, or explicit insults that make the content offensive]
> **Step 4:** Context and intent: [Discuss whether the content was likely meant to harm, insult, or demean someone, or if it might be interpreted as offensive even without harmful intent]
> **Step 5:** [Provide your final answer in the form of "offensive" or "not offensive"]

Prompt 3: CoT prompt to detect offense.

You are an expert at detecting metaphors in images and text. When given an image and text, analyze whether the content uses metaphorical language or not. Provide your reasoning process in the following format:

Step 1: What the image shows: [Detailed description of the image content]
Step 2: What the caption says: [Quote or paraphrase the caption]
Step 3: Metaphor cues: [Explain if there are figurative expressions, symbolic comparisons, or non-literal meanings that connect the caption and the image]
Step 4: Interpretation: [Discuss what abstract idea, concept, or meaning the metaphor might be conveying]
Step 5: [Provide your final answer in the form of "metaphorical" or "not metaphorical"]

Prompt 4: CoT prompt to detect metaphor.

You are an expert at detecting sarcasm in images and text. When given an image and text, analyze whether the content is sarcastic or not. Provide your reasoning process in the following format: Step 1: What the image shows: [Detailed description of the image content]. Step 2: What the caption says: [Quote or paraphrase the caption]. Step 3: Detecting mismatch: [Explain if there is a mismatch or congruence between the image and caption, and why]. Step 4: Inference of intent: [Conclude whether the intent is sarcastic or not based on the mismatch/congruence]. Your reasoning process and answer should be enclosed within <think> </think> and <answer> </answer> tags, respectively. Answer with either sarcastic or not sarcastic in the answer tags, i.e.,<think> reasoning process (Step 1 to Step 4) </think><answer> sarcastic/not sarcastic </answer>

Prompt 5: GRPO prompt to detect sarcasm.

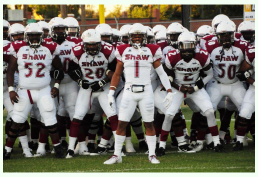
Table 7: Generated reasoning chains for four figurative language styles.

Style	Generated Chain-of-Thought
Sarcasm	Step 1: The image shows a map of Malaysia with a green area in the center, which appears to be a lake or a body of water. The map also shows the surrounding landmasses and the coastline. Step 2: The caption reads, “this is how our currency is shrinking #ringgitladesh.” It refers to the Malaysian currency (Ringgit) and implies that it is losing value. Step 3: There is a mismatch between the image and the caption. The image shows a map unrelated to currency, while the caption refers to economic value. Step 4: The mismatch suggests the post is sarcastic, using irony rather than serious commentary to highlight the shrinking value of the currency. Step 5: Sarcastic.
Humor	Step 1: The image depicts a man walking down the street with a woman on his arm while looking back at another woman. The man wears a blue plaid shirt; the woman on his arm is in a light blue top; the woman he looks at wears a red dress. Step 2: The caption reads “People of the future” above the man’s head, “Traditional human language” above the woman on his arm, and “Communicating entirely through the Distracted Boyfriend meme” above the woman he is looking at. Step 3: The humor arises from applying the well-known Distracted Boyfriend meme to a futuristic scenario where people communicate only through memes. The absurd combination creates incongruity and amusement. Step 4: The intent is to be humorous by exaggerating modern meme culture. Step 5: Humorous.
Offense	Step 1: The image shows two side-by-side photos of a man with a shaved head, wearing a tuxedo. The left photo looks natural; the right has exaggerated, cartoonish features. Step 2: “Stop doing this to your pics.” Step 3: The image has no hate speech, slurs, or derogatory content. The tone is slightly sarcastic, criticizing excessive photo editing. Step 4: The context is lighthearted social media humor, not malicious or harmful. Step 5: Not offensive.
Metaphor	Step 1: Two children stand behind a table with a red chair built from objects such as a bucket, a drum, and a book. The background has a bookshelf filled with books. Step 2:“HONDA – The Power of Dreams.” Step 3: The creative chair symbolizes imagination and innovation. The children’s presence implies creativity is accessible to everyone. Step 4: The metaphor expresses that creativity and imagination can lead to achieving dreams, encouraging viewers to think outside the box. Step 5: Metaphorical.

B Qualitative Examples

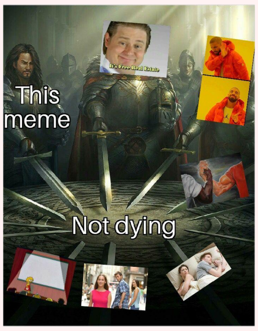
In this section, we showcase outputs from our optimal model (SFT followed by GRPO on combined dataset). We include one representative example per style to highlight the model's reasoning as revealed by its generated CoT. The sarcasm, humor, and metaphor examples reveal the model's ability to interpret image-text relationships. In sarcasm, the model identified consistency between the football team's image and the direct caption and accurately classified it as "Not Sarcastic." In humor, it recognized that combining incongruous images generated a comic effect. In metaphor, the model demonstrated abstract reasoning by detecting the symbolic link between the skier and the concept of sound. However, the offense example is notable because the model misclassified it, highlighting that defining offensive content is nuanced: While the model understood the joke about a character's nose, it erred on the side of caution and labeled it "Offensive" due to its reference to a physical trait. This example illustrates the challenge of aligning models with sensitive human judgments.

