MEANING BEYOND LANGUAGE: LEARNING COMPLEMENTARY SEMANTICS FOR MULTIMODAL METAPHOR DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Metaphor is widely used to express meanings beyond literal content. Multimodal metaphor detection aims to identify metaphorical meaning by jointly understanding text and images. Although existing methods have made notable progress, they mainly rely on a single language and overlook the complementary semantic information expressed in other languages. To address this limitation, we propose LAMCC, a framework that uses an auxiliary language to improve multimodal metaphor detection. First, we introduce a cross lingual semantic disentanglement module to separate shared semantics from language specific information. Contrastive learning, reconstruction, and orthogonality constraints are used to preserve useful information and reduce the effect of language differences. Second, since different vision language model layers capture different semantic cues, we fuse the representations of both languages at each layer through self attention. We then assign different weights to these layers based on the input sample, allowing the model to focus on the semantic cues most relevant to metaphor detection. Finally, to address the absence of translations during inference, we learn a mapping that predicts the semantic representation of the translation from that of the original text. Experiments on multiple benchmarks with different vision language models demonstrate the effectiveness of our framework.

1 INTRODUCTION

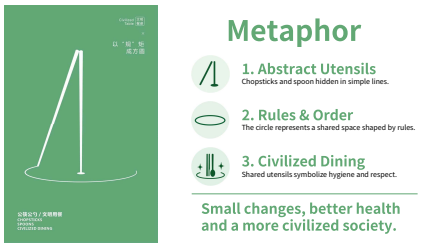
Metaphor is a common form of figurative language that uses one concept to describe another and convey an implicit meaning (Lakoff & Johnson, 2008). It is widely used in daily communication, social media, and advertising to express abstract ideas. Metaphor detection is therefore important for understanding meaning beyond literal content. Early studies mainly focus on metaphors expressed in text (Lakoff & Johnson, 2008; Forceville & Urios-Aparisi, 2009). However, with the growing use of multimedia, metaphors are often expressed through combinations of text and images. As shown in Figure 1(a), the image depicts a deer made of paper, while the text "SAVE PAPER, SAVE NATURE" connects paper use with environmental protection. Together, they convey that wasting paper harms forests and wildlife, whereas saving paper helps protect nature. In such cases, understanding the metaphor requires joint reasoning over textual and visual cues. Thus, multimodal metaphor detection received increasing research attention (Zhang et al., 2021; Meng et al., 2025; He et al., 2025).

Existing studies on multimodal metaphor detection explore the relations between text and images to identify metaphorical meaning (Xu et al., 2024; He et al., 2024; Su et al., 2021). Despite their progress, the potential of using complementary cues from different languages remains underexplored. Although a description and its translation express the same underlying meaning, a vision-language model (VLM) may capture different cues relevant to the metaphor when pairing each description with the same image. For example, Figure 1(b) contains the English description "CHOPSTICKS AND SPOONS, CIVILIZED DINING," which conveys the general theme of the image. However, the Chinese expression "守规矩" provides an additional clue. The word "规矩" means rules, while its characters originally refer to a compass and a square. This connection helps reveal that the chopsticks form a compass, which represents the rules of civilized dining. Combining information from both languages can therefore make the metaphor easier to recognize. Motivated by this observation, we introduce an auxiliary language to support multimodal metaphor detection.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) A deer made of paper, to indicate that we should save paper usage to protect nature.



(b) A pair of chopsticks forms a compass, to suggest civilized dining for people.

Figure 1: Two examples of multimodal metaphors.

However, incorporating an auxiliary language presents three challenges. 1) Differences in linguistic expression can lead to a gap between multimodal representations of the same sample across languages. Such differences may obscure the shared meaning relevant to metaphor detection, making auxiliary semantic cues difficult to use effectively. Hence, extracting shared semantic information while reducing the influence of language differences remains a key challenge. 2) Multimodal metaphor detection requires understanding how text and images jointly convey meaning beyond their literal content. Different VLM layers may capture this relation to different degrees, and their relevance to metaphor detection may vary across samples. Using a fixed layer or treating all layers equally may therefore overlook useful metaphor cues. Hence, combining semantic cues from both languages and identifying the most relevant layers for each sample remains a tough challenge. 3) Auxiliary translations provide additional semantic cues during training, whereas only the original text and image are available at inference time. This mismatch prevents the model from directly accessing the auxiliary cues it has learned to use for metaphor detection. Hence, preserving the benefit of auxiliary semantic information when predicting from the original text and image alone remains an essentital challenge.

To address these challenges, we propose LAMCC, a framework that uses complementary semantic cues from an auxiliary language for multimodal metaphor detection, as shown in Figure 2. The framework consists of three stages: cross-lingual semantic disentanglement, semantic-guided metaphor prediction, and semantic simulation for inference. In the first stage, we combine the original description and its translation with the same image to form two inputs for a shared VLM. To distinguish shared meaning from language differences, we disentangle the representation at each layer into semantic and language-specific representations. We align the semantic representations across languages through contrastive learning, while using reconstruction and orthogonality constraints to retain the original information and keep the two types of representations distinct. In the second stage, we use the extracted semantic representations to guide metaphor prediction. Since different samples may rely on metaphor cues from different VLM layers, we first fuse the semantic representations from both languages at each layer through self-attention. We then adaptively weight and combine the fused representations across layers to obtain a joint semantic representation for each sample. This representation is injected into the initial decoder layer of the main prediction path to support the final metaphor prediction. The prediction process above uses semantic representations from both languages, but auxiliary translations may be unavailable at inference time. To address this mismatch, the third stage estimates the auxiliary semantic representations from those of the original language through a linear mapping fitted on the training data after model training. The estimated representations replace those obtained from auxiliary translations, allowing the fusion and prediction process to proceed using only the original text and image. We evaluate LAMCC on multiple benchmarks using different VLMs, and the results demonstrate its effectiveness.

