M^{3}UCD: A Multi-task Multimodal Metaphor Understanding Challenge Dataset for LLMs

Tianlong Zheng[1,2,3], **Yating Yang**[1,2,3*], **Rui Dong**[1,2,3*], **Bo Ma**[1,2,3], **Lei Wang**[1,2,3], **Xi Zhou**[1,2,3], **Siru Miao**[1,2,3], **Turghun Osman**[1,2,3]

[1]Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011, China,
[2]University of Chinese Academy of Sciences, Beijing 100049, China,
[3]Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011, China
{yangyt, dongrui, mabo, wanglei, zhouxi, turghun}@ms.xjb.ac.cn, {zhengtianlong22, miaosiru23}@mails.ucas.ac.cn

Abstract

Understanding multimodal metaphors represents a crucial pathway for machines to comprehend human cognition. However, current research remains constrained by superficial dataset annotations, insufficient systematic evaluation of large language models, and fragmented task frameworks. To bridge these gaps, the paper proposes a systematic solution featuring: (I) We present the largest fine-grained **M**ulti-task **M**ultimodal **M**etaphor **U**nderstanding **C**hallenge **D**ataset (**M^{3}UCD**) built via multi-perspective collaborative annotation. It contains 15,345 samples, each annotated with 12 manual attribute labels. (II) Systematic benchmarking of LLMs' capacity boundaries in metaphor understanding. Evaluation results reveal the persistent challenges LLMs face in this domain while validating M^{3}UCD's effectiveness and potential. (III) A concise and unified multi-task baseline framework was developed and demonstrated its effectiveness in enhancing the metaphor understanding capabilities of MLLMs.

Datasets — https://github.com/DragonReed/M3UCD

Introduction

Metaphor, as a cognitive cornerstone of human thought expression, pervasively manifests in daily communication (Riggs 2024), artistic works (Dimova 2024), and multimodal scenarios (Jahameh and Zibin 2023). It operates through systematic mappings from source domains (vehicles) to target domains (tenors) (Lakoff and Johnson 2020, 2008), essentially enabling the concretization of abstract concepts via cross-domain conceptual projections. Traditional metaphor studies (Song et al. 2021; Zhang and Liu 2022; Tian et al. 2023; Reimann and Scheffler 2024) predominantly focus on unimodal textual paradigms, where linguistic symbols solely mediate conceptual mappings. However, the proliferation of multimodal metaphors constructed through the synergy of visual and linguistic symbols on digital platforms not only amplifies the rhetorical impact of information delivery but also introduces novel challenges for machines in comprehending human cognition.

Recent years have witnessed a paradigm shift in metaphor understanding research from unimodal to multimodal approaches. A series of multimodal metaphor detection datasets such as MultiMET (Zhang et al. 2021b), MET-Meme (Xu et al. 2022), and FigMemes (Liu et al. 2022) have been successively constructed, accompanied by corresponding cross-modal metaphor detection models (Su et al. 2021; He et al. 2024b,a; Zhang et al. 2024, 2025b; Yang et al. 2025a), achieving substantial progress in both data construction and algorithmic innovation. Nevertheless, current research still confronts multiple bottlenecks:

(1) Datasets with limited annotation dimensions, lacking deep semantic orientation and pragmatic figurativeness annotations related to metaphors, which hinders fine-grained metaphor analysis.

(2) Absence of systematic evaluation on LLMs' performance in multimodal metaphor understanding (MMU) tasks, coupled with the lack of fine-grained evaluation benchmarks for assessing LLMs in this domain.

(3) Imperfect task construction for MMU, where current research remains confined to basic detection tasks without delving into semantic orientation and pragmatic figurativeness in metaphorical mapping processes, nor establishing standardized paradigms for critical subtasks like target/-source domain extraction (tenor/vehicle extraction).

To tackle the aforementioned issues, this paper introduces a **M**ulti-task **M**ultimodal **M**etaphor **U**nderstanding **C**hallenge **D**ataset[1] (**M^{3}UCD**), with sample illustrations presented in Figure 1. We systematically evaluate the performance of LLMs on fine-grained MMU, while extending

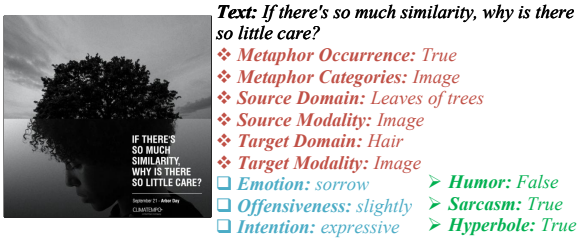

Figure 1: Samples from the M^{3}UCD with 12 manual labels.

*Corresponding Author

[1]***Disclaimer***: M^{3}UCD contains samples with potentially sensitive content (e.g., sarcasm, offensiveness, fake news, cultural references).

Metaphor Dataset	*Sample Size (%Metaphor)*	*Modality*	*Data Source*	*Annotation*	*Language*
TroFi [D1]	3,737 (43.5%)	Text	Wall Street Journal Corpus	Metaphor	English
VUA All [D2]	10,488 (11.6%)	Text	VU Amsterdam Metaphor Corpus	Metaphor	English
MOH-X [D3]	647 (48.7%)	Text	WordNet	Metaphor	English
CMDAG [D4]	27,989 (100%)	Text	Chinese literary sources	Metaphor	Chinese
META-ZH [D5]	5,491 (92.8%)	Text	Conference on Computational Linguistics	Metaphor	Chinese
VMCD [D6]	2,115 (33.3%)	Video	Adv	Metaphor	English
Ring That Bell [D7]	27 (6%)	Video & Audio	Social media	Metaphor	English
MetaCLUE [D8]	5,061 (100%)	Image	Adv	Metaphor	English
MultiMET [D9]	10,437 (58%)	Text, Image	Social media, Adv	Metaphor, Semantic	English
MET-Meme [D10]	10,045 (34%)	Text, Image	Social media	Metaphor, Semantic	Chinese, English
MultiCMET [D11]	13,820 (45.6%)	Text, Image	Search engine, Adv	Metaphor	Chinese
FigMemes [D12]	5,141 (20.4%)	Text, Image	4chan	Figurative	English
CM3D [D13]	6,108 (100%)	Text, Image	Search engine, Adv	Metaphor	Chinese
CII-Bench [D14]	698 (80.5%)	Text, Image	Illustration web	Rhetorical	Chinese
MultiMM [D15]	8461 (56.4%)	Text, Image	Adv	Metaphor, Semantic	Chinese, English
EmoMeta [D16]	5000 (100%)	Text, Image	Search engine, Adv	Metaphor, Semantic	Chinese
M^{3}UCD (Ours)	***15,345 (61.9%)***	***Text, Image***	***Social media, Search engine, Adv, Artwork***	***Metaphor, Semantic, Figurative***	***Chinese, English***

Table 1: A comparison between the M^{3}UCD and existing metaphor datasets. Dataset references: [D1] (Birke and Sarkar 2006, 2007); [D2] (Steen et al. 2010); [D3] (Mohammad, Shutova, and Turney 2016); [D4] (Shao et al. 2024); [D5] (Zheng et al. 2025b); [D6] (Rajakumar Kalarani, Bhattacharyya, and Shekhar 2024); [D7] (Alnajjar, Hämäläinen, and Zhang 2022); [D8] (Akula et al. 2023); [D9] (Zhang et al. 2021b); [D10] (Xu et al. 2022); [D11] (Zhang et al. 2023); [D12] (Liu et al. 2022); [D13] (Zhang et al. 2025c); [D14] (Zhang et al. 2025a); [D15] (Yang et al. 2025b); [D16] (Lu et al. 2025).

the research scope to metaphor-rich semantic orientation (including emotion analysis, offensiveness detection, and intention recognition), pragmatic figurativeness analysis (including humor, sarcasm, and hyperbole detection) and generative tasks such as tenor/vehicle extraction. Furthermore, we developed a concise and unified multi-task baseline that effectively enhances the metaphor understanding capabilities of multimodal large language models (MLLMs). The principal contributions are threefold:

(1) ***Dataset***: We present the largest and most richly annotated fine-grained multi-task MMU dataset, as illustrated in Table 1, featuring manual annotations across multiple dimensions: metaphor occurrence/categories, semantic orientation, pragmatic figurativeness, and tenors/vehicles. Rigorous quality control through annotation guidelines and inter-annotator agreement analysis ensures dataset reliability.

(2) ***LLMs Benchmarking***: We conduct systematic evaluation of LLMs' capabilities on fine-grained metaphor understanding. Benchmark results across various LLMs demonstrate M^{3}UCD's potential and generalizability, establishing valuable references for future MMU research.

(3) ***Task and Framework Formulation***: Beyond extending semantic orientation and pragmatic figurativeness analysis tasks inherent to metaphor-rich, we formalize the novel task of tenor/vehicle extraction in MMU and introduce a unified Multi-task Collaborative Learning Framework (MCLF) to enhance the metaphor understanding of MLLMs.

M^{3}UCD Overview

Data Collection and Filter

The construction of M^{3}UCD rigorously adheres to multiple principles including multi-source collection, privacy-first governance, and quality control protocols, aiming to establish a high-quality benchmark for systematically evaluating MMU capabilities of LLMs across diverse tasks. Publicly available data were collected from social media platforms (e.g., rednote), advertising repositories (e.g., ZCOOL), search engines (e.g., Bing, Baidu), and artistic competition archives (e.g., National Public Service Advertising Competition). To ensure textual accuracy and image quality, WeChat optical character recognition (OCR) was employed to extract textual elements from images for multimodal source text integration, complemented by manual correction of erroneous/repetitive texts and removal of blurred images. Throughout this process, strict privacy preservation protocols were enforced, with real-time anonymization of personally identifiable information (PII) including user IDs, pseudonyms, and geolocation data, while abstaining from persistent storage of any private user data. The statistical profile of M^{3}UCD is summarized in Table 2, with the data split into Train, Valid, and Test sets at a ratio of 7:1:2.

