

ImaRA: An Imaginative Frame Augmented Method for Low-Resource Multimodal Metaphor Detection and Explanation

Yuan Tian^{1,2}, Minzheng Wang^{2,1}, Nan Xu³, Wenji Mao^{1,2*}

¹State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences

²School of Artificial Intelligence, University of Chinese Academy of Sciences

³Beijing Wenge Technology Co., Ltd

{tianyuan2021, wangminzheng2023, wenji.mao}@ia.ac.cn; nan.xu@wenge.com

Abstract

Multimodal metaphor detection is an important and challenging task in multimedia computing, which aims to distinguish between metaphorical and literal multimodal expressions. Existing studies mainly utilize typical multimodal computing approaches for detection, neglecting the unique cross-domain and cross-modality characteristics underlying multimodal metaphor understanding. According to Conceptual Metaphor Theory (CMT), the inconsistency between source and target domains and their attribute similarity are essential to infer the intricate meanings implied in metaphors. In practice, the scarcity of annotated multimodal metaphorical contents in the real world brings additional difficulty to the detection task and further complicates the understanding of multimodal metaphors. To address the above challenges, in this paper, we propose a novel **Imaginative FFrame Augmented** (ImaRA) method for low-resource multimodal metaphor detection and explanation inspired by CMT. Specifically, we first identify *imaginative frame* as an associative structure to stimulate the imaginative thinking of multimodal metaphor detection and understanding. We then construct a cross-modal imagination dataset rich in multimodal metaphors and corresponding imaginative frames, and retrieve an augmented instance from this imagination dataset using imaginative frames mined from the input. This augmented instance serves as the demonstration exemplar to boost the metaphor reasoning ability of the multimodal large language model (MLLM) in low-resource multimodal scenarios. Experiments on two publicly available datasets show that our method consistently achieves robust results compared to MLLM-based methods for both multimodal metaphor detection and explanation in low-resource scenarios and meanwhile surpasses existing multimodal metaphor detection methods with full training data.

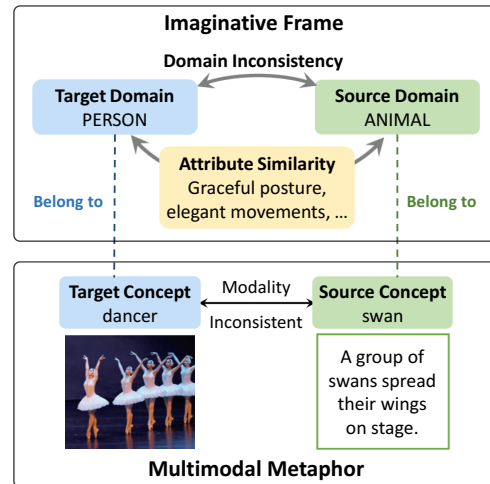


Figure 1: Illustration of the imaginative frame in the multimodal metaphor. The target concept *dancer* in the image and the source concept *swan* in the text encourage the imaginative thinking of corresponding domains and their relations. The higher-level imaginative frame of *domain inconsistency* and *attribute similarity* between source and target domains help infer the implicit meaning conveyed by this metaphor.

1 Introduction

Metaphor is an important figurative expression in literary works, advertisements, and online discussions. According to Merriam-Webster Dictionary, metaphor is “a figure of speech in which a word or phrase literally denoting one kind of object or idea is used in place of another to suggest a likeness or analogy between them”. Metaphor detection is a fundamental research topic in natural language processing (Li et al., 2013; Ge et al., 2023), which aims to distinguish between metaphorical and literal expressions. Traditional research on metaphor detection focuses on textual metaphors (Li et al., 2013; Ge et al., 2022; Tian et al., 2023). With the rapid development of social media, multimodal metaphors are widely used to implicitly convey users’ meanings and emotions online. Mul-

*Corresponding author

timodal metaphor detection, which aims to determine whether an image-text pair is metaphorical or literal, has attracted increasing attention in recent research (He et al., 2024; Xu et al., 2024). It can also benefit multiple multimodal applications in domains such as multimodal emotion recognition (Zhang et al., 2024) and hateful meme detection (Wang et al., 2024).

Existing studies on multimodal metaphor detection primarily focus on fusing visual and textual representations (Xu et al., 2022; He et al., 2024). Other studies incorporate additional lexicon knowledge (Zhang et al., 2023), implicit knowledge from other tasks (Wang et al., 2024) or commonsense knowledge distilled from a multimodal large language model (MLLM) (Xu et al., 2024) to enhance the performance. Despite the success of current methods, they mainly take the approaches typical in multimodal computing for multimodal metaphor detection, ignoring the unique cross-domain and cross-modality research challenges underlying multimodal metaphor detection and understanding. Unlike the aligned image-text pairs commonly used in multimodal computing, multimodal metaphors often utilize partially inconsistent yet implicitly related image and text contents to convey complicated meanings. Taking the multimodal metaphor in Figure 1 as an example, the dancer in the image and the swan in the text are seemingly inconsistent yet share similar characteristics.

To explain the cognitive mechanism underlying the characteristics of metaphor, Lakoff and Johnson (1980) introduced Conceptual Metaphor Theory (CMT), suggesting that a metaphor implies inconsistent source and target domains and their association of similar attributes in human cognition (Lakoff and Johnson, 1980). For a multimodal metaphor, a unique characteristic is that the source and target concepts are represented exclusively or predominantly in different modalities (Forceville and Urios-Aparisi, 2009). As illustrated in Figure 1, for understanding multimodal metaphors, the *imaginative thinking* is essential to conceive of source and target domains from source and target concepts as well as their associations in *domain inconsistency* and *attribute similarity* (Ricoeur, 1978).

Moreover, another critical challenge is the low-resource issue due to the scarcity of annotated multimodal metaphorical data compared to literal ones in practical applications. Existing research has not addressed the challenging issue inherent in the characteristics of multimodal metaphor detection and

understanding. In practice, the low-resource issue is also ignored by current research. Both issues require a deeper understanding of metaphor reasoning at the semantic and cognitive levels to bridge the cross-domain and cross-modality gaps, and develop the computational construct of associative imagination based on this understanding to help alleviate the low-resource situations.

To tackle the above issues, in this paper, we propose a novel **Imaginative FRame Augmented** (ImaRA) method for low-resource multimodal metaphor detection and explanation inspired by CMT. We first identify the imaginative frame as the computational construct of imaginative thinking for multimodal metaphor understanding. We then introduce an approach to construct a cross-modal imagination dataset rich in multimodal metaphors along with corresponding imaginative frames, and retrieve an augmented instance from this imagination dataset using imaginative frames mined from the multimodal input as the associative structure. The augmented multimodal metaphor instance is used as the demonstration exemplar, stimulating the imaginative thinking to bridge the cross-domain and cross-modality gaps and further boosting MLLM’s metaphor reasoning ability in low-resource scenarios. The main contributions of our work are summarized as follows:

- Based on the implications of conceptual metaphor theory, we identify the imaginative frame as an associative structure to bridge the cross-domain and cross-modality gaps inherent in multimodal metaphor understanding, and construct a cross-modal imagination dataset to enrich multimodal metaphor resource with the associated imaginative frames.
- To tackle the challenges in low-resource multimodal metaphor detection and explanation, we propose a novel imaginative frame augmented method, which mines imaginative frames from multimodal input and retrieves the augmented instance from cross-modal imagination dataset to stimulate MLLM’s imaginative thinking ability.
- Extensive experiments verify the effectiveness of our method for robust multimodal metaphor detection and explanation in low-resource settings, and also verify that our method surpasses previous SOTA multimodal metaphor detection methods with full training data.

