Cultural Bias Matters: A Cross-Cultural Benchmark Dataset and Sentiment-Enriched Model for Understanding Multimodal Metaphors

Senqi Yang[1], Dongyu Zhang[1], Jing Ren[2], Ziqi Xu[2],
Xiuzhen Zhang[2], Yiliao Song[3], Hongfei Lin[1], Feng Xia[2]

[1]Dalian University of Technology, China,
[2]RMIT University, Australia, [3]The University of Adelaide, Australia

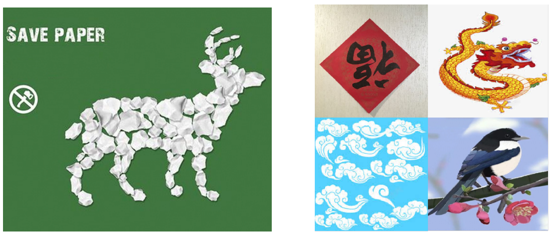
Correspondence: zhangdongyu@dlut.edu.cn, ziqi.xu@rmit.edu.au

Abstract

Metaphors are pervasive in communication, making them crucial for natural language processing (NLP). Previous research on automatic metaphor processing predominantly relies on training data consisting of English samples, which often reflect Western European or North American biases. This cultural skew can lead to an overestimation of model performance and contributions to NLP progress. However, the impact of cultural bias on metaphor processing, particularly in multimodal contexts, remains largely unexplored. To address this gap, we introduce MultiMM, a Multicultural Multimodal Metaphor dataset designed for cross-cultural studies of metaphor in Chinese and English. MultiMM consists of 8,461 text-image advertisement pairs, each accompanied by fine-grained annotations, providing a deeper understanding of multimodal metaphors beyond a single cultural domain. Additionally, we propose Sentiment-Enriched Metaphor Detection (SEMD), a baseline model that integrates sentiment embeddings to enhance metaphor comprehension across cultural backgrounds. Experimental results validate the effectiveness of SEMD on metaphor detection and sentiment analysis tasks. We hope this work increases awareness of cultural bias in NLP research and contributes to the development of fairer and more inclusive language models.[1]

1 Introduction

Metaphors appear in about one in three sentences and play a pivotal role in human cognition and communication (Steen et al., 2010; Shutova et al., 2010; Hu, 2023). In modern media, multimodal metaphors are more widely used than monomodal ones due to their superior ability to convey vivid and persuasive messages. A multimodal metaphor is a conceptual mapping from one source domain to a target domain, expressed through different combinations of modalities, such as text and image, text and sound, or image and sound (Forceville and Urios-Aparisi, 2009; Forceville, 2021; Zhang et al., 2025). For example, the text-image combination forming a deer shape in Figure 1(a) creates a metaphorical mapping from the source domain 'deer' to the target domain 'paper', symbolizing that preserving paper is akin to saving deer.



(a) A metaphor linking paper conservation to a deer.

(b) Illustrations of common Chinese metaphors.

Figure 1: Examples of metaphors across cultures.

The shift from monomodal to multimodal metaphors has created a growing need for improved comprehension of multimodal metaphors. However, comprehending them remains a significant challenge for machine-based systems. A key aspect of understanding multimodal metaphors is identifying the underlying mapping between source and target domains while also extracting the conveyed attributes (Su et al., 2021a; Ge et al., 2022a). Additionally, it involves deciphering implicit messages and recognizing semantic relationships between the source and target domains (Yang et al., 2013; Zhang et al., 2024).

Moreover, implicit messages and semantic relationships can change due to cultural variations, even though conceptual metaphors are deemed universal (Kövecses, 2010; Hong and Rossi, 2021). For instance, while the mapping from 'animal' to 'human' is universal, different cultures employ distinct linguistic and visual metaphors to express sim-

[1] Our dataset and code are available at `https://github.com/DUTIR-YSQ/MultiMM`.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26301–26317
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

ilar or even identical concepts. The phrase 'someone is a dinosaur' means 'old-fashioned' in English, while in Chinese, it conveys 'ugly'. To describe intoxication, Western cultures use animals such as 'newt' or 'skunk' (e.g., 'as drunk as a skunk'), whereas Chinese culture uses 'mud' (e.g., 'as drunk as mud'), highlighting a shift in the source domain. Furthermore, symbols of good fortune, such as 'fu', 'loong', 'auspicious clouds', and 'picapica' (as shown in Figure 1(b)), are prevalent in Chinese culture but absent in Western cultures.

Qualitative studies have investigated the interplay between metaphor and culture in fields such as cognitive linguistics (Richardson et al., 2021; Falck and Okonski, 2022; Kovaliuk, 2024), management (Piekkari et al., 2020; Musolff, 2022; Glaser et al., 2024), and psychology (Seering et al., 2022; Shokhrukhovna et al., 2024; Xu et al., 2024b). However, the impact of cultural biases on automatic metaphor processing has not been explored in depth. Moreover, the scarcity of metaphor resources from cultures outside Western Europe and North America inevitably leads to cultural bias in current automatic metaphor processing models.

To address this gap, we introduce a benchmark dataset and a sentiment-enriched model for cross-cultural research in multimodal metaphor processing. In summary, this paper makes the following contributions:

- We introduce MultiMM, a dataset of 8,461 text-image advertisement pairs, including 4,397 from Eastern cultures and 4,064 from Western cultures. To the best of our knowledge, MultiMM is the first dataset specifically designed for cross-cultural studies in multimodal metaphors.

- We propose a Sentiment-Enriched Metaphor Detection (SEMD) model that integrates sentiment embeddings to enhance metaphor comprehension across cultural contexts.

- We introduce two tasks: metaphor detection and sentiment analysis, to evaluate cross-cultural multimodal metaphor understanding. The evaluation results of eight textual, three visual, and seven multimodal baselines demonstrate the significant impact of cultural bias on metaphor processing.

- We present a new perspective on cultural bias in multimodal metaphor processing. Through the identification and analysis of these biases, our work aims to inspire future research that promotes awareness of cultural disparities and supports the creation of fairer, more inclusive NLP systems.

2 Related Work

2.1 Multimodal Metaphor Datasets

Most existing metaphor datasets are limited to text-only formats (Birke and Sarkar, 2006; Steen et al., 2010; Mohammad et al., 2016; Chen et al., 2023; Tong et al., 2024), resulting in a significant gap in multimodal metaphor research. To date, few multimodal metaphor datasets exist, and most originate from North American and Western European contexts (Shutova et al., 2016; Zhang et al., 2021)[2], with limited representation of other cultural perspectives (Xu et al., 2022; Lu et al., 2025; Zhang et al., 2023)[3]. In contrast, MultiMM provides diverse annotations for metaphor understanding across Eastern and Western cultures, addressing the limitations of existing benchmarks.

2.2 Metaphor Understanding

Early studies on metaphor comprehension in textual data primarily relied on manually crafted knowledge (Mason, 2004). Subsequent research explored distributional clustering (Shutova et al., 2013) and unsupervised learning approaches (Shutova et al., 2017; Mao et al., 2018). More recently, deep learning models have been employed to improve metaphor comprehension. With the advent of large language models (LLMs), further progress has been made in this domain (Liu et al., 2020; Choi et al., 2021; Stowe et al., 2021; Ge et al., 2022b; Aghazadeh et al., 2022; Wachowiak and Gromann, 2023; Tian et al., 2024). While some studies have explored multimodal metaphor processing, they have primarily focused on extracting and integrating textual and visual features (Shutova et al., 2016; Kehat and Pustejovsky, 2020; Su et al., 2021b; Xu et al., 2024c,a). Distinct from prior work, our model SEMD incorporates a cultural perspective by leveraging sentiment information, which is a universally recognized feature for multimodal metaphor understanding.