Item	*Train*	*Valid*	*Test*	*Total*
Metaphorical Samples	6,612	953	1,937	9,502
Literal Samples	4,129	582	1,132	5,843
Total Samples	10,741	1,535	3,069	15,345
Metaphorical Words	123,179	17,980	36,155	177,314
Literal Words	81,487	11,757	22,514	115,758
Total Words	204,666	29,737	58,669	293,072
Metaphorical Avg Words	18.63	18.87	18.67	18.66
Literal Avg Words	19.74	20.20	19.89	19.81
Total Avg Words	19.05	19.37	19.12	19.10

Table 2: The statistical profile of M^{3}UCD.

Data Annotation and Quality Control

To ensure the quality of dataset annotation, we implemented a rigorous annotation protocol. Prior to annotation, all annotators (carefully selected Natural Language Processing graduate students) underwent systematic training and assessment on metaphor cognitive theory and multimodal annota-

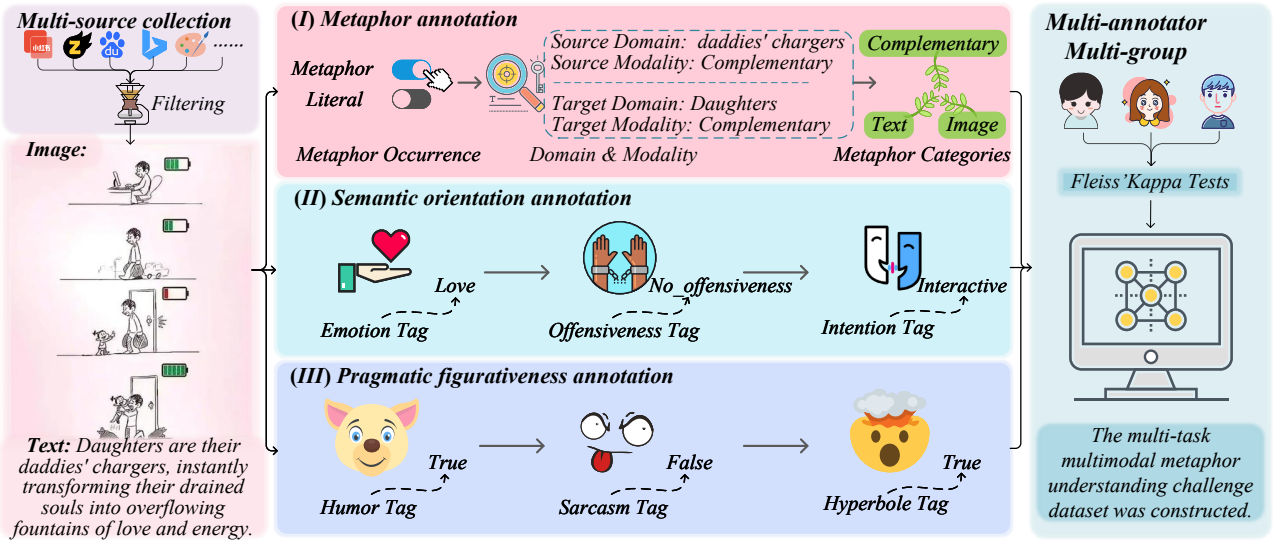

Figure 2: The annotation pipeline with the M^{3}UCD.

tion guidelines. The annotation process employed a multi-annotator, multi-group approach with kappa score, k, in Fleiss' Kappa Tests (Fleiss 1971) to evaluate quality control. As illustrated in Figure 2, the annotation pipeline comprises three core components: (I) Metaphor annotation (metaphor occurrence, source/target domain & modality, metaphor categories); (II) Semantic orientation annotation (emotion analysis, offensiveness level, intention recognition); (III) Pragmatic figurativeness annotation (humor, sarcasm, hyperbole detection). Comprehensive annotation guidelines (see Supplementary Material in Datasets URL) were established with strict definitions and selection criteria for each annotation category, including detailed explanations and exemplars. During pilot annotation, we iteratively refined the guidelines to ensure clarity and comprehensiveness before full-scale deployment. Discrepancies in annotations were resolved through multiple rounds of adjudication discussions, with majority voting employed when consensus remained unattainable, thereby enhancing annotation congruity and accuracy. All annotators received fair market-rate compensation throughout the process. The Fleiss' Kappa scores presented in Table 3 ($k \in [0.695, 0.965]$) demonstrate satisfactory inter-group agreement, indicating reliable annotation procedures.

Item	*Hyper-bole*	*Sar-casm*	*Hu-mor*	*Inten-tion*	*Offen-siveness*	*Emo-tion*	*Meta-phor*
Group 1&2	0.841	0.934	0.794	0.729	0.870	0.795	0.859
Group 2&3	0.881	0.964	0.871	0.767	0.882	0.965	0.845
Group 1&3	0.876	0.934	0.783	0.695	0.867	0.761	0.847

Table 3: Fleiss' Kappa score(k) on different annotation tasks.

Dataset Analysis

The statistical distribution of multi-task labels in M^{3}UCD is visualized in Figure 3, revealing the label distribution patterns across different tasks in MMU.

Metaphor Analysis To further investigate the distributional characteristics of multimodal metaphors, we conducted a meticulous analysis of textual and visual modality functions in metaphor understanding. As shown in Figure 3(a), the distribution of metaphor categories, source domain modality, and target domain modality exhibits two key patterns: (I) complementary metaphors (58.36%), where both modalities collaboratively contribute, dominate the distribution; (II) visually dominant (26.67%) and textually dominant (14.98%) metaphors follow in prevalence. This hierarchy suggests MMU relies on synergistic interactions between textual and visual information. Notably, significant discrepancies exist in modality distributions between source and target domains: visual modality dominates in the source domain (54.34%), while complementary patterns remain predominant in the target domain (42.83%). This distribution pattern reflects inherent cognitive preferences in human metaphor expression - a propensity for employing visual modality to convey vehicles, whereas understanding tenors requires multimodal information integration.

Semantic Orientation Analysis The metaphor-rich semantic orientation analysis comprises three core components: emotion analysis, offensiveness detection, and intention recognition. As shown in Figure 3(b), the semantic orientation distribution of M^{3}UCD is imbalanced. Specifically, in the emotion dimension, negative emotions (Sorrow 3.94%, Fear 5.97%, Hate 4.21%) are less prevalent than positive and neutral emotions; in the intent dimension, Expressive (27.03%) and Entertaining (37.45%) intents dominate;

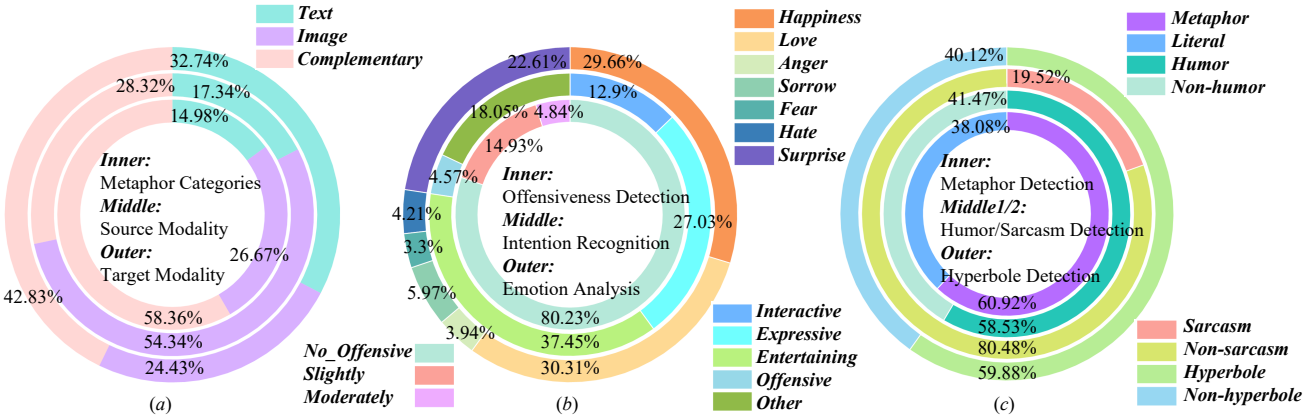

Figure 3: The distribution of labels for M^{3}UCD in the ring diagram.

and in the offensiveness dimension, No_offensive samples account for the majority (80.23%). This pattern mainly reflects the effective content moderation mechanisms in online communities and is considered a faithful representation of the real world online (Wang et al. 2023), indirectly validating M^{3}UCD as a robust simulation of practical scenarios.

Pragmatic Figurativeness Analysis The metaphor-rich pragmatic figurativeness analysis primarily encompasses humor, sarcasm, and hyperbole detection. As shown in Figure 3(c), the sarcasm distribution in the M^{3}UCD pragmatic figurativeness resembles the offensiveness dimension. Similarly to the metaphor dimension, Humor (58.53%) and Hyperbole (59.88%) categories have higher proportions than Non_humor (41.47%) and Non_hyperbole (40.12%). This correlational pattern suggests the co-occurrence tendency between metaphor expressions and humor/hyperbole figurativeness, indicating that metaphors often compound multiple pragmatic functions during information transmission. These findings provide empirical support for developing a multimodal collaborative learning framework.