Our main contributions are summarized as follows:

- We propose LAMCC, a framework that introduces an auxiliary language to enhance multimodal metaphor detection. Through semantic simulation, the framework uses auxiliary semantic information while requiring only the original text and image at inference time.
- We design a cross-lingual semantic disentanglement module with reconstruction, alignment, and orthogonality constraints to separate shared semantics from language-specific information. We further combine the extracted semantic cues through self-attention and adaptive layer selection to guide metaphor prediction for each sample.

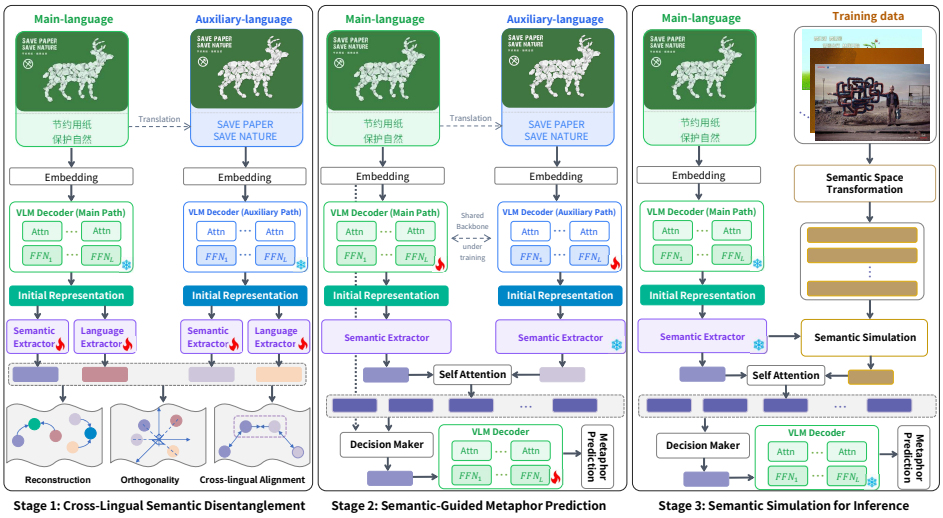


Figure 2: Illustration of the proposed LAMCC for multimodal metaphor detection.

- Experiments on multiple multimodal metaphor detection benchmarks demonstrate the effectiveness of LAMCC across different VLMs.

2 METHODOLOGY

2.1 PROBLEM FORMULATION

We study multimodal metaphor detection, which determines whether a given text-image pair conveys a metaphorical meaning by jointly reasoning over the textual and visual modalities. Let $\mathcal{D} = \{s_1, s_2, \cdots, s_N\}$ denote a set of N samples, where each sample $s_i = (\mathcal{X}_i, I_i, Y_i)$ consists of the textual description $\mathcal{X}_i$, the image I_i, and the ground-truth metaphor label Y_i. Here, $Y_i = 1$ indicates that the sample s_i contains a metaphorical expression, and $Y_i = 0$ otherwise. For readability, the sample index i is omitted in the following descriptions. Formally, we design a model $\mathcal{G}$ that predicts whether a sample expresses a metaphor from its textual description together with its image,

$$\mathcal{G}(\mathcal{X}, I \mid \Theta_F) \rightarrow Y, \tag{1}$$

where Θ_F denotes the set of parameters to be learned.

2.2 STAGE 1: CROSS-LINGUAL SEMANTIC DISENTANGLEMENT

Whether a text-image pair conveys a metaphor depends on the meaning that emerges when the text and the image are interpreted together, not on either modality alone Yang et al. (2025b). A vision-language model (VLM) that takes a text-image pair as input already produces a representation of this joint meaning. However, when the same image is paired with descriptions written in different languages, the resulting representation can differ, because each language tends to foreground different details of the scene that connect the image to the metaphor. Combining them can therefore capture cues that a single-language description alone might miss.

Motivated by this, beyond the original textual description, we introduce its translation into an auxiliary language as additional training information, to help the model capture richer cross-lingual semantic cues for metaphor detection. Specifically, we translate the original textual description $\mathcal{X}$ into another language to obtain $\bar{\mathcal{X}}$. In this way, each sample s corresponds to two image-text pairs: the main pair $(\mathcal{X}, I)$ and the auxiliary pair $(\bar{\mathcal{X}}, I)$, which use the same image but differ in the language of their textual descriptions. We then feed them separately into the same vision-language model (VLM) $\mathcal{V}(\cdot)$ and obtain the corresponding multimodal representations as follows,

$$\begin{cases} \mathbf{f}_l = \mathcal{V}(\mathcal{X}, I), \\ \bar{\mathbf{f}}_l = \mathcal{V}(\bar{\mathcal{X}}, I), \end{cases} \qquad l \in \{1, 2, \ldots, L\}, \tag{2}$$

where L is the total number of layers in the VLM, and $\mathbf{f}_l$ and $\bar{\mathbf{f}}_l$ denote the multimodal representations produced by the l-th layer for the main pair and the auxiliary pair, respectively. Since $\mathbf{f}_l$ and $\bar{\mathbf{f}}_l$ are obtained from texts in different languages, each of them contains not only the semantic content shared by the two pairs, but also variation caused by the language itself. To separate these two parts, we map each representation to a semantic representation and a language-specific representation. Taking the main pair as an example, this mapping is given as follows,

$$\begin{cases} \mathbf{m}_l = \mathcal{F}_S(\mathbf{f}_l), \\ \mathbf{r}_l = \mathcal{F}_L(\mathbf{f}_l), \end{cases} \qquad l \in \{1, 2, \ldots, L\}, \tag{3}$$

where $\mathcal{F}_S(\cdot)$ and $\mathcal{F}_L(\cdot)$ are two mapping functions, implemented as multilayer perceptrons (MLPs), which extract the semantic representation and the language-specific representation from a multimodal representation, respectively. $\mathbf{m}_l$ is the semantic representation of the main pair at layer l, which should carry the information most relevant to the metaphor label, and $\mathbf{r}_l$ is the language-specific representation of the main pair at layer l, which captures the part of $\mathbf{f}_l$ not explained by $\mathbf{m}_l$. Similarly, for the auxiliary pair, we obtain the semantic representation $\bar{\mathbf{m}}_l$ and the language-specific representation $\bar{\mathbf{r}}_l$ in the same way.