[2] https://github.com/DUTIR-YSQ/MultiMET

[3] https://github.com/DUTIR-YSQ/MultiCMET

Item	CN	EN	Total
Total	4,397	4,064	8,461
Metaphorical	2,583	2,189	4,772
Literal	1,814	1,875	3,689
Total Words	145,312	68,189	213,501
Average Words	33	15	24
Training Set Size	3,517	3,251	6,768
Validation Set Size	440	406	846
Test Set Size	440	407	847

Table 1: Statistical overview of MultiMM. CN refers to Chinese data, while EN refers to English data.

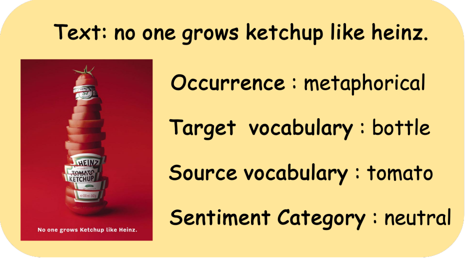

Text: no one grows ketchup like heinz.
Occurrence : metaphorical
Target vocabulary : bottle
Source vocabulary : tomato
Sentiment Category : neutral
HEINZ
TOMATO KETCHUP
No one grows Ketchup like Heinz.

Figure 2: Example of an advertisement with annotations. This figure shows an English advertisement in which stacked tomatoes visually form a ketchup bottle, accompanied by the text 'No one grows ketchup like Heinz'. The example is annotated as metaphorical, with 'tomato' as the source vocabulary, 'bottle' as the target vocabulary, and the sentiment labeled as neutral.

3 The MultiMM Dataset

3.1 Data Collection and Filter

MultiMM aims to provide a cross-cultural, labeled dataset for automated multimodal metaphor comprehension. Given that metaphorical content is primarily textual or visual (Steen et al., 2010; Akula et al., 2023), we collect data from commercial and public service advertisements incorporating both textual and visual elements to enable the analysis of visual and linguistic features for multimodal metaphor understanding. A statistical overview of MultiMM is presented in Table 1.

Chinese Advertisement Collection. MultiMM contains 4,397 Chinese samples, which native Chinese-speaking researchers collect by searching for Chinese keywords via Baidu [4]. We compile a set of keywords related to 'advertisement' and 'metaphor', encompassing three main categories: everyday products (e.g., 手机 [mobile], 汽车 [car]), public service topics (e.g., 吸烟 [smoking], 欺凌 [bullying]), and metaphorically relevant linguistic concepts (e.g., 愤怒 [anger], 颜色 [color]).

We also refer to the Master Metaphor List (Lakoff, 1994) to select target and source domains in conceptual metaphors, using keywords such as 变化 [change], 情感 [sentiment], 人们 [people], and 信念 [beliefs]. In addition, we collect potential Chinese metaphorical samples that contain both images and text from Chinese commercial advertisements released in 2021, following the iFlytek Advertising Picture Classification Competition [5].

English Advertisement Collection. The 4,064 English samples come from a public dataset containing product and public service advertisements with both images and text (Ye et al., 2021). We further clean the data by: (1) removing duplicate images based on their MD5 encoding (Rivest, 1992); (2) manually removing images that are not advertisements; (3) removing images that are blurry or smaller than 350 × 350 pixels; (4) extracting the embedded text using the optical character recognition (OCR) technique and manually correcting the results.

3.2 Annotation Model

We annotate the text-image advertisement pairs based on the following criteria: (1) the occurrence of metaphors (literal or metaphorical); (2) target and source domain relations, including target/source vocabulary in text and verbalized target/source vocabulary in images; (3) sentiment category, classified as negative, neutral, or positive.

The annotation model is defined as AnnotationModel = (Occurrence, Target, Source, SentimentCategory). Figure 2 illustrates an example of an advertisement with annotations. Additionally, we provide detailed annotation guidelines via the link in the Abstract.

3.3 Data Annotation

Metaphorical or literal. Following MultiMET (Zhang et al., 2021), we identify metaphorical and literal expressions in both verbal and visual forms. For each identified metaphor, annotators specify its domains: textual metaphors derive source and target domains from the original words, while visual metaphors require annotators to verbalize domain words inferred from the image. Adopting the approach of Šorm and Steen (2018), annotators detect metaphorical text-image pairs by identifying incongruous elements and explaining a non-reversible 'A is B' identity relation, which signifies two domains

[4] https://www.baidu.com/

[5] https://aistudio.baidu.com/aistudio/datasetdetail/102279

expressed across different modalities.

Sentiment categories. Understanding metaphors involves identifying domain mappings and analyzing linguistic properties, such as sentiment. Metaphors often have a stronger emotional impact than literal expressions (Mohammad et al., 2016; Schnepf and Christmann, 2022). To explore this, we annotated sentiment in MultiMM as negative, neutral, or positive to compare its impact on metaphorical and literal expressions across multicultural and multimodal contexts.

3.4 Annotation Process and Quality Control

We use an expert-based approach to annotate data with eight researchers: five Chinese native speakers (grouped into 2-2-1) annotate Chinese data, and three English native speakers (grouped into 2-1) annotate English data. The two-expert groups annotate data, while the one-expert groups resolve disagreements. If agreement is not reached, all experts discuss to finalize the annotation. To ensure quality, we implement strict annotation standards and provide detailed documentation with instructions, examples, and cautions. Training sessions are conducted before each annotation round, and all training materials and documents are continuously adjusted to address new challenges. Additionally, we pre-label a small dataset to identify and resolve potential issues early in the process.

Two strategies are used to mitigate cultural bias during annotation: Diverse Annotator Selection and Feedback Mechanisms. Diverse Annotator Selection ensures that annotators have multicultural backgrounds. For example, two of our five Chinese expert annotators have lived and studied in Western countries for over six years, while the other three have resided there for more than two years. Similarly, all three English annotators have over two years of experience living and working in China or other Eastern countries. This helps minimize the influence of any single cultural perspective. The Feedback Mechanisms allow annotators to share insights and flag potential cultural biases during the annotation process. The open communication channel also ensures prompt issue resolution and improves annotation quality.

To measure the consistency of the classification, we use the Fleiss' Kappa score (κ) (Fleiss, 1971), a statistical metric that evaluates the agreement among three or more raters while accounting for chance agreement. The Fleiss' kappa score ranges from -1 to 1, where $\kappa > 0.6$ is considered substantial agreement, while $\kappa > 0.8$ indicates near-perfect agreement. The score for metaphor identification is $\kappa = 0.73$, for target domain category identification is $\kappa = 0.70$, for source domain category identification is $\kappa = 0.66$, and for sentiment category identification is $\kappa = 0.82$. These results indicate that the annotation process is reliable.

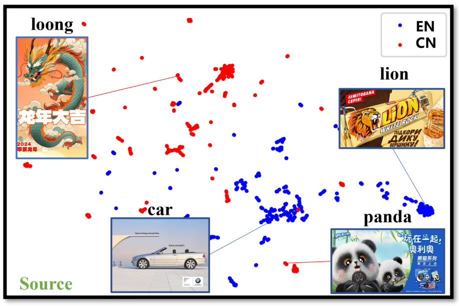


Figure 3: The distribution of source domain vocabulary in Chinese (CN) and English (EN) advertisements.