Task and Framework Formulation

Task Formulation

For the systematic evaluation of metaphor understanding and reasoning capabilities in LLMs, our research establishes 12 metaphor-related tasks through M^{3}UCD, primarily categorized into three types: metaphor reasoning, metaphor-rich semantic orientation, and metaphor-rich pragmatic figurativeness tasks.

The metaphor reasoning tasks comprise six subtasks: metaphor detection (2-class), source/target domain detection (3-class), metaphor classification (3-class), and tenor/vehicle extraction. Notably, the extraction of tenor/vehicle pairs adopts a generative approach evaluated using BERTScore (Zhang et al. 2019). The metaphor-rich semantic orientation tasks involve three subtasks: emotion analysis (7-class), offensiveness detection (3-class), and intention recognition (5-class), designed to assess LLMs' understanding of metaphorical semantics. The metaphor-rich pragmatic figurativeness tasks consist of three binary classification tasks: humor detection, sarcasm detection, and hyperbole detection, focusing on evaluating LLMs' understanding of metaphors with specific pragmatic functions.

Multi-task Collaborative Learning Framework (MCLF) Furthermore, to enhance MLLMs' metaphor understanding, this study proposes a concise and unified baseline framework—MCLF, as illustrated in Figure 4.

Specifically, we define a multi-task learning setup comprising T metaphor-related tasks. Let $\mathcal{D}^{(t)} = \{(x_i, p_i^{(t)}, y_i^{(t)})\}_{i=1}^{N_t}$ be the data corresponding to task t, where $x_i = (I_i, T_i)$ denotes a multimodal input consisting of an image-text pair, $p_i^{(t)}$ is the natural language prompt guiding task t, and $y_i^{(t)}$ is its corresponding label or target. Each task uses a head $f_{\theta+\Delta\theta}^{(t)}$ on top of a frozen base MLLM with parameters θ; only the low-rank adaptation parameters $\Delta\theta$ (LoRA; (Hu et al. 2021)) are trainable.

The training objective aggregates task-wise expected losses with equal weights ($\lambda_t \equiv 1$):

$$\mathcal{L}_{\text{total}} = \sum_{t=1}^{T} \lambda_t \, \mathbb{E}_{\substack{(x_i, p_i^{(t)}, y_i^{(t)}) \\ \sim \mathcal{D}^{(t)}}} \left[\mathcal{L}^{(t)}\big(f_{\theta+\Delta\theta}^{(t)}(x_i, p_i^{(t)}), y_i^{(t)}\big) \right] \tag{1}$$

where $\mathcal{L}^{(t)}$ denotes the loss function associated with task t. We adopt per-task losses with a piecewise definition; classification tasks use cross-entropy and generation tasks use a masked autoregressive language-model loss.

The MCLF is trained by optimizing the total loss with respect to the LoRA parameters, i.e., $\min_{\Delta\theta} \mathcal{L}_{\text{total}}$, while keeping the base model parameters θ frozen. The framework supports a unified optimization strategy for both classification and generation-based MMU tasks across multi-task setting.

Experiments

Benchmarking research evaluates both advanced open-source and closed-source LLMs. The open-source LLMs (Parameter size: 1B-76B) primarily include

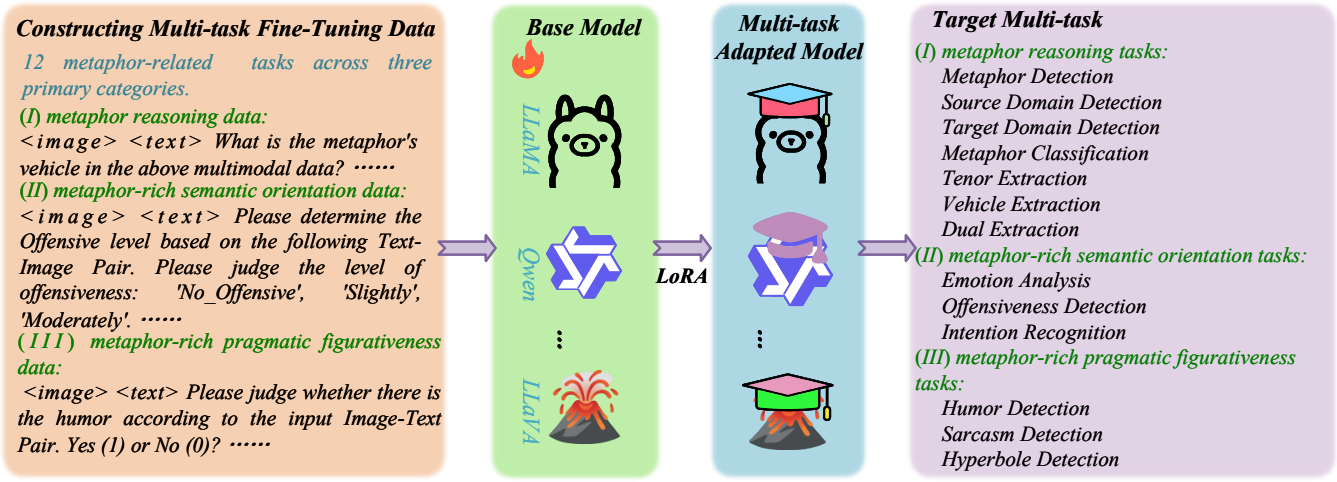

Figure 4: The multi-task collaborative learning framework.

InternVL2-1B/2B/4B/8B/26B/40B (Chen et al. 2024), ChatGLM3-6B-Base (Du et al. 2022), InternLM2.5-7B-Chat, LLaMA3-8B-Instruct, LLaMA3.1-8B-Instruct, Llama-3.2-11B-Vision-Instruct (AI@Meta 2024; Grattafiori et al. 2024), MiniCPM3-4B-Chat (Hu et al. 2024), LLaVA1.5-7B-Chat, LLaVA1.5-13B-Chat (Liu et al. 2024a), Qwen2-VL-7B-Instruct (Wang et al. 2024b), LLaVA-NeXT-7B-Chat, LLaVA-NeXT-13B-Chat (Li et al. 2024), Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct (Bai et al. 2025), Qwen2-VL-72B-Instruct (Wang et al. 2024b), and InternVL2-Llama3-76B (Chen et al. 2024). The closed-source LLMs encompass GPT-3.5-turbo (Achiam et al. 2023), GLM-4V-Plus (GLM et al. 2024), GPT-4o (Hurst et al. 2024), Gemini-1.5 Pro (Team et al. 2023), Claude-3.5-Sonnet (Anthropic 2024), o3-mini (OpenAI 2025b), and Claude-3-7-Sonnet-Thinking (Anthropic 2025), GPT-4.1 (OpenAI 2025a), and Gemini-2.5-flash-preview (Google DeepMind 2025). Furthermore, we have additionally conducted training and evaluation experiments on traditional Pre-trained Models (PTMs). These baseline models include Bi-LSTM (Yu et al. 2019), BERT (Devlin et al. 2018), ResNet50 (He et al. 2016), ViT (Dosovitskiy et al. 2020), MAE (He et al. 2022), DeiT (Touvron et al. 2021), SwinT (Liu et al. 2021), CLIP (Radford et al. 2021) and FLAVA (Singh et al. 2022) architectures. LoRA and MCLF were trained using the LLamaFactory (Zheng et al. 2024) framework under identical settings on 8× NVIDIA A100-80GB GPUs: 10 epochs, batch size 4, maximum sequence length 8192, learning rate 5e-5 with a cosine-annealing scheduler, maximum gradient norm 1, low-rank matrix dimension 8, and a weight-update ratio of 16.

Metaphor Understanding Explorations

To systematically delineate the capability boundaries of LLMs in MMU tasks, our metaphor understanding explorations encompasses the following research questions and experimental investigations:

Question I: To what extent are LLMs capable of metaphor detection and understanding (No LoRA and MCLF)?

Answer I: As evidenced in Tables 4 and 5, no one LLM demonstrates overall dominant superiority in MMU task. For metaphor detection, GPT-4o and o3-mini achieve optimal performance with accuracy and macro-F1 scores of 71.27% and 63.34% respectively. In tenor/vehicle extraction, GPT-4o and Gemini-1.5 Pro attain the highest BERTScore F1 values of 60.91% and 61.85%. Regarding source/target domain detection, although InternVL2-40B (Acc. 49.48%/44.84%), Gemini-1.5 Pro (macro-F1 42.24%), and o3-mini (macro-F1 36.62%) achieve the best performance, our analysis reveals that LLMs struggle to accurately identify the modality of tenors/vehicles. Additionally, the extraction task results in Table 5 demonstrate that dual extraction of both tenor and vehicle yields higher BERTScore F1 values compared to individual extraction, suggesting that LLMs tend to confuse these components during detection processes.

Question II: Do LoRA fine-tuning LLMs exhibit superior performance over traditional PTMs in MMU tasks?

Answer II: ***Not necessarily***. As evidenced by Table 4, LoRA fine-tuning LLMs exhibit superior performance compared to traditional PTMs son metaphor detection, metaphor-rich emotion analysis, offensiveness detection, and pragmatic figurativeness tasks. However, PTMs outperform LoRA fine-tuning LLMs in MMU tasks including metaphor-rich intention recognition, source/target domain detection, and metaphor classification.

Question III: Does LoRA fine-tuning lead to substantial improvements in downstream MMU tasks for MLLMs?

