Two Streams, One Sarcasm: Orthogonal Expert Tuning for Holistic Multimodal Sarcasm Understanding

Diandian Guo[1,2], Cong Cao[1,2,*], Fangfang Yuan[1,2], Pin Xu[1,2], Cheng Hu[1,2], Zhicheng Zhang[1,2], Yu Liu[1,2], Yanbing Liu[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{guodiandian, caocong}@iie.ac.cn

Abstract

Multimodal Sarcasm Understanding (MSU) comprises multiple subtasks, demanding both incongruity perception and intent reasoning. However, this progress is impeded by two bottlenecks. First, the lack of a unified benchmark for holistic satirical cognition hinders comprehensive evaluation of MSU. Second, jointly modeling these heterogeneous subtasks often leads to feature entanglement. Specifically, while subtasks share a dependence on incongruity, they diverge in granular focus, causing specific execution patterns to erode the fundamental perception capability. To address these challenges, we make two contributions. First, we introduce DocMSU-PLUS, a comprehensive benchmark covering five cognitive dimensions of MSU. All tasks are reformulated into multiple-choice questions (MCQs), enabling a unified accuracy-based evaluation. Second, we propose the **D**ual **O**rthogonal **S**tream **E**xperts (DOSE) framework. DOSE structurally decouples experts into orthogonal shared perception and private execution streams to physically block gradient interference between tasks. Experiments demonstrate that DOSE achieves superior performance on DocMSU-PLUS, effectively balancing general perception with task-specific adaptation.[1]

1 Introduction

Sarcasm, as a complex rhetorical device, extensively pervades multimodal content on social media. Since sarcasm typically implies a sentiment diametrically opposite to its literal meaning, accurate decoding is pivotal for robust sentiment analysis (Farias and Rosso, 2017; Khare et al., 2023), public opinion mining (Cai et al., 2019), and social AI systems (Balamurali et al., 2024). Theoretically grounded in the Incongruity-Resolution Theory (Yus, 2017), comprehending sarcasm operates as a hierarchical cognitive process that initiates with the perception of semantic dissonance between visual and textual modalities, and culminates in the reasoning of the underlying intent to rationalize such conflicts. Consequently, Multimodal Sarcasm Understanding (MSU) requires models to go beyond binary detection. Instead, it requires models to parse inter-modal dominance, identify sarcasm targets, and comprehend sarcasm logic.

Recently, Multimodal Large Language Models (MLLMs) have emerged as promising candidates to address complex cognitive demands. This is due to their exceptional capabilities in instruction following and reasoning. However, verifying these capabilities is challenging due to the fragmented nature of current benchmarks. Existing datasets are predominantly tailored for isolated single-task scenarios, most notably binary detection (Qin et al., 2023; Qiao et al., 2023; Guo et al., 2025). Even when studies attempt to incorporate diverse tasks such as explanation generation, they are forced to rely on heterogeneous evaluation metrics (Kumar et al., 2022; Zhuang et al., 2025). Specifically, perception-oriented tasks like detection utilize metrics like accuracy to measure deterministic correctness. In contrast, reasoning-oriented tasks like explanation rely on semantic similarity scores. This fundamental metric mismatch renders it inherently challenging to align the evaluation of reasoning with perception. Consequently, a unified standard remains absent to quantify the model's MSU capability across the complete cognitive chain, as semantic similarity does not necessarily equate to the logical validity required for this task.

However, achieving this through multi-task learning faces a critical feature entanglement challenge. Unlike general multi-task scenarios, sarcasm understanding involves an inherent conflict between low-level perception (capturing subtle cross-modal incongruity) and high-level execution

*Corresponding author.
[1]Warning: This paper contains content that may be disturbing to some readers.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7054–7072

(formulating complex responses for diverse downstream tasks). Mastering the complex execution patterns required for these diverse tasks often precipitates negative transfer. Gradient updates driven by execution objectives tend to erode the underlying, task-agnostic representations of sarcasm perception. While mixture-of-experts (MoE) architectures have shown promise in mitigating task interference, they rely on parameter-level isolation without explicit geometric constraints. Consequently, they fail to achieve feature decoupling, thus failing to prevent this catastrophic interference.

To bridge this evaluation gap, we introduce DocMSU-PLUS. Distinct from the original DocMSU (Du et al., 2024), which primarily facilitate perception-level tasks on a large-scale but highly imbalanced corpus, our benchmark shifts the focus towards cognitive depth and data quality. We meticulously distill a balanced corpus of over 14,000 high-quality samples from the original raw data and fundamentally expand the cognitive dimensions to include reasoning-centric tasks, specifically sarcasm description and mechanism analysis. Crucially, we establish the first standardized MCQ paradigm that unifies all five subtasks. This formulation transforms diverse output formats, including text generation and bounding box localization, into a consistent selection task. Capitalizing on this unified formulation, DocMSU-PLUS enables us to utilize accuracy as a single, rigorous metric. This is the first benchmark to assess the comprehensive MSU capability under unified indicators.

To address the feature entanglement challenge, we propose a **D**ual **O**rthogonal **S**tream **E**xperts (DOSE) framework. DOSE explicitly resolves the inherent conflict between perception and execution by implementing a structural disentanglement strategy. Specifically, we construct each expert into a dual-stream architecture, comprising a shared perception stream for universal mechanisms and a private execution stream for task-unique features. During the forward process, a task-adaptive hybrid gating mechanism dynamically modulates fusion weights. This effectively balances the model's reliance on global sarcasm logic versus local details. Furthermore, we impose orthogonal regularization during optimization to mathematically enforce disjoint feature subspaces. This design allows the model to master diverse execution needs while preserving its fundamental perception capabilities. The main contributions of this work are summarized as follows:

- We introduce DocMSU-PLUS, which unifies MSU subtasks into a standardized MCQ paradigm. This formulation bridges existing evaluation gaps, enabling the assessment of holistic sarcasm understanding with a single metric for the first time.
- We propose DOSE to resolve feature entanglement through structural decoupling. By imposing orthogonal regularization, it effectively disentangles universal sarcasm perception from task-specific representations.
- Experiments demonstrate that DOSE achieves state-of-the-art performance on DocMSU-PLUS. Further analysis confirms that DOSE effectively mitigates negative transfer between tasks, ensuring robust performance across varying data regimes. The code and data are available[2].

2 Related Work

2.1 Multimodal Sarcasm Understanding

Research on MSU has evolved from binary classification to fine-grained reasoning. Early works focused on sarcasm detection (Cai et al., 2019), aiming to capture the emotional or semantic incongruity between images and text. Prior studies adopted graph neural networks (Liang et al., 2021; Wei et al., 2024), and cross-modal interaction mechanisms (Qiao et al., 2023; Jia et al., 2024; Guo et al., 2025) to model this incongruity. Recent efforts have shifted toward more challenging tasks, including target identification (Wang et al., 2022; Chen et al., 2024; Lv et al., 2025) and explanation generation (Desai et al., 2022; Jing et al., 2023; Singh et al., 2024). However, existing datasets are typically constructed for single tasks, neglecting the intrinsic cognitive dependencies between different subtasks. Furthermore, various sarcasm understanding tasks lack unified evaluation metrics. To address these issues, we construct the first unified MCQ benchmark covering five sarcasm understanding subtasks.

2.2 Fine-Tuning MLLMs

Although pre-trained MLLMs demonstrate remarkable zero-shot capabilities, fine-tuning is typically required complex downstream tasks. Existing solutions can be primarily categorized into two types: reparameterization and partial fine-tuning. Reparameterization methods introduce low-rank matrices to approximate weight updates and reduce computation (Hu et al., 2022; Liu et al., 2024). Partial

[2] https://github.com/Remwlp/DOSE

Dataset	Modality	Volume	Task Form	Task Scope					Annotation
				Det.	Mec.	Des.	Tar.	Int.	
MMSD (Cai et al., 2019)	I + T	24,635	Classification	✓	-	-	-	-	Manual
MMSD 2.0 (Qin et al., 2023)	I + T	24,635	Classification	✓	-	-	-	-	Manual
MUStARD (Castro et al., 2019)	V + A + T	690	Classification	✓	-	-	-	-	Manual
MSTI (Wang et al., 2022)	I + T	5,015	Sequence Labeling	-	-	-	✓	-	Manual
MSTI-PLUS (Lv et al., 2025)	I + T	4,288	Sequence Labeling	-	-	-	✓	-	Manual
MORE (Desai et al., 2022)	I + T	3,510	Generation	-	-	✓	-	-	Manual
DocMSU (Du et al., 2024)	I + T	102,588	Classification, Sequence Labeling	✓	-	-	✓	-	Semi-Manual
DocMSU-PLUS	I + T	14,128	Multiple-Choice Questions	✓	✓	✓	✓	✓	Semi-Manual

Table 1: Comparison of multimodal sarcasm datasets.

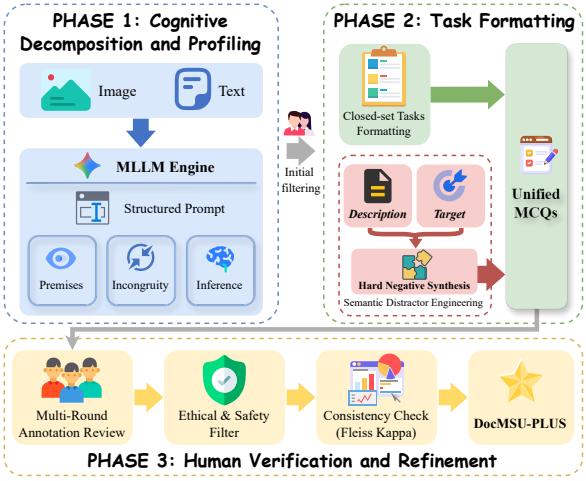


Figure 1: DocMSU-PLUS dataset annotation pipeline.

fine-tuning focuses on optimizing specific components or specific layers (Li et al., 2022; Zhu et al., 2024; Lu et al., 2024) to maintain architectural agnosticism and enhance transferability. Recently, MoE have been introduced into Parameter-Efficient Fine-Tuning (PEFT) for multi-task settings (Luo et al., 2024; Wu et al., 2024). Despite their success, these methods neglect the geometrical orthogonality of the feature space and gradient conflicts (Yu et al., 2020) between tasks. In contrast, our proposed DOSE achieves explicit feature stream decoupling via orthogonal projection.