We argue that $\mathbf{m}_l$ and $\mathbf{r}_l$ should jointly reconstruct $\mathbf{f}_l$, remain distinct from each other, and that the semantic representation should stay the same across the two languages of the same sample. We enforce these three properties with a reconstruction constraint, an orthogonality constraint, and a cross-lingual alignment constraint, respectively.

Reconstruction Constraint. We design a reconstruction constraint so that the semantic representation and the language-specific representation sum to the original representation, for both the main pair and the auxiliary pair, as follows,

$$\mathcal{L}_{\text{rec}} = \frac{1}{2BL} \sum_{i=1}^{B} \sum_{l=1}^{L} \left(\left\| \mathbf{f}_l^i - (\mathbf{m}_l^i + \mathbf{r}_l^i) \right\|_2^2 + \left\| \bar{\mathbf{f}}_l^i - (\bar{\mathbf{m}}_l^i + \bar{\mathbf{r}}_l^i) \right\|_2^2 \right), \tag{4}$$

where $\| \cdot \|_2$ denotes the Euclidean norm, and i indexes the samples in a minibatch of size B.

Orthogonality Constraint. We further introduce an orthogonality constraint so that the semantic representation and the language-specific representation are uncorrelated and do not carry the same information, as follows,

$$\mathcal{L}_{\text{orth}} = \frac{1}{2BL} \sum_{i=1}^{B} \sum_{l=1}^{L} \left(\cos^2\left(\mathbf{m}_l^i, \mathbf{r}_l^i\right) + \cos^2\left(\bar{\mathbf{m}}_l^i, \bar{\mathbf{r}}_l^i\right) \right). \tag{5}$$

Cross-Lingual Alignment Constraint. The semantic representation of the main pair and that of the auxiliary pair should correspond to the same underlying meaning, because whether a sample is a metaphor depends on its image and its meaning, not on the language of its text. We enforce this with contrastive learning, treating the semantic representations of the same sample as a positive pair and those of different samples in the same minibatch as negative pairs. Formally, we have,

$$\begin{cases} S_l^{ij} = \dfrac{\cos\left(\mathbf{m}_l^i, \bar{\mathbf{m}}_l^j\right)}{\tau}, \quad i, j \in \{1, \ldots, B\}, \\ \mathcal{L}_{\text{align}} = -\dfrac{1}{2BL} \displaystyle\sum_{i=1}^{B} \sum_{l=1}^{L} \left[\log \frac{\exp S_l^{ii}}{\sum_{j=1}^{B} \exp S_l^{ij}} + \log \frac{\exp S_l^{ii}}{\sum_{j=1}^{B} \exp S_l^{ji}} \right], \end{cases} \tag{6}$$

where S_l^{ij} is the scaled cosine similarity between the semantic representation of the main pair of sample s_i and that of the auxiliary pair of sample s_j at layer l, τ is a temperature parameter, and $\mathcal{L}_{\text{align}}$ is a symmetric InfoNCE loss (van den Oord et al., 2019). The overall training objective of this stage is defined as follows,

$$\mathcal{L}_{\text{dec}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{orth}} \mathcal{L}_{\text{orth}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \tag{7}$$

where λ_{rec}, λ_{orth}, and λ_{align} are hyperparameters that balance the three terms.

2.3 STAGE 2: SEMANTIC-GUIDED METAPHOR PREDICTION

Having extracted the semantic representation, we proceed to conduct metaphor prediction. We first attentively fuse the main and auxiliary semantic representations at each layer into a joint semantic representation. Since different samples may depend on different layers, we then let the model choose the layer whose joint semantic representation is most useful for the current sample. Finally, we add the selected semantic signal into the main prediction path, so the model predicts the metaphor label from the original image and text together with this joint semantic representation.

Self-Attention Fusion. Since the two languages may contribute differently to metaphor detection for each sample, we use self-attention to adaptively fuse their semantic representations at each layer. Both representations are projected into queries and keys using learnable matrices $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{D\times d}$ shared across languages and layers. Taking the main query as an example, we compute its attention weights over the two representations as follows,

$$\begin{cases} \mathbf{q}_l = \mathbf{m}_l\mathbf{W}_Q, \quad \mathbf{k}_l = \mathbf{m}_l\mathbf{W}_K, \quad \bar{\mathbf{k}}_l = \bar{\mathbf{m}}_l\mathbf{W}_K, \\ (p_l, 1-p_l) = \mathrm{Softmax}\left(\dfrac{\mathbf{q}_l\mathbf{k}_l^\top}{\tau\sqrt{d}}, \dfrac{\mathbf{q}_l\bar{\mathbf{k}}_l^\top}{\tau\sqrt{d}}\right), \end{cases} \tag{8}$$

where d is the attention dimension and τ is the temperature. We compute the attention weights $(r_l, 1-r_l)$ in the same way, using the auxiliary representation to form the query. We average the weights assigned by the two queries and obtain the joint semantic representation,

$$\begin{cases} \alpha_l = \dfrac{p_l + r_l}{2}, \\ \mathbf{z}_l = \alpha_l\mathbf{m}_l + (1-\alpha_l)\bar{\mathbf{m}}_l. \end{cases} \tag{9}$$