Answer III: ***Yes***. As quantitatively demonstrated in Tables 4 and 5, LoRA fine-tuning substantially enhances MLLMs' downstream metaphor understanding capabilities. In the tenor/vehicle extraction tasks, LoRA fine-tuning MLLMs with 3B-11B parameters outperform advanced closed-source MLLMs. This performance superiority ex-

Type	Model	Metaphor-rich Semantic Orientation Tasks						Metaphor-rich Pragmatic Figurativeness Tasks						Metaphor Reasoning Tasks (Detection Subtasks)									
		EA		OD		IR		HumorD		SD		HyperD		MD		SDD		TDD		MC			
		Acc.	*F1.*	*Acc.*	*F1.*	*Acc.*	*F1.*	*Acc.*	*F1.*	*Acc.*	*F1.*	*Acc.*	*F1.*	*Acc.*	*F1.*	*Acc.*	*F1.*	*Acc.*	*F1.*	*Acc.*	*F1.*		
	Random	14.04	11.30	33.52	25.39	19.35	17.33	50.85	50.79	50.98	44.99	50.29	49.89	50.49	49.56	30.65	29.15	31.79	31.59	34.98	32.42		
PTMs	Bi-LSTM	48.53	41.97	82.80	47.67	53.03	46.04	72.64	70.09	87.88	77.60	76.29	74.91	74.14	69.83	63.73	56.38	53.35	52.57	65.17	58.14		
	BERT	51.07	43.92	84.63	59.46	57.46	51.34	76.38	75.60	88.57	80.79	79.09	77.68	78.47	76.99	68.63	61.28	58.88	57.87	67.49	63.81		
	ResNet50	40.39	28.94	75.34	48.43	40.26	34.77	69.32	67.79	80.46	71.18	77.49	76.15	69.77	67.08	53.30	47.85	47.01	44.92	56.50	47.23		
	ViT	37.79	20.10	81.14	42.15	41.40	29.11	70.55	67.07	80.91	68.69	74.40	73.89	70.68	68.38	57.07	45.90	45.36	42.76	58.31	49.33		
	MAE	37.39	24.96	70.09	43.52	42.21	28.82	69.54	65.27	80.62	66.79	73.71	72.79	70.29	65.56	56.86	48.67	46.39	43.12	60.63	43.84		
	DeiT	41.34	30.97	80.26	51.65	45.18	35.06	70.98	70.25	84.63	68.52	74.04	73.81	72.31	69.16	55.37	49.00	50.00	47.89	60.22	49.95		
	SwinT	40.20	26.19	80.94	48.33	44.07	32.78	71.73	70.42	82.54	73.14	76.09	75.42	70.52	69.74	60.58	49.60	50.05	43.19	62.28	49.73		
	CLIP	53.49	44.69	84.98	60.45	59.77	54.20	76.78	75.97	89.64	82.78	79.58	78.66	77.20	75.29	66.77	62.44	58.93	58.77	67.65	63.15		
	FLAVA	38.53	23.11	82.67	43.10	46.22	29.01	67.59	66.24	83.94	69.38	70.68	68.26	69.87	66.75	56.19	36.65	47.47	42.63	59.55	24.88		
LLMs	ChatGLM3-6B-Base	30.39	11.57	73.88	35.67	16.19	12.83	60.00	52.64	47.46	44.64	55.57	50.32	61.92	46.05	20.59	14.49	32.30	19.85	17.65	14.16		
	LLaMA3-8B-Instruct	22.38	15.97	69.54	40.20	28.66	11.67	60.33	54.84	54.69	49.79	53.97	52.10	60.52	50.81	21.21	20.02	34.93	28.26	33.28	25.86		
	InternLM2.5-7B-Chat	35.21	19.95	79.28	41.50	26.84	21.06	57.23	56.80	67.04	56.29	53.45	53.02	58.66	54.28	18.68	12.08	32.97	18.74	15.94	10.02		
	LLaMA3.1-8B-Instruct	30.62	10.59	71.07	33.73	14.95	7.83	58.89	42.36	21.89	20.48	58.60	41.93	61.76	42.91	19.56	14.60	33.38	21.78	18.94	13.64		
	MiniCPM3-4B-Chat	31.43	15.84	76.51	45.38	27.23	23.90	49.38	48.26	78.93	60.43	48.11	45.44	57.39	54.89	18.68	11.55	32.97	17.43	15.58	9.48		
	GPT-3.5-turbo	37.07	27.57	70.29	39.98	24.20	21.68	57.17	56.95	63.19	52.09	48.73	45.55	59.74	57.44	18.47	11.52	33.13	17.53	15.38	8.96		
MLLMs	InternVL2-1B	32.77	19.09	79.22	34.48	14.36	11.13	38.79	31.12	81.34	44.85	40.26	28.20	36.84	33.02	27.30	14.80	25.28	13.53	25.44	14.12		
	InternVL2-2B	33.62	21.14	35.80	29.43	33.00	25.49	58.11	50.59	81.89	54.03	49.87	45.91	52.87	52.25	51.29	31.49	25.44	13.59	51.96	31.94		
	InternVL2-4B	35.21	23.43	79.54	45.47	33.71	22.50	59.35	59.31	23.29	22.01	58.96	49.47	63.42	39.88	54.08	26.18	40.20	22.30	58.46	27.26		
	LLaVA1.5-7B-Chat	17.98	11.78	46.58	32.11	29.09	19.09	58.37	41.11	36.22	35.83	52.44	50.50	60.65	43.07	18.68	14.75	35.29	24.09	20.95	15.86		
	Qwen2-VL-7B-Instruct	29.80	17.85	72.44	37.95	37.23	18.19	54.07	53.59	75.73	59.24	47.17	44.64	52.25	51.85	39.22	33.24	39.16	34.28	46.13	35.54		
	InternVL2-8B	38.86	27.41	80.72	30.40	39.25	28.60	61.24	61.20	76.86	69.17	57.72	56.94	63.39	50.69	52.53	30.81	26.16	17.27	38.85	34.53		
	Llama-3.2-11B-Vision-Instruct	31.60	21.09	75.02	47.48	35.90	23.58	51.99	51.18	64.95	56.85	54.10	54.03	63.16	39.29	35.91	31.78	37.41	33.01	54.54	30.05		
	InternVL2-26B	39.71	30.50	81.07	35.09	34.46	28.84	70.13	67.66	77.82	70.22	66.94	65.98	66.32	49.94	34.98	25.24	42.26	32.62	59.13	29.36		
	InternVL2-40B	39.61	30.78	80.23	43.61	38.60	27.15	70.39	69.78	71.07	66.33	60.00	58.93	65.96	48.93	49.48	34.37	44.84	31.73	58.00	35.51		
	Qwen2-VL-72B-Instruct	37.70	25.32	80.84	41.76	42.03	21.25	67.74	67.06	82.70	58.57	55.29	54.56	66.89	53.64	34.59	30.37	33.76	32.62	59.27	26.89		
	InternVL2-Llama3-76B	37.56	27.04	65.83	42.28	34.40	25.91	59.15	38.04	50.62	48.71	59.22	37.71	61.50	48.82	34.11	24.09	43.45	31.29	53.87	32.45		
	GLM-4V-Plus	35.86	20.36	82.21	47.80	34.04	26.36	61.43	59.73	75.73	67.81	55.31	55.21	63.13	38.78	19.31	12.39	32.58	25.69	24.88	19.26		
	GPT-4o	37.65	26.65	82.18	48.74	37.72	19.56	71.01	70.78	81.92	72.72	67.30	58.30	71.27	62.32	30.08	21.76	42.52	24.43	58.77	26.49		
	Gemini-1.5 Pro	39.54	29.44	81.04	35.52	41.14	21.16	64.43	64.21	84.50	77.33	64.85	55.97	66.94	51.42	48.56	42.24	39.63	35.44	59.34	28.29		
	Claude-3.5-Sonnet	37.92	28.26	80.78	38.48	40.00	25.54	63.55	63.43	83.81	76.97	66.35	64.92	65.57	47.13	47.24	39.50	40.22	35.27	56.93	36.23		
	o3-mini	40.29	29.71	80.72	32.79	36.35	26.26	58.86	57.30	83.13	58.65	59.74	59.64	67.65	63.34	46.96	41.97	39.63	36.62	55.26	41.44		
	Claude-3-7-Sonnet-Thinking	39.09	28.75	80.78	38.27	40.59	24.46	63.16	62.67	83.65	76.59	67.72	65.63	67.26	52.26	46.54	36.44	40.71	33.00	58.15	34.04		
	GPT-4.1	38.37	25.74	81.89	40.48	29.54	11.30	66.55	66.54	78.24	64.26	66.51	55.29	66.68	49.43	39.83	31.10	42.47	26.55	59.29	28.10		
	Gemini-2.5-flash-preview	40.52	26.25	80.55	30.81	37.56	23.17	68.01	67.97	83.68	67.05	72.41	70.35	68.93	57.53	39.32	37.63	42.11	33.24	60.06	31.73		
LLMs (LoRA)	ChatGLM3-6B-Base (LoRA)	42.28	35.53	84.14	57.92	53.00	46.22	68.14	66.44	88.27	80.34	70.78	69.01	76.06	74.12	55.57	45.93	43.29	40.63	55.01	44.57		
	LLaMA3-8B-Instruct (LoRA)	46.12	38.82	84.10	57.33	53.32	46.02	75.47	74.53	89.38	82.12	74.01	72.52	73.68	71.56	60.22	51.76	46.85	45.46	57.84	47.91		
	InternLM2.5-7B-Chat (LoRA)	42.38	35.02	83.65	52.51	48.34	39.26	67.30	65.46	88.83	81.07	70.85	69.47	70.20	67.75	45.30	42.40	45.30	42.40	54.54	44.31		
	LLaMA3.1-8B-Instruct (LoRA)	43.26	37.26	83.36	53.35	46.64	39.15	71.37	69.99	89.15	81.52	69.09	67.42	68.31	65.72	53.77	39.45	42.83	40.25	51.50	40.05		
	MiniCPM3-4B-Chat (LoRA)	39.84	32.71	82.90	51.51	46.03	38.57	68.11	65.54	87.98	79.72	69.41	67.49	67.52	63.75	52.63	41.32	42.47	39.85	54.23	41.78		
MLLMs (LoRA)	LLaVA1.5-7B-Chat (LoRA)	42.54	36.49	84.27	55.34	47.95	39.95	74.20	72.90	90.07	82.98	77.23	76.02	74.66	72.43	56.97	48.00	43.70	42.54	54.33	44.58		
	LLaVA-1.5-13B-Chat (LoRA)	53.42	46.61	86.51	63.80	60.07	54.26	80.39	79.69	91.17	85.59	80.94	80.24	79.15	77.48	62.80	56.13	52.68	51.79	61.40	56.06		
	LLaVA-NeXT-7B-Chat (LoRA)	45.44	38.97	84.30	56.14	49.25	44.19	78.70	77.87	91.04	85.02	79.51	78.40	76.25	74.27	57.12	49.58	43.55	41.67	59.34	51.99		
	LLaVA-NeXT-13B-Chat (LoRA)	54.01	47.49	86.22	64.70	57.65	52.35	80.23	79.46	91.37	85.79	81.82	81.11	80.26	78.72	64.04	57.60	51.75	51.05	61.82	55.74		
	Qwen2-VL-7B-Instruct (LoRA)	42.15	36.32	84.07	57.73	52.35	44.86	78.99	78.34	91.40	85.88	81.14	80.17	80.13	78.25	62.69	55.36	49.07	46.56	61.51	54.44		
	Qwen2.5-VL-3B-Instruct (LoRA)	43.55	36.65	84.01	57.29	51.82	45.00	76.91	75.99	90.85	85.17	82.31	81.59	79.67	78.20	64.96	59.00	53.51	52.57	62.95	56.72		
	Qwen2.5-VL-7B-Instruct (LoRA)	53.62	48.60	86.45	64.69	59.48	53.96	81.60	80.92	92.35	87.27	82.74	82.04	80.98	79.68	63.05	56.61	50.05	48.60	60.63	53.03		
	Llama-3.2-11B-Vision-Instruct (LoRA)	48.44	42.21	85.08	59.63	54.36	48.99	79.80	79.09	90.49	84.70	80.26	79.56	76.78	75.48	64.81	58.18	49.95	48.33	64.50	55.36		
MLLMs (MCLF)	LLaVA1.5-7B-Chat (MCLF)	54.14	47.57	83.45	61.47	58.01	52.22	80.42	79.60	90.78	84.93	80.36	79.55	78.31	76.60	68.01	62.51	54.02	53.70	66.15	61.61		
	LLaVA-1.5-13B-Chat (MCLF)	53.94	47.71	85.44	60.10	59.38	53.84	80.39	79.60	90.68	84.58	81.37	80.48	78.86	77.15	68.32	62.93	55.16	54.87	67.91	63.49		
	LLaVA-NeXT-7B-Chat (MCLF)	52.80	45.52	85.21	61.07	57.49	51.66	79.41	78.51	90.36	84.01	81.34	80.59	77.75	76.04	68.11	63.33	53.72	53.24	67.29	62.78		
	LLaVA-NeXT-13B-Chat (MCLF)	54.98	47.77	86.61	64.13	60.16	54.79	79.90	78.87	91.27	85.64	81.56	80.75	80.16	78.52	68.42	63.32	55.37	55.12	66.41	62.00		
	Qwen2-VL-7B-Instruct (MCLF)	54.36	48.29	86.74	63.99	58.86	52.63	81.63	80.90	92.28	87.30	83.49	82.73	80.94	79.45	68.73	64.73	56.55	55.92	66.20	62.64		
	Qwen2.5-VL-3B-Instruct (MCLF)	51.92	45.56	85.90	62.11	56.42	51.49	81.01	80.31	90.75	84.91	80.65	79.94	77.17	75.48	65.89	60.28	52.48	51.68	63.98	59.33		
	Qwen2.5-VL-7B-Instruct (MCLF)	53.97	46.08	86.48	63.99	59.22	53.31	82.28	81.55	91.79	86.70	82.57	81.78	80.07	78.50	69.09	64.45	56.09	55.34	66.31	61.96		
	Llama-3.2-11B-Vision-Instruct (MCLF)	54.63	48.09	86.25	62.84	59.64	53.09	80.72	79.83	91.76	86.36	81.56	80.80	79.45	77.99	67.08	61.74	59.49	58.60	68.73	63.87		
	Avg. Improvement of MCLF over LoRA	+5.95	+5.41	+0.65	+2.55	+4.53	+4.93	+1.87	+1.86	+0.12	+0.25	+0.87	+0.94	+0.60	+0.65	+5.90	+7.85	+6.08	+6.92	+5.81	+8.72		

