MetaGPT: A Large Vision-Language Model for Meme Metaphor Understanding

Bo Xu[1], Chenyuan Wang[1], Xinyu Chen[1], Hongfei Lin[1], Feng Xia[2*]

[1]Dalian University of Technology, China
[2]RMIT University, Australia
{boxu, hflin}@dlut.edu.cn, {chenyuanwang, 20212241068}@mail.dlut.edu.cn, f.xia@ieee.org

Abstract

Meme is an expressive medium that often conveys rich emotions and intentions. Recent studies have confirmed the critical role of metaphors in meme understanding. However, existing metaphor research heavily relies on manual annotations, and mainstream vision-language models (VLMs) still struggle with the recognition and comprehension of metaphors. To address these challenges, we introduce MetaGPT, the first vision-language model specifically designed for meme metaphor understanding. MetaGPT is capable of identifying and extracting metaphors in memes, and generating accurate meme interpretations. Furthermore, we construct a dedicated dataset for meme understanding, MUnd, which comprises approximately 32,000 high-quality question-answer (QA) pairs across three core tasks: metaphor detection, metaphor domain extraction, and meme interpretation. Based on MUnd, we further propose an evaluation benchmark for meme understanding and conduct a comprehensive assessment of existing VLMs. Experimental results reveal that current models still face challenges in metaphor comprehension, while MetaGPT consistently outperforms them across all tasks, highlighting its potential in advancing meme understanding.

1 Introduction

Memes have become a prevalent medium on social media platforms. Users can convey authentic emotions and thoughts solely through memes (Tanaka et al. 2022). This concise and efficient form of interaction makes meme interpretation crucial for capturing and analyzing users' true intentions and emotional expressions (Castaño Díaz 2013). Moreover, accurately understanding meme semantics is key to various downstream tasks, such as MemeMQA (Agarwal et al. 2024), harmful content detection (Lu et al. 2024b), and sentiment analysis (French 2017).

However, memes often contain implicit metaphorical content (Piata 2016), which limits the performance of existing models in meme understanding. For instance, in Figure 1(a), the source domains "what I say" and "what I think" are metaphorically mapped to the target domains—the visible and submerged parts of an iceberg, respectively. This multi-modal metaphor highlights the gap between spoken expression and true thoughts. Similarly, in Figure 1(b), "my phone" and "screenshots I'll never use" are metaphorically represented as an overloaded truck and a massive pile of cargo, humorously illustrating the storage burden caused by excessive unused screenshots. Such cross-modal metaphors pose challenges for existing models in meme understanding.

(a) What I say; What I think (b) Screenshots I'll never use; my phone

Figure 1: Examples of metaphorical memes.

Recently, researchers have investigated the capability of VLMs (Li et al. 2023; DeepMind 2022; Zhu et al. 2023; OpenAI 2024) in understanding metaphors. For example, Hwang and Shwartz (2023) provide metaphor annotations for memes and find that existing VLMs struggle to interpret metaphor-rich memes, significantly underperforming compared to humans. Akula et al. (2023) propose four types of visual metaphor tasks and conduct systematic evaluations, revealing that both visual and language models perform poorly. Saakyan et al. (2024) identify five distinct types of multimodal metaphor phenomena, further demonstrating the limitations of VLMs in generalizing from literal to metaphorical meanings.

The above studies indicate that existing VLMs are not effective in recognizing and understanding metaphors. Moreover, current metaphor annotations in memes still heavily rely on manual labeling. Therefore, how to automatically and accurately extract metaphors from memes remains a key challenge in this field.

To address the above challenges, we propose **MetaGPT**, the first vision-language model designed for **me**me me**ta**phor understanding. Specifically, we thoroughly de-

*Corresponding author

velop MetaGPT across three essential aspects: dataset, model, and benchmark. Firstly, we construct MUnd, a meme understanding dataset consisting of approximately 32,000 high-quality QA pairs. Next, we present the MetaGPT model, which incorporates a MetaClean module to enhance performance on both metaphor domain extraction and meme interpretation. Finally, we meticulously create a benchmark to comprehensively assess the meme understanding capabilities of existing VLMs. Notably, metaphor domain extraction is a newly introduced task that evaluates a model's ability to understand metaphorical mappings, which constitutes a novel aspect not covered in existing VLM benchmarks. Experimental results show that MetaGPT consistently outperforms state-of-the-art (SOTA) open-source VLMs, demonstrating its effectiveness in meme understanding. In summary, our contributions are summarized as follows:

- We propose MetaGPT, the first large vision-language model specifically designed for metaphor understanding and meme interpretation. MetaGPT effectively identifies and extracts complex cross-modal metaphors in memes and provides coherent meme interpretations.
- We construct a large-scale, high-quality dataset for meme understanding, named MUnd. MUnd comprises approximately 32,000 high-quality QA pairs across three core tasks, providing a reliable foundation for training and evaluating models in meme understanding.
- We define a new task, metaphor domain extraction, as a crucial step toward deeper meme understanding with VLMs. Furthermore, extensive experiments validate the strong potential of MetaGPT in this domain.