Adaptive Layer Selection. Different layers of the VLM capture different kinds of information. To adaptively weight their contributions for each sample, we design a Decision Maker to score the joint semantic representations from all layers. It projects each representation together with the main-path input embedding, then averages the resulting vectors to obtain one score per layer,

$$\begin{cases} \mathbf{h} = \dfrac{1}{L}\displaystyle\sum_{j=1}^{L}\mathbf{W}_{\mathrm{dm}}(\mathbf{z}_j + \mathbf{e}), \\ \boldsymbol{\pi} = \mathrm{Softmax}(\mathbf{h}), \end{cases} \tag{10}$$

where $\mathbf{W}_{\mathrm{dm}} \in \mathbb{R}^{L\times D}$ is a learnable projection matrix, and $\mathbf{e} \in \mathbb{R}^D$ is the input embedding of the last main-path prompt token. The vectors $\mathbf{h}, \boldsymbol{\pi} \in \mathbb{R}^L$ contain the layer scores and weights, respectively. During training, we replace the Softmax in Eq. equation 10 with Gumbel-Softmax (Jang et al., 2017) at temperature τ_{dm}. Its random perturbation encourages exploration of different layers, while a lower temperature concentrates the weights toward one layer. During inference, we use the deterministic Softmax. We then form the final joint semantic representation $\hat{\mathbf{z}}$ as a weighted sum,

$$\hat{\mathbf{z}} = \sum_{l=1}^{L}\pi_l\mathbf{z}_l. \tag{11}$$

Metaphor Prediction. We inject $\hat{\mathbf{z}}$ into the initial decoder layer of the main path, so that later layers can gradually combine this cross-lingual semantic representation with the model's own reasoning about the metaphor. Specifically, we update the initial decoder layer representation $\mathbf{f}_{\mathrm{init}}$ as follows,

$$\widetilde{\mathbf{f}}_{\mathrm{init}} = \mathbf{f}_{\mathrm{init}} + \hat{\mathbf{z}}, \tag{12}$$

where $\widetilde{\mathbf{f}}_{\mathrm{init}}$ is the updated initial decoder layer representation, which will be fed into the remaining decoder layers. Finally, we can obtain the final prediction of the metaphor. The training objective is,

$$\mathcal{L}_{\mathrm{task}} = -\frac{1}{B}\sum_{i=1}^{B}\log P\left(y_i \mid \mathcal{X}, \hat{\mathbf{z}}_i\right). \tag{13}$$

Overall, the model is trained in two stages. In the first stage, we train $\mathcal{V}(\cdot)$, $\mathcal{F}_S(\cdot)$, and $\mathcal{F}_L(\cdot)$ with $\mathcal{L}_{\text{dec}}$, so that the semantic representation becomes language-invariant. And in this second stage, we freeze $\mathcal{V}(\cdot)$, $\mathcal{F}_S(\cdot)$, and $\mathcal{F}_L(\cdot)$, and train the fusion, layer-selection, and injection modules, together with LoRA adapters inserted into the backbone, so that the model learns to use the semantic representations $\mathbf{m}_l$ and $\bar{\mathbf{m}}_l$ for metaphor prediction.

2.4 Stage 3: Semantic Simulation for Inference

At test time, however, only the main pair $(\mathcal{X}, I)$ is available, since no auxiliary translation is provided. Hence, we estimate the auxiliary semantic representation $\bar{\mathbf{m}}_l$ from the main-language semantic representation $\mathbf{m}_l$ through a calibration step performed once after training.

Specifically, this calibration step uses the trained, frozen model to extract the semantic representations of the training samples, and fits a linear map from the main-language semantic space to the auxiliary-language semantic space. We denote the main-language and auxiliary-language semantic representations of the l-th layer of the k-th training sample as $\mathbf{m}_l^k$ and $\bar{\mathbf{m}}_l^k$, where $k \in \{1, \ldots, K\}$ and K is the number of training samples used for calibration. Collecting these representations over all samples and all layers gives two matrices $\mathbf{M} = [\mathbf{m}_l^k] \in \mathbb{R}^{P \times D}$ and $\bar{\mathbf{M}} = [\bar{\mathbf{m}}_l^k] \in \mathbb{R}^{P \times D}$, where $P = K \times L$ is the total number of sample-layer pairs, and the same row of $\mathbf{M}$ and $\bar{\mathbf{M}}$ corresponds to the same sample and the same layer. We then fit a linear map from $\mathbf{M}$ to $\bar{\mathbf{M}}$ by solving the following ridge regression problem for $\mathbf{W}_T \in \mathbb{R}^{D \times D}$, and denote its optimal solution as $\mathbf{W}_T^*$,

$$\mathbf{W}_T^* = \arg\min_{\mathbf{W}_T} \left\| \mathbf{M}\mathbf{W}_T - \bar{\mathbf{M}} \right\|_F^2 + \rho \left\| \mathbf{W}_T \right\|_F^2, \tag{14}$$

which admits the closed-form solution,

$$\mathbf{W}_T^* = \left(\mathbf{M}^\top \mathbf{M} + \rho \mathbf{I}_D \right)^{-1} \mathbf{M}^\top \bar{\mathbf{M}}. \tag{15}$$

$\mathbf{W}_T^*$ is obtained directly from this closed-form solution and then kept fixed. Using $\mathbf{W}_T^*$, we simulate the auxiliary semantic representation of each layer from the main-language semantic representation,

$$\hat{\bar{\mathbf{m}}}_l = \mathbf{m}_l \mathbf{W}_T^*. \tag{16}$$

The simulated representation $\hat{\bar{\mathbf{m}}}_l$ then takes the place of $\bar{\mathbf{m}}_l$ in Eq. (8), so that the fusion, layer selection, and injection steps in Stage 2 proceed unchanged and the model still produces a metaphor prediction from the main pair $(\mathcal{X}, I)$ alone.