Sarcasm Example



Text: blessed to have received my first offer from shaw university
Ground Truth: Not Sarcastic
CoT: <think> Step 1: The image shows a group of football players in white uniforms with red accents, standing together on a field. They appear to be part of a team, possibly a college football team, given their uniforms and the setting. Step 2: The caption reads, B̈lessed to have received my first offer from Shaw University.T̈his statement expresses gratitude for receiving an offer from Shaw University, which is likely a college or university where the player is interested in attending. Step 3: There is no apparent mismatch between the image and the caption. The image depicts a group of football players, and the caption directly relates to them, indicating that the player is expressing gratitude for receiving an offer from Shaw University. Step 4: Given that the caption is a direct statement about the player's feelings towards Shaw University, and there is no indication of sarcasm in the image or the caption, it can be inferred that the intent is not sarcastic.</think><answer>not sarcastic</answer>"
Prediction: Not Sarcastic

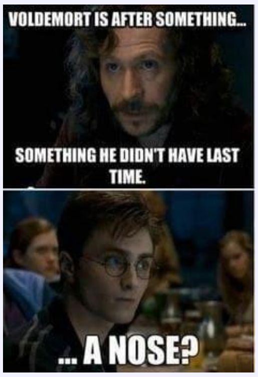
Humor Example



Text: It's Free Real Estate This meme Not dying
Ground Truth: Humorous
CoT: <think> Step 1: The image is a collage of various memes, including a knight in armor holding a sword, a man with a caption, It's Free Real Estate, and several other images that appear to be unrelated to each other. The background of the image is dark and foggy, with a faint light shining down from above. Step 2: The caption reads ,It's Free Real Estate This meme Not dying. Step 3: The humor in this image comes from the juxtaposition of the serious-looking knight with the absurd and unrelated captions and images. The use of the phrase, It's Free Real Estate, as a caption for a knight holding a sword is unexpected and nonsensical, creating a sense of incongruity that is humorous. Additionally, the inclusion of other memes in the image, such as the man with the caption, It's Free Real Estate, adds to the absurdity and playfulness of the image. Step 4: Based on the humor cues present in the image, it appears that the intent of the creator is to be humorous. The use of absurd and unrelated captions and images creates a sense of playfulness and lightheartedness, suggesting that the creator is trying to make the viewer laugh.</think><answer> humor-

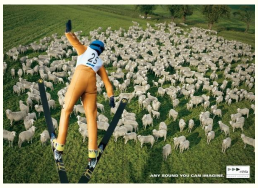
ous </answer>",
Prediction: Humorous

Offense Example


VOLDEMORT IS AFTER SOMETHING...
SOMETHING HE DIDN'T HAVE LAST TIME.
... A NOSE?

Text: VOLDEMORT IS AFTER SOMETHING. SOMETHING HE DIDN'T HAVE LAST TIME. ...A NOSE?
Ground Truth: Not offensive
CoT: <think>Step 1: The image is a meme featuring two characters from the Harry Potter series, Voldemort and Harry Potter. Voldemort is depicted with a serious expression, while Harry Potter looks concerned. Step 2: The caption reads, VOLDEMORT IS AFTER SOMETHING... SOMETHING HE DIDN'T HAVE LAST TIME. ...A NOSE? Step 3: The caption contains a joke about Voldemort not having a nose, which could be perceived as making fun of individuals with physical disabilities or deformities. Step 4: The meme is likely intended to be humorous and lighthearted, rather than meant to offend or demean individuals with physical disabilities. However, the joke may still be perceived as insensitive or offensive by some viewers. </think><answer> offensive </answer>",
Prediction: Offensive

Metaphor Example



Text: any sound you can imagine
Ground Truth: Metaphorical
CoT: <think>Step 1: The image depicts a person dressed in a yellow ski suit, complete with a blue helmet and black skis, standing in a field of sheep. The individual is positioned centrally in the frame, with their arms outstretched and their legs spread apart, as if they are about to take off or have just landed from a jump. The surrounding environment is characterized by a lush green field, dotted with numerous white sheep grazing peacefully. Step 2: The caption reads, any sound you can imagine. This phrase is concise and enigmatic, leaving room for interpretation. Step 3: Upon closer examination, it becomes apparent that the image and caption are connected through metaphorical language. The juxtaposition of the skier amidst the sheep, coupled with the caption's reference to sound, suggests a symbolic comparison between the skier's actions and the auditory experience. The use of the word imagine further reinforces this interpretation, implying that the viewer should consider the skier's movements as a form of sound or music. Step 4: The metaphorical connection between the image and caption conveys the idea that the skier's actions can be perceived as a form of creative expression, akin to music or sound. The caption's emphasis on imagination encourages the viewer to consider the skier's movements as a means of generating sound or music, thereby blurring the lines between physical activity and auditory experience. </think><answer> metaphorical </answer>"
Prediction: Metaphorical