2 Related Work

2.1 Meme Understanding

Early research on meme understanding primarily focuses on harmful content detection. Kiela et al. (2020) introduce the Hateful Memes Challenge (HMC), a dataset for detecting hateful memes. Fersini et al. (2022) construct MAMI, a benchmark for misogynistic meme detection. Based on these datasets, various detection models (Suryawanshi et al. 2020; Lee et al. 2021; Pramanick et al. 2021; Cao et al. 2022) are proposed to improve the identification of harmful memes. However, due to the inherent complexity and abstraction of memes, traditional detection models still face significant challenges. To address these limitations, recent studies begin to explore interpretability as a means of enhancing detection performance. For example, Huang et al. (2024) leverage VLMs to generate semantic explanations of memes to support better detection.

In recent years, researchers have introduced multimodal metaphors to enhance meme understanding. Xu et al. (2022) construct a metaphor-rich meme dataset, MET-Meme. Hwang and Shwartz (2023) annotate memes with explanations and metaphorical content to improve model comprehension. The creation of these metaphor datasets advances research on multimodal metaphor detection. Xu et al. (2024) adopt a chain-of-thought (CoT) strategy for detecting multimodal metaphors. Zheng et al. (2025) integrate multimodal cues to improve metaphor detection performance. Zhang et al. (2023) explore the capability of understanding multimodal metaphors in Chinese memes. Despite progress in metaphor detection, research on metaphor understanding remains limited, particularly in the generation of source and target domains.

2.2 VLMs in the Domain of Memes

The multimodal reasoning capabilities of VLMs provide new directions for meme-related research. For harmful meme detection, Lin et al. (2023) employ lightweight fine-tuning of VLMs to assess the harmfulness of memes. Cao, Lee, and Jiang (2024) progressively enhance the understanding of different types of hateful memes through staged fine-tuning. Lin et al. (2024) propose a multimodal debate mechanism for interpretable detection of harmful memes. Rizwan et al. (2024) study the performance of VLMs under various prompting strategies and find that textual content in memes plays a key role in hate detection.

For meme understanding, Jha et al. (2024) leverage VLMs models to automatically generate intervention content aimed at mitigating the potential risks posed by harmful memes. Agarwal et al. (2024) introduce MemeMQA, a meme-based QA dataset that focuses on enhancing VLMs' interpretability of harmful memes. Cao et al. (2025) construct a large-scale dataset of locally-sourced memes from Singapore and fine-tune VLMs to improve their recognition of locally offensive content. While VLMs perform well in harmfulness detection, they struggle to comprehend and accurately interpret the overall semantics of memes.

3 Method

3.1 The MUnd Dataset

Existing meme datasets primarily focus on annotating visual content, while the metaphorical mappings embedded in memes remain underexplored. However, effective meme understanding requires not only surface-level visual analysis but also the ability to capture implicit mappings between source and target domains. To address this gap, we extend two high-quality metaphor-rich datasets, MET-Meme (Xu et al. 2022) and MEMECAP (Hwang and Shwartz 2023), to enhance VLMs' performance in meme understanding. The data construction pipeline of MUnd is illustrated in Figure 2. MUnd consists of three core QA tasks: metaphor detection, metaphor domain extraction, and meme interpretation.

Metaphor Detection. Before extracting metaphorical mappings, it is essential to determine whether a meme contains metaphorical content. To this end, we categorize each meme based on existing metaphor annotations to identify the presence or absence of metaphor. Given that metaphors in memes often span multiple modalities and are implicitly embedded in the overall semantics, we employ diverse prompts to guide the model's attention toward metaphor detection. This task serves as a foundation for the subsequent task of metaphor domain extraction.

Metaphor Domain Extraction. We are the first to propose Metaphor Domain Extraction, a crucial task for metaphor understanding that explicitly extracts the source