3 The DocMSU-PLUS Dataset

To bridge the evaluation gap in existing multimodal sarcasm understanding benchmarks, we present DocMSU-PLUS, a refined and expanded version of DocMSU (Du et al., 2024). This benchmark serves as a comprehensive evaluation suite covering **Sarcasm Detection**, **Mechanism Analysis**, **Sarcasm Description**, **Target Identification**, and **Intent Reasoning**. Crucially, we adopt a unified MCQ paradigm, enabling standardized cross-task evaluation. A comparison with existing multimodal sarcasm datasets is presented in Table 1.

Task	Consistency	Fleiss Kappa
Sarcasm Detection	96.3%	0.8507
Mechanism Analysis	91.9%	0.8085
Sarcasm Description	93.6%	0.8097
Target Identification	92.6%	0.7819
Intent Reasoning	94.9%	0.8774
Overall	93.9%	0.8256

Table 2: Annotation consistency analysis.

3.1 Cognitive-Driven Annotation Pipeline

To ensure accuracy and cognitive depth, we move beyond simple crowdsourcing and design a standardized pipeline consisting of cognitive decomposition, profiling, task formatting, and expert verification. The pipeline is shown in Figure 1.

Phase 1: Cognitive Decomposition and Profiling. Understanding sarcasm requires a full grasp of the *Premises* → *Incongruity* → *Inference* reasoning chain. Leveraging the multimodal capabilities of Gemini-2.5-pro (Comanici et al., 2025), we perform cognitive decomposition on the raw data. Through structured instructions, the model generates a comprehensive sarcasm profile for each sample: **(1) Premises:** Extract objective visual scenes and literal textual statements. **(2) Incongruity:** Locate the semantic contrast between modalities to determine the specific sarcasm mechanism. **(3) Inference:** Derive the attack target and ultimate intent based on the logic of the conflict.

Phase 2: Task Formatting. To ensure data quality, we perform a preliminary manual screening to discard obviously unreasonable annotations. To unify evaluation metrics across different subtasks, we convert all five subtasks into a MCQ format. This unified paradigm allows us to use accuracy as the consistent metric for cross-task comparison. Prompt templates are provided in Appendix F. Based on the generated profiles, tasks are categorized into two processing streams: **(1)**

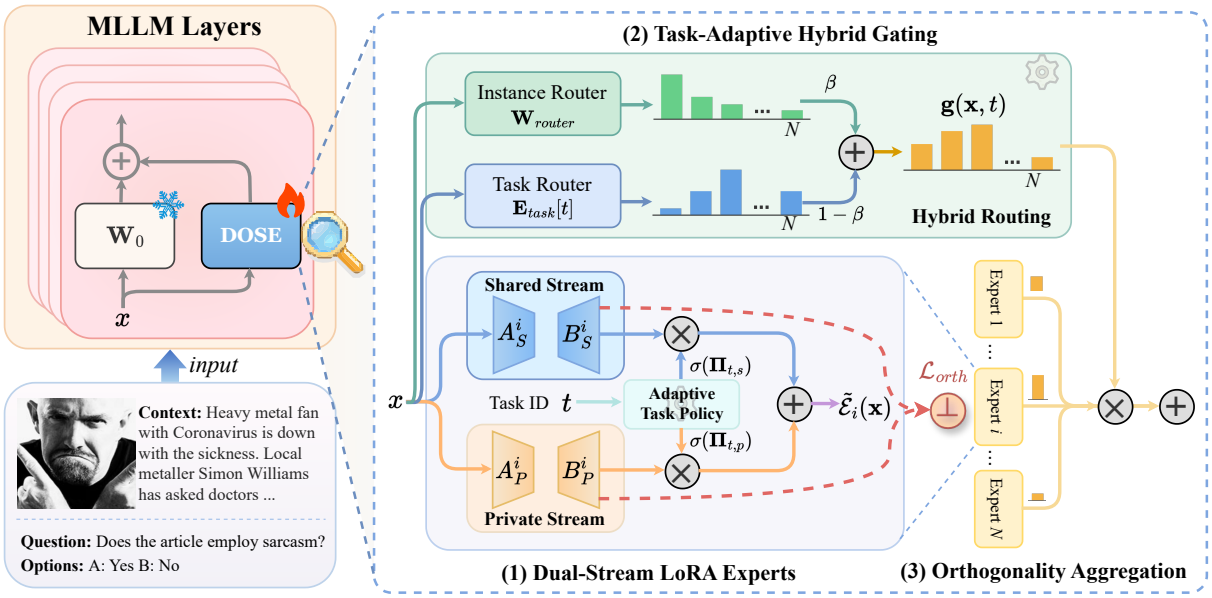


Figure 2: The overview of DOSE architecture. It comprises three core components: **(1) Dual-Stream LoRA Experts** for structural feature disentanglement; **(2) Task-Adaptive Hybrid Gating** for dynamic expert selection and weight modulation; and **(3) Orthogonality Aggregation** for enforcing geometric constraints.

Closed-set Tasks: For *Sarcasm Detection*, the binary label is derived from the profiling results in Phase 1. Samples where a valid incongruity and mechanism are successfully identified are labeled as sarcastic. For *Mechanism Analysis* and *Intent Reasoning*, we employ predefined fixed options. **(2) Open-ended Tasks:** For *Sarcasm Description* and *Target Identification*, where answers involve natural language descriptions or specific text spans or bounding boxes, we employ semantic distractor engineering (detailed in Section 3.2) to construct distinguishing options.

Phase 3: Human Verification and Refinement. To eliminate potential hallucinations and logical loopholes in model generationand ensure gold-standard quality, we conduct a rigorous human verification process. Five annotators perform a secondary review of each sample, filtering out ambiguous instances and correcting approximately 12% of label noise from the original source. Since sarcasm often involves sensitive content, we manually exclude samples containing extreme hate speech or high political sensitivity to maintain ethical compliance for academic research. To validate annotation consistency, we randomly select 1,000 samples for overlap annotation and calculate Fleiss Kappa coefficient, as shown in Table 2. For the most subjective task, sarcasm description, the Kappa score reaches 0.81, which indicates that our annotation quality meets high standards.

3.2 Semantic Distractor Engineering

The core challenge in converting open-ended understanding tasks into MCQs lies in constructing distinguishing distractors. Therefore, we introduce hard negatives to design diagnostic options that probe specific cognitive deficiencies.

For *Sarcasm Description*, we generate specific error types to diagnose cognitive pitfalls: **(1) Literal Interpretation Trap.** Options that reference factual terms from the text but strip away the sarcastic context. **(2) Over-interpretation Lure.** Options that introduce plausible but irrelevant grand narratives or social contexts.

For *Target Identification*, the correct answers typically consist of the key text spans or image regions (represented as normalized coordinates $[x_1, y_1, x_2, y_2]$). We construct cross-modal distractors with boundary confusion: **(1) Truncation & Redundancy.** Options containing core vocabulary but with incorrect boundary definitions, diagnosing the model's grasp of semantic unit integrity. **(2) Non-core Entity Confusion.** Selecting entities present in the text but irrelevant to sarcasm. **(3) Spatial Shift.** Selecting image regions logically unrelated to the sarcasm intent.

4 Methodology

This section details our proposed Dual Orthogonal Stream Experts framework. As illustrated in Figure 2, DOSE aims to structurally disentangle *invariant*

perception from *task-specific execution* through three core components: (1) Dual-Stream LoRA Experts, (2) Task-Adaptive Hybrid Gating, and (3) Orthogonality Aggregation.

4.1 Preliminaries: LoRA-based MoE

Given an input hidden state $\mathbf{x} \in \mathbb{R}^d$, the standard Transformer Feed-Forward Network (FFN) layer typically consists of fully connected layers. To achieve PEFT, LoRA decomposes the weight update ΔW into the product of two low-rank matrices $\mathbf{BA}$. In the MoE architecture, the FFN layer is replaced by a set of expert networks $\{\mathcal{E}_i\}_{i=1}^N$. The output of a LoRA-based MoE can be expressed as:

$$\mathbf{y} = \mathbf{W}_0\mathbf{x} + \sum_{i=1}^{N} G(\mathbf{x})_i \cdot \mathcal{E}_i(\mathbf{x}), \tag{1}$$

where $\mathbf{W}_0$ is the frozen pre-trained weight, $G(\cdot)$ represents the router, and $\mathcal{E}_i(\mathbf{x}) = \mathbf{B}_i\mathbf{A}_i\mathbf{x}$ denotes the i-th LoRA expert, with $\mathbf{B}_i \in \mathbb{R}^{r\times d}, \mathbf{A}_i \in \mathbb{R}^{d\times r}$, and rank $r \ll d$.

4.2 Dual-Stream LoRA Experts

Traditional LoRA-MoE approaches typically map the input hidden state $\mathbf{x} \in \mathbb{R}^{d_{in}}$ into a single low-rank stream. To achieve feature disentanglement, we design each expert $\mathcal{E}_i$ (where $i \in \{1, \ldots, N\}$) as a dual projection expert, comprising two parallel processing streams:

$$\mathcal{E}_i(\mathbf{x}) = \underbrace{\mathbf{B}^i_{\mathcal{S}}\mathbf{A}^i_{\mathcal{S}}\mathbf{x}}_{\text{Shared Stream}} + \underbrace{\mathbf{B}^i_{\mathcal{P}}\mathbf{A}^i_{\mathcal{P}}\mathbf{x}}_{\text{Private Stream}}, \tag{2}$$

where $\mathbf{B}^i_{\mathcal{S}} \in \mathbb{R}^{r_s\times d}$ and $\mathbf{A}^i_{\mathcal{S}} \in \mathbb{R}^{d\times r_s}$ define the shared stream with rank r_s. $\mathbf{B}^i_{\mathcal{P}} \in \mathbb{R}^{r_p\times d}$ and $\mathbf{A}^i_{\mathcal{P}} \in \mathbb{R}^{d\times r_p}$ define the private stream with rank r_p. Specifically, the shared stream is designed to capture universal sarcasm cues across tasks, such as cross-modal incongruity, which are fundamental basis for all subtasks. In parallel, the private stream is responsible for adapting to task-specific reasoning logic. This dual-stream architecture physically isolates the feature channels, allowing the model to flexibly adjust task-specific learning while preserving general sarcasm perception.