Table 4: Experimental results (Part 1) of PTMs, LLMs, and MLLMs on metaphor-rich semantic orientation, metaphor-rich pragmatic figurativeness, and metaphor reasoning tasks (detection subtasks). Cyan means the second best. Purple means the best. Green indicates the average improvement of MCLF over LoRA. Task abbreviations are as follows: Emotion Analysis (EA), Offensiveness Detection (OD), Intention Recognition (IR), Humor Detection (HumorD), Sarcasm Detection (SD), Hyperbole Detection (HyperD), Metaphor Detection (MD), Source Domain Detection (SDD), Target Domain Detection (TDD), Metaphor Classification (MC). *macro-Prec.* and *macro-Rec.* are detailed in the Supplementary Material.

Model	Metaphor Reasoning Tasks (Extraction Subtasks)								
	Tenor Extraction			Vehicle Extraction			Dual Extraction		
	F1.	Prec.	Rec.	F1.	Prec.	Rec.	F1.	Prec.	Rec.
LLMs									
ChatGLM3-6B-Base	56.32	56.55	58.26	54.28	54.29	56.35	60.57	60.75	61.14
LLaMA3-8B-Instruct	50.06	46.55	55.56	51.63	49.50	55.26	56.80	54.48	59.76
InternLM2.5-7B-Chat	54.31	50.73	59.89	52.54	49.40	57.61	59.94	57.53	63.13
LLaMA3.1-8B-Instruct	45.48	39.31	55.90	45.73	39.85	55.79	53.53	48.76	59.93
MiniCPM3-4B-Chat	57.82	62.04	55.36	57.72	62.82	54.78	61.12	64.68	58.31
GPT-3.5-turbo	57.35	58.41	57.49	54.66	55.24	55.39	60.77	62.57	59.51
MLLMs									
InternVL2-1B	53.08	48.80	59.72	53.68	50.67	58.83	60.60	57.89	64.10
InternVL2-2B	55.80	53.36	60.18	55.00	53.72	58.14	61.38	61.09	62.31
InternVL2-4B	53.98	49.54	60.77	52.64	48.95	58.43	60.13	58.39	62.47
LLaVA1.5-7B-Chat	55.27	54.38	57.92	55.61	64.36	50.99	61.11	62.09	60.74
Qwen2-VL-7B-Instruct	59.40	61.78	58.68	58.77	63.12	56.54	63.57	66.65	61.32
InternVL2-8B	56.00	52.98	60.84	54.01	51.96	57.69	61.33	60.78	62.44
Llama-3.2-11B-Vision-Instruct	54.68	55.32	56.23	53.72	49.05	61.24	61.46	58.27	65.70
InternVL2-26B	57.02	54.50	61.27	55.85	54.09	59.22	62.92	62.24	64.23
InternVL2-40B	58.73	56.83	62.15	59.15	59.99	59.75	64.40	64.87	64.49
Qwen2-VL-72B-Instruct	60.19	61.61	60.14	59.01	60.56	59.00	66.44	67.63	65.84
InternVL2-Llama3-76B	57.54	57.71	59.50	54.48	55.63	55.37	61.17	62.69	60.47
GLM-4V-Plus	58.48	60.05	58.67	55.55	58.10	54.76	62.56	64.09	61.66
GPT-4o	60.91	62.49	60.70	60.05	62.77	58.81	66.30	68.28	64.92
Gemini-1.5 Pro	60.67	61.49	61.28	61.85	62.16	62.94	66.71	67.00	66.93
Claude-3.5-Sonnet	58.00	56.97	60.46	57.50	56.71	59.79	63.97	63.21	65.23
o3-mini	59.70	60.35	60.43	60.39	62.03	60.32	65.35	66.29	64.98
Claude-3-7-Sonnet-Thinking	57.78	57.18	59.72	58.22	57.87	60.05	64.25	63.88	65.09
GPT-4.1	58.83	58.57	60.40	60.14	60.53	61.14	66.36	66.61	66.63
Gemini-2.5-flash-preview	60.60	60.00	62.62	60.16	60.06	61.70	66.62	66.25	67.53
LLMs (LoRA)									
ChatGLM3-6B-Base	60.55	62.94	59.83	59.69	62.82	58.39	63.93	66.00	62.47
LLaMA3-8B-Instruct	61.31	64.26	60.26	60.14	63.49	58.80	64.52	66.77	62.97
InternLM2.5-7B-Chat	61.09	63.68	60.38	59.54	62.28	58.77	64.27	66.29	62.95
LLaMA3.1-8B-Instruct	60.47	62.98	59.79	59.32	62.40	58.27	64.03	66.03	62.67
MiniCPM3-4B-Chat	61.20	64.31	59.70	58.95	62.62	57.28	64.15	66.92	62.07
MLLMs (LoRA)									
LLaVA1.5-7B-Chat	59.83	62.47	58.96	59.35	62.65	58.05	64.08	66.11	62.67
LLaVA-1.5-13B-Chat	64.46	66.22	64.27	62.99	65.14	62.59	67.75	68.92	67.13
LLaVA-NeXT-7B-Chat	61.76	64.13	61.01	60.24	63.11	59.30	65.01	66.79	63.90
LLaVA-NeXT-13B-Chat	65.15	67.10	64.80	63.49	65.84	62.87	68.21	69.51	67.50
Qwen2-VL-7B-Instruct	65.14	67.66	64.34	64.62	68.08	63.06	68.62	70.94	67.03
Qwen2.5-VL-3B-Instruct	65.34	67.61	64.83	65.33	68.07	64.47	69.26	70.92	68.25
Qwen2.5-VL-7B-Instruct	65.83	68.22	65.12	65.83	69.01	64.50	69.63	71.69	68.24
Llama-3.2-11B-Vision-Instruct	64.14	66.78	63.26	64.53	67.58	63.37	68.32	70.12	67.19
MLLMs (MCLF)									
LLaVA1.5-7B-Chat	64.87	66.31	64.97	64.12	65.96	63.88	68.26	69.31	67.76
LLaVA-1.5-13B-Chat	66.05	67.34	66.18	64.75	66.15	64.89	69.06	69.88	68.80
LLaVA-NeXT-7B-Chat	64.59	66.22	64.51	64.23	66.02	64.02	68.34	69.36	67.87
LLaVA-NeXT-13B-Chat	66.73	68.00	66.88	64.90	66.35	64.90	69.49	70.19	69.49
Qwen2-VL-7B-Instruct	68.33	69.86	68.38	67.18	68.75	67.14	71.13	72.06	70.81
Qwen2.5-VL-3B-Instruct	66.23	67.98	66.12	65.66	67.78	65.23	69.86	71.12	69.22
Qwen2.5-VL-7B-Instruct	67.28	68.51	67.59	66.94	68.53	66.99	70.66	71.42	70.49
Llama-3.2-11B-Vision-Instruct	67.32	69.00	67.20	66.39	68.14	66.26	70.44	71.40	70.05
Avg. Improvement of MCLF over LoRA	+2.47	+1.63	+3.16	+2.22	+1.03	+3.14	+2.05	+1.22	+2.82