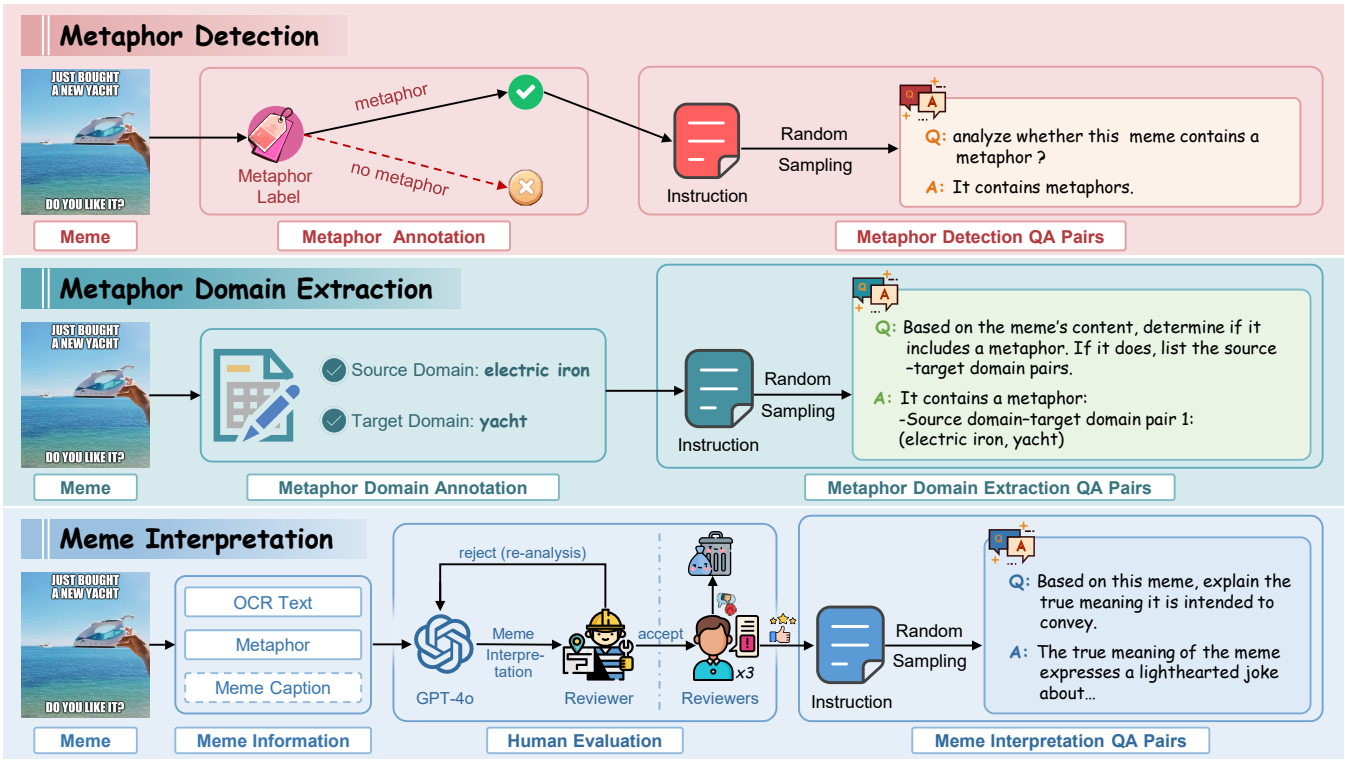

Figure 2: Data construction pipeline of MUnd.

and target domains in memes. Each metaphor is standardized as a source-target domain pair. As shown in the middle of Figure 2, the meme's metaphor is represented as (electric iron, yacht). This structured representation guides the model to explicitly learn metaphorical mappings in memes. For the MEMECAP dataset, we deduplicate the annotated metaphor domains and retain the top three most frequent source-target pairs in each meme.

Meme Interpretation. This task requires accurately describing the semantic intent conveyed by the meme. While MEMECAP provides human-written annotations, these descriptions are typically brief summaries focused on the meme poster's intended message, lacking detailed and in-depth semantic explanations. To address this limitation, we leverage GPT-4o to enrich and refine meme interpretations, aiming for more detailed and accurate semantic descriptions. We apply this process to all memes by first providing initial annotations to GPT-4o to enhance interpretation accuracy. A primary author reviews the output, identifies incorrect or ambiguous parts, and iteratively refines it with GPT-4o until the explanation is accepted. Each explanation is then independently validated by three additional authors. A sample is retained only if all reviewers agree on its correctness; otherwise, it is discarded. Ultimately, this process yields a 94% acceptance rate, demonstrating the reliability and consistency of our annotation pipeline.

Dataset Statistics. The dataset statistics of MUnd are illustrated in Figure 3. The data distribution across the three

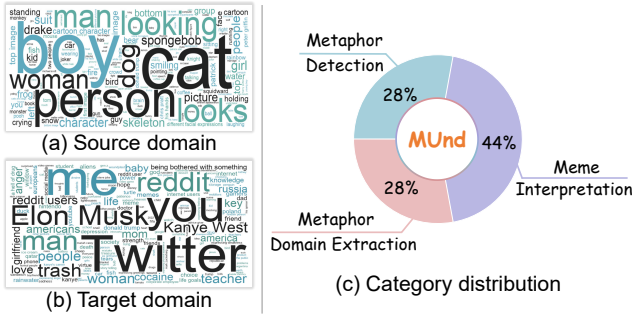

Figure 3: Data statistics of MUnd.

tasks is relatively balanced. To gain deeper insights into the semantic structure of metaphors, we further analyze the metaphors in MUnd. Specifically, for each meme containing metaphors, we extract source-target domain pairs and compute their frequency to observe overall distribution patterns. Figure 3(a) and Figure 3(b) show the word clouds of source and target domains, respectively. In the source domain, high-frequency entities such as "person", "cat", and "boy" are frequently observed and primarily correspond to characters depicted in meme images. This suggests that metaphors in memes are often grounded in visual entities to convey cultural or everyday scenarios. In contrast, the target domain contains frequent mentions of "Twitter", "Elon Musk", and "Reddit", which are embedded in online cultural and social contexts. This indicates that the metaphorical intent of

memes often centers on internet culture, public figures, or trending events. Such metaphor analysis not only reveals structural patterns of metaphors in memes but also provides a semantic foundation for the downstream metaphor domain extraction task.

3.2 MetaGPT

Overall Framework. The overall training framework of MetaGPT is illustrated in Figure 4. It consists of three main components: a visual encoder, a visual projector, and a large language model (LLM). Given a meme image X from the MUnd dataset, the visual encoder first extracts visual features $H = f_{vis}(X)$. These features are then projected into the language space via the visual projector, resulting in $V = f_{proj}(H)$. Finally, V is concatenated with the instruction prompt T_{inst} and fed into the LLM to generate the response:

$$Y = \Phi(V, T_{\text{inst}}) \tag{1}$$

where $\Phi(\cdot)$ denotes LLM. After training, MetaGPT is capable of identifying and extracting metaphors from memes and generating corresponding semantic interpretations, thereby enhancing the understanding and interpretability of memes for VLMs.