4.3 Task-Adaptive Hybrid Gating

To precisely control expert activation and handle task heterogeneity, we propose a task-adaptive hybrid gating mechanism. This mechanism comprises two core components: hybrid routing and task policy adaptation.

Hybrid Routing. Unlike traditional routers that rely solely on input content, DOSE incorporates task priors to guide expert selection. The final router $\mathbf{g} \in \mathbb{R}^N$ is a weighted fusion of instance-level routing and task-level routing:

$$\begin{aligned}\mathbf{g}(\mathbf{x}, t) =& \beta \cdot \text{Softmax}(\mathbf{W}_{router}\mathbf{x}) \\ &+ (1-\beta) \cdot \text{Softmax}(\mathbf{E}_{task}[t]),\end{aligned} \tag{3}$$

where $\mathbf{W}_{router} \in \mathbb{R}^{N\times d}$ is the learnable instance routing projection, $\mathbf{E}_{task} \in \mathbb{R}^{T\times N}$ is the learnable embedding matrix for task ID t, and β is a balancing factor. This design ensures that the model can select experts based on the current image-text content while rapidly locking onto relevant expert groups using the task ID.

Adaptive Task Policy. Sarcasm understanding consists of several subtasks, which exhibit varying dependencies on perception versus execution. To dynamically regulate this dependency, we introduce a learnable task policy matrix $\mathbf{\Pi} \in \mathbb{R}^{T\times 2}$, where T denotes the total number of task types. For an input belonging to task type t, the final output $\tilde{\mathcal{E}}_i(\mathbf{x})$ of an activated expert $\mathcal{E}_i$ is modulated as:

$$\tilde{\mathcal{E}}_i(\mathbf{x}) = \sigma(\mathbf{\Pi}_{t,s})\cdot(\mathbf{B}^i_{\mathcal{S}}\mathbf{A}^i_{\mathcal{S}}\mathbf{x}) + \sigma(\mathbf{\Pi}_{t,p})\cdot(\mathbf{B}^i_{\mathcal{P}}\mathbf{A}^i_{\mathcal{P}}\mathbf{x}), \tag{4}$$

where $\sigma(\cdot)$ is the Softplus activation function ensuring positive scaling. Through this mechanism, the model automatically learns the optimal perception-execution ratio for each subtask.

4.4 Orthogonality Aggregation

Feature Aggregation. Based on the dual-stream experts $\tilde{\mathcal{E}}_i(\mathbf{x})$ and the router $\mathbf{g}(\mathbf{x}, t)$, the final output $\mathbf{y}$ of DOSE is computed as:

$$\mathbf{y} = \mathbf{W}_0\mathbf{x} + \sum_{i=1}^{N} \mathbf{g}(\mathbf{x}, t)_i \cdot \tilde{\mathcal{E}}_i(\mathbf{x}). \tag{5}$$

Orthogonality Regularization. To mathematically enforce feature disentanglement and prevent gradient updates from the private stream from causing catastrophic interference to the shared stream, we impose orthogonality constraints between the up-projection matrices. The regularization term $\mathcal{L}_{orth}$ is defined as:

$$\mathcal{L}_{orth} = \sum_{i=1}^{N} \left\| (\mathbf{B}^i_{\mathcal{S}})^\top \mathbf{B}^i_{\mathcal{P}} \right\|_F^2, \tag{6}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Minimizing this term forces the column vectors of $\mathbf{B}^i_S$ and

Method	Size	Detection	Mechanism	Description	Target	Intent	Overall
			Zero-shot				
Gemma-3 (Team et al., 2025)	4B	71.10	16.37	79.88	18.02	67.51	50.58
Phi-4-Multimodal (Abouelenin et al., 2025)	5.6B	90.72	29.05	79.43	21.99	20.08	48.25
InternVL-3 (Zhu et al., 2025)	8B	89.93	36.98	89.49	40.58	28.40	57.08
Molmo-D-0924 (Deitke et al., 2024)	7B	52.59	32.22	77.17	32.55	51.83	49.27
LLaVA-1.6 (Li et al., 2024)	7B	48.66	32.28	51.18	26.59	68.88	45.52
Qwen3-VL (Team, 2025)	8B	92.14	36.25	90.56	36.81	18.62	54.88
Gemma-3 (Team et al., 2025)	12B	85.33	18.27	84.55	46.74	31.84	53.35
InternVL-3 (Zhu et al., 2025)	14B	94.68	26.39	88.75	34.00	33.74	55.51
LLaVA-1.6 (Li et al., 2024)	13B	80.20	32.51	61.64	27.71	13.69	43.15
Gemma-3 (Team et al., 2025)	27B	75.88	37.49	87.22	38.73	46.35	57.13
Qwen3-VL (Team, 2025)	32B	85.17	31.00	90.91	46.16	22.68	55.18
Gemini-2.5-Flash (Comanici et al., 2025)	-	94.33	32.65	92.13	47.45	43.44	68.33
GPT-5-mini (OpenAI, 2025)	-	86.49	27.63	93.42	34.85	63.16	65.84
			Fine-tuned				
LLaVA-1.6							
w/ LoRA (Hu et al., 2022)	7B	92.96	33.85	92.28	71.33	69.15	71.91
w/ MoE-LoRA (Luo et al., 2024)	7B	48.63	32.48	29.92	26.31	68.99	41.27
w/ CL-MoE (Huai et al., 2025)	7B	90.82	33.88	92.56	67.92	68.93	70.82
w/ DOSE (ours)	**7B**	**94.71**	**39.92**	**93.65**	**77.34**	**74.35**	**75.99**
Qwen3-VL							
w/ LoRA (Hu et al., 2022)	8B	94.76	39.03	94.98	85.27	**74.45**	77.70
w/ MoE-LoRA (Luo et al., 2024)	8B	91.51	36.11	93.62	33.46	15.37	54.01
w/ CL-MoE (Huai et al., 2025)	8B	94.65	38.73	94.81	81.36	71.99	76.31
w/ DOSE (ours)	**8B**	**95.86**	**40.14**	**95.69**	**88.69**	74.14	**78.90**

Table 3: Main results on DocMSU-PLUS (Accuracy, %). All fine-tuned models are trained on 1% of the DocMSU-PLUS dataset and evaluated on the remaining data.

$\mathbf{B}_P^i$ to be geometrically orthogonal. This implies that the feature directions learned by the private stream are perpendicular to those of the shared stream. According to gradient dynamics analysis (Yu et al., 2020), this orthogonality minimizes the projection of gradient updates for private tasks onto the shared subspace, thereby fundamentally mitigating negative transfer.

Optimization Objective. DOSE employs an end-to-end multi-task joint optimization strategy. The total loss function $\mathcal{L}_{total}$ comprises the cross-entropy loss for the generative tasks $\mathcal{L}_{CE}$ and the weighted orthogonality regularization term:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{orth}, \tag{7}$$

where $\mathcal{L}_{CE}$ maximizes prediction accuracy across all subtasks, and λ is a hyperparameter controlling the strength of disentanglement.

5 Experiments

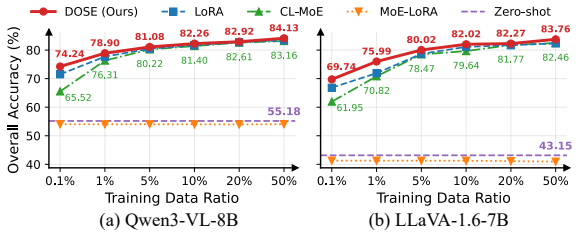
5.1 Experimental Setup

Baselines. To comprehensively evaluate DOSE, we compare it against two categories of methods: **(1) General MLLMs (Zero-Shot):** We evaluate leading proprietary and open-source models, including Gemma-3 (4B/12B/27B) (Team et al., 2025), InternVL-3 (8B/14B) (Zhu et al., 2025), LLaVA-1.6 (7B/13B) (Li et al., 2024), Molmo-D-0924 (7B) (Deitke et al., 2024), Phi-4-Multimodal (5.6B) (Abouelenin et al., 2025), Qwen3-VL (8B/32B) (Team, 2025), Gemini-2.5-Flash (Comanici et al., 2025), and GPT-5-mini (OpenAI, 2025). This allows us to probe the intrinsic sarcasm understanding capabilities of existing MLLMs. **(2) Fine-Tuning Methods:** We apply LoRA (Hu et al., 2022), MoE-LoRA (Luo et al., 2024), CL-MoE (Huai et al., 2025), and our DOSE on two representative backbones, LLaVA-1.6-7B and Qwen3-VL-8B. MoE-LoRA performs PEFT using the MoE framework. CL-MoE serves as a strong baseline for multi-task MoE. All fine-tuning experiments are strictly constrained to 1% of the training data. This setting is designed to compare performance in low-resource scenarios.

Evaluation Metrics. Following the standardized MCQ paradigm of DocMSU-PLUS, we employ Accuracy as the unified metric across all five subtasks. Detailed implementation details and settings are provided in the Appendix E.

Method	Detection	Mechanism	Description	Target	Intent	Overall
w/o Private Stream	92.02	38.13	91.85	83.35	71.17	75.30
w/o Shared Stream	89.14	38.94	92.81	85.13	70.43	75.49
w/o $\mathcal{L}_{orth}$	90.10	36.13	90.89	**91.34**	73.61	76.41
w/o Adaptive Gating	88.18	35.32	88.98	90.46	**74.87**	75.56
DOSE	**95.86**	**40.14**	**95.69**	88.69	74.14	**78.90**

Table 4: Ablation study on Qwen3-VL-8B, trained with 1% data (%).