4 Data Analysis

4.1 Metaphor Distribution

As shown in Figure 3, the distribution of source domain vocabulary in English and Chinese samples differs significantly. English advertisements often utilize concrete, universally recognized source vocabulary to evoke immediate associations. Animals such as lions and eagles symbolize strength and freedom, while vehicles like rockets and sports cars represent innovation and luxury. In contrast, Chinese advertisements frequently draw on source vocabulary that reflects traditional symbolism and societal values. Animals such as loong and pandas symbolize power and national pride, while natural elements like bamboo and lotus flowers represent resilience and purity.

Despite these cultural differences, certain source-target pairs appear in both English and Chinese advertisements, reflecting universal human experiences or the homogenizing effects of globalization. For instance, the cheetah is frequently associated with automotive products in both cultures to emphasize speed. Similarly, water often symbolizes cleaning agents or bottled beverages to convey purity, though its visual representation may vary across cultures. However, many metaphorical relationships remain culture-specific, shaped by distinct social and linguistic contexts. Natural elements such as fire exemplify this divergence: in English advertisements, fire is often linked to luxury perfumes to evoke passion, as seen in campaigns featuring fiery

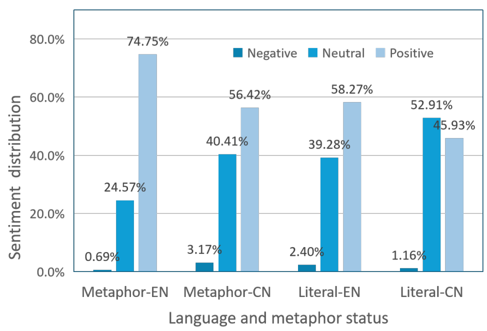


Figure 4: Sentiment category distribution in metaphorical and literal data from Chinese (CN) and English (EN) advertisements.

motifs, whereas, in Chinese contexts, fire is rarely used for such purposes.

4.2 Sentiment Category Distribution

Figure 4 compares sentiment distributions in English and Chinese metaphorical and literal advertisement data. In metaphorical advertisements, English data exhibits a pronounced emphasis on positive sentiment (74.75%), with neutral sentiment at 24.57% and no negative sentiment (0.69%). This contrasts with Chinese metaphorical advertisements, where positive sentiment remains dominant (56.42%) but is tempered by a higher level of neutrality (40.41%) and a small presence of negativity (3.17%). These patterns suggest that while metaphors in both languages are leveraged to evoke positivity, English advertisements employ figurative language more assertively to amplify optimism, whereas Chinese advertisements strike a balance between positivity and neutrality. The complete absence of negative sentiment in English metaphorical advertisements may reflect cultural taboos against associating metaphors with negativity in advertising. In contrast, the minimal negative sentiment in Chinese metaphorical advertisements could serve as a strategic rhetorical device, such as juxtaposing challenges with solutions to enhance persuasive impact.

In literal contexts, sentiment dynamics shift significantly. In English advertisements, positive sentiment declines sharply to 58.27%, while neutral sentiment rises to 39.28%, and negative sentiment emerges at 2.40%. Chinese literal advertisements exhibit an even more pronounced shift: neutral sentiment dominates (52.91%), surpassing positive sentiment (45.93%), with negative sentiment remaining minimal (1.16%). The data confirms that metaphors consistently convey richer emotional resonance than literal expressions across both languages. These findings align with prior studies (Mohammad et al., 2016; Stanley et al., 2021), which argue that metaphors amplify emotional engagement by activating deeper cognitive and affective associations.

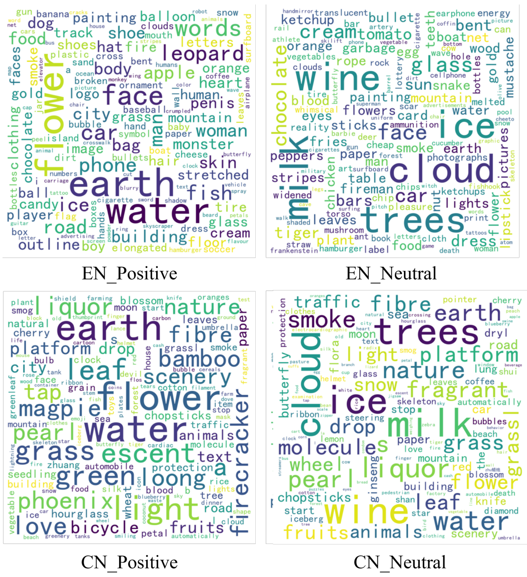


Figure 5: Word clouds of source vocabulary across sentiment categories in Chinese (CN) and English (EN) advertisements. The Chinese word cloud is translated into English for better understanding.

4.3 Sentiment Vocabulary Analysis

Our word clouds, shown in Figure 5, reveal significant overlaps between Chinese and English advertisements. Both languages frequently convey positive sentiments by utilizing words linked to universal symbols, such as ‘earth’ (environment) and ‘water’ (purity). These words tap into shared human experiences and cultural associations. Neutral sentiments are often conveyed through vocabulary such as ‘trees’ in both English and Chinese advertisements, with subtle variations influenced by cultural nuances.

However, it is important to note that the same sentiment may be expressed through different vocabulary in the two languages. For example, a metaphor used to express love or passion in English (such as ‘fire’) may not carry the same association in Chinese, where different symbols might evoke similar emotional responses. Conversely, similar vocabulary can be employed in both languages to convey different sentiments. For instance, "storm" might be used in English to express tur-

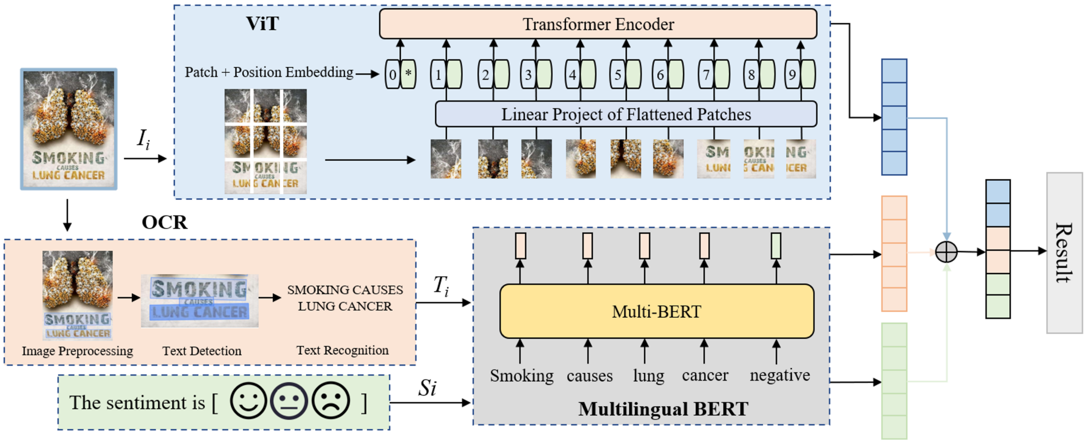


Figure 6: The SEMD framework consists of three main branches: image features extracted by ViT, text features extracted from OCR and encoded by BERT, and emotional features from sentiment analysis. These are fused for final metaphor prediction.

moil, whereas in Chinese, the same word could evoke feelings of renewal or transformation.