2 Related Work

Traditional metaphor detection mainly identifies the metaphorical information in texts (Tsvetkov et al., 2014; Ge et al., 2022; Tian et al., 2024a). With the prevalence of multimodal data in social media, many people often use metaphors to express their thoughts and emotions within multimodal messages. Although some studies (Shutova et al., 2016; Kehat and Pustejovsky, 2020; Su et al., 2021) have incorporated visual features to enhance textual metaphor detection, they struggle with multimodal metaphor detection that requires a deep understanding of cross-modal relations.

To advance research on **multimodal metaphor detection**, some researchers have constructed multimodal metaphor datasets from social media platforms and advertisement resources (Zhang et al., 2021; Xu et al., 2022; Zhang et al., 2023). After that, He et al. (2024) propose a multi-interactive cross-modal residual network to improve the iterative information fusion between modalities. In addition, Wang et al. (2024) adopt inter-modality attention to capture the metaphorical features between image and text, and meanwhile exploits a multi-task framework to boost the performance. The work by Xu et al. (2024) achieves SOTA results via distilling commonsense knowledge from MLLMs with a chain-of-thought method to improve the pretrained model’s ability on multimodal metaphor detection. Despite the success of current studies, they ignore multimodal metaphor detection in **low-resource settings**. Moreover, none of them have addressed another important task of **multimodal metaphor explanation**, which aims to explain the underlying meanings conveyed by multimodal metaphors.

Multimodal metaphor explanation has not been explored in previous studies, though several works have focused on building datasets for cartoon joke understanding (Hessel et al., 2023) and figurative meme captioning (Hwang and Schwartz, 2023). In addition, Saakyan et al. (2024) propose a task of explainable visual entailment and provide a dataset containing memes employing figures of speech and paired captions, requiring models to determine whether the paired caption explains the meme’s figurative meaning. However, these datasets only contain image caption information to explain the figurative meanings.

Conceptual metaphor theory (CMT) (Lakoff and Johnson, 1980) is the most influential work in

metaphor research, widely accepted and utilized in computational metaphor research (Stowe et al., 2021; Ge et al., 2022; Tian et al., 2024b). CMT provides a fundamental basis for metaphor explanation going beyond merely descriptive captions. However, existing research only employs CMT for the analysis of textual metaphors, neglecting its potential for assisting the detection and explanation of multimodal metaphors. Thus, in this paper, we take advantage of the imaginative frame that contains the source and target domains along with their relations of the *domain inconsistency* and *attribute similarity* implied by CMT, and develop a computational method to mitigate the low-resource issue in multimodal metaphor detection and explanation.

3 Problem Definition

Formally, $\mathcal{D}_{tr} = \{(v_k, t_k), l_k, (src_k, tgt_k)\}_{k=1}^{N_{tr}}$ denotes the training dataset with N_{tr} instances, where (v_k, t_k) is an image-text pair and l_k is the label (metaphorical/literal) for the k -th instance. For metaphorical instances, the source and target concepts (src_k, tgt_k) are provided, while for literal instances, they are labeled as *None*. The test dataset is $\mathcal{D}_{te} = \{(v_k, t_k), l_k, (src_k, tgt_k)\}_{k=1}^{N_{te}}$. The goal of multimodal metaphor detection and explanation is to predict the label of each image-text pair and the source and target concepts for metaphorical instances in \mathcal{D}_{te} by training a model on \mathcal{D}_{tr} .

4 Proposed Method

We propose a novel imaginative frame augmented method ImaRA, which can detect multimodal metaphors and generate source and target concepts as explanations for identified metaphors in low-resource scenarios. Figure 2 shows its overview, containing four components: (1) *Cross-Modal Imagination Data Construction*, which constructs a cross-modal imagination dataset rich in multimodal metaphors and corresponding imaginative frames; (2) *Cross-Modal Imaginative Frame Mining*, which mines the imaginative frames in image-text input; (3) *Imaginative Frame Augmented Retrieval*, which uses the mined imaginative frames in image-text input as queries to retrieve the augmented instance from cross-modal imagination dataset; and (4) *Instruction Fine-Tuning*, which uses the augmented instance and the input to fine-tune an MLLM for multimodal metaphor detection and explanation.

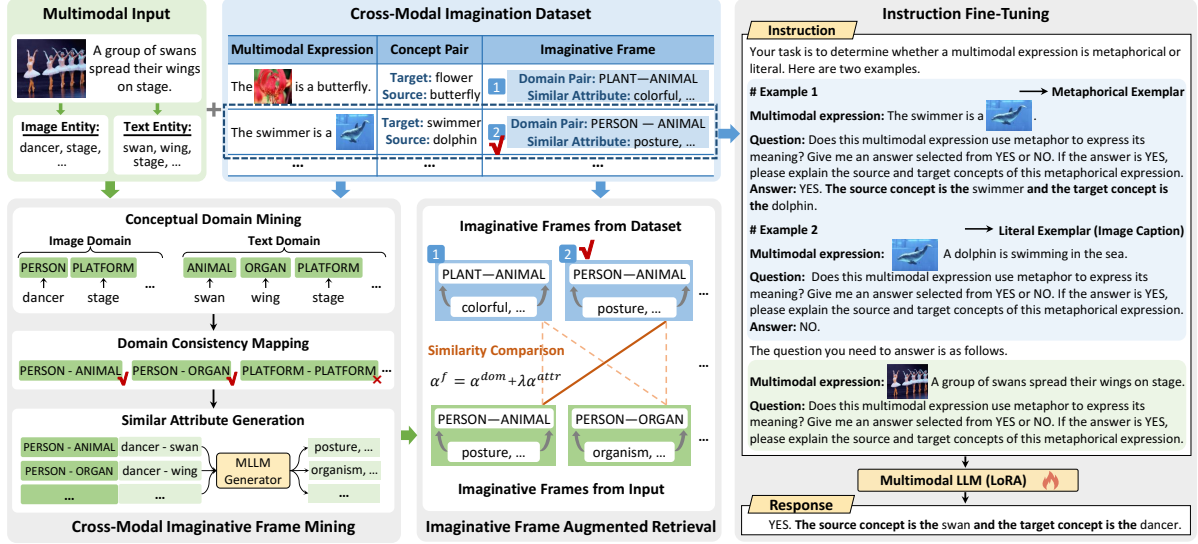


Figure 2: Overview of our proposed ImaRA for low-resource multimodal metaphor detection and explanation.

4.1 Cross-Modal Imagination Dataset

To mitigate the low-resource issue in multimodal metaphor research and take advantage of imaginative frames, we introduce an approach to construct a cross-modal imagination dataset containing multimodal metaphors paired with their imaginative frames. In this section, we explain how we convert large textual simile datasets into a multimodal format, detail the transformation of these multimodal similes into a cross-modal imagination dataset containing multimodal metaphors, and give the quality control and data analysis.

Multimodal Simile Data Construction Different from metaphor, which makes an *implicit* comparison between two objects, the simile compares two things *explicitly* using the words “like” or “as”. A simile has three explicit components: (1) *tenor*, which is the subject of the comparison; (2) *vehicle*, which is the object of the comparison; and (3) *comparator*, which is the trigger word such as “like” or “as”. A metaphor has two important components: (1) *target*, which we try to understand, and (2) *source*, which we implicitly draw metaphorical expressions from. The *tenor-vehicle* pair in simile and the *target-source* pair in metaphor are similar and interchangeable (Cope, 1877). They both allow us to understand one thing in terms of another and can stimulate imaginative frames (Lakoff and Johnson, 1980).¹ Thus, textual similes can serve

¹For example, the simile “Time is as valuable as money” and the metaphor “I have invested a lot of time in her” reflect the same comparison between the tenor/target *time* and the vehicle/source *money*. They evoke an imaginative frame where

as resources to construct a metaphor dataset rich in imaginative frames.