(a) Qwen3-VL-8B (b) LLaVA-1.6-7B

Figure 3: Performance scaling with training data size.

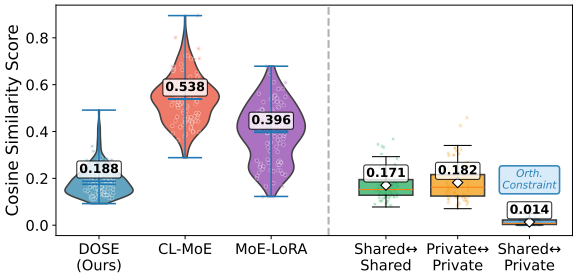


Figure 4: Analysis of expert diversity (left) and subspace orthogonality (right).

5.2 Main Results

Table 3 reports the main results on DocMSU-PLUS. We demonstrate the superiority of DOSE from three perspectives: **(1) DOSE bridges the perception-execution gap inherent in general MLLMs.** As shown in the zero-shot block, even advanced proprietary models (e.g., GPT-5-mini) exhibit severe capability skew. They score high on sarcasm detection but fail on reasoning tasks like mechanism analysis and target identification. DOSE effectively bridges this gap, transforming general sarcasm perception into fine-grained execution capabilities. **(2) DOSE resolves task conflicts to achieve holistic superiority over standard fine-tuning.** LoRA exhibits performance divergence. It nearly matches DOSE on global sarcasm detection, yet the gap widens on local target identification. This indicates that shared parameters hit a bottleneck when simultaneously accommodating conflicting granularities. DOSE's decoupled design effectively breaks this limit, maintaining consistent superiority across both dimensions. **(3) DOSE ensures robust low-resource adaptation, outperforming generic MoE architectures.** DOSE demonstrates superior stability compared to generic MoEs. MoE-LoRA collapses on LLaVA, and CL-MoE fails to surpass the simple LoRA baseline. This proves that simply introducing MoE to adapt to MTL is insufficient. DOSE's orthogonal regularization provides the essential geometric constraints to prevent expert redundancy and ensure stable convergence where other MoEs fail.

5.3 Data Efficiency and Ablation Study

Data Efficiency. To investigate the adaptability of models, we evaluate performance across training data ratios from 0.1% to 50% on Qwen3-VL and LLaVA-1.6, as shown in Figure 3. Results show that DOSE significantly outperforms baselines. Specifically on LLaVA-1.6, DOSE achieves an accuracy of 75.99% with only 1% of the data, surpassing CL-MoE by over 5%. Notably, MoE-LoRA consistently underperforms the zero-shot baseline, indicating that unconstrained experts introduce optimization noise. In contrast, DOSE avoids such negative transfer via orthogonal regularization. Overall, DOSE consistently achieves the best performance across all settings, verifying its robustness under varying resource constraints.

Component Ablation. We conduct ablation study on Qwen3-VL-8B trained with 1% data. As shown in Table 4, single-stream variants exhibit extreme imbalance. Specifically, retaining only the shared stream collapses local targeting, whereas relying solely on the private stream impairs global detection. This confirms the need for dual-stream. Furthermore, removing $\mathcal{L}_{orth}$ induces feature homogenization and degrades fine-grained reasoning, which validates the necessity of the orthogonality constraint. Finally, adaptive gating ensures dynamic synergy, while fixed weights fail to optimize the task trade-off. Consequently, every component of DOSE is indispensable.

5.4 Analysis

All analyses are based on Qwen3-VL-8B trained with 1% of the data.

Feature Orthogonality and Diversity. To verify whether orthogonal regularization effectively isolates perception from execution, we analyze the cosine similarity between output features as shown in Figure 4. DOSE exhibits significantly lower inter-expert similarity compared to CL-MoE, indicating reduced redundancy and higher expert diversity. Furthermore, the similarity between DOSE's shared and private streams is suppressed to a negligible 0.014. This geometric orthogonality confirms that $\mathcal{L}_{orth}$ successfully forces the two streams to encode instruction-invariant perception and task-specific execution in complementary subspaces.

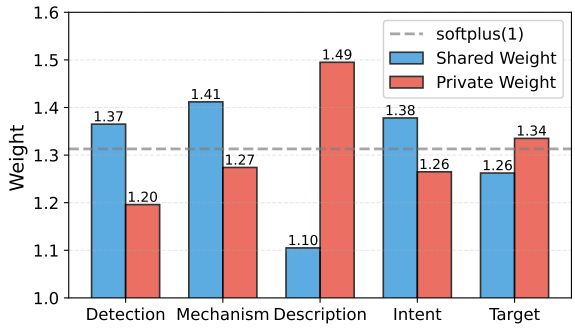


Figure 5: Visualization of learned task-adaptive policies.

Visualization of Adaptive Task Policy. To interpret how DOSE balances diverse cognitive demands, we visualize the learned adaptive task policy across five distinct task types in Figure 5. Here, we define $\alpha_{\mathcal{S}} = \sigma(\mathbf{\Pi}_s)$ and $\alpha_{\mathcal{P}} = \sigma(\mathbf{\Pi}_p)$ for short. Results reveal a clear difference in execution. For abstract reasoning tasks such as sarcasm detection, mechanism analysis, and intent reasoning, the policy consistently assigns higher weights to the shared subspace. For instance, mechanism analysis yields $\alpha_{\mathcal{S}} = 1.41$ compared to $\alpha_{\mathcal{P}} = 1.27$, which indicates that understanding sarcasm prioritizes holistic incongruity features over detail patterns. In contrast, the model shifts focus to the private subspace for grounding-oriented tasks like sarcasm description and target identification. Notably, sarcasm description exhibits the strongest execution bias ($\alpha_P = 1.49$ vs. $\alpha_S = 1.10$). This suggests that detailed sarcasm identification relies on fine-grained coordinates.Consequently, DOSE effectively acts as a dynamic router to modulate the perception-execution trade-off.

Mitigation of Negative Transfer. To evaluate robustness against task conflicts, we evaluate DOSE under three out-of-distribution settings in Table 5. We observe two key findings: (1) Optimization conflicts exist between tasks. For instance, when trained on sarcasm description and target identification, LoRA sacrifices mechanism analysis performance. (2) DOSE mitigates these issues via orthogonal decoupling. It achieves positive transfer on sarcasm detection and intention reasoning when trained on sarcasm description and target identification while preserving mechanism analysis ability. This verifies that our approach fosters constructive task synergy.

Method	Det.	Mec.	Des.	Tar.	Int.
Zero-shot	92.14	36.25	90.56	36.81	18.62
Trained on Detection					
LoRA	-	38.75↑	92.65↑	33.60↓	14.40↓
MoE-LoRA	-	36.08↓	93.69↑	33.38↓	15.36↓
CL-MoE	-	37.75↑	89.73↓	32.71↓	13.92↓
DOSE	-	38.20↑	92.57↑	33.96↓	17.14↓
Trained on Description					
LoRA	94.11↑	31.36↓	-	42.21↑	27.48↑
MoE-LoRA	91.53↓	36.22↓	-	33.51↓	15.32↓
CL-MoE	93.36↑	31.55↓	-	43.40↑	28.89↑
DOSE	92.51↑	31.76↓	-	43.67↑	36.88↑
Trained on Description & Target					
LoRA	91.51↓	32.82↓	-	-	56.33↑
MoE-LoRA	91.58↓	36.28↑	-	-	15.30↓
CL-MoE	92.99↑	32.11↓	-	-	49.77↑
DOSE	93.46↑	36.27↑	-	-	55.60↑
DOSE(1%)	**95.86**	**40.14**	**95.69**	**88.69**	**74.14**

Table 5: Task conflict and transfer analysis(%). ↑ and ↓ represent improvements and decreases in the effectiveness of zero-shot, respectively.

6 Conclusion

This paper bridges the long-standing evaluation gap in MSU. By establishing DocMSU-PLUS, we unify diverse subtasks into a standardized MCQ paradigm. This shift provides the first benchmark for holistic MSU cognitive assessment. To tackle feature entanglement, we introduce the DOSE framework, which mathematically isolates instruction-invariant perception from task-specific execution. Extensive experiments demonstrate that DOSE achieves state-of-the-art performance by effectively balancing commonality with specificity. Beyond specific performance gains, this work sets a precedent for aligning complex cognitive tasks within a discrete decision space. We hope our findings inspire future research to explore deeper geometric constraints for multimodal reasoning.

Acknowledgments

This research was supported by the National Key R&DProgram of China (No. 2023YFC3303800).

Limitation

Despite the promising results, our work has limitations. First, sarcasm is highly culture-dependent. Our dataset is primarily sourced from Western social media (e.g., News Thump), which may limit the model's generalization to non-English contexts or diverse cultural nuances. Second, while DOSE effectively captures pixel-text incongruity, it may struggle with sarcasm grounded in specific external knowledge (e.g., political events or pop culture references) that is absent from the immediate visual context. Third, our experiments are conducted on 7B/8B-parameter models. Investigating whether the benefits of orthogonal decoupling persist or diminish in larger-scale foundation models (e.g., >70B) is a subject for future research.

Ethical Statement

Our DocMSU-PLUS benchmark builds upon the foundational data collection of the original DocMSU, inheriting its rigorous adherence to the copyright regulations of source platforms (e.g., TheOnion, UNNews). Since these platforms grant copyright usage to compliant users, we meticulously conform to their regulations during data collection and annotation. Beyond standard compliance, we have enhanced the ethical framework by integrating a strict human-in-the-loop screening process during our new annotation phase to manually filter out potential toxicity and safeguard user privacy. All data is for research purposes only and includes only publicly available information; no personal privacy information is included. Upon publication, we will maintain full transparency regarding data provenance and commit to a responsive takedown policy for any content concerns. We will implement a strict license agreement requiring all downstream researchers to abide by the original intellectual property rights and ethical guidelines of the source websites. We made limited use of AI-assisted tools (e.g., ChatGPT) for text polishing and minor code assistance. All research ideas, methods, and experimental analyses were independently conducted and verified by the authors.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Ommi Balamurali, AM Abhishek Sai, and Sruthy Anand. 2024. Advancing tourism chatbots: Understanding irony, sarcasm, and negative emotions of users. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE.