5 Methodology

While most metaphor understanding models are developed for English datasets, their effectiveness may decline on Chinese datasets. Existing studies indicate that sentimental cognition is influenced by our shared neurophysiological structure (Mauro et al., 1992; Ortony et al., 2022) and is thus universal across different cultures (Wallbott and Scherer, 1986). Motivated by this, we develop a generalized model, Sentiment-Enriched Metaphor Detection (SEMD), which integrates sentiment information as an auxiliary feature. The model architecture is illustrated in Figure 6.

The model consists of three main components: two encoders, a fusion layer, and a final classifier. First, the model employs BERT (Devlin et al., 2019) as the text encoder to process textual content and sentiment information, and ViT (Dosovitskiy et al., 2020) as the image encoder to process representations of text and image modalities. Both encoders output 768-dimensional feature vectors, denoted as T_i, S_i, and I_i for text, sentiment, and image, respectively. The fusion layer then applies a cascading strategy to combine these feature vectors. Once the fused vectors are obtained, a feed-forward network extracts intrinsic information from the text and images. Finally, these vectors pass through a fully connected layer, and the model generates the final results by applying the sigmoid function to the output vector.

For the metaphor detection task, our goal is to determine whether a textual-visual pair contains a metaphor. By extracting text, image, and sentiment features, concatenating them, and passing them through the feed-forward network, we obtain the prediction results after applying the activation function:

$$P_{Meta} = \text{Sigmoid}(\text{Fusion}(\text{concat}(I_i, T_i, S_i))).$$

For the sentiment analysis task, our objective is to classify the sentimental inclination of textual-visual pairs into negative, neutral, or positive sentiments. In this task, feature extraction is performed solely on text and image inputs, and the subsequent network framework is utilized to obtain the sentiment classification results:

$$P_{Senti} = \text{Sigmoid}(\text{Fusion}(\text{concat}(I_i, T_i))).$$

6 Experiments

6.1 Experimental Settings

The proposed dataset, MultiMM, is evaluated using 18 existing baselines and one proposed multimodal model (i.e., SEMD). All models are built with the PyTorch framework (Paszke et al., 2019) and implemented using the Hugging Face library. The base networks of the model include multilingual BERT [6] and ViT [7]. We employ AdamW (Loshchilov and Hutter, 2018) as the optimizer and use the cross-entropy loss as the primary loss function. During training, the parameters of the pre-trained models

[6] https://huggingface.co/bert-base-multilingual-cased

[7] https://huggingface.co/google/vit-base-patch16-224

are frozen. Model performance is evaluated based on accuracy, macro precision, and macro F1-score.

The following hyperparameters are applied for optimal performance: an embedding size of 768 to ensure robust text representation, a dropout rate of 0.3 to prevent overfitting, a maximum text length of 30 tokens to balance context capture and computational efficiency, 10 training epochs for convergence, a batch size of 64 for efficient mini-batch processing, and a learning rate ranging from 3e-5 to 5e-4 for fine-tuning weight updates.

6.2 Baselines

In this section, we briefly introduce the 18 baselines used in our experiment, spanning three modalities: textual (8 models), visual (3 models), and multimodal (7 models).

For the text modality, we select mBERT (Devlin et al., 2019), Unsupervised Cross-lingual Representation Learning at Scale (XLM-R) (Conneau et al., 2020), SixTP (Chen et al., 2022), Language-agnostic BERT Sentence Embedding (LaBSE) (Ansell et al., 2022), GPT-3.5-TURBO (Gao et al., 2023), LLaMA2-8B (Touvron et al., 2023), LLaMA3.1-70B (Dubey et al., 2024) and Deepseek-R1-70B(DeepSeek-AI et al., 2025). For the imaging modality, we apply VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), and ViT (Dosovitskiy et al., 2020) to extract image features. For multimodal analysis, we select models designed for metaphorical language and adapt their text encoders to multilingual versions. These models include mBERT-Res (Zhang et al., 2021), which integrates multilingual BERT and ResNet50; Cross-Modal Graph Convolutional Networks (CMGCN) (Liang et al., 2022), which leverage a cross-modal graph structure; Caption Enriched Samples (CES) (Blaier et al., 2021), which incorporates image captions; MET (Xu et al., 2022), which integrates annotated vocabularies; and MFC (Maity et al., 2022), designed for sentiment and irony recognition. Additionally, we consider multimodal LLMs, including LLaVA (Liu et al., 2023), GPT-4o (Hurst et al., 2024), Gemini (Team et al., 2023), and Qwen2.5-VL (Bai et al., 2025) models at the 3B, 7B, and 72B parameter scales.

6.3 Performance Analysis

Table 2 presents the comparative results for metaphor detection and sentiment analysis. The main text reports F1 scores, with detailed results in Appendix A.1. The *Random* entry denotes predictions made without training, using random guesses.

Model	Metaphor Detection F1 (%)		Sentiment Analysis F1 (%)	
	EN	CN	EN	CN
Random	46.70	41.20	40.41	35.02
mBERT	64.00	65.52	62.83	58.43
XLM_R	69.90	62.38	69.44	66.21
SixTP	68.61	65.60	64.43	65.59
LaBSE	68.15	66.41	69.93	69.02
GPT-3.5-TURBO	30.95	15.49	42.06	52.08
Llama2-8B	22.58	19.73	40.11	50.72
Llama3.1-70B	44.04	49.70	33.46	30.90
Deepseek R1-70B	44.06	47.95	53.38	49.01
VGG16	70.64	67.85	59.90	59.02
ResNet50	71.46	68.56	57.49	58.86
ViT	71.67	69.04	61.46	62.17
mBERT-Res	77.83	74.40	69.75	68.84
CMGCN	79.04	74.91	72.79	70.19
CES	78.45	75.53	73.24	70.26
MET	76.04	72.51	71.95	67.17
MFC	75.84	73.25	71.93	69.60
GPT-4O	64.00	67.00	36.00	29.00
LLaVA	73.31	69.84	56.72	53.21
Gemini	64.19	68.51	53.61	43.48
Qwen2.5-VL-3b	57.30	41.69	52.50	55.26
Qwen2.5-VL-7b	56.62	54.51	50.71	55.15
Qwen2.5-VL-72B	59.12	67.66	50.44	57.17
SEMD	**80.16**	**77.79**	**75.69**	**70.51**

Table 2: Experimental results on metaphor detection and sentiment analysis tasks. The best overall results are highlighted in **bold**, while the best results within each modality are underlined. Models are categorized by modality, with textual, visual, and multimodal models highlighted in cyan, green, and orange, respectively.

Metaphor Detection. In the textual modality, XLM-R achieves the best performance in English tasks, while LaBSE excels in Chinese tasks, benefiting from its multilingual alignment capabilities. However, GPT-3.5 TURBO and the Llama series underperform in both metaphor and sentiment recognition, particularly in Chinese tasks, where F1 scores are notably low, highlighting their limitations in handling complex metaphors. In the imaging modality, ViT outperforms ResNet50 and VGG models in both languages, demonstrating strengths in global information extraction and visual representation. The imaging modality generally surpasses the textual modality, suggesting that rich metaphorical features in advertising images enhance recognition. Among multimodal models, CMGCN and CES perform best in English and Chinese tasks, respectively, proving the effectiveness

Sentiment Feature	Fusion Method	EN F1 (%)	CN F1 (%)
w/o	add	75.95	72.05
	max	76.29	74.09
	concat	78.64	74.84
w/	add	77.76	74.37
	max	78.59	74.37
	concat	**80.88**	**77.39**

Table 3: Results of the ablation study. The best and second-best results are highlighted in **bold** and underline, respectively.

		mBERT	mBERT-Res	SEMD
EN → CN	Acc (%)	64.29	76.35	78.57
	Pre (%)	63.72	76.73	78.19
	F1 (%)	59.70	75.63	77.64
CN → EN	Acc (%)	65.22	73.86	75.45
	Pre (%)	67.46	75.90	77.73
	F1 (%)	65.31	73.96	75.53

Table 4: The results of metaphor detection tasks after Chinese-English translation. Experimental results on metaphor detection task via bidirectional Chinese-English translation.

of modeling interactions between images and text. In contrast, GPT-4o, LLaVA, and Gemini exhibit relatively weak performance on multimodal tasks, particularly on Chinese tasks. Notably, Qwen2.5-VL demonstrates a corresponding improvement in metaphor detection accuracy as its parameter size increases.