We collect publicly available simile datasets where each sample is labeled with its tenor, vehicle and comparator. For each textual representation of a tenor or vehicle (denoted as T), we utilize the Bing Image Search API to retrieve n images, represented as $\mathcal{I} = [I_1, I_2, \dots, I_n]$. To ensure the relevance and concreteness of the collected data, we design a filtering module that distinguishes concrete tenors/vehicles (e.g., “flower”) from abstract ones (e.g., “love”). The image set \mathcal{S}_{img} for T is

$$\mathcal{S}_{img} = \{I_i \in \mathcal{I} \mid \text{CLIP}(T, I_i) \geq \theta_{sim}\}, \quad (1)$$

where $\text{CLIP}(T, I)$ is a model to calculate the similarity score between a text and an image (Radford et al., 2021), and θ_{sim} is the similarity threshold. If $|\mathcal{S}_{img}| \geq n_{sim}$, we classify T as a concrete concept, and regard \mathcal{S}_{img} as the paired image set for T ; otherwise, T is categorized as an abstract concept. We determine the optimal values for θ_{sim} and n_{sim} in Eq. (1) by maximizing classification accuracy for distinguishing concrete and abstract concepts on a randomly sampled dataset of tenors and vehicles pre-annotated with concrete/abstract labels. Finally, we obtain a multimodal simile dataset by selecting simile instances with concrete tenors/vehicles and their corresponding paired image sets.

Multimodal Metaphor Data Construction We convert each simile into a metaphor by apply-

TIME is the target domain, MONEY is the source domain, and *valuable commodity* is the similar attribute shared by source and target domains.

Samples	Bilingual	English	Chinese
#Total Instances	6071	2328	3743
#Instances with Visual Src & Textual Tgt	3415	1949	1466
#Instances with Visual Tgt & Textual Src	2656	379	2277
#Unique Sentences	6066	2328	3738
#Mean Words per Instance	32	25	36

Table 1: Statistics of cross-modal imagination dataset. *Src/Tgt* denotes the abbreviation of *Source/Target*.

ing a rule-based method. Details of these rules are provided in Appendix A. We then replace the tenor/vehicle with a randomly sampled image from its paired set while keeping the vehicle/tenor as text for each instance in the multimodal simile dataset. The tenor and vehicle are regarded as the target concept and source concept in the cross-modal metaphor, respectively. To construct the imaginative frame for each instance, we mine the conceptual target and source domains using the conceptual domain mining algorithm illustrated in Section 4.2. We then employ ChatGPT (OpenAI et al., 2024) to generate similar attributes for each instance using source concept, target concept and corresponding domains as the inputs. More details of the attribute generator are provided in Appendix A.

Quality Control and Data Analysis To ensure the quality of the concrete tenor/vehicle concept identified by CLIP using Eq. (1), we randomly sampled 100 English and 100 Chinese tenor/vehicle concepts from simile datasets, and invited two PhD students to annotate these concepts as either concrete or abstract. The Cohen’s kappa coefficient κ (Cohen, 1960) of the inter-rater agreement is 0.74 (note that $0.6 \leq \kappa \leq 0.8$ means substantial agreement). We use this sampled annotated dataset to select optimal values of θ_{sim} and n_{sim} in Eq. (1). To ensure the quality of the similar attributes generated by ChatGPT, we employ a self-verification mechanism. Specifically, we prompt ChatGPT whether the generated attribute is applicable to the source/target concept. Samples that pass this verification are included in the final dataset. We finally obtain a bilingual cross-modal imagination dataset with 6K multimodal metaphors and corresponding imaginative frames, where the source and target concepts of each metaphor are in different modalities. Its statistics are provided in Table 1.

4.2 Cross-Modal Imaginative Frame Mining

To obtain imaginative frames for the image-text input, we first develop a conceptual domain mining algorithm to mine the conceptual domains of en-

tities from different modalities. We then consider inconsistent domains from different modalities as domain pairs, and train a model to generate similar attributes between domains in each domain pair.

Conceptual Domain Mining Previous metaphor research (Lakoff and Johnson, 1980; Forceville and Urios-Aparisi, 2009) indicates that, in a multimodal metaphor, the source and target concepts are represented separately in image and text, and imply inconsistency between source and target domains. Inspired by this, we propose an approach to mine the conceptual domains for entities in the multimodal input. We first utilize an MLLM-based image captioning method to generate a textual description of the image, and then extract nouns and pronouns from both the image description and text to form the image entity set $E_I = \{e_I^i\}_{i=1}^{n_I}$ and the text entity set $E_T = \{e_T^i\}_{i=1}^{n_T}$, respectively.

To establish a solid foundation for conceptual domain mining, we first create a conceptual domain set \mathcal{S}_d , based on the master metaphor list (Lakoff et al., 1991). This list includes the conceptual source and target domains found in representative metaphors developed by cognitive linguists. Details of \mathcal{S}_d are provided in Appendix A. The large lexical semantic database WordNet (Miller, 1995) organizes words into hierarchical structures through conceptual relations. The richness and transitivity of hypernym relations in WordNet make it a valuable resource for identifying conceptual domains. Our conceptual domain mining algorithm leverages \mathcal{S}_d and WordNet to find an appropriate conceptual domain for each entity in image entity set E_I and text entity set E_T , resulting in image domain set $D_I = \{d_I^i\}_{i=1}^{n_I}$ and text domain set $D_T = \{d_T^i\}_{i=1}^{n_T}$. For the term with multiple word senses, we treat each word sense as a distinct entity. For each entity, we traverse its hypernym path in WordNet. If one or more hypernyms are present \mathcal{S}_d , the lowest hypernym is assigned as the entity’s conceptual domain. If no hypernyms belong to \mathcal{S}_d , we regard the first hypernym in the hypernym path as its conceptual domain.

Domain Consistency Mapping We use path similarity to measure the consistency/inconsistency between the image and text domains:

$$\text{sim}(d_I, d_T) = \frac{1}{1 + l(d_I, d_T)}, \quad (2)$$

where $l(d_I, d_T)$ calculates the number of edges in the shortest path that connects d_I and d_T in

WordNet. If $\text{sim}(d_I, d_T) \leq \theta_{incon}$, the pair (d_I, d_T) is considered as a cross-modal domain pair, where θ_{incon} is a hyperparameter. We then construct the cross-modal domain pair set $\mathcal{P}_d = \{(d_I^i, d_T^i)\}_{i=1}^{n_d}$, where d_I^i and d_T^i are the inconsistent conceptual domains for the i -th image-text entity pair (e_I^i, e_T^i) .

Similar Attribute Generation We use concept pairs and domain pairs along with similar attributes in cross-modal imagination dataset as training data to fine-tune an MLLM as the similar attribute generator that can employ the concept pair (e_I, e_T) and its domain pair (d_I, d_T) as inputs and then generate the similar attributes $A = \{a_i\}_{i=1}^{n_a}$ shared between the concepts in concept pair. Details of this attribute generator are shown in Appendix A. We label the similar attributes for each cross-modal domain pair in \mathcal{P}_d and obtain the candidate cross-modal imaginative frames for the multimodal input $\mathcal{F} = \{(d_I^i, d_T^i, A_{I-T}^i)\}_{i=1}^{n_f}$, where d_I^i, d_T^i and A_{I-T}^i denote the image domain, text domain and similar attribute set in the i -th imaginative frame.

4.3 Imaginative Frame Augmented Retrieval

To improve MLLM’s performance on multimodal metaphor detection and explanation, we obtain an augmented instance from the cross-modal imagination dataset via leveraging the imaginative frame as the associative structure. We first introduce how we compare the *similarity between imaginative frames* and then illustrate the retrieval process of augmented exemplar based on the similarity between imaginative frames in multimodal input and instances from cross-modal imagination dataset.