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*, pages 2506–2515.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629.

Zixin Chen, Hongzhan Lin, Ziyang Luo, Mingfei Cheng, Jing Ma, and Guang Chen. 2024. Cofipara: A coarse-to-fine paradigm for multimodal sarcasm target identification with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9663–9687.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 32 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

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.

Hang Du, Guoshun Nan, Sicheng Zhang, Binzhu Xie, Junrui Xu, Hehe Fan, Qimei Cui, Xiaofeng Tao, and Xudong Jiang. 2024. Docmsu: A comprehensive benchmark for document-level multimodal sarcasm

understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17933–17941.

DI Hernández Farias and Paolo Rosso. 2017. Irony, sarcasm, and sentiment analysis. In *Sentiment Analysis in Social Networks*, pages 113–128. Elsevier.

Diandian Guo, Cong Cao, Fangfang Yuan, Yanbing Liu, Guangjie Zeng, Xiaoyan Yu, Hao Peng, and Philip S Yu. 2025. Multi-view incongruity learning for multimodal sarcasm detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1754–1766.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. 2025. Clmoe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19608–19617.

Mengzhao Jia, Can Xie, and Liqiang Jing. 2024. Debiasing multimodal sarcasm detection with contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18354–18362.

Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia, and Liqiang Nie. 2023. Multi-source semantic graph-based multimodal sarcasm explanation generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11349–11361.

Arpit Khare, Amisha Gangwar, Sudhakar Singh, and Shiv Prakash. 2023. Sentiment analysis and sarcasm detection in indian general election tweets. In *Research Advances in Intelligent Computing*, pages 253–268. CRC Press.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. *arXiv preprint arXiv:2203.06419*.

Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: What else influences visual instruction tuning beyond data?

Yuchao Li, Fuli Luo, Chuanqi Tan, Mengdi Wang, Songfang Huang, Shen Li, and Junjie Bai. 2022. Parameter-efficient sparsity for large language models fine-tuning. *arXiv preprint arXiv:2205.11005*.

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. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4707–4715.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. 2024. Twin-merging: Dynamic integration of modular expertise in model merging. *Advances in Neural Information Processing Systems*, 37:78905–78935.

Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. 2024. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *arXiv preprint arXiv:2402.12851*.

Fengmao Lv, Mengting Xiong, Junlin Fang, Lingli Zhang, Tianze Luo, Weichao Liang, and Tianrui Li. 2025. Msti-plus: Introducing non-sarcasm reference materials to enhance multimodal sarcasm target identification. In *Proceedings of the ACM on Web Conference 2025*, pages 614–624.

Alice Myers Roy. 1981. The function of irony in discourse. *Text-Interdisciplinary Journal for the Study of Discourse*, 1(4):407–423.

OpenAI. 2025. Gpt-5 system card. Technical report, OpenAI. Technical report. Accessed: 2025-08-10.

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*, pages 1383–1392.

Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. 2023. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9507–9515.

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.

Graeme Ritchie. 1999. Developing the incongruity-resolution theory. Technical report.

Gopendra Vikram Singh, Mauajama Firdaus, Dushyant Singh Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2024. Well, now we

know! unveiling sarcasm: Initiating and exploring multimodal conversations with reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18981–18989.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jiquan Wang, Lin Sun, Yi Liu, Meizhi Shao, and Zengwei Zheng. 2022. Multimodal sarcasm target identification in tweets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8164–8175.

Yiwei Wei, Shaozu Yuan, Hengyang Zhou, Longbiao Wang, Zhiling Yan, Ruosong Yang, and Meng Chen. 2024. Gˆ2sam: Graph-based global semantic awareness method for multimodal sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9151–9159.

Haoyuan Wu, Haisheng Zheng, Zhuolun He, and Bei Yu. 2024. Parameter-efficient sparsity crafting from dense to mixture-of-experts for instruction tuning on general tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 737–749.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836.

Francisco Yus. 2017. Incongruity-resolution cases in jokes. *Lingua*, 197:103–122.

Didi Zhu, Zhongyisun Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Chao Wu, and Kun Kuang. 2024. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. In *International Conference on Machine Learning*, pages 62581–62598. PMLR.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Xingjie Zhuang, Zhixin Li, Canlong Zhang, and Huifang Ma. 2025. A cross-modal collaborative guiding network for sarcasm explanation in multi-modal multi-party dialogues. *Engineering Applications of Artificial Intelligence*, 142:109884.

A Dataset Construction Details

A.1 Detailed Statistics and Distribution

Table 6 presents the comprehensive statistics of the constructed dataset. Our dataset maintains a balanced distribution between positive (sarcastic) and negative (non-sarcastic) samples, with a ratio of approximately 1:1, preventing model bias towards the majority class. The corpus covers a diverse range of 8 distinct topics, spanning from high-frequency domains like Entertainment (23.8%) to specialized fields such as Science (4.9%) and Environment (5.8%).

Crucially, unlike previous datasets that rely on simple binary labels, DocMSU-PLUS provides fine-grained reasoning annotations for a significant portion of the data. As shown in the "Dataset Overview" section of Table 6, 100% of sarcastic samples are equipped with Sarcasm Description QAs, and 80.6% contain Target Identification QAs. This rich annotation density supports the training of comprehensive multimodal reasoning capabilities.

A.2 Cognitive Taxonomy Definitions

To ensure methodological rigor and minimize annotator subjectivity, our fine-grained categories for sarcasm mechanisms and intents are grounded in established linguistic and cognitive theories.

Sarcasm Mechanism. Based on the Incongruity-Resolution Theory (Ritchie, 1999; Yus, 2017) and multimodal discourse analysis, we classify the logic of sarcasm into five types:

(1) Image-Text Contradiction: Represents the canonical definition of multimodal irony where the visual content directly contradicts the literal proposition of the text. For example, a caption praising "beautiful weather" paired with an image of a thunderstorm.

(2) Text-led Contextualization: Here, the text contains the primary sarcastic marker (e.g., hyperbole), but it requires the image to serve as the necessary context to trigger the pragmatic insincerity. As shown in Table 6, "Text-dominant" samples account for 61.9% of the dataset, challenging the model's ability to decode subtle textual nuances grounded in vision.

(3) Situational Mismatch: Corresponds to situational irony, where the incongruity arises not from a direct semantic clash, but from a discrepancy between the expected outcome of a situation depicted and the actual reality shown.

Dataset Overview			**Topic Distribution**		
Total Records	14,128	100.0%	Entertainment	3,360	23.8%
Sarcastic Samples	6,874	48.7%	Health	2,107	14.9%
Non-Sarcastic Samples	7,254	51.3%	Sport	1,921	13.6%
With Description QA	6,874	48.7%	Politic	1,412	10.0%
A: 1638 B: 1830 C: 1752 D: 1654			Education	1,345	9.5%
With Target QA	5,539	39.2%	Business	1,273	9.0%
A: 1457 B: 1313 C: 1381 D: 1388			Technology	1,198	8.5%
With Both QA	5,539	39.2%	Environment	825	5.8%
			Science	687	4.9%
Sarcasm Mechanism*			**Modality Contribution***		
Image-Text Contradiction	2,235	32.5%	Text-dominant	4,251	61.9%
Text-led Contextualization	2,143	31.2%	Equal contribution	1,550	22.5%
Situational Mismatch	1,015	14.8%	Image-dominant	1,073	15.6%
Image-led Punchline	779	11.3%			
Single-modality Dominant	702	10.2%			
Sarcasm Intent*					
To criticize/expose	4,753	69.1%	To entertain readers	1,026	14.9%
To express discontent	989	14.4%	To create absurdity	104	1.5%
To provoke thought	3	0.0%			

*Percentage calculated among 6,874 sarcastic samples only.

Table 6: Complete Statistics of MyDocMSU Dataset

(4) Image-led Punchline: Inspired by visual humor structure, where the text acts as a setup and the visual content delivers the punchline or revelation that reinterprets the message as sarcastic.

(5) Single-modality Dominant: Cases where sarcasm is self-contained within a single modality (e.g., a purely sarcastic text caption) while the other modality serves merely as background, reflecting unimodal sarcasm in a multimodal context.

Sarcasm Intents. Following the pragmatic function theory of irony (Myers Roy, 1981), we categorize the underlying purpose:

(1) To criticize/expose: The aggressive function of sarcasm, used to attack a specific target or expose a flaw/vice.

(2) To express discontent: A milder form of criticism focusing on the speaker's emotional state (frustration/complaint) rather than attacking an external target.

(3) To entertain readers: The primary goal is humor or social bonding rather than aggression.

(4) To create absurdity: Cases where the incongruity is used to highlight the inherent irrationality or surreal nature of a situation.

(5) To provoke thought: Rare instances where sarcasm serves a deeper philosophical or reflective purpose, transcending mere criticism or humor to prompt introspection about societal issues

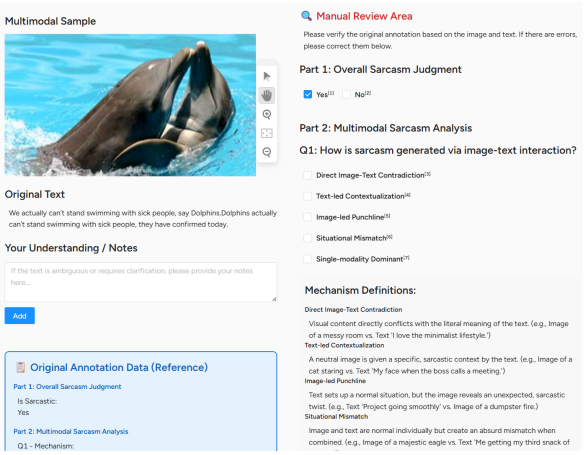


Figure 6: User interface used by human annotators.