Sentiment Analysis. In the textual modality, LaBSE leads in both Chinese (69.02%) and English (69.93%), possibly due to the longer average length of Chinese texts, which enhances contextual understanding. However, advanced LLMs such as GPT-3.5, Deepseek, and the Llama series perform poorly. Among them, GPT-3.5 TURBO achieves an F1 score of only 52.08% in Chinese, DeepseekR1-70B reaches 49.01%, Llama2-8B scores 50.72%, and Llama3.1-70B performs even worse, with a score of just 33.46%. In the visual modality, ViT leads in both Chinese and English, highlighting the importance of visual information in sentiment prediction. Multimodal models excel, particularly in English tasks, as CMGCN and CES achieve high F1 scores in both languages, demonstrating the effectiveness of multimodal approaches. In contrast, GPT-4o, LLaVA, Gemini, and Qwen2.5-VL struggle with sentiment analysis.

Our proposed model, SEMD, outperforms all baselines in metaphor detection and sentiment analysis by incorporating sentiment information as an auxiliary feature. As shown in Table 2, SEMD achieves the highest F1 scores, particularly in Chinese tasks, demonstrating its robustness in cross-cultural and multimodal metaphor understanding.

6.4 Ablation Study

To evaluate the impact of fusion methods and sentiment features, we conduct an ablation study under metaphor detection tasks. We compare three fusion methods (add, max, and concat) and test the effects of removing sentiment features. As shown in Table 3, the concat fusion method effectively integrates text and image features, achieving the best performance across both Chinese and English datasets. Furthermore, including sentiment features significantly enhances recognition performance compared to their absence. Note that the proposed SEMD uses the concat fusion method with sentiment features, which corresponds to the last entry in the table.

6.5 Metaphor Detection via Bidirectional Chinese-English Translation

To explore issues related to language understanding and expression in cross-cultural communication, we conduct supplementary experiments using Chinese-English translated texts for metaphor detection tasks. We evaluate the performance of the pure text model mBERT, the multimodal model mBERT-Res, and our proposed model on this task. As shown in Table 4, the results indicate a performance drop in metaphor detection when using translated texts compared to the original versions. We attribute this decline to two main factors.

First, significant differences in language structure and grammar often lead to the loss of semantic and rhetorical nuances of metaphors during direct translation, thereby reducing the accuracy of metaphor detection.

Second, the understanding and expression of metaphors heavily rely on cultural background and contextual factors. Variations in values, belief systems, customs, and historical backgrounds across cultures result in different interpretations and perceptions of the same metaphor. As a result, simple language conversion often fails to fully capture and reproduce the subtle cultural and pragmatic meanings embedded in metaphors.

6.6 Error Analysis

We conduct an error analysis of the test set prediction results for metaphor detection and sentiment analysis tasks. The error analysis of metaphor detection shows that the accuracy of Chinese metaphor detection is 73%, lower than the 86% accuracy for English. This discrepancy may result from cultural and ideological influences in the advertising data. Chinese metaphors tend to be more subtle, while English metaphors are more straightforward. The error analysis of sentiment analysis reveals that although the Chinese dataset has a higher proportion of neutral sentiment, its recognition accuracy is only 55%, compared to 62% for English. The low performance in Chinese is due to lexical ambiguity and polysemy. Many misclassified cases contain positive words (e.g., cherish [珍爱], civilization [文明]) but lack an overall positive sentiment, leading to errors. Due to page limitations, detailed results and discussion are provided in Appendix A.2.

In addition, to vividly illustrate the profound impact of cultural context on metaphor understanding, we also provides a case study in Appendix B.

7 Conclusion

Metaphors play a fundamental role in communication, yet NLP research predominantly relies on English datasets, introducing cultural biases that limit cross-cultural applicability. To address this, we introduce MultiMM, a benchmark dataset designed for multimodal metaphor studies in Chinese and English. MultiMM consists of 8,461 text-image advertisement pairs with annotations, enabling a more comprehensive evaluation of metaphor processing across cultures. We also propose SEMD, a sentiment-enriched model that integrates sentiment embeddings to enhance multilingual and multimodal metaphor comprehension. Experimental results highlight the significant impact of cultural bias on metaphor detection and sentiment analysis. SEMD outperforms all baselines, demonstrating its robustness in cross-cultural and multimodal metaphor understanding. We hope MultiMM serves as a valuable resource for multicultural multimodal metaphor processing and that SEMD offers insights for improving cross-cultural NLP models.

Limitations

While MultiMM is a valuable resource for studying multimodal metaphors, it has limitations that present opportunities for future research. First, the dataset is currently limited to advertising, as annotating multimodal metaphors across genres (e.g., social media, news, literature) poses challenges in consistency and scalability. Expanding to these domains would enable a more comprehensive analysis of multimodal metaphors in diverse contexts.

Second, our work focuses on English and Chinese, representing Western and Eastern cultures, respectively. While this provides a foundational comparison, it does not capture the full diversity of global languages and cultures. Future research could extend MultiMM to more languages and cultural frameworks, deepening insights into how multimodal metaphors vary across linguistic and cultural landscapes.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful and valuable comments. This work was supported by the Key Program of the National Language Commission (No. ZDI145-80) and the National Natural Science Foundation of China (NSFC) (No. 62076051).

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, and 1 others. 2023. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23201–23211.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

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

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336, Trento, Italy. Association for Computational Linguistics.

Efrat Blaier, Itzik Malkiel, and Lior Wolf. 2021. Caption enriched samples for improving hateful memes detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9350–9358, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2022. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 142–157, Dublin, Ireland. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

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 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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

Marlene Johansson Falck and Lacey Okonski. 2022. Procedure for identifying metaphorical scenes (pims): A cognitive linguistics approach to bridge theory and practice. *Cognitive semantics*, 8(2):294–322.

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

Charles Forceville. 2021. Multimodality. In *The Routledge handbook of cognitive linguistics*, pages 676–687. Routledge.

Charles J. Forceville and Eduardo Urios-Aparisi, editors. 2009. *Multimodal Metaphor*, volume 11. De Gruyter Mouton, Berlin, New York.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Mengshi Ge, Rui Mao, and Erik Cambria. 2022a. Explainable metaphor identification inspired by conceptual metaphor theory. In *Proceedings of the AAAI conference on artificial intelligence*, 10, pages 10681–10689.

Mengshi Ge, Rui Mao, and Erik Cambria. 2022b. Explainable metaphor identification inspired by conceptual metaphor theory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10681–10689.