Frame Similarity Calculation The structure of an imaginative frame consists of two key components: *a domain pair and the attribute*. To calculate the similarity between two imaginative frames, we compute the similarity of these components individually and balance their contributions in the overall frame similarity. $F_{I-T} = (d_I, d_T, A_{I-T})$ represents an imaginative frame in the multimodal input. Similarly, $F_{s-t} = (d_{src}, d_{tgt}, A_{s-t})$ denotes an imaginative frame in cross-modal imagination dataset, where d_{src}, d_{tgt} and A_{s-t} denote the source domain, target domain and similar attribute set, respectively.

To calculate the similarity score α^{dom} between two *domain pairs* in F_{I-T} and F_{s-t} , we leverage the path similarity in WordNet to measure the simi-

ilarity of two domains, which is computed as

$$\alpha^1 = \text{sim}(d_I, d_{src}) + \text{sim}(d_T, d_{tgt}), \quad (3)$$

$$\alpha^2 = \text{sim}(d_T, d_{src}) + \text{sim}(d_I, d_{tgt}), \quad (4)$$

$$\alpha^{dom} = \max(\alpha^1, \alpha^2), \quad (5)$$

where $\text{sim}(\cdot)$ is the path similarity function as illustrated in Eq. (2) and $\max(\alpha^1, \alpha^2)$ returns the maximum value between α^1 and α^2 .

To calculate the similarity score α^{attr} between *attribute sets* F_{I-T} and F_{s-t} , we compute their semantic similarity, which is computed as

$$\alpha^{attr} = \cos(\text{emb}(A_{I-T}), \text{emb}(A_{s-t})), \quad (6)$$

where $\text{emb}(A)$ embeds the concatenation of similar attributes in A using a pretrained model XLM-RoBERTa (Conneau et al., 2020) and $\cos(\cdot)$ is the cosine similarity function.

Finally, we balance the different components in the imaginative frames and obtain the similarity score α^f between F_{I-T} and F_{s-t} :

$$\alpha^f = \alpha^{dom} + \lambda \alpha^{attr}, \quad (7)$$

where λ is the balance parameter.

Retrieval Process After calculating the similarity scores between the imaginative frames in the multimodal input and cross-modal imagination dataset, we retrieve the instance \mathcal{X}_m from the cross-modal imagination dataset that can calculate the highest imaginative frame similarity score with imaginative frames in the multimodal input:

$$(i_{max}, k_{max}) = \arg \max_{(i,k)} (\alpha_{(i,k)}^f), \quad (8)$$

where $\alpha_{(i,k)}^f$ is the similarity score between the i -th imaginative frame in the multimodal input and the imaginative frame of k -th instance in cross-modal imagination dataset, and k_{max} is the index of the instance \mathcal{X}_m in cross-modal imagination dataset. We use this multimodal metaphor \mathcal{X}_m as the exemplar.

4.4 Instruction Fine-Tuning

We treat the retrieved instance \mathcal{X}_m as the metaphorical exemplar. To help the MLLM differentiate between metaphorical and literal expressions, we construct its literal multimodal exemplar \mathcal{X}_l by connecting the same image from \mathcal{X}_m and a literal image caption generated by an MLLM-based image

captioning model. The instruction for an image-text input \mathcal{X} is denoted as \mathcal{T} . Our training objective is represented as

$$\mathcal{L}(\theta) = - \sum_{i=1}^N \log p(\mathcal{R}_i | \mathcal{T}, \mathcal{X}_m, \mathcal{X}_l, \mathcal{X}, \mathcal{R}_{<i}; \theta), \quad (9)$$

where \mathcal{R} is the ground truth response, N is its length and θ denotes the parameters of the MLLM. We fine-tune the MLLM using low-rank adaptation (LoRA) method (Hu et al., 2022).

5 Experiments

5.1 Datasets

Multimodal Metaphor Dataset We evaluate our method using two publicly available multimodal metaphor datasets, which are as follows:

- **MET-Meme** (Xu et al., 2022) contains 4,000 English and 6,045 Chinese multimodal instances. Each instance is labeled as either metaphorical or literal, with each metaphorical instance further labeled with its source and target concepts. Following the convention of previous research (Xu et al., 2024), we randomly divided it into training, validation and test sets with a ratio of 6:2:2.
- **MultiCMET** (Zhang et al., 2023) is a Chinese multimodal metaphor detection dataset, where each text-image pair is labeled as either metaphorical or literal. This dataset comprises two parts: one collected from public service announcements, and the other collected from commercial advertisements on e-commerce platforms. Due to the low quality of the latter, we only use the former for evaluation, excluding samples with inaccessible images. We randomly divided it into training, validation and test sets with a ratio of 7:1:2.

We employ MET-Meme to evaluate our method for both multimodal metaphor detection and explanation tasks, while MultiCMET is utilized only for multimodal metaphor detection. Table 2 shows the statistics of these datasets.

Textual Simile Datasets To construct cross-modal imagination dataset, we use five publicly available English and Chinese textual simile datasets, including Xiang (Liu et al., 2018), SPGC (He et al., 2022), CMC (Li et al., 2022), MAPS (He et al., 2023), and CMRE (Chen et al., 2023).

Samples	MET-Meme		MultiCMET
	English	Chinese	Chinese
#Total	4000	6045	6645
%Metaphorical Sample	28%	39%	46%
#Avg. Words per Sample	12	7	30

Table 2: Statistics of MET-Meme and MultiCMET datasets. Avg. denotes the abbreviation of *average*.

These datasets comprise simile sentences labeled with tenor, vehicle and comparator components. More details are illustrated in Appendix B.

5.2 Baselines

Unimodal Methods We employ several representative visual baseline models, which only use visual information as the input, including **VGG** (Simonyan and Zisserman, 2015), **ViT** (Dosovitskiy et al., 2021) and **Swin Transformer** (Liu et al., 2021). We employ several multilingual pre-trained baseline models, which only use textual information as the input, including **mBART** (Liu et al., 2020), **mT5** (Xue et al., 2021) and **M-BERT** (Papadimitriou et al., 2021).

Multimodal Methods We use representative methods for multimodal metaphor detection as multimodal baselines, including (1) **Fusion** (Xu et al., 2022), which fuses the text, image feature and metaphor features for multimodal metaphor detection; (2) **M3F** (Wang et al., 2024), which employs a multi-task framework with inter-modality attention to capture the features between image and text; and (3) **C4MMD** (Xu et al., 2024), the SOTA method for multimodal metaphor detection, which distills knowledge from the MLLM with a chain-of-thought method. We use representative MLLMs as baselines in low-resource multimodal metaphor detection and explanation, including (1) **GPT-4o** (OpenAI et al., 2024), a representative MLLM that achieves the SOTA performance on multiple multimodal tasks, for which we apply zero-shot and few-shot prompting strategies; and (2) **LLaVA** (Liu et al., 2023), a publicly available MLLM with versions of different sizes, for which we apply low-rank adaptation (LoRA) (Hu et al., 2022) fine-tuning. The prompt design of MLLM-based baselines is illustrated in Appendix F.

5.3 Implementation Details

We use LLaVA-1.5 (7B and 13B) (Liu et al., 2023) as the MLLM in our method and LLaVA baselines. More details are provided in Appendix C.