A.3 Annotation Quality Control and Case Examples

To ensure the reliability of the initial profiles generated by the MLLM engine (Phase 1) and to reduce the workload for the subsequent expert verification (Phase 3), we implement a rigorous preliminary manual screening process. This phase specifically targets obviously unreasonable annotations by applying the following rejection criteria. (1) **Format Violations**, where outputs failed to adhere to the

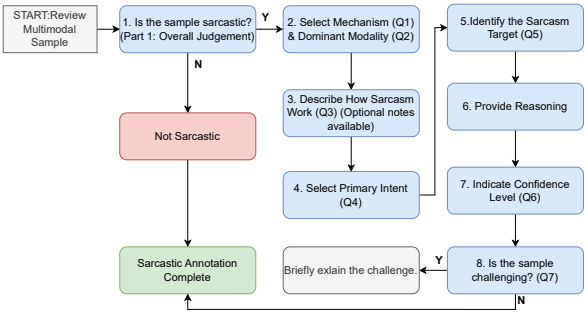


Figure 7: The annotation workflow.

strict JSON schema or contained syntax errors; (2) **Object Hallucinations**, where the reasoning explicitly referenced entities factually absent from the image or text; (3) **Logical Inconsistencies**, such as contradictions between the identified sarcasm mechanism and the dominant modality; and (4) **Safety Refusals**, where the model declined analysis due to internal guardrails. Approximately 9% of raw samples are discarded in this phase.

As described in Table 2 of the main text, we validate the reliability of our taxonomy via a human-in-the-loop verification pipeline using a customized Label Studio interface. The screenshot of the user interface is shown in Figure 6. As illustrated in Figure 7, the human annotation process follows a strict seven-step decision protocol to ensure consistency and efficiency. The overall inter-annotator agreement (Fleiss' Kappa) reached 0.82, indicating strong consistency.

Figure 8 illustrates a representative sample from DocMSU-PLUS. The central multimodal input is surrounded by five distinct MCQ tasks. This visualization demonstrates how we transform diverse cognitive dimensions into a standardized evaluation format, requiring the model to not only detect if an image is sarcastic but also articulate why, who, and for what purpose.

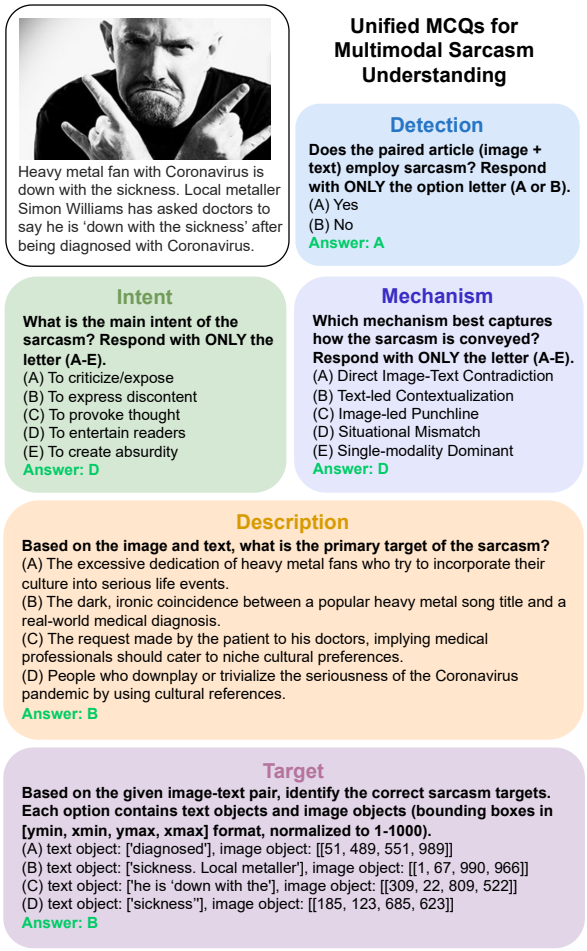


Figure 8: An example from DocMSU-PLUS.

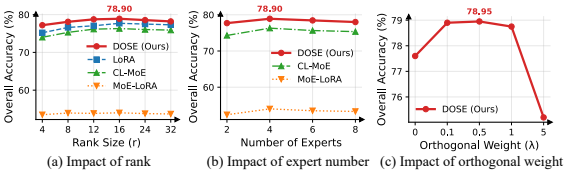


Figure 9: Hyperparameter sensitivity analysis.

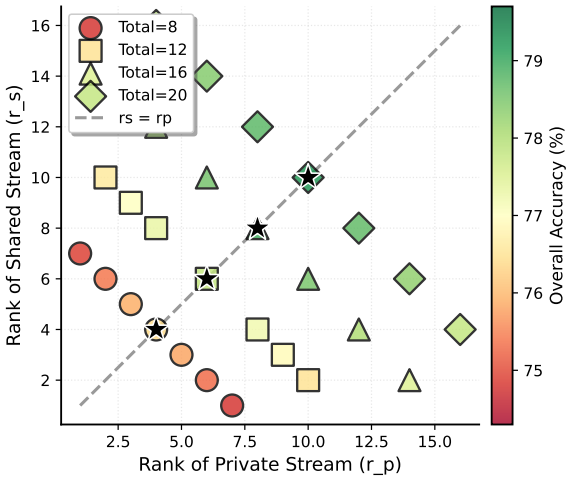


Figure 10: Rank distribution analysis.

B Additional Analysis

B.1 Hyperparameter Sensitivity

To understand how different configurations affect DOSE's performance, we conduct a series of ablation studies focusing on three key hyperparameters: the rank r, the number of experts N, and the orthogonality weight λ. The results are shown in Figure 9.

Impact of Rank Configuration. We compare DOSE against baselines under same total rank setting. For DOSE, we set $r_s = r_p = r/2$. As illus-

Method	Detection	Mechanism	Description	Target	Intent	Overall
Trained on health						
LoRA	**94.19**	39.78	98.57	95.14	74.77	80.49
MoE-LoRA	91.36	36.70	93.50	32.63	14.76	53.79
CL-MoE	93.24	42.16	98.62	95.00	**75.16**	80.84
DOSE (ours)	92.66	**43.53**	**98.91**	**95.33**	74.79	**81.04**
Trained on health&technology						
LoRA	**96.05**	40.18	98.87	95.88	75.44	81.28
MoE-LoRA	91.76	37.05	93.47	32.87	15.03	54.04
CL-MoE	95.34	43.12	98.99	95.74	75.36	81.71
DOSE (ours)	95.58	**44.53**	**99.09**	**95.90**	**75.64**	**82.15**

Table 7: Cross-domain generalization experiments. (%)

trated in the results, while increasing the rank generally benefits all methods by providing larger capacity, DOSE consistently outperforms baselines. Notably, DOSE demonstrates superior parameter efficiency at lower ranks ($r = 4$ and $r = 8$), achieving competitive results where other MoE variants struggle. As the rank increases, DOSE's performance climbs steadily and reaches its peak at $r = 16$. Beyond this point, further increasing the parameter budget to $r = 24$ or $r = 32$ does not yield additional gains and even shows slight fluctuations, suggesting that a rank of 16 provides sufficient capacity for capturing the necessary sarcasm features without overfitting. This confirms that DOSE's superiority stems from its structural disentanglement rather than high-parameter settings.

Impact of Expert Number. We investigate the scalability of the Mixture-of-Experts architecture by varying the number of experts N. The performance exhibits a distinct "inverted-U" trend. With a small number of experts, the model suffers from limited capacity, failing to adequately capture the diverse patterns of sarcasm mechanisms. As N increases, the performance improves significantly, converging to an optimal point at $N = 4$, where the model effectively balances expert specialization and training stability. However, scaling N further to 6 or 8 leads to a gradual decline in accuracy. We argue that an excessive number of experts dilutes the training signal for each individual expert, making optimization more difficult. Thus, $N = 4$ represents robust configuration.

Impact of Orthogonality Weight. We examine the influence of the orthogonality regularization weight λ, which controls the strength of the disentanglement constraint. The model exhibits high robustness within a reasonable range. Specifically, performance remains stable and optimal when λ is set between 0.1 and 1. When $\lambda = 0$ (i.e., removing the constraint), the model degenerates into a standard unconstrained MoE, resulting in a noticeable performance drop due to the lack of explicit feature separation. Conversely, an aggressively large weight leads to a sharp decline in accuracy, indicating that over-regularization constrains the subspace and hinders the learning of task-specific features.

B.2 Impact of Rank Distribution

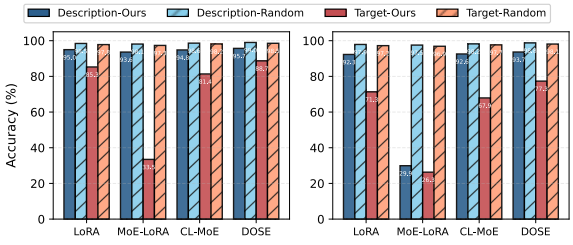
Beyond the total parameter budget, we investigate the optimal allocation of capacity between the Shared and Private streams. As visualized in Figure 10, we evaluate various split configurations across different total ranks. The model consistently achieves optimal performance when the rank is evenly distributed between the two streams (e.g., $r_s = 8, r_p = 8$ for r=16). Deviating from this balance leads to suboptimal results. Specifically, a high r_s with low r_p setting degrades fine-grained reasoning capabilities due to insufficient execution capacity, whereas a low r_s with high r_p setting fails to capture universal incongruity patterns, impairing cross-task transfer. This performance confirms that multimodal sarcasm understanding requires a equilibrium between perception and execution.