Table 5: Experimental results (Part 2) of LLMs and MLLMs on metaphor reasoning extraction tasks: tenor, vehicle, and dual extraction. Cyan means the second best. Purple means the best. Green indicates the average improvement of MCLF over LoRA. *macro-Prec.* and *macro-Rec.* are detailed in the Supplementary Material.

tends comprehensively to MMU tasks.

Task	Human		MLLMs 1		MLLMs 2		MLLMs 3	
	Acc.	F1.	Acc.	F1.	Acc.	F1.	Acc.	F1.
EA	77.00	75.43	37.00	23.22	42.00	27.09	39.00	24.20
OD	90.50	75.37	80.00	37.94	80.05	29.73	84.00	47.61
IR	77.50	75.85	47.50	29.16	35.50	20.90	33.00	12.86
HumorD	89.00	90.18	67.50	66.47	74.00	73.68	74.50	74.36
SD	93.50	81.69	87.50	80.28	83.00	63.61	80.50	65.04
HyperD	88.50	87.96	66.50	66.26	66.50	66.13	60.50	54.43
MD	93.00	93.97	64.50	52.98	66.00	57.37	62.50	48.96
SDD	79.31	77.73	40.50	34.07	41.00	39.53	39.50	32.76
TDD	68.97	69.66	38.00	27.83	42.50	30.45	46.00	28.83
MC	83.62	80.60	60.50	36.65	63.00	32.16	62.00	29.91
Tenor Extraction	–	75.51	–	57.91	–	61.98	–	59.68
Vehicle Extraction	–	76.58	–	56.78	–	61.71	–	59.99
Dual Extraction	–	76.27	–	63.76	–	67.29	–	66.63

Table 6: Comparison results between MLLMs and human baselines. *MLLMs 1/2/3* correspond to Claude-3-7-Sonnet-Thinking, Gemini-2.5-flash-preview, and GPT-4.1.

Question IV: Are LLMs capable of discerning both semantic orientation and pragmatic figurativeness in metaphors?

Answer IV: ***Not necessarily***. As evidenced in Table 4, LLMs demonstrate competent understanding of metaphor-laden offensive semantic orientation and sarcastic pragmatic figurativeness, yet exhibit limitations in handling other semantic and pragmatic tasks. This phenomenon is primarily attributed to the safety alignment process where the annotation and filtering of offensive/sarcastic content constitute core training objectives (Gong et al. 2025). Through exposure to massive violation cases, LLMs develop acute sensitivity to metaphorical expressions frequently employed as covert attack vectors for circumventing sensitive content detection mechanisms.

Question V: Does parameter scaling in LLMs lead to improved performance in MMU tasks?

Answer V: ***Not necessarily***. In general, parameter scaling of LLMs typically enhances both accuracy and macro-F1 in MMU tasks. The expanded parameter space strengthens the model's capacity to encode implicit semantic features of complex metaphors. However, this relationship is not strictly linear or universally applicable. For instance, as shown in Tables 4 and 5, the InternVL2-26B/40B mod-

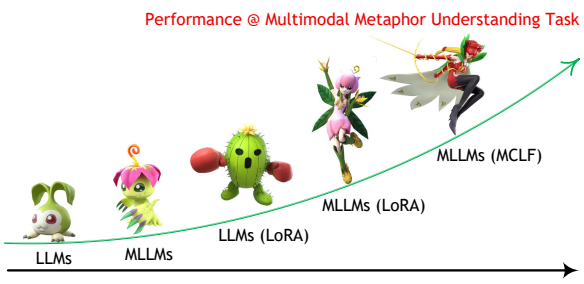

Figure 5: The performance @ multimodal metaphor understanding task.

Model	*binary-average*					
	HumorD			*SD*		
	Prec.	*Rec.*	*F1.*	*Prec.*	*Rec.*	*F1.*
InternVL2-2B	60.81	82.10	69.87	57.41	10.84	18.24
InternVL2-4B	71.04	52.81	60.58	19.16	96.85	32.00
LLaVA-NeXT-7B-Chat	59.22	95.10	72.99	19.31	76.22	30.81
Qwen2-VL-7B-Instruct	66.74	48.56	56.21	71.55	37.11	48.88
InternVL2-8B	77.12	49.01	59.93	42.81	72.38	53.80
Llama-3.2-11B-Vision-Instruct	69.88	33.09	44.92	28.42	58.04	38.16
InternVL2-26B	71.37	82.65	76.60	44.24	73.25	55.17
InternVL2-40B	76.85	71.48	74.07	38.26	90.03	53.70
Qwen2-VL-72B-Instruct	74.35	69.44	71.81	63.23	17.13	26.96
InternVL2-Llama3-76B	59.23	99.34	74.21	25.24	84.09	38.82
GLM-4V-Plus	66.76	69.33	68.02	41.13	70.10	51.84
GPT-4o	80.30	67.57	73.39	51.19	63.99	56.88
Gemini-1.5 Pro	85.77	47.80	61.39	56.22	75.87	64.58
Claude-3.5-Sonnet	82.24	48.95	61.37	54.55	78.67	64.42
o3-mini	91.45	33.59	49.13	69.85	16.61	26.84
Claude-3-7-Sonnet-Thinking	87.93	43.72	58.40	54.31	77.10	63.73
GPT-4.1	85.58	55.07	66.07	41.70	42.13	41.91
Gemini-2.5-flash-preview	86.39	54.52	66.85	61.20	33.92	43.64
LLaVA-NeXT-7B-Chat (LoRA)	81.47	82.82	82.14	77.15	73.78	75.42
LLaVA-NeXT-13B-Chat (LoRA)	82.73	84.14	83.43	76.70	77.10	76.90
Qwen2-VL-7B-Instruct (LoRA)	82.76	81.44	82.10	76.64	77.45	77.04
Qwen2.5-VL-7B-Instruct (LoRA)	84.12	84.91	84.52	80.14	78.32	79.22
Llama-3.2-11B-Vision-Instruct (LoRA)	82.86	83.04	82.95	72.95	77.80	75.30
LLaVA-NeXT-7B-Chat (MCLF)	81.46	84.42	82.91	74.47	73.43	73.94
LLaVA-NeXT-13B-Chat (MCLF)	81.05	86.18	83.53	76.39	76.92	76.66
Qwen2-VL-7B-Instruct (MCLF)	83.76	85.52	84.63	79.13	79.55	79.34
Qwen2.5-VL-7B-Instruct (MCLF)	84.08	86.40	85.23	76.76	80.24	78.46
Llama-3.2-11B-Vision-Instruct (MCLF)	82.28	85.90	84.05	78.13	77.45	77.79
	HyperD			*MD*		
	Prec.	*Rec.*	*F1.*	*Prec.*	*Rec.*	*F1.*
InternVL2-2B	57.14	64.39	60.55	66.58	50.88	57.68
InternVL2-4B	61.18	85.61	71.36	63.35	99.79	77.50
LLaVA-NeXT-7B-Chat	60.14	60.47	60.30	62.86	92.05	74.71
Qwen2-VL-7B-Instruct	34.13	32.52	33.30	68.28	21.59	32.81
InternVL2-8B	66.26	59.54	62.72	65.13	90.40	75.71
Llama-3.2-11B-Vision-Instruct	69.16	41.82	52.12	63.20	99.69	77.36
InternVL2-26B	73.39	70.07	71.69	65.65	97.83	78.57
InternVL2-40B	67.49	63.74	65.56	65.37	97.99	78.42
Qwen2-VL-72B-Instruct	77.30	35.66	48.81	66.63	95.30	78.43
InternVL2-Llama3-76B	59.57	98.75	74.31	64.21	88.13	74.29
GLM-4V-Plus	71.15	42.37	53.11	63.14	99.95	77.39
GPT-4o	65.59	95.20	77.67	70.09	95.05	80.68
Gemini-1.5 Pro	64.44	91.88	75.75	66.10	97.78	78.88
Claude-3.5-Sonnet	71.57	72.46	72.701	64.96	98.71	78.35
o3-mini	71.54	54.14	61.64	71.62	80.75	75.92
Claude-3-7-Sonnet-Thinking	71.15	77.32	74.11	66.35	97.68	79.02
GPT-4.1	64.53	97.60	77.69	65.64	99.07	78.96
Gemini-2.5-flash-preview	74.13	82.66	78.16	68.07	95.61	79.53
LLaVA-NeXT-7B-Chat (LoRA)	81.17	85.55	83.30	80.30	82.66	81.46
LLaVA-NeXT-13B-Chat (LoRA)	84.79	84.79	84.79	84.01	84.88	84.45
Qwen2-VL-7B-Instruct (LoRA)	82.77	86.42	84.56	82.65	86.74	84.64
Qwen2.5-VL-7B-Instruct (LoRA)	85.43	85.71	85.57	85.48	84.16	84.82
Llama-3.2-11B-Vision-Instruct (LoRA)	84.04	82.66	83.34	83.34	79.00	81.11
LLaVA-NeXT-7B-Chat (MCLF)	84.32	84.46	84.39	82.13	82.77	82.45
LLaVA-NeXT-13B-Chat (MCLF)	83.98	85.44	84.70	83.54	85.40	84.46
Qwen2-VL-7B-Instruct (MCLF)	85.31	87.40	86.35	84.53	85.45	84.99
Qwen2.5-VL-7B-Instruct (MCLF)	84.60	86.59	85.58	83.83	84.78	84.30
Llama-3.2-11B-Vision-Instruct (MCLF)	84.31	84.95	84.63	84.02	83.28	83.65