MetaClean Module. Inspired by the statistical analysis of metaphors, we observe that the source domain in memes typically corresponds to concrete visual entities, while the target domain relies more on the embedded textual semantics and socio-cultural context. However, meme text often overlaps with key visual regions (e.g., Figure 1(b)), which can hinder the model's ability to extract metaphors and comprehend the overall meme semantics.

To address this issue, we design a MetaClean module that processes the image both before and after visual encoder. Specifically, we first apply OpenCV inpainting (Navier–Stokes method) to the meme image x^v, yielding a cleaned version $x^c = Inpaint(x^v)$ with text removed. Then, x^c is symmetrically divided into four regions based on the image center:

$$x^{split} = \{x_i^c\}_{i=1}^4 \tag{2}$$

The above images collectively form the image set:

$$X^{img} = \{x^v, x^c, x^{split}\} \tag{3}$$

The image set X^{img} is first processed by the visual encoder f_{vis} to obtain the visual features $h^v = f_{vis}(x^v)$, $h^c = f_{vis}(x^c)$, and $h^{split} = f_{vis}(x^{split})$. We then apply adaptive pooling to the features of the text-removed image and its sub-regions, and concatenate them with the original image feature before feeding into the visual projector.

$$V = f_{proj}\left(Concat\left(h^v, AdaptPooling\left(h^c, \{h_i^c\}_{i=1}^4\right)\right)\right) \tag{4}$$

This mechanism preserves the original meme context while enabling the model to effectively capture information from key visual regions, thereby enhancing the consistency and robustness of metaphor and overall meme modeling.

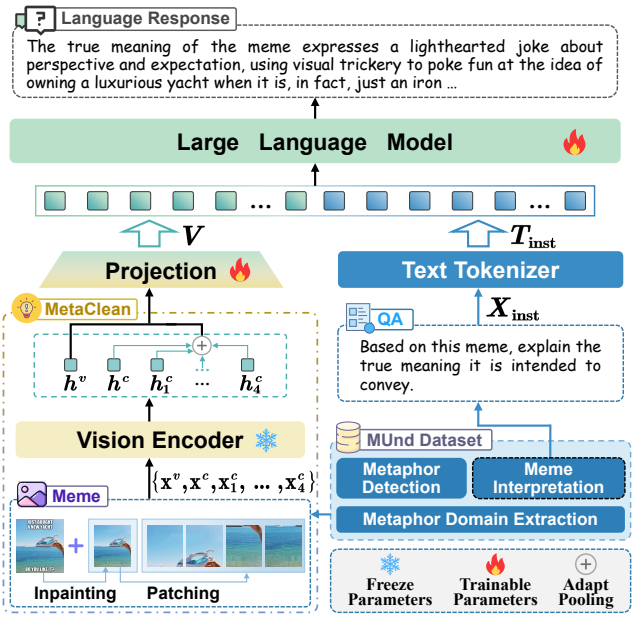

Figure 4: Training framework of MetaGPT.

In addition to the model design, we further conduct a series of exploratory experiments to investigate how different semantic elements in memes (e.g., shallow and mid-level visual features) affect the model's performance on metaphor understanding and meme interpretation. Detailed analyses are provided in Section 4.6.

4 Experiments

4.1 Experimental Setup

LLAVA-v1.5 (Liu et al. 2024) demonstrates strong visual understanding and extensive knowledge capabilities. Thus, we adopt LLAVA-v1.5-7B as the backbone and apply LoRA (Hu et al. 2021) for efficient fine-tuning. We use Vicuna-7B (Chiang et al. 2023) as the language model and CLIP-L/14 (Radford et al. 2021) as the visual encoder. Specifically, we extract the second-to-last hidden layer from CLIP-L/14 as the fine-grained visual representation. The input image is resized to 336×336 pixels, and the visual projector consists of a two-layer MLP. Our experiments are conducted on two RTX 4090 GPUs. Fine-tuning is completed within two training epochs. We use the AdamW optimizer with an initial learning rate of 2e-4, applying cosine annealing for learning rate scheduling and setting the first 3% of steps as the warm-up phase. For the LoRA configuration, the rank r is set to 128, α is set to 256, and the dropout rate is 0.1.

4.2 Baseline Models

We adopt a suite of SOTA open-source models with varying parameter scales as evaluation baselines. For 7B-scale models, we select LLaVA-v1.5 (Liu et al. 2024), MiniGPT-v2 (Chen et al. 2023), DeepSeek-VL (Lu et al. 2024a), and Qwen-VL-Chat (Bai et al. 2023). In addition, we further compare with four larger-scale models: Fuyu (Bavishi et al.