Modality	Method	MET-Meme				MultiCMET			
		P	R	F1	Acc	P	R	F1	Acc
Image	VGG16 (Simonyan and Zisserman, 2015)	62.72	72.36	67.20	75.30	62.94	72.18	67.24	68.60
	ViT-base (Dosovitskiy et al., 2021)	65.14	64.67	64.90	75.55	65.53	77.57	71.04	71.76
	Swin Transformer (Liu et al., 2021)	67.87	72.51	70.11	78.39	65.72	77.91	71.30	71.99
Text	M-T5-base (Xue et al., 2021)	68.39	64.10	66.18	77.09	49.04	56.16	52.36	54.37
	M-BERT-base (Papadimitriou et al., 2021)	76.70	72.22	74.39	82.62	51.86	65.94	58.05	57.45
	M-BART-large (Liu et al., 2020)	77.94	74.50	76.18	83.72	53.69	72.34	61.64	59.79
Image + Text	Fusion (Xu et al., 2022)	75.97	76.07	76.02	-	62.37	62.30	62.34	58.64
	M3F (Wang et al., 2024)	78.11	83.36	-	79.80	60.91	67.87	64.20	58.42
	C4MMD (Xu et al., 2024)	83.33	81.58	82.44	87.70	66.28	77.61	71.22	72.00
	ImaRA-7B	<u>85.04</u>	<u>83.38</u>	<u>84.20</u>	<u>89.06</u>	<u>67.74</u>	<u>78.25</u>	<u>72.61</u>	<u>73.64</u>
	ImaRA-13B	86.59	83.97	84.82	89.49	70.98	79.17	74.75	76.13

Table 3: Comparison between our method and baselines on multimodal metaphor detection in full training data. The best results are in bold font and the second-best results are underlined.

Method	Training Data (%)											
	Percent = 60%				Percent = 40%				Percent = 20%			
	F1	Acc	Src	Tgt	F1	Acc	Src	Tgt	F1	Acc	Src	Tgt
M3F	67.40	78.30	-	-	66.01	76.35	-	-	66.01	75.11	-	-
C4MMD	79.14	85.54	-	-	76.36	83.41	-	-	70.89	79.27	-	-
LLaVA-7B	79.82	86.55	48.01	55.13	74.74	82.92	42.02	48.53	70.46	81.87	30.13	42.24
LLaVA-13B	80.59	86.90	51.99	59.54	79.66	86.01	44.94	54.70	74.09	83.07	34.05	47.44
ImaRA-7B	81.89	<u>87.25</u>	52.28	<u>60.61</u>	80.56	86.60	46.44	55.32	76.43	83.17	35.75	48.29
ImaRA-13B	83.13	88.20	54.13	62.68	81.05	86.75	47.22	59.54	78.16	84.86	39.03	54.42

Table 4: Results of our method and baselines for low-resource multimodal metaphor detection and explanation on MET-Meme. *Src/Tgt* represents the accuracy score of source/target concept predictions for metaphorical instances.

Method	Training Data (%)					
	Percent=60%		Percent=40%		Percent=20%	
	F1	Acc	F1	Acc	F1	Acc
M3F	63.96	56.60	60.22	55.20	55.26	54.25
C4MMD	69.37	71.41	68.49	71.31	65.33	68.80
LLaVA-7B	72.89	72.21	72.52	71.01	69.81	70.37
LLaVA-13B	73.33	73.53	72.11	73.49	69.28	72.89
ImaRA-7B	<u>73.90</u>	<u>75.15</u>	<u>73.45</u>	<u>74.60</u>	<u>71.04</u>	<u>73.98</u>
ImaRA-13B	74.28	75.25	73.94	74.62	72.25	74.32

Table 5: Comparison between our method and baselines for low-resource multimodal metaphor detection on MultiCMET dataset.

Method	Detection				Explanation	
	MET-Meme		MultiCMET		MET-Meme	
	F1	Acc	F1	Acc	Src	Tgt
GPT-4o (0-shot)	54.75	47.81	63.86	49.55	16.52	22.79
GPT-4o (5-shot)	58.47	71.40	67.04	60.32	18.23	26.21
ImaRA-7B	<u>70.80</u>	<u>79.93</u>	<u>67.73</u>	<u>68.00</u>	<u>25.93</u>	<u>40.53</u>
ImaRA-13B	74.69	81.67	70.57	70.93	28.21	46.58

Table 6: Comparison between our method using 10% training data and GPT-4o for multimodal metaphor detection and explanation in the low-resource scenario. *Src/Tgt* represents the accuracy score of source/target concept predictions for metaphorical instances.

5.4 Main Results

Comparison in Full Training Data Following previous research (Xu et al., 2024), we use accuracy, precision, recall and F1 score as evaluation metrics for multimodal metaphor detection. The experimental results in Table 3 show that multimodal methods perform better compared with unimodal methods, indicating that cross-modal interaction is important to capture the implicit mean-

ings conveyed by multimodal metaphors. Although C4MMD achieves previous SOTA performance by distilling commonsense knowledge from the MLLM to enhance the detection performance of the pretrained model, it fails to improve the MLLM’s capability in multimodal metaphor detection. In contrast, our methods surpass existing multimodal metaphor detection methods on both datasets via retrieving the augmented exemplar with the imag-

Variant	Detection (F1)						Explanation (Avg. Acc)		
	MET-Meme			MultiCMET			MET-Meme		
	60%	40%	20%	60%	40%	20%	60%	40%	20%
ImaRA-7B	81.89	80.56	76.43	73.90	73.45	71.04	56.45	50.88	42.02
- ImagFrame	81.37	76.84	74.98	73.48	73.30	70.37	52.07	45.48	40.81
- Retrieval	79.82	74.74	70.46	72.89	72.52	69.81	51.57	45.28	36.18
- LoRA	43.41	43.41	43.41	44.50	44.50	44.50	5.62	5.62	5.62
ImaRA-13B	83.13	81.05	78.16	74.28	73.94	72.25	58.40	53.38	46.72
- ImagFrame	82.42	80.24	75.61	73.37	72.95	71.40	56.70	51.07	41.77
- Retrieval	80.59	79.66	74.09	73.33	72.11	69.28	55.77	49.82	40.74
- LoRA	44.15	44.15	44.15	48.90	48.90	48.90	13.17	13.17	13.17

Table 7: Experimental results of ablation study on multimodal metaphor detection and explanation in low-resource scenarios. Avg. Acc represents the average accuracy score for source and target concept predictions.

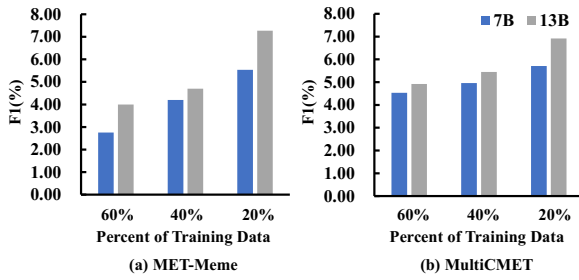


Figure 3: Performance improvements achieved by our proposed methods compared to C4MMD for low-resource multimodal metaphor detection.

inative frame as the associative structure, verifying the effectiveness of our method on multimodal metaphor detection in full training data.

Comparison in Low-Resource Scenarios We compare our method with existing SOTA methods for multimodal metaphor detection (M3F and C4MMD), a fine-tuned MLLM (LLaVA) and a prompting-based MLLM (GPT-4o) on low-resource metaphor detection and explanation. We compare our method with SOTA methods for multimodal metaphor detection and the fine-tuned LLaVA by reducing the percentage of training data from 60% to 20%. The experimental results in Table 5 and Table 4 show that our methods exhibit more gradual performance declines compared to baselines and consistently outperform them when training data are decreasing. Figure 3 further illustrates that our method achieves increasing performance gains over C4MMD on both MET-Meme and MultiCMET datasets as training data is reduced. These results verify the robustness and effectiveness of our imaginative frame augmented method on multimodal metaphor detection and explanation in low-resource scenarios. Experimental results in Table 6 show that our methods, fine-tuned

with only 10% training data, significantly outperform GPT-4o methods using different prompting strategies on both metaphor detection and explanation, verifying the effectiveness of our proposed method in the extremely low-resource scenario.

5.5 Ablation Study

The experimental results of ablation study in Table 7 show that replacing the retrieved instance from our imaginative frame augmented retrieval module with a randomly sampled one reduces the performance of our methods (- ImagFrame). Directly removing the retrieved instance further leads to significant drops in the performance of our methods (- Retrieval). These results verify the effectiveness of our method for increasing the robustness of the model on low-resource multimodal metaphor detection and explanation. Removing the LoRA fine-tuning significantly reduces our method’s performance, showing that the MLLM in a zero-shot setting struggles with multimodal metaphor tasks.