Vern L Glaser, Jennifer Sloan, and Joel Gehman. 2024. Organizations as algorithms: A new metaphor for advancing management theory. *Journal of Management Studies*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Wenjie Hong and Caroline Rossi. 2021. The cognitive turn in metaphor translation studies: A critical overview. *Journal of Translation Studies*, 5(2):83–115.

Zhuanglin Hu. 2023. *Metaphor and cognition*. Springer.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Gitit Kehat and James Pustejovsky. 2020. Improving neural metaphor detection with visual datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5928–5933, Marseille, France. European Language Resources Association.

Yurii Kovaliuk. 2024. Connecting idioms and metaphors: Where cognitive linguistics meets cognitive stylistics. *British and American Studies Journal*, 30:157–170.

Zoltán Kövecses. 2010. Metaphor, language, and culture. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 26:739–757.

George Lakoff. 1994. *Master metaphor list*. University of California.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multimodal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777, Dublin, Ireland. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Jerry Liu, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin. 2020. Metaphor detection using contextual word embeddings from transformers. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 250–255, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Xingyuan Lu, Yuxi Liu, Dongyu Zhang, Zhiyao Wu, Jing Ren, and Feng Xia. 2025. Emometa: A multimodal dataset for fine-grained emotion classification in chinese metaphors. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 3080–3083.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1739–1749, New York, NY, USA. Association for Computing Machinery.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia. Association for Computational Linguistics.

Zachary J Mason. 2004. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational linguistics*, 30(1):23–44.

Robert Mauro, Kaori Sato, and John Tucker. 1992. The role of appraisal in human emotions: a cross-cultural study. *Journal of personality and social psychology*, 62(2):301.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Andreas Musolff. 2022. War against covid-19": Is the pandemic management as war metaphor helpful or hurtful. *A. Musolff, R. Breeze, K. Kondo, & S. Vilar-Lluch (Eds.), Pandemic and crisis discourse: Communicating Covid-19 and public health strategy*, pages 307–320.

Andrew Ortony, Gerald L Clore, and Allan Collins. 2022. *The cognitive structure of emotions*. Cambridge university press.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Rebecca Piekkari, Susanne Tietze, and Kaisa Koskinen. 2020. Metaphorical and interlingual translation in moving organizational practices across languages. *Organization Studies*, 41(9):1311–1332.

Peter Richardson, Charles M Mueller, and Stephen Pihlaja. 2021. *Cognitive Linguistics and religious language: An introduction*. Routledge.

Ronald Rivest. 1992. The md5 message-digest algorithm. Technical report.

Julia Schnepf and Ursula Christmann. 2022. "it's a war! it's a battle! it's a fight!": Do militaristic metaphors increase people's threat perceptions and support for covid-19 policies? *International Journal of Psychology*, 57(1):107–126.

Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. *New Media & Society*, 24(3):621–640.

Ulugova Shokhida Shokhrukhovna, Uktamova Malika Khasanovna, and Abdullayeva Parvina. 2024. Metaphorical insights into human psychology in jane eyre. *SPAST Reports*, 1(5).

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California. Association for Computational Linguistics.

Ekaterina Shutova, Lin Sun, Elkin Darío Gutiérrez, Patricia Lichtenstein, and Srini Narayanan. 2017. Multilingual Metaphor Processing: Experiments with Semi-Supervised and Unsupervised Learning. *Computational Linguistics*, 43(1):71–123.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010, Beijing, China. Coling 2010 Organizing Committee.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical Metaphor Processing. *Computational Linguistics*, 39(2):301–353.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Esther Šorm and Gerard Steen. 2018. Towards a method for visual metaphor identification. *Visual metaphor: Structure and process*, 18:47–88.

B Liahnna Stanley, Alaina C Zanin, Brianna L Avalos, Sarah J Tracy, and Sophia Town. 2021. Collective emotion during collective trauma: A metaphor analysis of the covid-19 pandemic. *Qualitative Health Research*, 31(10):1890–1903.

Gerard Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, Trijntje Pasma, and 1 others. 2010. A method for linguistic metaphor identification. *Amsterdam: Benjamins*.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6724–6736, Online. Association for Computational Linguistics.

Chang Su, Weijie Chen, Ze Fu, and Yijiang Chen. 2021a. Multimodal metaphor detection based on distinguishing concreteness. *Neurocomputing*, 429:166–173.

Chang Su, Kechun Wu, and Yijiang Chen. 2021b. Enhanced metaphor detection via incorporation of external knowledge based on linguistic theories. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1280–1287, Online. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yuan Tian, Ruike Zhang, Nan Xu, and Wenji Mao. 2024. Bridging word-pair and token-level metaphor detection with explainable domain mining. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13311–13325, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Harald G Wallbott and Klaus R Scherer. 1986. How universal and specific is emotional experience? evidence from 27 countries on five continents. *Social Science Information*, 25(4):763–795.

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 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2887–2899, New York, NY, USA. Association for Computing Machinery.

Bo Xu, Junzhe Zheng, Jiayuan He, Yuxuan Sun, Hongfei Lin, Liang Zhao, and Feng Xia. 2024a. Generating multimodal metaphorical features for meme understanding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 447–455.

Xiaobing Xu, Miaolei Jia, and Rong Chen. 2024b. Time moving or ego moving? how time metaphors influence perceived temporal distance. *Journal of Consumer Psychology*, 34(3):466–480.

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024c. Exploring chain-of-thought for multimodal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101.

Fan-Pei Gloria Yang, Kailyn Bradley, Madiha Huq, Dai-Lin Wu, and Daniel C. Krawczyk. 2013. Contextual effects on conceptual blending in metaphors: An event-related potential study. *Journal of Neurolinguistics*, 26(2):312–326.

Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. 2021. Interpreting the rhetoric of visual advertisements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1308–1323.

Dongyu Zhang, Xingyuan Lu, Mulin Zhuang, Senqi Yang, and Hongjun Chen. 2025. Multimodal metaphor recognition based on chain-of-cognition prompting. *Cognitive Systems Research*, 91:101356.

Dongyu Zhang, Jingwei Yu, Senyuan Jin, Liang Yang, and Hongfei Lin. 2023. Multicmet: A novel chinese benchmark for understanding multimodal 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 multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225, Online. Association for Computational Linguistics.

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*, 8, pages 9341–9349.

A Supplementary Experimental Results

A.1 Performance Analysis

In this section, we provide detailed results for metaphor detection and sentiment analysis on 18 baselines. We use accuracy (Acc), precision (Pre), and F1 score (F1) as metrics to compare performance.

Metaphor Detection. We present the detailed results of metaphor detection in Table 5. The Random entry denotes predictions made by the model without training, relying purely on random guessing. Overall, multimodal approaches significantly outperform both text-only and image-only models, demonstrating their ability to leverage complementary information from both modalities. Additionally, English datasets yield better results than their Chinese counterparts, likely due to the abundance of pre-trained English-language data, which facilitates model learning.

Among text-based models, XLM-R achieves the highest performance for English, while LaBSE demonstrates superior results for Chinese, highlighting its robust multilingual alignment capabilities. The strong performance of these models underscores the importance of cross-lingual knowledge transfer in metaphor comprehension.

In the visual domain, ViT outperforms both ResNet50 and VGG models in both languages, reinforcing the superiority of Transformer-based architectures in capturing global image features. The generally stronger performance of image models over text models suggests that advertising images contain rich metaphorical elements, which models can effectively leverage for metaphor detection.