Models	Parameter	$\tau = 0.5$			$\tau = 0.6$			$\tau = 0.7$			$\tau = 0.8$		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
LLaVA-v1.5	7B	11.13	7.06	8.11	0.98	0.70	0.76	-	-	-	-	-	-
MiniGPT-v2	7B	5.01	2.06	2.77	0.72	0.34	0.43	0.18	0.04	0.07	-	-	-
DeepSeek-VL	7B	2.47	1.37	1.60	0.27	0.13	0.16	-	-	-	-	-	-
Qwen-VL-Chat	7B	8.97	5.19	6.06	1.43	0.85	1.03	0.18	0.18	0.18	-	-	-
MiniCPM-V2.6	8B	8.02	5.86	6.38	0.69	0.58	0.62	0.36	0.31	0.33	0.18	0.13	0.15
LLaVA-v1.6	13B	7.05	4.48	4.99	0.54	0.30	0.35	0.18	0.04	0.07	0.18	0.04	0.07
CogVLM	17B	2.15	0.75	1.09	0.18	0.04	0.07	-	-	-	-	-	-
MetaGPT(ours)	7B	**41.95**	**24.05**	**29.13**	**24.63**	**13.55**	**16.65**	**19.02**	**10.20**	**12.62**	**15.32**	**8.27**	**10.17**
△	-	↑30.82	↑16.99	↑21.02	↑23.20	↑12.70	↑15.62	↑18.66	↑9.89	↑12.29	↑15.14	↑8.14	↑10.02

Table 1: Results of metaphor domain extraction. We use BERTScore F1 to compute the similarity between the predicted source-target domain pairs and the references, under different thresholds $\tau \in \{0.5, 0.6, 0.7, 0.8\}$. A prediction is considered correct if the score exceeds the threshold. '-' denotes that the model fails to perform this task and achieves a score of zero.

Models	Parameter	P	R	F1
LLaVA-v1.5	7B	54.21	29.04	16.69
MiniGPT-v2	7B	66.24	36.92	32.09
DeepSeek-VL	7B	71.92	28.79	14.20
Qwen-VL-Chat	7B	70.16	39.55	35.62
Fuyu	8B	60.54	49.06	51.56
MiniCPM-V2.6	8B	69.51	67.83	68.53
LLAVA-v1.6	13B	58.55	30.16	19.19
CogVLM	17B	7.79	27.91	12.18
MetaGPT(ours)	7B	**88.35**	**86.98**	**87.33**
△	-	↑16.43	↑19.15	↑18.80

Table 2: Results of metaphor detection. The top-2 scores are marked in bold and underlined, respectively. △ indicates the performance gap between our method and the best baseline.

Models	Parameter	Human Evaluation		
		P	R	F1
LLaVA-v1.5	7B	1.50	2.50	1.83
MiniGPT-v2	7B	6.00	3.50	4.33
DeepSeek-VL	7B	-	-	-
Qwen-VL-Chat	7B	3.50	4.00	3.33
MiniCPM-V2.6	8B	1.50	2.00	1.67
LLaVA-v1.6	13B	1.50	3.00	1.92
CogVLM	17B	-	-	-
MetaGPT(ours)	7B	**48.17**	**41.50**	**42.63**
△	-	↑42.17	↑37.50	↑38.30

Table 3: Human evaluation on metaphor domain extraction.

2023), MiniCPM-V2.6 (Yao et al. 2024), LLaVA-v1.6 (Li et al. 2024), and CogVLM (Wang et al. 2023). These models possess strong multimodal understanding capabilities and serve as representative references for assessing the current landscape of meme understanding performance.

4.3 Benchmark Tasks

We design three tasks to assess model performance in the domain of meme understanding, and all evaluation samples are strictly excluded from the MUnd training data.

Metaphor Detection. This binary classification task aims to determine whether a given meme contains metaphorical content. We adopt the MET-Meme dataset as the evaluation benchmark and follow an 80/20 split for training and testing.

Metaphor Domain Extraction. For memes identified as metaphorical, this generative task requires the model to extract all corresponding source-target domain pairs. We use the MEMECAP test set for evaluation.

Meme Interpretation. This generative task evaluates the model's ability to infer and articulate the underlying intent conveyed by a meme. The MEMECAP test set is used as the benchmark dataset.

4.4 Evaluation Metrics

Metaphor Detection. Evaluated by weighted precision, recall, and F1-score.

Metaphor Domain Extraction. We evaluate performance using macro-averaged precision, recall, and F1-score. The metrics are computed as follows:

- True Positive (TP): The number of source-target domain pairs correctly predicted by the model.
- False Positive (FP): The number of incorrectly predicted source-target pairs.
- False Negative (FN): The number of gold source-target pairs that the model failed to identify.

To ensure fair evaluation, we first deduplicate the reference source–target domain pairs in the MEMECAP test set to remove redundant annotations. During evaluation, duplicate predictions generated by the model are also removed to prevent inflated counts. For each predicted pair, we calculate BERTScore, implemented with the *microsoft / deberta-xlarge-mnli* model, against all reference pairs and select the one with the highest similarity as the final match. If the score exceeds a predefined threshold τ, the prediction is considered correct; otherwise, it is marked as incorrect.

Meme Interpretation. We adopt BLEU, CHRF, ROUGE-L, and BERTScore as evaluation metrics. Instruction prompts follow the original MEMECAP setup.