6 Conclusion

Inspired by CMT, we propose a novel method ImaRA for low-resource multimodal metaphor detection and explanation, which leverages imaginative frame mined from the input as the associative structure to stimulate imaginative thinking for metaphor understanding. The retrieved exemplar based on it from a cross-modal imagination dataset we construct assists the MLLM to understand metaphor for the detection and explanation tasks. Experiments on two publicly available datasets verify the effectiveness of our method for robust multimodal metaphor detection and explanation in low-resource settings and against existing multimodal detection methods in full training data.

Limitations

Our work has some limitations. Firstly, due to the costs of experimentation with MLLMs, we are unable to evaluate our method on MLLMs in larger sizes. Thus, for this task, the performance gains achieved by our method on MLLMs with larger sizes deserve further exploration. In addition, since our method is designed to identify metaphors in multimodal mode, it performs less effectively in text-centric cases involving wordplay, which is worth further exploration in the future.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grants #72293575, #72225011, #72434005 and #62206287, and Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park under Grant #Z231100007423016.

References

- Guihua Chen, Tiantian Wu, MiaoMiao Cheng, Xu Han, Jiefu Gong, Shijin Wang, and Wei Song. 2023. [Chinese metaphorical relation extraction: Dataset and models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9085–9095.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- Edward M. Cope. 1877. *The Rhetoric of Aristotle*. University Press.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, et al. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *Proceedings of International Conference on Learning Representations*, pages 1–22.
- Charles Forceville and Eduardo Urios-Aparisi. 2009. *Multimodal Metaphor*. Mouton de Gruyter Berlin.
- Mengshi Ge, Rui Mao, and Erik Cambria. 2022. [Explainable metaphor identification inspired by conceptual metaphor theory](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10681–10689.
- Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56(Suppl 2):1829–1895.
- Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022. Can pre-trained language models interpret similes as smart as human? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7875–7887.
- Qianyu He, Xintao Wang, Jiaqing Liang, and Yanghua Xiao. 2023. MAPS-KB: A million-scale probabilistic simile knowledge base. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6398–6406.
- Xiaoyu He, Long Yu, Shengwei Tian, Qimeng Yang, Jun Long, and Bo Wang. 2024. [VIEMF: Multimodal metaphor detection via visual information enhancement with multimodal fusion](#). *Information Processing and Management*, 61(3).
- Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. [Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 688–714.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *Proceedings of International Conference on Learning Representations*, pages 1–26.
- EunJeong Hwang and Vered Shwartz. 2023. [MemeCap: A dataset for captioning and interpreting memes](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445.
- Gitit Kehat and James Pustejovsky. 2020. [Improving neural metaphor detection with visual datasets](#). In *Proceedings of the Language Resources and Evaluation Conference*, pages 5928–5933.
- George Lakoff, Jane Espenson, and Alan Schwartz. 1991. The master metaphor list. Technical report, University of California at Berkeley.
- George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago press.
- Hongsong Li, Kenny Q Zhu, and Haixun Wang. 2013. Data-driven metaphor recognition and explanation. *Transactions of the Association for Computational Linguistics*, 1:379–390.
- Yucheng Li, Chenghua Lin, and Frank Guerin. 2022. [CM-Gen: A neural framework for Chinese metaphor generation with explicit context modelling](#). In *Proceedings of the International Conference on Computational Linguistics*, pages 6468–6479.

- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 1–25.
- Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. [Neural multitask learning for simile recognition](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.
- George A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, et al. 2024. [GPT-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. [Deep Subjecthood: Higher-order grammatical features in multilingual BERT](#). In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 2522–2532.
- Alec Radford, Jong W. Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763.
- Paul Ricoeur. 1978. The metaphorical process as cognition, imagination, and feeling. *Critical inquiry*, 5(1):143–159.
- Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. [V-FLUTE: Visual figurative language understanding with textual explanations](#). *Preprint*, arXiv:2405.01474.
- Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. [Black holes and white rabbits: Metaphor identification with visual features](#). In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 160–170.
- Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of International Conference on Learning Representations*, pages 1–14.
- Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. [Metaphor generation with conceptual mappings](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 6724–6736.
- Chang Su, Weijie Chen, Ze Fu, and Yijiang Chen. 2021. [Multimodal metaphor detection based on distinguishing concreteness](#). *Neurocomputing*, 429:166–173.
- Yuan Tian, Nan Xu, and Wenji Mao. 2024a. [A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning](#). In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 7738–7755.
- Yuan Tian, Nan Xu, Wenji Mao, and Daniel Zeng. 2023. Modeling conceptual attribute likeness and domain inconsistency for metaphor detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7736–7752.
- Yuan Tian, Ruike Zhang, Nan Xu, and Wenji Mao. 2024b. [Bridging word-pair and token-level metaphor detection with explainable domain mining](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 13311–13325.
- Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. [Metaphor detection with cross-lingual model transfer](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 248–258.
- Bingbing Wang, Shijue Huang, Bin Liang, Geng Tu, Min Yang, and Ruifeng Xu. 2024. [What do they “meme”? a metaphor-aware multi-modal multi-task framework for fine-grained meme understanding](#). *Knowledge-Based Systems*, 294:111778.
- Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. [MET-Meme: A multimodal meme dataset rich in metaphors](#). In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2887–2899.
- Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024. [Exploring chain-of-thought for multi-modal metaphor detection](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 91–101.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 483–498.
- Dongyu Zhang, Jingwei Yu, Senyuan Jin, Liang Yang, and Hongfei Lin. 2023. [MultiCMET: A novel Chinese benchmark for understanding multimodal](#)

Language	Simile Comparator	Metaphor Comparator
English	BE like as ADJ as	BE ADJ
Chinese	宛如(是) / 好比(是) / 仿佛(是) / (好)像(是) / (好/恰)似(是) / 犹同(是) / 犹如(是) 如...一样/一般	是 是...一样/一般

Figure 4: Rules for converting simile comparators into metaphor comparators.

English Prompt Template
The domain of “[source concept]” is “[source domain]”. The domain of “[target concept]” is “[target domain]”. Please give me the similar attributes shared between “[source concept]” and “[target concept]”. Please response with the following format: similar attribute 1; similar attribute 2; ... ; similar attribute n.
Chinese Prompt Template
“[source concept]”的领域是“[source domain]”。“[target concept]”的领域是“[target domain]”。请给出“[source concept]”与“[target concept]”之间的相似属性。请按照以下格式回答：相似属性 1;相似属性 2; ... ;相似属性 n。

Figure 5: Prompt design for ChatGPT to label similar attributes. Here, [source concept], [source domain], [target concept] and [target domain] denote the input slots for the source concept, source domain, target concept and target domain in an instance from cross-modal imagination dataset.

metaphor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6141–6154.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. **MultiMET: A multi-modal dataset for metaphor understanding**. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3214–3225.

Linhao Zhang, Li Jin, Guangluan Xu, Xiaoyu Li, Cai Xu, Kaiwen Wei, Nayu Liu, and Haonan Liu. 2024. **CAMEL: Capturing metaphorical alignment with context disentangling for multimodal emotion recognition**. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9341–9349.

A More Details on Proposed Method

We provide more details on our proposed method.

Cross-Modal Imagination Data Construction

We convert each simile into a metaphor using a rule-based method by replacing simile comparators with metaphor comparators. Figure 4 shows the detailed rules. We use ChatGPT to label similar attributes between source and target concepts for

English Prompt Template
Attributes: [similar attributes]. Does these attributes belong to the “[source/target concept]”? Please give me an answer selected from YES or NO.
Chinese Prompt Template
属性: [similar attributes]. 上述属性是“[source/target concept]”的属性吗? 请回答我 “是”或“否”。

Figure 6: Prompt design for the self-verification of similar attributes generated by ChatGPT. Here, [similar attributes] and [source/target concept] denote the input slots for the similar attributes generated by ChatGPT and the source or target concept in an instance from the cross-modal imagination dataset, respectively.

the instances in cross-modal imagination dataset, whose prompt design is shown in Figure 5. To ensure the quality, we do self-verification on the attributes generated by ChatGPT, whose prompt design is shown in Figure 6.