3 Experiment

3.1 Experimental Setup

Datasets and Evaluation. We evaluate our method on two bilingual multimodal metaphor benchmarks: MultiMM (Yang et al., 2025b), which contains 8,461 Chinese and English advertisement pairs, and MET-MEME (Xu et al., 2022), which contains 10,045 Chinese and English meme pairs. We formulate metaphor detection as a binary classification task on both datasets. Only the training samples are paired with translations, while validation and test samples remain in their original languages. For MultiMM, we retain the sentiment field but do not use its source and target domain annotations. Following the evaluation settings of each benchmark, we report accuracy, precision, and F1 score, together with recall when available.

Backbones. We evaluate four vision-language backbones with different model families and scales: Qwen3-VL-4B, Qwen3-VL-8B, Qwen2.5-VL-3B, and Qwen2.5-VL-7B. We evaluate four vision language backbones from different model families and scales: Qwen3-VL-4B, Qwen3-VL-8B, Qwen2.5-VL-3B, and Qwen2.5-VL-7B. Each backbone is trained and evaluated separately on the Chinese and English subsets of the dataset, resulting in eight backbone and language combinations. During both training and inference, we use the same fixed task instruction for multimodal metaphor detection. This instruction remains unchanged across all methods and language settings, ensuring that any performance differences are not caused by variations in the prompt.

Table 1: Performance comparison (%) on the Chinese and English splits of MultiMM. The best result is shown in **bold**, and the second-best result is underlined.

	Chinese			English		
Method	**Acc**	**Pre**	**F1-M**	**Acc**	**Pre**	**F1-M**
LLaVA(Liu et al., 2023b)	56.81	58.20	69.84	59.11	59.06	73.31
SEMD(Yang et al., 2025b)	77.75	77.84	77.79	79.50	82.72	80.16
MAMMD(Zhang et al., 2026)	78.36	80.29	83.49	80.85	85.20	84.69
Qwen2.5-VL-3B	79.32	80.68	77.44	75.18	78.07	73.72
Qwen3-VL-4B	80.45	82.13	78.65	79.61	79.52	79.45
Qwen2.5-VL-7B	82.27	81.66	81.76	77.64	77.74	77.33
Qwen3-VL-8B	81.59	81.79	80.48	76.41	76.90	76.41
CC-Tuning + Multimodal Aux	87.73	87.29	87.35	87.22	87.15	87.19
CC-Tuning	90.68	90.41	90.37	88.94	88.88	88.92
Semantic disentangled CC-Tuning	91.59	91.29	91.33	90.17	90.10	90.12
LAMCC	**92.27**	**92.02**	**92.02**	**90.41**	**90.37**	**90.35**

Table 2: Performance comparison (%) on the Chinese and English splits of MET-MEME. "–" indicates that the corresponding result is not reported.

	Chinese			English		
Method	**Acc**	**Pre**	**F1-W**	**Acc**	**Pre**	**F1-W**
MGMCF (Zheng et al., 2025)	78.39	78.16	–	87.51	88.58	–
MetaCRN (Meng et al., 2025)	85.92	87.67	85.98	84.87	86.20	84.84
SFVE (Yang et al., 2025a)	86.27	87.02	86.05	85.47	86.27	84.87
CDGLT (Qian et al., 2025)	–	–	–	91.38	–	91.34
iAMeta (He et al., 2025)	89.01	88.94	88.95	89.48	89.26	89.19
LAMCC	**92.72**	**92.73**	**92.72**	**92.49**	**92.40**	**92.40**

Baselines. We compare LAMCC with SFT and existing multimodal metaphor detection methods. **SFT** fine-tunes each backbone using only the original-language input. On MultiMM, we compare with **LLaVA** (Liu et al., 2023a), **SEMD** (Yang et al., 2025b), and **MAMMD** (Zhang et al., 2026). On MET-MEME, we compare with **MGMCF** (Zheng et al., 2025), which fuses multimodal clues at different levels; **MetaCRN** (Meng et al., 2025), which uses language augmentation and dynamic feature replacement; **SFVE** (Yang et al., 2025a), which enhances and fuses visual features; **CDGLT** (Qian et al., 2025), which combines concept drift with LayerNorm tuning; and **iAMeta** (He et al., 2025), which uses MLLMs and information debiasing. We also include three variants on MultiMM. **CC-Tuning** (Ye et al., 2025) transfers selected raw activations from a text-only auxiliary-language path. **CC-Tuning + Multimodal Aux** further adds the image to this path. **CC-Tuning + Semantic Disentanglement** replaces the raw activations with shared semantic representations while retaining the text-only auxiliary path.

3.2 Main Results

Tables 1 and 2 provide detailed comparisons with existing methods on MultiMM and MET-MEME, respectively. We discuss the metrics reported for each benchmark separately.

Comparison on MultiMM. In Table 1, LAMCC achieves the highest accuracy, macro precision, and macro F1 in both languages. Its Chinese and English accuracies reach 92.27% and 90.41%, exceeding CC-Tuning by 1.59 and 1.47 percentage points, respectively. Language Agnostic CC-Tuning improves over CC-Tuning, and the full model provides further accuracy gains of 0.68 and 0.24 points. In contrast, adding an image to the raw-activation auxiliary path lowers performance in both languages. This comparison suggests that effectively using auxiliary information depends on the semantic connection design, rather than simply providing an additional visual input.