B.3 Distractor Quality Verification

To quantify the discriminative power of our MCQ benchmark and refute the concern that the task relies on simple process of elimination, we conduct an ablation study on distractor difficulty. We construct a control group dataset where the hard negative options generated by SDE are replaced with random negatives. As shown in Figure 11, the model achieves an accuracy of nearly 98% on

Method	Detection	Mechanism	Description	Target	Intent	Overall
Trained with 1% data						
LoRA	94.76	39.03	94.98	85.27	**74.45**	77.70
MoE-LoRA	91.51	36.11	93.62	33.46	15.37	54.01
CL-MoE	94.65	38.73	94.81	81.36	71.99	76.31
LoRA (specialist)	94.01	30.45	93.09	79.83	69.51	73.38
DOSE (ours)	**95.86**	**40.14**	**95.69**	**88.69**	74.14	**78.90**
Trained with 0.1% data						
LoRA (specialist)	88.27	26.64	91.86	57.92	70.96	67.13
DOSE	**92.71**	**27.13**	**95.20**	**83.94**	**72.19**	**74.24**
Trained with 5% data						
LoRA (specialist)	95.89	37.30	95.79	89.06	73.82	78.37
DOSE	**96.85**	**42.23**	**97.69**	**93.56**	**75.06**	**81.08**
Trained with 10% data						
LoRA (specialist)	97.01	37.08	97.40	92.77	75.20	79.89
DOSE	**97.56**	**45.13**	**98.40**	**94.20**	**76.02**	**82.26**
Trained with 20% data						
LoRA (specialist)	97.46	41.19	97.89	94.51	76.15	81.44
DOSE	**97.92**	**46.01**	**98.96**	**95.55**	**76.17**	**82.92**
Trained with 50% data						
LoRA (specialist)	97.88	44.60	98.86	96.66	76.39	82.88
DOSE	**98.19**	**48.82**	**99.18**	**96.69**	**77.76**	**84.13**

Table 8: Compared with specialist LoRA. LoRA (specialist) refers to LoRA trained separately for each task.



(a) Qwen3-VL-8B (b) LLaVA-1.6-8B

Figure 11: Distractor quality verification.

the random setting of both task, indicating that distinguishing sarcasm from irrelevant content is trivial. However, performance drops under our SDE setting. This substantial performance gap quantitatively proves that our SDE method successfully constructs a non-trivial decision boundary, forcing the model to engage in deep semantic reasoning rather than superficial matching.

B.4 Cross-Domain Generalization

To verify whether the model captures universal sarcasm logic rather than overfitting to domain-specific keywords, we conduct domain out-of-distribution experiments. The results are shown in Table 7. We observe a trade-off between surface-level detection and deep reasoning. While the LoRA baseline achieves a higher score on binary detection, it suffers a performance drop in reasoning-intensive tasks, specifically mechanism and intent reasoning. This phenomenon suggests that LoRA likely overfits to domain-specific shortcuts or keywords within the data, which aids simple classification but fails to transfer the underlying logic of sarcasm to unseen domains. In contrast, DOSE demonstrates superior generalization on deep reasoning tasks, achieving the highest overall accuracy. These results confirm that DOSE successfully enables robust transfer to unseen domains even when training data is conceptually limited.

B.5 Mitigation of Negative Transfer

To investigate the efficacy of DOSE in resolving the task conflict phenomenon in MTL, we compare DOSE against a series of single-task specialists LoRA in Table 8. The results strongly validate our original motivation regarding task conflict. As observed in the main experiments, standard multi-task baselines like MoE-LoRA suffer from catastrophic negative transfer, yielding an accuracy of only 54.01%, which is far inferior to the single-task specialists. This confirms that without explicit disentanglement, the heterogeneous objectives of MSU sub-tasks severely interfere with each other. In contrast, DOSE not only eliminates this interference but successfully reverses the trend. This proves that DOSE turns the inherent task conflict into positive synergy.

B.6 Performance on Generation Task

To verify that our model's superiority is not an artifact of the discriminative MCQ format, we reformulate the sarcasm description task as a purely open-ended generation problem. Instead of selecting

Model	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-4
zero-shot	0.8914	0.3489	0.1029	0.2806	0.3017	0.1402	0.0511
LoRA	0.9048	0.4231	0.2014	0.3698	0.3756	0.2378	0.1453
MoE-LoRA	0.8652	0.2891	0.0742	0.2318	0.2534	0.1086	0.0352
CLMoE	0.8897	0.3412	0.0987	0.2753	0.2956	0.1368	0.0489
DOSE	**0.9065**	**0.4298**	**0.2077**	**0.3750**	**0.3821**	**0.2424**	**0.1496**

Table 9: Experiments on generative sarcasm description task.

Model	Detection	Mechanism	Intent
RoBERTa (Liu et al., 2019)	94.33	35.07	69.71
Att-BERT (Pan et al., 2020)	94.76	36.83	70.42
Multi-CLIP (Qin et al., 2023)	**96.39**	37.43	71.46
DOSE	95.86	**40.14**	**74.14**

Table 10: Comparison with specialized models.

Method	MMSD 2.0	WITS
Zero-shot	67.78	29.46
LoRA	68.36	29.01
MoE-LoRA	67.78	29.46
CL-MoE	68.24	27.23
DOSE	**69.21**	**30.35**

Table 11: Cross dataset experiments (%).

from options, the model is prompted to generate the explanation directly. The results in Table 9 demonstrate that DOSE maintains its leadership position, consistently generating more coherent and logically sound explanations than the baselines. This confirms that DOSE does not merely learn to hack the MCQ format through elimination strategies, but genuinely acquires a robust, format-independent capability to comprehend and articulate the nuances of multimodal sarcasm.

B.7 Comparison with Specialized Models

We compare DOSE against traditional, specialized small-scale models trained with 20% data. Despite being lightweight and specifically designed for classification tasks, these small models struggles to capture the deep semantic incongruity required for sarcasm understanding, often overfitting to surface-level textual or visual features. DOSE exhibits superiority on sarcasm mechanism and intent reasoning tasks over these traditional methods. This highlights a critical limitation of small-scale specialists. They typically require full-scale supervision to establish complex reasoning boundaries. For complex cognitive tasks, the combination of a large-scale foundation model with our effective parameter decoupling mechanism offers a decisive advantage over varying specialized small models.

B.8 Cross Dataset Performance

To verify that DOSE captures universal sarcasm logic rather than overfitting, we evaluate the models of main experiments on external benchmarks MMSD 2.0 (Qin et al., 2023) and WITS (Kumar et al., 2022) by reformatting them to match our prompt schema without further fine-tuning. As shown in Table 11, DOSE demonstrates superior generalization, achieving the highest accuracy on both datasets (69.21% and 30.35%). A critical finding is observed on the challenging WITS dataset where LoRA (29.01%) and CL-MoE (27.23%) underperform even the Zero-shot baseline (29.46%), indicating negative transfer caused by domain-specific overfitting. In contrast, DOSE successfully avoids this pitfall, confirming that its orthogonal disentanglement enables the learning of domain-invariant sarcasm mechanics.

B.9 Model Scale

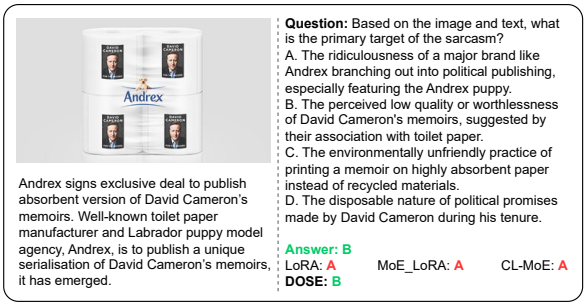
To verify if feature entanglement is merely an artifact of smaller models, we conducted scale-up experiments on Qwen3-VL-32B and LLaVA-1.6-13B in Table 12. Our results explicitly confirm two points: (1) Entanglement persists at scale. The zero-shot Qwen3-VL-32B still exhibits severe capability skew (e.g., 85.17% on Detection vs. only 22.68% on Intent). Parameter scale alone does not naturally resolve this cognitive conflict. (2) DOSE is scalable. DOSE consistently outperforms standard LoRA across both 13B and 32B scales. This validates that DOSE provides fundamental geometric constraints that benefit 30B+ models, proving its necessity far beyond edge devices.

C Case Study

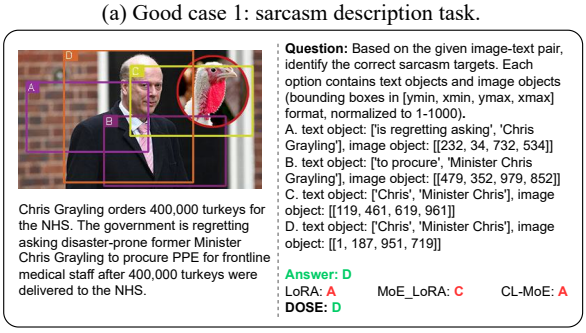
Good Cases. To provide a more intuitive understanding of DOSE's capabilities and limitations, we present qualitative visualizations. Figure 12(a) illustrates a representative case requiring deep sar-

Model	Detection	Mechanism	Description	Target	Intent	Overall
Qwen3-VL-32B	85.17	31.00	90.91	46.16	22.68	55.18
+LoRA	95.97	39.70	93.91	86.79	74.82	78.24
+CL-MoE	95.46	39.79	93.25	76.23	74.64	75.87
+DOSE	**96.17**	**40.60**	**94.78**	**87.98**	**75.57**	**79.02**
LLaVA-1.6-13B	80.20	32.51	61.64	27.71	13.69	43.15
+LoRA	93.87	33.42	93.03	72.04	71.07	72.69
+CL-MoE	94.16	32.52	93.16	69.63	70.02	71.90
+DOSE	**94.66**	**35.72**	**93.87**	**75.24**	**71.11**	**74.12**

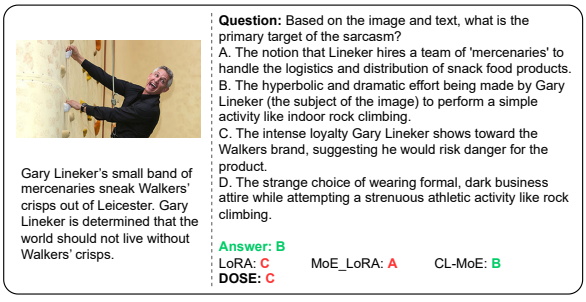
Table 12: Scale-up experiments (%).



(a) Good case 1: sarcasm description task.



(b) Good case 2: target identification task.



(c) Failure case: target identification task.