Among multimodal approaches, CMGCN and CES stand out as top performers for English and Chinese, respectively. CMGCN excels at fine-grained feature extraction from images, while CES enhances metaphor comprehension by incorporating image captions as global auxiliary features, effectively modeling the interplay between text and images. Notably, mBERT-Res, which replaces VGG with ResNet, outperforms MET, highlighting the critical role of strong visual feature extraction in improving overall performance. In contrast, MFC, a multitask model jointly trained on metaphor and sentiment recognition, struggles to achieve competitive results, likely due to interference between the two learning objectives.

Our proposed model, SEMD, achieves the best overall results, validating the effectiveness of sentiment-aware approaches in multimodal metaphor processing.

Sentiment Analysis. Table 6 presents the results of sentiment recognition on the test set. As in metaphor detection tasks, multimodal approaches generally outperform unimodal ones, and English results surpass those in Chinese. However, unlike metaphor detection, the text modality outperforms

	EN			CN		
Model	Acc (%)	Pre (%)	F1 (%)	Acc (%)	Pre (%)	F1 (%)
Random	50.25	46.21	46.70	45.95	52.88	41.20
mBERT	66.50	66.89	64.00	65.65	65.44	65.52
XLM_R	69.70	70.53	69.90	65.43	65.84	62.38
SixTP	68.71	68.55	68.61	66.74	66.30	65.60
LaBSE	68.72	68.31	68.15	66.96	66.49	66.41
GPT-3.5-TURBO	42.85	52.00	30.95	45.45	68.75	15.49
Llama2-8B	40.74	47.29	22.58	44.54	57.69	19.73
Llama3.1-70B	45.18	53.40	44.04	49.70	57.80	49.70
Deepseek R1-70B	45.00	52.81	44.06	52.00	61.34	47.95
VGG16	70.44	71.60	70.64	68.27	67.84	67.85
ResNet50	72.91	73.92	71.46	71.55	75.35	68.56
ViT	73.40	75.14	71.67	71.99	75.99	69.04
mBERT-Res	77.83	80.61	77.83	75.49	76.46	74.40
CMGCN	**79.56**	79.96	79.04	75.93	76.65	74.91
CES	78.57	78.46	78.45	76.15	76.24	75.53
MET	76.15	76.01	76.04	74.18	75.74	72.51
MFC	77.34	79.97	75.84	74.84	76.74	73.25
GPT-4O	54.00	63.00	64.00	57.00	62.00	67.00
LLaVA	59.11	59.06	73.31	56.81	58.20	69.84
Gemini	64.75	64.19	64.19	69.00	68.72	68.15
Qwen2.5-VL-3b	56.5	59.01	57.30	32.78	32.78	41.69
Qwen2.5-VL-7b	63.00	60.08	56.62	63.00	72.14	54.51
Qwen2.5-VL-72b	60.00	58.83	59.12	70.25	72.95	67.66
SEMD	79.50	**82.72**	**80.16**	**77.75**	**77.84**	**77.79**

Table 5: Experimental results on metaphor detection task. The best overall results are highlighted in **bold**, while the best results within each modality are underlined. Models are categorized by modality, with textual, visual, and multimodal models highlighted in cyan, green, and orange, respectively.

the image modality in sentiment recognition. This is likely because textual information more directly conveys sentiment.

Within the text modality, although LaBSE achieves higher precision and F1 scores on English data, XLM-R still attains higher accuracy. LaBSE also yields the best results on Chinese textual data. In the image modality, ViT consistently outperforms VGG and ResNet. Among multimodal models, CES achieves the highest accuracy in both Chinese and English, suggesting that image captions are particularly helpful for sentiment recognition. This may be because captions directly describe the overall advertisement, providing more clues about sentiment and theme.

Notably, the MET model shows improvement in both English and Chinese when introducing source and target domains as auxiliary features, outperforming mBERT-Res and highlighting its effectiveness. Our proposed model, SEMD, achieves the best results in both English and Chinese, further demonstrating the benefits of incorporating sentiment features.

A.2 Error Analysis

In analyzing the results of the metaphor task, as shown in Table 7, we examine the accuracy of metaphorical samples and observe a notable disparity between English and Chinese metaphorical image-text recognition. The accuracy for Chinese is only 73%, whereas English data achieves 86%. The lower accuracy in Chinese metaphor recognition is likely due to cultural and ideological influences in advertising data. Specifically, Chinese metaphors tend to be more implicit, while in English advertising, metaphors are often more direct.

	EN			CN		
Model	Acc (%)	Pre (%)	F1 (%)	Acc (%)	Pre (%)	F1 (%)
Random	39.41	45.82	40.41	34.48	41.47	35.02
mBERT	70.27	68.97	62.83	67.76	64.26	58.43
XLM_R	71.50	69.34	69.44	68.71	65.00	66.21
SixTP	69.46	66.50	64.43	68.49	64.53	65.59
LaBSE	70.44	69.44	69.93	70.02	68.04	69.02
GPT-3.5-TURBO	42.36	56.7	42.06	51.36	52.91	52.08
Llama2-8B	40.64	54.49	40.11	50.22	51.39	50.72
Llama3.1-70B	52.49	34.61	33.46	51.00	37.90	30.90
Deepseek R1-70B	53.50	54.25	53.83	53.50	49.90	49.01
VGG16	65.60	59.66	59.90	65.52	58.66	59.02
ResNet50	68.55	66.70	57.49	68.49	66.19	58.86
ViT	69.04	65.84	61.46	68.64	65.55	62.17
mBERT-Res	73.46	71.58	69.75	72.21	68.95	68.84
CMGCN	74.40	71.59	72.79	73.03	**70.84**	70.19
CES	74.56	72.52	73.24	73.30	69.76	70.26
MET	73.68	71.46	71.95	71.55	68.06	67.17
MFC	73.30	70.76	71.93	72.43	68.91	69.60
GPT-4O	61.00	40.00	36.00	50.00	39.00	29.00
LLaVA	55.41	58.61	56.72	52.95	53.59	53.21
Gemini	51.75	56.61	53.61	37.75	53.24	43.48
Qwen2.5-VL-3b	56.00	53.11	52.50	60.25	63.51	55.26
Qwen2.5-VL-7b	59.00	59.81	50.17	60.75	58.85	55.15
Qwen2.5-VL-72b	53.75	51.39	50.44	55.75	64.33	57.17
SEMD	**76.60**	**74.83**	**75.69**	**73.40**	70.66	**70.51**

Table 6: Experimental results on sentiment analysis task. The best overall results are highlighted in **bold**, while the best results within each modality are underlined. Models are categorized by modality, with textual, visual, and multimodal models highlighted in cyan, green, and orange, respectively.