Comparison on MET-MEME. Table 2 shows that LAMCC achieves the highest reported values across all four metrics in both languages. Its accuracy reaches 92.72% on Chinese and 92.49% on English, exceeding iAMeta and CDGLT, the strongest accuracy baselines on the respective splits,

Table 3: Accuracy comparison (%) between SFT and LAMCC across different backbones on MET-MEME and MultiMM. Mean denotes the average accuracy over the Chinese and English splits.

		MET-MEME			**MultiMM**		
Backbone	Method	ZH	EN	Mean	ZH	EN	Mean
Qwen3-VL-8B	SFT	91.48	90.73	91.11	81.59	76.41	79.00
	LAMCC	**92.72**	**92.49**	**92.61**	**92.27**	**90.41**	**91.34**
Qwen2.5-VL-7B	SFT	90.81	90.98	90.90	82.27	77.64	79.96
	LAMCC	**92.05**	**92.74**	**92.40**	**88.86**	**90.36**	**89.61**
Qwen3-VL-4B	SFT	90.65	89.98	90.32	80.45	79.61	80.03
	LAMCC	**92.22**	**91.48**	**91.85**	**90.79**	**90.65**	**90.72**
Qwen2.5-VL-3B	SFT	89.23	91.11	90.17	79.32	75.18	77.25
	LAMCC	**90.15**	**91.48**	**90.82**	**86.02**	**88.61**	**87.32**

by 3.71 and 1.11 points. The corresponding weighted F1 gains are 3.77 and 1.06 points, indicating improvements beyond accuracy alone. Together with the MultiMM results, this comparison supports the effectiveness of LAMCC across both benchmarks, with larger gains on some language and dataset settings than others.

Effectiveness across backbones. Table 3 shows that LAMCC improves over SFT in all 16 backbone–language–dataset settings. Averaged over the eight settings per dataset, accuracy increases from 90.62% to 91.92% on MET-MEME and from 79.06% to 89.94% on MultiMM. These consistent gains across both Qwen families and four model sizes support the effectiveness of incorporating cross-lingual semantic information. The improvement is larger on MultiMM, while LAMCC also improves on the already strong SFT results on MET-MEME. Qwen3-VL-8B achieves the highest mean accuracy on both datasets, although the best individual result varies with the language and dataset.

3.3 Ablation Study

Table 4: Ablation results (%) in terms of accuracy. Δ denotes the change compared with LAMCC.

	MET-MEME				**MultiMM**			
Setting	ZH	EN	Avg.	Δ	ZH	EN	Avg.	Δ
w/o DM	92.39	90.11	91.25	-1.36	90.45	80.09	85.27	-6.07
w/o Disen.	91.81	91.65	91.73	-0.88	90.45	88.20	89.33	-2.01
w/o Self-Attn.	91.89	92.11	92.00	-0.61	91.13	**90.90**	91.02	-0.32
LAMCC	**92.72**	**92.49**	**92.61**	–	**92.27**	90.41	**91.34**	–

Table 4 shows the ablation study results. 1) **Adaptive Layer Selection.** Removing the Decision Maker causes the largest average accuracy drops: 1.36 points on MET-MEME and 6.05 points on MultiMM. The effect is particularly pronounced on English MultiMM, where accuracy falls to 80.09%. These results support adapting the layer contributions to each input when combining cross-lingual semantic cues. 2) **Semantic Disentanglement.** Without disentanglement, average accuracy decreases by 0.88 points on MET-MEME and 1.99 points on MultiMM, with lower performance in all four dataset–language settings. This consistent decrease supports separating shared semantics from language-specific variation before fusion, helping the model use auxiliary information for metaphor prediction. 3) **Self-Attention Fusion.** Removing self-attention reduces average accuracy by 0.61 points on MET-MEME and 0.30 points on MultiMM. Although the full model performs better in three settings, the ablated variant reaches 90.90% on English MultiMM. Thus, adaptive fusion contributes to the best average performance on both datasets, but its benefit is not uniform across languages.



(a) Three-example prediction comparison (b) Layer-wise semantic fusion

Figure 3: Three case studies shown in one row. (a) Predictions of Qwen3-VL-8B SFT and Qwen-on three samples; check and cross marks denote correct and incorrect predictions, respectively. (b) A drunk-driving example with its bilingual input and layer-wise main-path mixing coefficients.

3.4 Case Study

On Comparison with SFT backbone. As shown in Figure 3(a), our LAMCC model correctly detects the metaphors in Cases 1 and 2, while the Qwen3-VL-8B SFT baseline fails. In Case 1, the word “Beer” and the car symbol on the deformed bottle cap jointly imply drunk driving and its consequences. In Case 2, the skull embedded in the Chinese character for “sea” associates dangerous waters with death. These two cases require the model to combine visual structure, input text, and the relation between them; neither the object nor the text alone states the metaphor directly. SFT misclassifies both examples, suggesting that it does not sufficiently integrate these cross-modal and cross-lingual cues. Unlike the first two examples, Case 3 does not introduce an unusual object relation or a visual transformation. The metaphor depends on an additional social association between donating blood and expressing love. Both SFT and LAMCC predict this sample incorrectly, indicating that the cross-lingual connection is insufficient when the literal visual frame is much stronger than the abstract text-to-value relation.

On auxiliary-language semantic fusion In Figure 3(b), the English translation, “Strive to be a civilized ‘driver’; do not become a traffic drunk.” preserves the warning in the Chinese slogan while losing its wordplay. Together with the image of death riding beside a drinking driver, the bilingual input supports the metaphor of drunk driving as inviting death. The heatmap shows the varying layer-wise mixing weights, indicating that the main and auxiliary representations participate in semantic fusion in different proportions across layers. Because different layers encode information at different levels of abstraction, this variation is consistent with the two language views playing different roles at different information scales. It also motivates the use of adaptive fusion and the Decision Maker to adjust both the relative contribution of the main and auxiliary paths and the weights assigned to different layers.