Models	Parameter	N-gram Matching					Embedding-based			Average
		BLEU-1	BLEU-2	BLEU-4	ROUGE-L	CHRF	BERT-P	BERT-R	BERT-F	
LLAVA-v1.5	7B	15.33	9.37	4.56	17.49	37.78	48.57	63.97	54.60	31.46
MiniGPT-v2	7B	20.71	12.89	7.15	25.50	39.83	53.81	64.08	57.87	35.23
DeepSeek-VL	7B	6.98	4.15	1.35	8.82	26.51	39.22	61.21	47.20	24.43
Qwen-VL-Chat	7B	17.59	10.95	5.13	20.04	40.68	50.52	67.05	56.97	33.62
Fuyu	8B	19.13	9.82	4.48	16.80	22.99	58.88	57.59	57.28	30.87
MiniCPM-V2.6	8B	21.19	11.68	4.25	21.12	39.66	53.65	68.98	59.70	35.03
LLAVA-v1.6	13B	12.00	7.21	2.90	14.17	34.57	45.44	64.37	52.63	29.16
CogVLM	17B	51.96	36.57	20.41	42.12	44.93	70.72	71.53	70.38	51.08
MetaGPT(ours)	7B	**58.63**	**43.56**	**29.37**	**45.28**	**46.77**	**74.57**	**74.64**	**74.27**	**55.89**
+OIS	-	52.50	39.19	26.26	44.41	46.33	74.12	74.24	73.53	53.82
+SMV(1)	-	57.46	42.70	28.73	44.53	46.54	74.13	74.50	73.64	55.28
+SMV(2)	-	55.46	41.20	27.47	44.80	46.40	74.15	74.54	73.68	54.71
+SMV(3)	-	56.71	41.99	28.21	44.08	46.08	73.75	74.29	73.27	54.80
+IMT	-	51.14	37.40	25.93	40.92	41.40	72.33	70.82	70.93	51.36
△	-	**↑6.67**	**↑6.99**	**↑8.96**	**↑3.16**	**↑1.84**	**↑3.85**	**↑3.11**	**↑3.89**	**↑4.81**

Table 4: Comparison of performance and exploratory results on meme interpretation.

4.5 Results

Metaphor Detection. Results for metaphor detection are presented in Table 2. MetaGPT achieves the best performance on this task. Compared with the strongest baseline, MetaGPT outperforms MiniCPM-V2.6 by 18.8% in F1 score, demonstrating a significant advantage in cross-modal metaphor recognition. However, most existing VLMs underperform in this task, indicating their limited ability to comprehend complex metaphors embedded in memes. Moreover, we observe substantial performance gaps across models of different parameter scales. For example, despite their larger sizes, LLaVA-v1.6-13B and CogVLM-17B yield only 19.19% and 12.18% F1 scores, respectively, suggesting that increasing model size alone does not guarantee better metaphor detection capabilities. Even among models with similar parameter scales, performance can vary significantly—for instance, Qwen-VL-Chat and DeepSeek-VL differ by 21.42% in F1 score. These results highlight the instability of current VLMs on metaphor detection. Overall, the findings underscore that existing VLMs remain insufficient for this task, whereas MetaGPT provides a promising solution with markedly improved performance.

Metaphor Domain Extraction. Table 1 presents the results for the metaphor domain extraction task. Given that deberta-xlarge-mnli applies a relatively strict criterion for semantic similarity, we evaluate model performance under four different thresholds, $\tau \in \{0.5, 0.6, 0.7, 0.8\}$. MetaGPT achieves the best performance across all thresholds, indicating its strong ability to accurately identify metaphorical mappings within memes. In contrast, most baseline models exhibit an imbalance between precision and recall, which reflects the inherent complexity of this task. Specifically, existing models often struggle to comprehensively detect all source-target domain pairs, leading to notably low recall scores. Overall, metaphor domain extraction proves to be more challenging than metaphor detection. In addition to determining whether a meme contains metaphors, the model must also ensure that the quantity and quality of extracted domain pairs align with the reference annotations. These findings highlight the complexity and challenges of meme metaphor understanding.

Setting	MD			MDE		
	P	R	F1	P	R	F1
MetaGPT(Ours)	88.35	86.98	87.33	41.95	24.05	29.13
+OIS	86.90	85.10	85.77	39.26	22.77	27.46
+SMV(1)	87.96	86.11	86.54	40.13	23.50	28.33
+SMV(2)	87.79	86.48	86.83	40.85	23.53	28.63
+SMV(3)	88.31	86.56	86.96	40.73	22.94	28.03
+IMT	86.76	85.48	85.85	39.42	21.41	26.44

Table 5: Exploratory results on metaphor detection (MD) and metaphor domain extraction (MDE) at $\tau = 0.5$.

To more accurately assess the performance of VLMs on the metaphor domain extraction task, we conduct a human evaluation on 100 meme samples randomly selected from the MEMECAP test set, each containing no more than three metaphorical mappings. During evaluation, we manually counted the number of correctly predicted, incorrectly predicted, and missed metaphor pairs. The results are reported in Table 3. MetaGPT significantly outperformed its automatic evaluation performance under the threshold $\tau = 0.5$, demonstrating stronger recognition capabilities in memes with fewer metaphors. In contrast, baseline models such as LLaVA-v1.5 and Qwen-VL-Chat yielded noticeably lower F1 scores in the human evaluation than in the automatic setting ($\tau = 0.5$), indicating notable limitations of existing VLMs in this task. Qualitative analysis further revealed that most baseline models tend to rely heavily on shallow visual cues to infer metaphors, lacking a deeper cross-modal understanding of metaphorical mappings.