Figure 12: Case study.

casm reasoning, where DOSE outperforms the baseline. Other baselines merely understand the surface content. They incorrectly choose option A (the absurdity of brand expansion), interpreting the text as an objective description of a bizarre business move. They successfully extract the entity but fail to deduce the satirical intent regarding the book's quality. Our DOSE correctly identify option B, accurately pinpointing the satire of the memoir's worthlessness. We attribute this to the dual-stream architecture. The shared stream effectively encodes the semantic inconsistency between the high-status concept of political memoir and the low-status utility of toilet paper. Simultaneously, the private stream leverages this conflict to deduce specific pragmatic meanings, preventing the model from overfitting to the literal narrative of publishing deals. Similarly, Figure 12(b) also demonstrates the effectiveness of DOSE in utilizing MLLMs grounding capabilities.

Failure Case. Figure 12(c) illustrates a failure case. We identify a failure pattern associated with a lack of external context. Taking a sample involving British popular culture as an example, the text describes football legend Gary Lineker using mercenaries to smuggle Walkers chips, accompanied by dramatic footage of him indoor rock climbing. DOSE incorrectly predicts option C, interpreting keywords like "mercenaries" and "smuggling" too seriously as a promise. However, the correct interpretation requires recognizing Gary Lineker's long-standing advertising persona for the brand, characterized by a comedic, exaggerated greed for chips. Because the model lacks this specific cultural contextual knowledge, its perception processes the text information too literally, missing the inherent absurdity. This confirms that efficient parameter fine-tuning alone cannot infuse the model with sufficient world knowledge.

D Theoretical Analysis of Subspace Orthogonality

In this section, we provide a theoretical analysis of the proposed DOSE framework from the perspective of gradient dynamics. We aim to rigorously demonstrate how the orthogonality regularization physically blocks negative transfer in MTL.

D.1 Geometric Decoupling of Feature Representations

For the i-th expert, the output is defined as the superposition of two parallel streams:

$$E_i(x) = \underbrace{B^i_{\mathcal{S}} A^i_{\mathcal{S}} x}_{\text{Shared Stream}} + \underbrace{B^i_{\mathcal{P}} A^i_{\mathcal{P}} x}_{\text{Private Stream}}, \tag{8}$$

where the matrices $B^i_{\mathcal{S}}$ and $B^i_{\mathcal{P}}$ span the shared subspace $\mathcal{S}$ and the private subspace $\mathcal{P}$, respectively. To enforce structural disentanglement, we introduce the regularization term based on the Frobenius norm:

$$\mathcal{L}_{orth} = \sum_{i=1}^{N} \left| (B^i_{\mathcal{S}})^\top B^i_{\mathcal{P}} \right|_F^2. \tag{9}$$

Minimizing this term enforces $\mathcal{S} \perp \mathcal{P}$. Geometrically, this implies that the features captured by the shared stream (e.g., "incongruity patterns") and those captured by the private stream (e.g., "syntactic styles") are orthogonal, thereby achieving feature disentanglement.

D.2 Gradient Dynamics and Interference Mitigation

To analyze the parameter update dynamics during training, we examine the gradient of the total objective function $\mathcal{L} = \mathcal{L}_{task} + \lambda \mathcal{L}_{orth}$ with respect to the shared basis matrix $B_{\mathcal{S}}$.

Using $\frac{\partial}{\partial X}\text{Tr}(X^\top A X) = (A + A^\top)X$, the gradient of the regularization term can be derived as:

$$\nabla_{B_{\mathcal{S}}} \mathcal{L}_{orth} = \frac{\partial}{\partial B_{\mathcal{S}}} \text{Tr}(B_{\mathcal{S}}^\top B_{\mathcal{P}} B_{\mathcal{P}}^\top B_{\mathcal{S}}) = 2 B_{\mathcal{P}} B_{\mathcal{P}}^\top B_{\mathcal{S}}. \tag{10}$$

Consequently, the stochastic gradient descent update rule for $B_{\mathcal{S}}$ at step $t+1$ becomes:

$$B_{\mathcal{S}}^{(t+1)} \leftarrow B_{\mathcal{S}}^{(t)} - \eta \underbrace{\nabla B_{\mathcal{S}} \mathcal{L}_{task}}_{\text{Task Adaptation}} - 2\eta\lambda \underbrace{B_{\mathcal{P}}(B_{\mathcal{P}}^\top B_{\mathcal{S}})}_{\text{Projection Constraint}}. \tag{11}$$

The term $-2\eta\lambda B_{\mathcal{P}}(B_{\mathcal{P}}^\top B_{\mathcal{S}})$ in the update acts as a soft gradient projection operation. In each iteration, this term attempts to subtract the component of $B_{\mathcal{S}}$ that projects onto the subspace spanned by $B_{\mathcal{P}}$. This mimics a soft Gram-Schmidt process, dynamically constraining the shared parameters to evolve towards the space of the private parameters. In a multi-task scenario, if the gradient from an explanation task (Task B) attempts to modify $B_{\mathcal{S}}$ to fit a specific linguistic style, and this modification conflicts with the detection task (Task A), the orthogonality regularization generates a counterforce. This force neutralizes the drift of $B_{\mathcal{S}}$ towards the direction of $B_{\mathcal{P}}$. This mechanism aligns mathematically with the core principles of gradient surgery and domain separation networks (Yu et al., 2020). Through this mechanism, DOSE theoretically guarantees that new task learning primarily updates the private subspace, while maximally preserving the universal perception patterns acquired in the shared subspace.

Hyperparameter	Value
Training Dynamics	
Optimizer	AdamW
Learning Rate	1e-4
Iteration	2
Warmup Ratio	0.03
Batch Size	2
Weight Decay	0.05
Maximum Sequence Length	960
LoRA & DOSE Architecture	
LoRA Rank (r) for Baselines	16
LoRA Alpha	32
LoRA Dropout	0.05
Target Modules	up_proj,down_proj,gate_proj
Expert Numbers	4
DOSE Shared Rank (r_s)	8
DOSE Private Rank (r_p)	8
Orthogonality Weight	1e-3

Table 13: Hyperparameter configurations.

E Implementation Details

We implemented DOSE and all baselines using the PyTorch framework on a single NVIDIA A100 (80GB) GPUs. To ensure a fair comparison, all methods were fine-tuned under the same foundational settings. The detailed hyperparameter configurations are listed in Table 13. Training time for all methods on 1% of the training data is around 10 minutes. For evaluation time, using Qwen3-VL-8B as the base model, LoRA takes approximately 1 hour, MoE-LoRA approximately 1.5 hours, CL-MoE approximately 2 hours, and DOSE approximately 2.5 hours. For LLaVA-1.6-7B, LoRA and MoE-LoRA take approximately 3 hours, CL-MoE approximately 5 hours, and DOSE approximately 6 hours. For the experiments in B.6, we

Prompt for semantic distractor

Instructions:
You are an expert AI assistant specializing in multimodal sarcasm research. Your task is to generate a multiple-choice question based on a provided sarcastic image-text pair and its correct target explanation.

Your Goal:
Given an image, its accompanying text, and the Correct Sarcasm Description, you must generate three (3) incorrect but plausible "distractor" explanations.

Distractor Requirements:
1. Plausible: They should seem reasonable at a glance.
2. Related: They must be related to the general themes present in the image or text.
3. Incorrect: They must *not* be the actual intended target of the sarcasm.

Process:
1. Analyze the provided Image, Text, and Correct Target.
2. Create three unique distractors.
3. Combine and shuffle to create options A, B, C, D.
4. Construct a JSON output.

Output Format (Must be JSON):
{ "question": "...", "answer": "[Correct Letter]" }

used the punkt library from NLTK, along with the rouge_score and bert_score libraries, to perform the metric evaluations. The models in the main experiment are all trained independently three times, and the average value is taken.

F Prompt Templates

In this section, we provide the specific prompt templates used in our data construction pipeline.

Prompt for Cognitive-Driven Sarcasm Decomposition & Profiling

Role & Methodology:
You are an expert cognitive analyst specializing in multimodal sarcasm. Your goal is to move beyond simple classification and perform a Cognitive Decomposition of the provided image-text pair. You must strictly follow the Premises → Incongruity → Inference reasoning chain to construct a comprehensive sarcasm profile.

Phase 1: Cognitive Decomposition (The Reasoning Chain)

Step 1: Premises (Objective Extraction)
- Analyze the text literal meaning and, based on the context/URL, infer the objective visual scene (as you cannot directly "see" images).
- Goal: Establish the factual baseline before identifying sarcasm.

Step 2: Incongruity (Mechanism Analysis)
- Locate the semantic contrast between the visual and textual modalities.
- Determine the Sarcasm Mechanism (e.g., Contradiction, Contextualization) using the definitions below.
- Identify the Dominant Modality that drives the conflict.

Step 3: Inference (Description & Intent)
- Derive the Sarcastic Description (who/what is being attacked) based on the logic of the conflict.
- Determine the Sarcastic Intent (e.g., to criticize, to entertain).

Phase 2: Sarcasm Profiling (Definitions & Constraints)

1. Mechanism Definitions (Choose for 'sarcasm_mechanism'):
- Direct Image-Text Contradiction: Image contradicts literal text (e.g., messy room vs. "minimalist life").
- Text-led Contextualization: Neutral image given sarcastic meaning by text.
- Image-led Punchline: Text sets up normal situation, image provides the twist.
- Situational Mismatch: Normal components combine to create absurd mismatch.
- Single-modality Dominant: Sarcasm is contained in one modality; the other is supportive.

2. Modality Definitions (Choose for 'dominant_modality'):
- Text-dominant, Image-dominant, or Equal contribution.

3. Intent Definitions (Choose for 'sarcastic_intent'):
- To criticize/expose, To express discontent, To provoke thought, To entertain readers, To create absurdity.

Phase 3: Input & Output Generation
Input Data: Text: {text}
Output Format (JSON Only):
{ "is_sarcastic": "Yes/No", "sarcasm_mechanism": "[Selection from Phase 2.1]", "dominant_modality": "[Selection from Phase 2.2]", "sarcasm_description": "[Derived from Phase 1 Step 3 - Inference]", "sarcastic_intent": "[Selection from Phase 2.3]", "confidence": "High/Medium/Low", "is_challenging": [], "Yes..." or empty list "challenging_reason": null, "reasoning": "[Summary of your Premises -> Incongruity -> Inference chain]" }