4 Conclusion

In this work, we presented **LAMCC**, a framework that uses an auxiliary-language view to support multimodal metaphor detection. It combines cross-lingual semantic decomposition with attention-based fusion and adaptive layer weighting to construct an instance-specific semantic signal for prediction. A transformation mapping fitted after training approximates the auxiliary representations, to remove the need for auxiliary-language translation at inference time. Experiments on MultiMM and MET-MEME with four Qwen backbones suggest that this approach can improve accuracy over supervised fine-tuning in the evaluated Chinese and English settings. Ablations indicate the value of semantic decomposition and adaptive layer weighting, while the benefit of attention-based fusion varies across settings. Future work will examine performance without these annotations, extend the evaluation to more languages, and investigate how translation quality and auxiliary-representation approximation affect metaphor detection.

AI USE STATEMENT

We used deepseek-v4-flash model API to translate the text in datasets into auxiliary languages used in our experiments. All translated content was checked by human. And we used generative AI tools to polish writing and assist code development. All the generated content was reviewed, verified, and tested by human. We take full responsibility for the final content and artifacts presented in this work.

REFERENCES

Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. LLaVA steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15230–15250. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.739. URL https://aclanthology.org/2025.acl-long.739/.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1347–1356. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-eacl.90. URL https://aclanthology.org/2024.findings-eacl.90/.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. AltCLIP: Altering the language encoder in CLIP for extended language capabilities. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8666–8682. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.552. URL https://aclanthology.org/2023.findings-acl.552/.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3576–3588. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.280. URL https://aclanthology.org/2021.naacl-main.280/.

Charles Forceville and Eduardo Urios-Aparisi. *Multimodal metaphor*, volume 11. Walter de Gruyter, 2009.

Keita Fukushima, Tomoyuki Kajiwara, and Takashi Ninomiya. Reversible disentanglement of meaning and language representations from multilingual sentence encoders. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pp. 265–270. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.mrl-main.18. URL https://aclanthology.org/2025.mrl-main.18/.

Xiaoyu He, Long Yu, Shengwei Tian, Qimeng Yang, Jun Long, and Bo Wang. Viemf: Multimodal metaphor detection via visual information enhancement with multimodal fusion. *Information Processing & Management*, 61(3):103652, 2024.

Xiaoyu He, Long Yu, and Shengwei Tian. iameta: Advancing multimodal metaphor detection using mllms and information debiasing. *Information Fusion*, 124:103348, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. MURAL: Multimodal, multitask representations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3449–3463. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.293. URL https://aclanthology.org/2021.findings-emnlp.293/.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017. URL https://arxiv.org/abs/1611.01144.

George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b. URL https://arxiv.org/abs/2304.08485.

Hao Meng, Qimeng Yang, Jingwen Ma, and Qixing Wei. Metacrn: Language-augmented multimodal metaphor detection using cross-modal dynamic replacement. In *International Conference on Neural Information Processing*, pp. 151–165. Springer, 2025.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3P: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3977–3986, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Ni_M3P_Learning_Universal_Representations_via_Multitask_Multilingual_Multimodal_Pre-Training_CVPR_2021_paper.html.

Wenhao Qian, Zhenzhen Hu, Zijie Song, and Jia Li. Concept drift guided layernorm tuning for efficient multimodal metaphor identification. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pp. 1100–1108, 2025.

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7961–7973. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.473. URL https://aclanthology.org/2024.findings-acl.473/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

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

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7764–7774, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.612. URL https://aclanthology.org/2021.emnlp-main.612/.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Zetian Wu, Zhongkai Sun, Zhengyang Zhao, Sixing Lu, Chengyuan Ma, and Chenlei Guo. Fine-grained multi-lingual disentangled autoencoder for language-agnostic representation learning. In *Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22)*, pp. 12–24. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.mmnlu-1.2. URL https://aclanthology.org/2022.mmnlu-1.2/.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 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, pp. 2887–2899, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3532019. URL https://doi.org/10.1145/3477495.3532019.

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 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)*, pp. 91–101, 2024.

Qimeng Yang, Hao Meng, Yuanbo Yan, Shisong Guo, and Qixing Wei. Sfve: visual information enhancement metaphor detection with multimodal splitting fusion: Q. yang et al. *The Journal of Supercomputing*, 81(3):467, 2025a.

Senqi Yang, Dongyu Zhang, Jing Ren, Ziqi Xu, Xiuzhen Zhang, Yiliao Song, Hongfei Lin, and Feng Xia. Cultural bias matters: A cross-cultural benchmark dataset and sentiment-enriched model for understanding multimodal metaphors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26301–26317, Vienna, Austria, 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1275. URL https://aclanthology.org/2025.acl-long.1275/.

Yangfan Ye, Xiaocheng Feng, Zekun Yuan, Xiachong Feng, Libo Qin, Lei Huang, Weitao Ma, Yichong Huang, Zhirui Zhang, Yunfei Lu, Xiaohui Yan, Duyu Tang, Dandan Tu, and Bing Qin. Cc-tuning: A cross-lingual connection mechanism for improving joint multilingual supervised fine-tuning, 2025. URL https://arxiv.org/abs/2506.00875.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 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)*, pp. 3214–3225, 2021.

Yujie Zhang, Zhihao Chen, Qingsong Guo, and Jing Wan. Mammd: A multi-agent collaboration framework for multimodal metaphor detection. *Neurocomputing*, 681:133381, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2026.133381. URL https://www.sciencedirect.com/science/article/pii/S0925231226007782.

Li Zheng, Hao Fei, Ting Dai, Zuquan Peng, Fei Li, Huisheng Ma, Chong Teng, and Donghong Ji. Multi-granular multimodal clue fusion for meme understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26057–26065, 2025.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. UC^2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4155–4165, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhou_UC2_Universal_Cross-Lingual_Cross-Modal_Vision-and-Language_Pre-Training_CVPR_2021_paper.html.