Table 7: Experimental results of MLLMs on metaphor-rich pragmatic figurativeness, and metaphor detection tasks. Purple means that MLLMs prefer the Rec. Orange means that MLLMs prefer the Prec.

els outperform InternVL2-Llama3-76B and GLM-V-Plus across multiple metaphor-related tasks. Moreover, the small-parameter models trained with MCLF surpass Claude-3-7-Sonnet-Thinking and so on. These results indicate that metaphor understanding in MLLMs is influenced not only by model scale but also critically depends on architecture, training methodology, and data.

Question VI: Can MLLMs trained with MCLF significantly enhance metaphor understanding?

Answer VI: ***Yes***. As shown in Tables 4, 5 and Figure 5, compared to LoRA fine-tuning, MLLMs trained with MCLF consistently achieve improved performance on MMU tasks. Furthermore, gains in Source/Target Domain Detection, Emotion Analysis, Intention Recognition, and Tenor/Vehicle Extraction surpass those in other surface-level detection tasks, indicating that MCLF effectively strengthens MLLMs' deep metaphor understanding capabilities.

Question VII: Is there a significant gap between MLLMs and humans in MMU tasks?

Answer VII: ***Yes***. Given the cost of human evaluation we randomly selected 200 test samples using stratified sampling and had multiple Natural Language Processing researchers independently annotate them. The lowest human performance is reported in Table 6. It can be observed that MLLMs achieve Acc. close to the human baseline only in offensiveness and sarcasm detection, while underperforming on all other MMU tasks.

Question VIII: Do MLLMs exhibit metric preferences in MMU tasks?

Answer VIII: ***Yes***. Taking the metaphor-rich pragmatic figurativeness and metaphor detection tasks as examples, we set the preference threshold as $\tau = |Prec. - Rec.| = 30$. The binary-average results in Table 7 demonstrate that MLLMs generally exhibit metric preferences, which vary across models and tasks. Furthermore, the experimental results show that MCLF not only enhances model performance but also effectively mitigates the metric biases of MLLMs.

Related Work

Multimodal Metaphor Dataset

Compared to textual metaphor research (Birke and Sarkar 2006, 2007; Steen et al. 2010; Mohammad, Shutova, and Turney 2016; Shao et al. 2024; Zheng et al. 2025b), the construction of multimodal metaphor datasets remains in its nascent stage. Xu et al. (2022) pioneered the first multimodal metaphor detection (MMD) dataset MET-Meme, followed by Zhang et al. (2021b) who subsequently established the currently largest English dataset MultiMET. However, existing multimodal metaphor datasets commonly suffer from three critical limitations: coarse-grained annotation schemes, insufficient diversity in data sources, and constrained scale of metaphorical instances. These shortcomings collectively impede the advancement of downstream tasks in multimodal metaphor understanding. To address these challenges, this study introduces M^3UCD, which features finer-grained metaphor annotation, broader data source coverage, and substantially enriched metaphorical samples.

Multimodal Metaphor Understanding

Current research in multimodal metaphor exhibits a paradigm shift from elementary detection to advanced understanding. Early studies primarily focused on MMD. Su et al. (2021) first proposed MMD based on distinguishing concreteness, pioneering the field of MMD. He et al. (2024b) developed a visual-enhanced multi-interactive cross-modal residual network that significantly improved inter-modal congruity and complementary fusion. Subsequently, He et al. (2024a) drew on linguistic metaphor recognition theories to detect metaphors by mapping features into different subspaces to capture conflicts. Researchers later attempted synergistic optimization of metaphor detection with other multimodal learning tasks. Zhang et al. (2021a) used facial expression features to analyze metaphoric emotions by examining changes in expressions of subjects exposed to multimodal metaphors, detecting metaphoric emotions. Wang et al. (2024a) introduced cross-modal attention and multi-interactive encoders to propose a metaphor-aware fine-grained meme understanding multimodal multi-task framework. Zhang et al. (2024) addressed metaphor alignment in multimodal sentiment recognition through conditional generative modeling of metaphorical analogies. Zheng et al. (2025c) proposed the intention-semantic incongruity perception network for MMD.

Recent investigations have progressively shifted focus toward profound metaphor understanding. Zhang et al. (2025c) established a chain-of-thought (CoT) based metaphor mapping recognition model simulating human cognitive processes for interpretable metaphor mapping. Zheng et al. (2025a) proposed a multi-granular multimodal clue fusion model that enhances both fine-grained metaphor feature extraction and semantic orientation analysis. Furthermore, preliminary attention has emerged regarding MLLMs' metaphor detection capabilities. Xu et al. (2024) first employed MLLMs with CoT prompting for metaphor detection, while Liu et al. (2024b) identified significant limitations in MLLMs' capacity for advanced semantic understanding and image detail capture, demonstrating that incorporating affective polarity cues in prompts substantially improves abstract image comprehension. However, existing studies lack systematic exploration of LLMs' performance in MMU, particularly in analyzing metaphor-rich semantic orientation and pragmatic figurativeness.

Conclusion

Our research established the largest fine-grained multi-task multimodal metaphor understanding dataset (M^3UCD) to date. This dataset comprises 15,345 multi-label annotation instances. Through meticulous annotation of deep metaphors, it substantially transcends conventional datasets limited to single-task metaphor detection. We conducted systematic evaluations of advanced PTMs, LLMs and MLLMs on this annotated content. Additionally, we developed a concise and unified multi-task collaborative learning framework (MCLF) for metaphor understanding. Experimental results reveal that M^3UCD's complexity presents significant challenges for MMU tasks, with even cutting-edge models like GPT-4o and Claude-3-7-Sonnet-Thinking struggling with metaphor understanding. MCLF effectively enhances MLLMs' metaphor understanding. This underscores the necessity for developing novel metaphor understanding paradigms. M^3UCD will be released under an access license agreement as a comprehensive multimodal benchmark to drive fundamental progress in LLMs' metaphor understanding capabilities.

Acknowledgments

We extend our gratitude to the annotators for their diligent work, as well as to the anonymous reviewers for their insightful feedback and constructive comments. This research was supported by the Natural Science Foundation of Xinjiang Uyghur Autonomous Region (Grant No. 2022D01D04), the Tianshan Talent Training Program (No. 2023TSYCCX0044, 2023TSYCCX0041, 2022TSYCLJ0046), the Natural Science Foundation of Xinjiang Uyghur Autonomous Region (No. 2022D01D81, 2023D01D17, 2024D01D29), the Outstanding Member Program of the Youth Innovation Promotion Association of Chinese Academy of Sciences (No. Y2023118), the Key Research and Development Program of Xinjiang Uyghur Autonomous Region (No. 2023B03024, 2024B03026), the Major Scientific and Technological Projects of Xinjiang Uygur Autonomous Region (No. 2023A01006).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 Model Card.

Akula, A. R.; Driscoll, B.; Narayana, P.; Changpinyo, S.; Jia, Z.; Damle, S.; Pruthi, G.; Basu, S.; Guibas, L.; Freeman, W. T.; et al. 2023. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23201–23211.

Alnajjar, K.; Hämäläinen, M.; and Zhang, S. 2022. Ring That Bell: A Corpus and Method for Multimodal Metaphor Detection in Videos. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, 24–33.

Anthropic. 2024. Introducing Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Birke, J.; and Sarkar, A. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European chapter of the association for computational linguistics*, 329–336.