Cross-Modal Imaginative Frame Mining In Conceptual Domain Mining (Section 4.2), the complete *conceptual domain set* \mathcal{S}_d extracted from the master metaphor list in Lakoff et al. (1991) is summarized in Figure 7. In Similar Attribute Generation (Section 4.2), the *similar attribute generator* is obtained by fine-tuning LLaVA-7B (Liu et al., 2023) with concept pairs, domain pairs and similar attributes from cross-modal imagination dataset as the training data, as well as the prompts in Figure 5, using LoRA fine-tuning method (Hu et al., 2022).

B More details on Datasets and Baselines

Datasets The five publicly available textual simile datasets we use in cross-modal imagination data construction are as follows: (1) **Xiang** (Liu et al., 2018), which comprises 5088 Chinese simile sentences labeled with their simile components, including tenor, vehicle and comparator; (2) **SPGC** (He et al., 2022), an English simile dataset containing 775 sentences, where all sentences uniformly use the structure “(tenor) is as (property) as (vehicle)” to present English similes; (3) **CMC** (Li et al., 2022), encompassing 8027 Chinese nominal simile sentences, where each is labeled with its tenor, vehicle, and comparator; (4) **MAPS** (He et al., 2023), consisting of 0.5 million English simile sentences extracted from 70 GB of corpus using syntactic patterns (e.g. Noun₁ BE like Noun₂) and subsequently annotated with simile components (topic and vehicle) via predefined rules; (5) **CMRE** (Chen et al.,

ABILITY	ANGER	ARGUMENT
BATTERY	BELIEF	BODY
BURDEN	CAREER	CHANGE
CHILD	CLOTH	COMMODITY
COMPETITION	CONTAINER	DEATH
EMOTION	FAILURE	FIRE
FIGHT	FLUID	FOOD
HARM	HOPE	IDEA
IMPORTANCE	INJURY	INFORMATION
JOURNEY	LIFE	LIGHT
LIQUID	LOVE	MACHINE
MONEY	MOTION	OBLIGATION
PATH	PEOPLE	PRECEDENCE
PROBLEM	RACE	RESOURCE
RESPONSIBILITY	SCALE	SOCIETY
THEORY	TIME	WAR
WATER	WEAPON	WORD

Figure 7: Conceptual domain set extracted from the master metaphor list in Lakoff et al. (1991).

2023), a Chinese dataset that includes both similes and nominal metaphors, totalling 8494 sentences, where each sentence is annotated with its target (tenor), source (vehicle), and comparator.

Baseline Methods VIEMF (He et al., 2024) is another baseline for multimodal metaphor detection, which introduces a multi-interactive cross-modal residual network. Since the code of VIEMF is not publicly available and it only evaluates on MET-Meme dataset in its original paper, we are unable to evaluate its performance on the MultiCMET dataset. Thus we only compare it with our method on MET-Meme dataset in Table 8.

C More Implementation Details

In *Cross-Modal Imagination Data Construction* (Section 4.1), the number of images n we collect for each tenor/vehicle is 10. The similarity thresholds θ_{sim} in Eq. (1) for Chinese and English similes are 0.65 and 0.61, respectively. The threshold numbers of paired images n_{sim} for Chinese and English similes are 7 and 8, respectively. We use GPT-3.5 (gpt-3.5-turbo)² as the implementation of ChatGPT for attribute generation through the OpenAI API.

In *Cross-Modal Imaginative Frame Mining* (Section 4.2), we utilize LLaVA-1.5 13B as the image caption model and the threshold for incon-

sistency domains θ_{incon} is 0.25. We use NLTK Python package³ as the implementation of WordNet in our method, which supports both English and Chinese. We employ the en_core_web_sm and zh_core_web_sm models in spaCy Python package⁴ to label nouns and pronouns in English and Chinese texts, respectively.

In our experiments, we use *llava-1.5-7b-hf*⁵ and *llava-1.5-13b-hf*⁶ as the implementations of LLaVA-1.5 7B and LLaVA-1.5 13B, respectively. We fine-tune LLaVA with the learning rate of $2e^{-4}$, the batch size of 8 and the epoch of 5. The rank of the update matrices and the scaling factor of LoRA are 128 and 256, respectively. GPT-4o is implemented through the OpenAI API, utilizing GPT-4o-0513⁷ model. We use *clip-vit-large-patch14-336* and *chinese-clip-vit-large-patch14-336px* models on the huggingface platform as the implementations of CLIP(\cdot) function in Eq. (1) for English and Chinese texts, respectively. The balance parameter λ is 1. We use a multilingual pretrained model xlm-roberta-base⁸ as the implementation of XLM-RoBERTa in our method. The model achieving the best performance of F1 in the validation set is used for the test set. All experimental results reported are the averaged scores of five runs with different random seeds. We use rule-based methods to extract detection results along with the corresponding source and target concepts from the generated response. All the experiments are conducted on NVIDIA GeForce RTX 3090 GPUs and NVIDIA A100 SXM4 80GB GPUs.⁹

D More Details on Experimental Results

We conducted *statistical tests* (Wilcoxon signed-rank tests) on the experimental results of our methods and baselines. The results of our method are all statistically significantly different from the best results of the baselines with $p \leq 0.05$ in full training data and low-resource scenarios. Table 8 further shows the detailed results of our method and baselines on MET-Meme dataset.

³<https://www.nltk.org/howto/wordnet.html>

⁴<https://spacy.io>

⁵<https://huggingface.co/llava-hf/llava-1.5-7b-hf>

⁶<https://huggingface.co/llava-hf/llava-1.5-13b-hf>

⁷<https://platform.openai.com/docs/models>

⁸<https://huggingface.co/FacebookAI/xlm-roberta-base>

⁹Our code is available at <https://github.com/TIAN-viola/ImaRA>.

²<https://platform.openai.com/docs/models/gpt-3-5>

Modality	Method	Bilingual				English				Chinese			
		P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
Image	VGG16	62.72	72.36	67.20	75.30	74.66	72.69	73.66	85.25	58.23	72.21	64.47	68.71
	ViT-base	65.14	64.67	64.90	75.55	76.30	70.93	73.52	85.50	60.29	61.68	60.98	68.96
	Swin Transformer	67.87	72.51	70.11	78.39	71.88	81.06	76.19	85.62	65.79	68.42	67.08	73.59
Text	M-BERT	76.70	72.22	74.39	82.62	70.76	53.30	60.80	80.50	78.78	81.26	80.00	84.02
	M-T5-base	68.39	64.10	66.18	77.09	57.04	33.92	42.54	74.00	71.32	78.53	74.75	79.14
	M-BART-large	77.94	74.50	76.18	83.72	69.38	63.88	66.51	81.75	81.82	79.58	80.68	85.02
Image + Text	Fusion	75.97	76.07	76.02	-	82.69	83.33	82.39	-	72.80	73.30	72.90	-
	VIEMF	83.07	83.33	83.20	-	85.13	<u>83.38</u>	<u>83.92</u>	-	83.58	83.43	83.50	-
	M3F	78.11	83.36	-	79.80	85.86	84.38	-	83.98	72.94	82.68	-	77.01
	C4MMD	79.45	83.33	81.29	86.59	78.87	79.44	79.05	88.04	79.74	85.19	82.34	85.62
	ImaRA-7B	<u>85.04</u>	<u>83.38</u>	<u>84.20</u>	<u>89.06</u>	<u>87.60</u>	79.74	83.45	<u>91.04</u>	<u>83.94</u>	85.12	<u>84.53</u>	<u>87.75</u>
	ImaRA-13B	86.59	83.97	84.82	89.49	89.05	80.62	84.62	91.69	84.29	85.58	84.91	88.04

Table 8: Comparison between our method and baselines for multimodal metaphor detection on MET-Meme dataset in full training data. The results of C4MMD are reproduced using the code released by Xu et al. (2024). The best results are in bold font and the second-best results are underlined.