Meme Interpretation. Results on the meme interpretation task are presented in Table 4. MetaGPT achieves

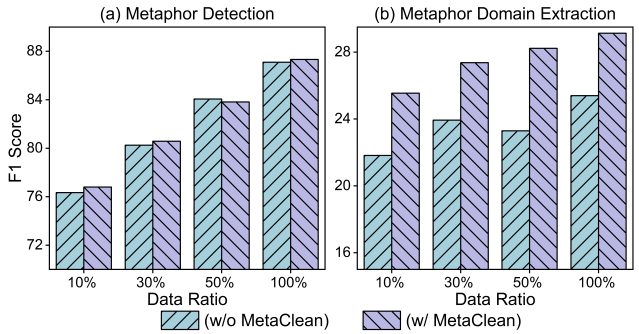

Figure 5: Ablation study on the MetaClean module under varying data ratios.

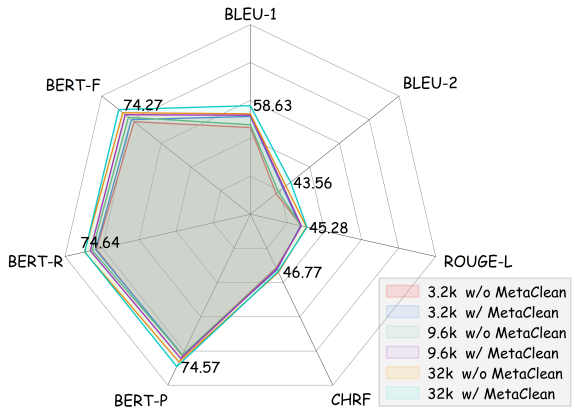

Figure 6: Ablation study on meme interpretation.

the best performance across all N-gram matching metrics. Specifically, it achieves a BLEU-1 score of 58.63%, outperforming CogVLM by 6.67%, and scores 45.28% on ROUGE-L and 46.77% on CHRF, indicating that MetaGPT's generated explanations better align with reference texts in lexical overlap and sequence structure. In embedding-based semantic matching, MetaGPT achieves the best performance, closely followed by CogVLM. Notably, although memes often contain complex metaphors and MetaGPT has significantly fewer parameters than CogVLM, it still captures the deep semantic intent behind memes with high accuracy, highlighting its strong cross-modal comprehension capabilities.

4.6 Exploratory Experiments on Meme Semantics

In this section, we conduct additional experiments to investigate how different semantic elements of memes affect the model's performance in meme understanding. The study comprises three components.

Original Image Segmentation vs. Inpainting (OIS). We replace the inpainted and patched images in the MetaClean module with directly segmented regions from the original meme image.

Integrating Shallow and Mid-level Visual Features (SMV). We design three experiments: (1) We concatenate the 8th-layer and 16th-layer features with the original and inpainted image features. (2) We aggregate features from layers 1–8 and 9–16 into shallow and mid-level representations, then concatenate them with the original and inpainted features. (3) We further adjust the concatenation order in (2) to: original, shallow, mid-level, and inpainted features.

Incorporating Meme Text (IMT). The OCR-extracted text is used as additional input to the model.

Exploratory results are presented in Tables 4 and 5. We observe that all exploratory settings lead to performance degradation compared to the MetaClean module. We believe this is related to the unique characteristics of meme understanding. For OIS, directly segmenting the original meme image may interfere with the overall understanding of the meme and fail to provide meaningful signals to the model. For SMV, the results show that introducing multi-level features does not bring improvements, suggesting that the high-level visual representations of the current model already cover the essential information. In addition, IMT performs worst, indicating that the additional meme text may introduce semantic conflicts and hinder the holistic understanding of memes.

4.7 Ablation Study

This section presents ablation studies on the training ratio of the MUnd dataset and the proposed MetaClean module. For each training ratio, we further evaluate model performance with and without MetaClean. Figure 5 illustrates the ablation results on the metaphor detection and metaphor domain extraction tasks ($\tau = 0.5$). Figure 6 reports the corresponding performance changes on the meme interpretation task. All improvements are statistically significant according to a paired two-tailed t-test ($p < 0.05$).

As the size of the MUnd dataset increases, performance consistently improves across all tasks, which aligns with the findings in (Liu et al. 2024). For the MetaClean module, we observe a substantial improvement in the metaphor domain extraction task, validating its effectiveness in enhancing the model's capability to identify metaphorical mappings. In contrast, its impact on the metaphor detection task is relatively minor, suggesting that this task may rely more on coarse-grained features. For the meme interpretation task, MetaClean helps strengthen the model's understanding of global meme semantics and improves its generation quality.

5 Conclusion and Future Work

In this paper, we propose MetaGPT, a vision-language model specifically designed for meme metaphor understanding. In addition, we construct a large-scale and high-quality dataset for meme understanding, named MUnd. Trained on MUnd, MetaGPT demonstrates capabilities in cross-modal metaphor reasoning and meme interpretation. Experimental results demonstrate the effectiveness of MetaGPT in meme understanding and highlight a viable path for multimodal metaphor research. In future work, we plan to expand MUnd to cover more diverse meme expressions and enhance model generalization.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62072073, 62106034, in part by the Fundamental Research Funds for the Central Universities under Grant DUT24ZD124, the Science and Technology Project of Liaoning Province, 2023JH2/101700363 and the Dalian Innovation Fund 2021JJ12GX016.