A Related Work

Multimodal Metaphor Detection. Multimodal metaphor detection extends the conventional text-based metaphor analysis by requiring models to jointly interpret linguistic expressions and visual context. Earlier studies mainly investigated how visual evidence could complement textual signals, such as selecting modality-specific information according to lexical concreteness (Su et al., 2021). The release of MultiMET also provided a large-scale text–image benchmark annotated with metaphor occurrence, source–target domain relations, sentiment, and author intent, allowing a more systematic study of multimodal metaphor detection (Zhang et al., 2021). Subsequent work has increasingly moved beyond direct feature fusion toward richer forms of cross-modal reasoning. VIEMF strengthens visual representations before multimodal fusion (He et al., 2024), while C4MMD introduces chain-of-thought knowledge generated by multimodal large language models (Xu et al., 2024). Recent methods further explore auxiliary reasoning signals and model debiasing: iAMeta incorporates knowledge generated by MLLM together with emotion control and causal debiasing (He et al., 2025), while MAMMD decomposes metaphor detection into modality-specific analysis, synthesis, and reflective reasoning through multiple agents (Zhang et al., 2026). MultiMM provided a new benchmark for Chinese–English multimodal metaphor detection and reveals substantial cross-cultural variation in metaphorical expression (Yang et al., 2025b). These developments improve how models exploit information within a given image–text pair, but the textual input is still typically treated as a single linguistic realization of the underlying meaning. LAMCC considers an additional source of evidence: auxiliary linguistic realizations of the same image-associated semantics. By introducing an auxiliary-language view, it uses cross-lingual semantic complementarity to support metaphor detection.

Cross-Lingual and Language-Agnostic Representation Learning. Cross-lingual representation learning is largely concerned with identifying semantic structure that can be shared across languages despite differences in linguistic form. InfoXLM, for example, treats bilingual sentence pairs as alternative views of the same meaning and improves their alignment through contrastive learning (Chi et al., 2021). Other studies suggest that multilingual representations contain both language-agnostic semantic content and language-dependent information. Tiyajamorn et al. distill language-agnostic meaning representations from multilingual sentence encoders (Tiyajamorn et al., 2021), and FMDA explicitly separates semantic factors from language-specific ones (Wu et al., 2022). Reversible disentanglement further considers how such separation can preserve sufficient information to reconstruct the original multilingual representation (Fukushima et al., 2025). Similar principles appear in multilingual vision–language learning, where visual context provides an additional anchor for cross-lingual correspondence. M3P learns a shared semantic space across languages and modalities (Ni et al., 2021); UC^2 uses translated image captions and common visual context to strengthen cross-lingual alignment (Zhou et al., 2021); MURAL combines image–caption and translation supervision (Jain et al., 2021); and AltCLIP aligns a multilingual text encoder with a pretrained visual representation space (Chen et al., 2023). While these methods mainly emphasize semantic invariance across languages, we concern that two linguistic realizations may express the same underlying meaning while emphasizing different task-relevant cues. So LAMCC separates shared semantics from language-specific variation and combines the semantic evidence extracted from both language views for metaphor detection.

Multilingual Fine-Tuning and Latent Representation Intervention. Most multilingual instruction-tuning approaches improve cross-lingual transfer at the data or supervision level, for example through translated instruction data, multilingual mixtures, or translation-oriented objectives (Chen et al., 2024; Ranaldi et al., 2024). CC-Tuning is particularly relevant to our work because it instead establishes cross-lingual interaction directly in latent space (Ye et al., 2025). It combines feed-forward activations from English and non-English inputs, employs a trainable Decision Maker to identify useful activations, and learns a transformation that approximates this bilingual connection when only one language is available at inference time. LAMCC adopts the broader idea that cross-lingual information can be exchanged inside the model rather than only through multilingual training data, but the transferred information is defined differently. Instead of connecting raw multilingual activations, LAMCC first extracts language-agnostic semantic representations, performs layer-wise self attention fusion between the two language views, and uses the resulting instance-dependent semantic signal to modify the main prediction path.

This design also makes LAMCC closer to the broader literature on activation steering and representation intervention. Rimsky et al. derives directions from differences between positive and negative examples and adds them to residual-stream activations during inference (Rimsky et al., 2024). Related interventions have been introduced in multimodal models: LLaVA Steering applies layer-wise linear transformations to rebalance visual and textual representations (Bi et al., 2025). Rather than learning or applying a reusable direction associated with a fixed behavior, LAMCC's intervention is constructed separately for each image–text instance from the semantic representations of two language views and is gathered adaptively across layers. The resulting signal therefore functions as instance-specific cross-lingual semantic supplementation rather than conventional behavior steering.

B Implementation Details

We use the same optimization configuration across comparable variants. In Stage 1, the disentangler is trained for 3 epochs with a learning rate of 1×10^{-4}, a per-device batch size of 2, and gradient accumulation of 4. The coefficients of the reconstruction, meaning-alignment, and orthogonality objectives are set to 1.0, 1.0, and 0.1, respectively, and the temperature used by the meaning-alignment objective is 0.05. In Stage 2, we train LoRA (Hu et al., 2021) and the Decision Maker for 3 epochs with a learning rate of 1×10^{-4}, a per-device batch size of 1, and gradient accumulation of 8. We set the LoRA rank to 8, the scaling factor to 32, and the dropout rate to 0.05. The Decision Maker temperature is set to 1.0. During inference, the meaning-space transformation matrix is estimated from 1,000 paired calibration samples using ridge regularization with coefficient 10^{-3}. We use bfloat16 computation, 4-bit model loading, and a fixed random seed of 20.