Multimodal Input	Retrieved Exemplar from Cross-Modal Imagination Dataset	Paired Literal Exemplar	Our Response
 <p>下班走在路上的我 (After-work me on the way back)</p> <p>Imaginative Frame</p> <p>Image Domain: Canine Text Domain: Person</p> <p>Similar Attributes: relaxation, ...</p>	 <p>他觉得他的对手是一只 (He thinks that his opponent is a fox)</p> <p>源概念是狐狸，目标概念是他的对手 (The source concept is the fox and the target concept is his opponent)</p> <p>Imaginative Frame</p> <p>Source Domain: Canine Target Domain: Person</p> <p>Similar Attributes: Slyness, ...</p>	 <p>这是一只红色的狐狸。狐狸的脸部充满了细细的毛发。 (This is a red fox. The fox's face is covered with fine fur.)</p>	<p>是。源概念是狗，目标概念是我。 (YES. The source concept is the dog and the target concept is me.)</p> <p>✓</p>
<p>BEST FISHES ON YOUR BIRTHDAY</p>  <p>Imaginative Frame</p> <p>Image Domain: Water Text Domain: Aquatic Vertebrate</p> <p>Similar Attributes: Movement, ...</p>	 <p>The in that room always was hard Rock of Gibraltar. The source concept is the Rock and the target concept is the bed.</p> <p>Imaginative Frame</p> <p>Source Domain: Natural Object Target Domain: Furniture</p> <p>Similar Attributes: Solid, ...</p>	 <p>The image features a bedroom with a bed placed on a wooden floor.</p>	<p>NO.</p> <p>✗ (Wordplay)</p>

Figure 8: Two cases where our ImaRA-7B makes both correct and incorrect predictions on MET-Meme.

E Case Study

Figure 8 gives cases where our ImaRA makes both correct and incorrect predictions. We also provide the imaginative frame pairs that can calculate the highest frame similarity score used to retrieve an augmented exemplar from cross-modal imagination dataset in the imaginative frame augmented retrieval module. In the first case, our ImaRA retrieves a metaphorical exemplar that stimulates an imaginative frame with “Canine” as the source domain and “Person” as the target domain. This exemplar aids the MLLM in recognizing the metaphorical usage within the multimodal input, as it evokes a similar imaginative frame. We also find that our ImaRA performs less effectively in some cases involving wordplay. For example, the second case in

Figure 8 expresses a humorous birthday greeting that plays on the phrase “best wishes” by replacing “wishes” with “fishes”. Since our method focuses on capturing metaphorical meanings in multimodal contexts, it struggles to recognize such wordplay.

F Prompt Design of Baselines

Figure 9 provides prompt design details of GPT-4o and LLaVA baselines, respectively.

G Licenses of Scientific Artifacts

WordNet’s license is WordNet 3.0 license, while MET-Meme, CLIP, XLM-RoBERTa and spaCy are all licensed under the MIT license. LLaVA is released under the Apache-2.0 license, and ChatGPT operates under its respective API license.

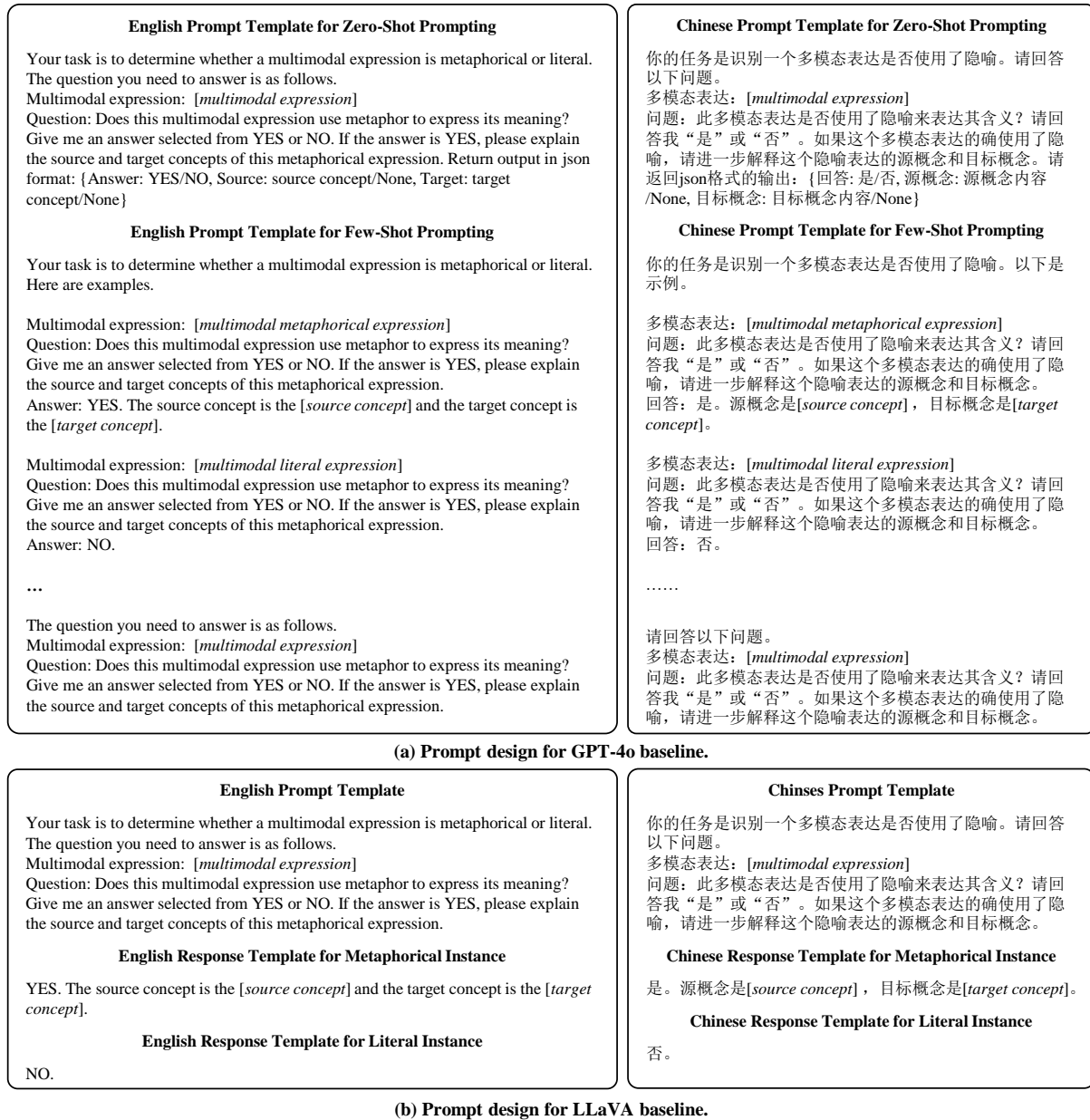


Figure 9: Prompt design for MLLM-based baselines. We randomly sampled three metaphorical instances and two literal instances from the training set as the few-shot exemplars for GPT-4o. Here, *[multimodal expression]* denotes the input slot for the multimodal expression of an instance in the multimodal input. The notations *[multimodal metaphorical expression]*, *[source concept]* and *[target concept]* denote the input slots for the multimodal expression, source concept and target concept in a metaphorical exemplar, respectively, and *[multimodal literal expression]* denotes the input slot for the multimodal expression in a literal exemplar.