Task	EN	CN
	Acc (%)	Acc (%)
Sentiment Analysis	62.00	55.00
Metaphor Detection	86.00	73.00

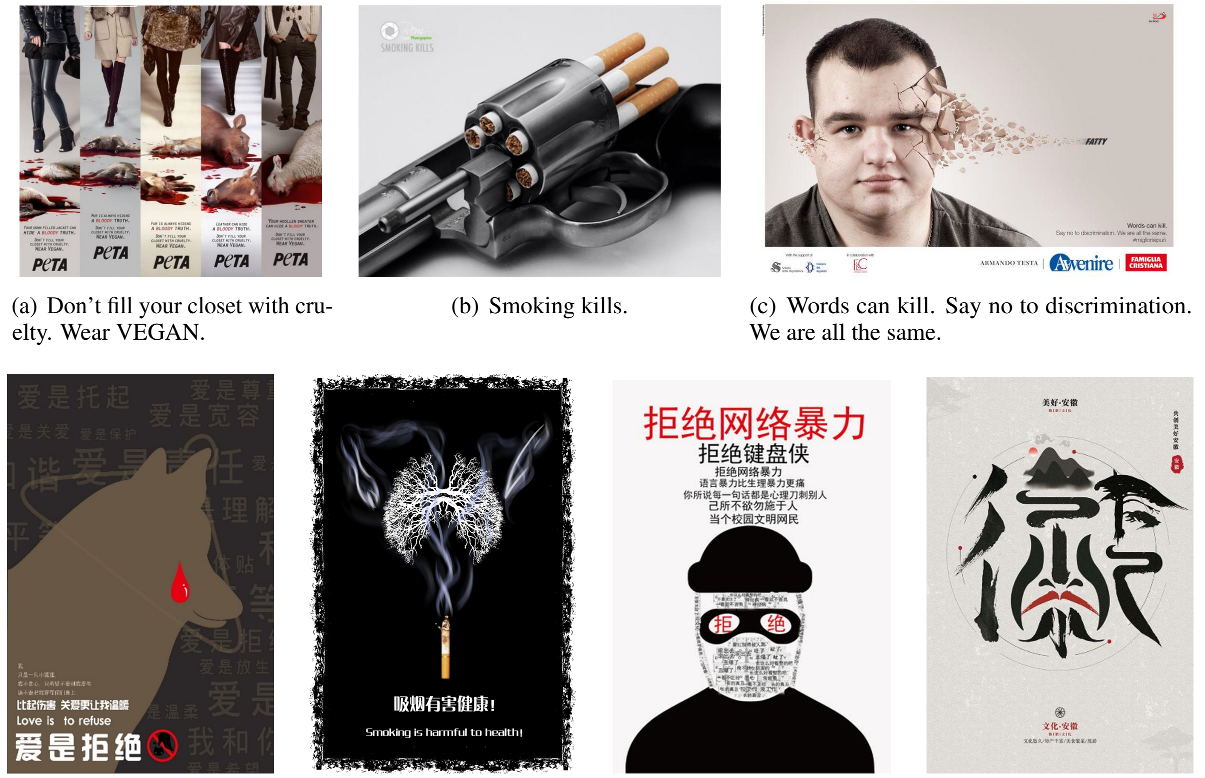
Table 7: The accuracy differences between Chinese and English data in sentiment analysis and metaphor detection tasks.

Figure 7 provides examples of shared metaphor themes in Chinese and English data. For instance, Figures 7(a) and 7(d) depict metaphors related to animal protection. In English data, the comparison between deceased animals and clothing implies that leather garments contribute to animal deaths. In contrast, Chinese data conveys harm to animals through imagery, where tears resembling blood flow from animal figures, while text in the image expresses the concept of love and protection.

Figures 7(b) and 7(e) revolve around the theme of smoking as harmful to health. English data depicts cigarettes as bullets, signifying that smoking can lead to death. In contrast, Chinese data subtly conveys the harm of smoking by shaping smoke patterns to resemble lungs, emphasizing its impact on lung health.

Figures 7(c) and 7(f) illustrate themes of verbal harm and discrimination. English data depicts the damage of language through fractured facial expressions and bullet trajectories, whereas Chinese data uses harmful words to form the image of a "bad person," criticizing verbal attacks.

Furthermore, Chinese data tends to rely more on cultural connotations, such as family values, collectivism, and traditional culture. In Figure 7(g), local



(a) Don't fill your closet with cruelty. Wear VEGAN.

(b) Smoking kills.

(c) Words can kill. Say no to discrimination. We are all the same.

(d) 爱是拒绝, 爱是责任 (Love is rejection, love is responsibility.)

(e) 吸烟有害健康 (Smoking is harmful to health.)

(f) 拒绝网络暴力, 拒绝键盘侠 (Reject online violence, reject keyboard warriors.)

(g) 共创美好安徽 (Let's work together to create a better Anhui.)

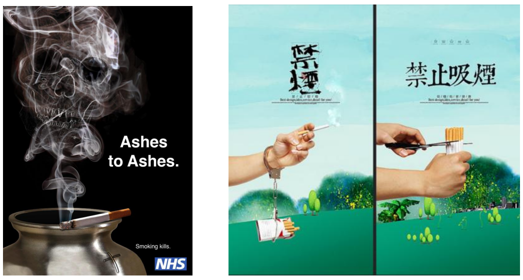
Figure 7: Examples of metaphors.

culture is promoted by incorporating drama masks and landscapes to form the character '徽' (an abbreviation for Anhui Province in China), which frequently appears in Chinese advertisements. In conclusion, cultural and linguistic differences increase the difficulty and complexity of Chinese metaphor detection, making it more challenging for models to distinguish metaphorical data.

In the sentiment analysis task, we observe that in the Chinese dataset, despite a higher proportion of neutral sentiment compared to the English dataset, the accuracy of neutral sentiment recognition is relatively low, at only 55%. In contrast, English data achieves an accuracy of 62%. An analysis of the results reveals that the pure text modality significantly outperforms the image modality, with text playing a predominant role in sentiment recognition tasks.

We examine cases where Chinese predictions of neutral sentiment are incorrect and identify a common pattern. Specifically, these instances contain words with positive sentiment (e.g., '珍爱' (cherish), '文明' (civilization), '珍视' (treasure)), yet the overall message does not express positive sentiment. Clearly, these words lead to model misjudgments. This relates to the polysemy and ambiguity of Chinese vocabulary, where the same word can carry entirely different meanings depending on context. For example, in Chinese, '好" (good) is typically positive, but in certain contexts, it can convey negative sentiment, as in '好烦' (very annoying). Moreover, Chinese grammatical structures and particles significantly impact sentiment interpretation, further affecting the model's ability to accurately recognize sentiment.

Due to the limited proportion of negative sentiment in the dataset, the model lacks sufficient negative samples during training, limiting its ability to recognize and capture negative sentiment. Therefore, the challenges in multimodal sentiment recognition primarily stem from the scarcity of negative sentiment data, the polysemy of neutral sentiment, and cultural influences. Future research can explore strategies to address these challenges and enhance sentiment recognition across diverse cultural datasets.



(a) Ashes to Ashes. (b) 禁止吸烟 (No smoking.)

Figure 8: Metaphor comparison across cultures for a common theme.

B Case Study

Cultural bias significantly affects model performance in cross-cultural metaphor understanding. Western advertisements often use shocking visuals to convey anti-smoking messages. For example, the ‘Ashes to Ashes’ advertisement in Figure 8(a) features a visual metaphor where smoke forms a skull, creating a direct and intense visual impact that evokes negative emotions like fear and despair. These advertisements clearly illustrate the deadly consequences of smoking through a combination of images and text, making it easier for the model to recognize the negative sentiment.

In contrast, Chinese advertisements present anti-smoking messages more subtly, often using warning symbols instead of shocking imagery. For instance, Figure 8(b) depicts ‘handcuffs’ and ‘scissors’ to symbolize restraint and quitting smoking, emphasizing caution rather than shock. While this effectively conveys an anti-smoking message, its emotional tone is milder, making it more challenging for models to capture the emotional intensity from the image alone. As a result, sentiment classification models tend to perform worse on Chinese advertisements compared to English advertisements.

These examples highlight the profound influence of cultural context on metaphor understanding. Western culture favors direct and emotionally intense expressions, while Chinese advertisements rely on cautious symbols and milder emotional tones. Consequently, models face challenges in accurately interpreting metaphors across such cultural differences.