Birke, J.; and Sarkar, A. 2007. Active learning for the identification of nonliteral language. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, 21–28.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dimova, P. 2024. *At the crossroads of the senses: The synaesthetic metaphor across the arts in European Modernism*. Penn State Press.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; Wang, H.; Sun, J.; Zhang, J.; Cheng, J.; Gui, J.; Tang, J.; Zhang, J.; Li, J.; Zhao, L.; Wu, L.; Zhong, L.; Liu, M.; Huang, M.; Zhang, P.; Zheng, Q.; Lu, R.; Duan, S.; Zhang, S.; Cao, S.; Yang, S.; Tam, W. L.; Zhao, W.; Liu, X.; Xia, X.; Zhang, X.; Gu, X.; Lv, X.; Liu, X.; Liu, X.; Yang, X.; Song, X.; Zhang, X.; An, Y.; Xu, Y.; Niu, Y.; Yang, Y.; Li, Y.; Bai, Y.; Dong, Y.; Qi, Z.; Wang, Z.; Yang, Z.; Du, Z.; Hou, Z.; and Wang, Z. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Gong, Y.; Ran, D.; He, X.; Cong, T.; Wang, A.; and Wang, X. 2025. Safety misalignment against large language models. In *Proceedings of the 2025 Annual Network and Distributed System Security Symposium (NDSS)*.

Google DeepMind. 2025. Gemini 2.5 Flash. https://deepmind.google/models/gemini/flash/.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, X.; Yu, L.; Tian, S.; Yang, Q.; and Long, J. 2024a. SC-Net: Multimodal metaphor detection using semantic conflicts. *Neurocomputing*, 594: 127825.

He, X.; Yu, L.; Tian, S.; Yang, Q.; Long, J.; and Wang, B. 2024b. VIEMF: Multimodal metaphor detection via visual information enhancement with multimodal fusion. *Information Processing & Management*, 61(3): 103652.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Hu, S.; Tu, Y.; Han, X.; He, C.; Cui, G.; Long, X.; Zheng, Z.; Fang, Y.; Huang, Y.; Zhao, W.; et al. 2024. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. *arXiv preprint arXiv:2404.06395*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jahameh, H.; and Zibin, A. 2023. The use of monomodal and multimodal metaphors in advertising Jordanian and American food products on Facebook: A comparative study. *Heliyon*, 9(5).

Lakoff, G.; and Johnson, M. 2008. *Metaphors we live by*. University of Chicago press.

Lakoff, G.; and Johnson, M. 2020. Conceptual metaphor in everyday language. In *Shaping entrepreneurship research*, 475–504. Routledge.

Li, B.; Zhang, K.; Zhang, H.; Guo, D.; Zhang, R.; Li, F.; Zhang, Y.; Liu, Z.; and Li, C. 2024. LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild.

Liu, C.; Geigle, G.; Krebs, R.; and Gurevych, I. 2022. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, 7069–7086.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, Z.; Fang, F.; Feng, X.; Du, X.; Zhang, C.; Wang, N.; Zhao, Q.; Fan, L.; GAN, C.; Lin, H.; et al. 2024b. Ii-bench: An image implication understanding benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 37: 46378–46480.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lu, X.; Liu, Y.; Zhang, D.; Wu, Z.; Ren, J.; and Xia, F. 2025. Emometa: A multimodal dataset for fine-grained emotion classification in chinese metaphors. In *Companion Proceedings of the ACM on Web Conference 2025*, 3080–3083.

Mohammad, S.; Shutova, E.; and Turney, P. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the fifth joint conference on lexical and computational semantics*, 23–33.

OpenAI. 2025a. Introducing GPT-4.1 in the API (2025). https://openai.com/index/gpt-4-1/.

OpenAI. 2025b. OpenAI o3-mini System Card (2025). https://openai.com/index/o3-mini-system-card/.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rajakumar Kalarani, A.; Bhattacharyya, P.; and Shekhar, S. 2024. Seeing the Unseen: Visual Metaphor Captioning for Videos. *arXiv e-prints*, arXiv–2406.

Reimann, S.; and Scheffler, T. 2024. When is a metaphor actually novel? annotating metaphor novelty in the context of automatic metaphor detection. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, 87–97.

Riggs, A. 2024. Verbal and visual communication in constructive news across cultures: A case study of a bilingual English-Spanish corpus with a focus on metaphor. *Language & Communication*, 96: 26–41.

Shao, Y.; Yao, X.; Qu, X.; Lin, C.; Wang, S.; Huang, S. W.; Zhang, G.; and Fu, J. 2024. CMDAG: a Chinese metaphor dataset with annotated grounds as cot for boosting metaphor generation. *arXiv preprint arXiv:2402.13145*.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15638–15650.

Song, W.; Zhou, S.; Fu, R.; Liu, T.; and Liu, L. 2021. Verb metaphor detection via contextual relation learning. 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)*, 4240–4251.

Steen, G. J.; Dorst, A. G.; Krennmayr, T.; Kaal, A. A.; and Herrmann, J. B. 2010. A method for linguistic metaphor identification.

Su, C.; Chen, W.; Fu, Z.; and Chen, Y. 2021. Multimodal metaphor detection based on distinguishing concreteness. *Neurocomputing*, 429: 166–173.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tian, Y.; Xu, N.; Mao, W.; and Zeng, D. 2023. Modeling conceptual attribute likeness and domain inconsistency for metaphor detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7736–7752.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Wang, B.; Huang, S.; Liang, B.; Tu, G.; Yang, M.; and Xu, R. 2024a. What do they "meme"? A metaphor-aware multi-modal multi-task framework for fine-grained meme understanding. *Knowledge-Based Systems*, 294: 111778.

Wang, H.; Wang, Y.; Song, X.; Zhou, B.; Zhao, X.; and Xie, F. 2023. Quantifying controversy from stance, sentiment, offensiveness and sarcasm: a fine-grained controversy intensity measurement framework on a Chinese dataset. *World Wide Web*, 26(5): 3607–3632.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024b. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Xu, B.; Li, T.; Zheng, J.; Naseriparsa, M.; Zhao, Z.; Lin, H.; and Xia, F. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 2887–2899.

Xu, Y.; Hua, Y.; Li, S.; and Wang, Z. 2024. Exploring Chain-of-Thought for Multi-modal Metaphor Detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 91–101.

Yang, Q.; Meng, H.; Yan, Y.; Guo, S.; and Wei, Q. 2025a. SFVE: visual information enhancement metaphor detection with multi-modal splitting fusion. *The Journal of Supercomputing*, 81(3): 467.

Yang, S.; Zhang, D.; Ren, J.; Xu, Z.; Zhang, X. J.; Song, Y.; Lin, H.; and Xia, F. 2025b. Cultural Bias Matters: A Cross-Cultural Benchmark Dataset and Sentiment-Enriched Model for Understanding Multimodal Metaphors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 26301–26317.

Yu, Y.; Si, X.; Hu, C.; and Zhang, J. 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*, 31(7): 1235–1270.

Zhang, C.; Feng, X.; Bai, Y.; Du, X.; Hou, J.; Deng, K.; Han, G.; Li, Q.; Wang, B.; Liu, J.; et al. 2025a. Can MLLMs Understand the Deep Implication Behind Chinese Images? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14369–14402.

Zhang, D.; Lu, X.; Zhuang, M.; Yang, S.; and Chen, H. 2025b. Multimodal metaphor recognition based on chain-of-cognition prompting. *Cognitive Systems Research*, 91: 101356.

Zhang, D.; Yin, S.; Yu, J.; Wu, Z.; Li, Z.; Xu, C.; Wang, X.; and Xia, F. 2025c. Towards Multimodal Metaphor Understanding: A Chinese Dataset and Model for Metaphor Mapping Identification. *arXiv preprint arXiv:2501.02434*.

Zhang, D.; Yu, J.; Jin, S.; Yang, L.; and Lin, H. 2023. Multicmet: A novel chinese benchmark for understanding multimodal metaphor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6141–6154.

Zhang, D.; Zhang, M.; Guo, T.; Peng, C.; Saikrishna, V.; and Xia, F. 2021a. In Your Face: Sentiment Analysis of Metaphor with Facial Expressive Features. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Zhang, D.; Zhang, M.; Zhang, H.; Yang, L.; and Lin, H. 2021b. MultiMET: A multimodal 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)*, 3214–3225.

Zhang, L.; Jin, L.; Xu, G.; Li, X.; Xu, C.; Wei, K.; Liu, N.; and Liu, H. 2024. CAMEL: capturing metaphorical alignment with context disentangling for multimodal emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 9341–9349.

Zhang, S.; and Liu, Y. 2022. Metaphor detection via linguistics enhanced Siamese network. In *Proceedings of the 29th international conference on computational linguistics*, 4149–4159.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zheng, L.; Fei, H.; Dai, T.; Peng, Z.; Li, F.; Ma, H.; Teng, C.; and Ji, D. 2025a. Multi-Granular Multimodal Clue Fusion for Meme Understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 26057–26065.

Zheng, T.; Dong, R.; Yang, Y.; Ma, B.; Wang, L.; and Zhou, X. 2025b. Metaphor detection model based on linguistic multi-incongruity. *Journal of Computer Applications*, 45(12).

Zheng, T.; Yang, Y.; Dong, R.; Ma, B.; Wang, L.; and Zhou, X. 2025c. ISIPN: Intention-Semantic Incongruity Perception Network for Multimodal Metaphor Detection. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 549–562. Springer.

Zheng, Y.; Zhang, R.; Zhang, J.; YeYanhan, Y.; and Luo, Z. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 400–410.