References

Agarwal, S.; Sharma, S.; Nakov, P.; and Chakraborty, T. 2024. MemeMQA: Multimodal Question Answering for Memes via Rationale-Based Inferencing. *arXiv preprint arXiv:2405.11215*.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23201–23211.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Bavishi, R.; Elsen, E.; Hawthorne, C.; Nye, M.; Odena, A.; Somani, A.; and Taşırlar, S. 2023. Introducing our Multimodal Models.

Cao, R.; Lee, R. K.-W.; Chong, W.-H.; and Jiang, J. 2022. Prompting for Multimodal Hateful Meme Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 321–332.

Cao, R.; Lee, R. K.-W.; and Jiang, J. 2024. Modularized Networks for Few-shot Hateful Meme Detection. In *Proceedings of the ACM Web Conference 2024*, 4575—-4584.

Cao, Y.; Wu, J.; Cheong, A. L. C.; Guanrong, B. S.; Lee, T. C. J.; and Chann, S. Z. S. 2025. Detecting Offensive Memes with Social Biases in Singapore Context Using Multimodal Large Language Models. *arXiv preprint arXiv:2502.18101*.

Castaño Díaz, C. 2013. Defining and characterizing the concept of Internet Meme. *Revista CES Psicología*, 6: 82–104.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

DeepMind, R. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198*.

Fersini, E.; Gasparini, F.; Rizzi, G.; Saibene, A.; Chulvi, B.; Rosso, P.; Lees, A.; and Sorensen, J. 2022. SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 533–549.

French, J. H. 2017. Image-based memes as sentiment predictors. In *2017 International Conference on Information Society*, 80–85.

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 preprint arXiv:2106.09685*.

Huang, J.; Lin, H.; Ziyan, L.; Luo, Z.; Chen, G.; and Ma, J. 2024. Towards Low-Resource Harmful Meme Detection with LMM Agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2269–2293.

Hwang, E.; and Shwartz, V. 2023. MemeCap: A Dataset for Captioning and Interpreting Memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1433–1445.

Jha, P.; Jain, R.; Mandal, K.; Chadha, A.; Saha, S.; and Bhattacharyya, P. 2024. MemeGuard: An LLM and VLM-based Framework for Advancing Content Moderation via Meme Intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 8084–8104.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2611–2624.

Lee, R. K.-W.; Cao, R.; Fan, Z.; Jiang, J.; and Chong, W.-H. 2021. Disentangling Hate in Online Memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, 5138—-5147.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *arXiv preprint arXiv:2407.07895*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*.

Lin, H.; Luo, Z.; Gao, W.; Ma, J.; Wang, B.; and Yang, R. 2024. Towards Explainable Harmful Meme Detection through Multimodal Debate between Large Language Models. In *Proceedings of the ACM Web Conference 2024*, 2359—-2370.

Lin, H.; Luo, Z.; Ma, J.; and Chen, L. 2023. Beneath the Surface: Unveiling Harmful Memes with Multimodal Reasoning Distilled from Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9114–9128.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26286–26296.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; Sun, Y.; Deng, C.; Xu, H.; Xie, Z.; and Ruan, C. 2024a. DeepSeek-VL: Towards Real-World Vision-Language Understanding. *arXiv preprint arXiv:2403.05525*.

Lu, J.; Xu, B.; Zhang, X.; Wang, H.; Zhu, H.; Zhang, D.; Yang, L.; and Lin, H. 2024b. Towards Comprehensive Detection of Chinese Harmful Memes. *arXiv preprint arXiv:2410.02378*.

OpenAI, R. 2024. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Piata, A. 2016. When metaphor becomes a joke: Metaphor journeys from political ads to internet memes. *Journal of Pragmatics*, 106: 39–56.

Pramanick, S.; Sharma, S.; Dimitrov, D.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4439–4455.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*.

Rizwan, N.; Bhaskar, P.; Das, M.; Majhi, S. S.; Saha, P.; and Mukherjee, A. 2024. Exploring the Limits of Zero Shot Vision Language Models for Hate Meme Detection: The Vulnerabilities and their Interpretations. *arXiv preprint arXiv:2402.12198*.

Saakyan, A.; Kulkarni, S.; Chakrabarty, T.; and Muresan, S. 2024. Understanding Figurative Meaning through Explainable Visual Entailment. *arXiv preprint arXiv:2405.01474*.

Suryawanshi, S.; Chakravarthi, B. R.; Arcan, M.; and Buitelaar, P. 2020. Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 32–41.

Tanaka, K.; Yamane, H.; Mori, Y.; Mukuta, Y.; and Harada, T. 2022. Learning to Evaluate Humor in Memes Based on the Incongruity Theory. In *In Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, 81–93.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; Xu, J.; Xu, B.; Li, J.; Dong, Y.; Ding, M.; and Tang, J. 2023. CogVLM: Visual Expert for Pretrained Language Models. *arXiv preprint arXiv:2311.03079*.

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*, 91–101.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; Chen, Q.; Zhou, H.; Zou, Z.; Zhang, H.; Hu, S.; Zheng, Z.; Zhou, J.; Cai, J.; Han, X.; Zeng, G.; Li, D.; Liu, Z.; and Sun, M. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*.

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.

Zheng, L.; Fei, H.; Dai, T.; Peng, Z.; Li, F.; Ma, H.; Teng, C.; and Ji, D. 2025. Multi-Granular Multimodal Clue Fusion for Meme Understanding. *arXiv preprint arXiv:2503.12560*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*.