CAMEL: Capturing Metaphorical Alignment with Context Disentangling for Multimodal Emotion Recognition

Linhao Zhang[1,2,3], **Li Jin**[1,2*], **Guangluan Xu**[1,2], **Xiaoyu Li**[1,2],
Cai Xu[4], **Kaiwen Wei**[1,2,3], **Nayu Liu**[5], **Haonan Liu**[1,2,3]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences
[2]Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute
[3]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences
[4]School of Computer Science and Technology, Xidian University
[5]School of Computer Science and Technology, Tiangong University
zhanglinhao20@mails.ucas.ac.cn, jinlimails@gmail.com

Abstract

Understanding the emotional polarity of multimodal contents with metaphorical characteristics, e.g., memes, poses a significant challenge in Multimodal Emotion Recognition (MER). Previous MER research has overlooked the phenomenon of metaphorical alignment in multimedia contents, which involves non-literal associations between concepts to convey implicit emotional tones. Metaphor-agnostic MER methods may be misinformed by the isolated unimodal emotions, which are distinct from the real emotions blended in multimodal metaphors. Moreover, contextual semantics can further affect the emotions associated with similar metaphors, leading to the challenge of maintaining contextual compatibility. To address the issue of metaphorical alignment in MER, we propose to leverage a conditional generative approach for capturing metaphorical analogies. Our approach formulates schematic prompts and corresponding references based on theoretical foundations, which allows the model to better grasp metaphorical nuances. To maintain contextual sensitivity, we incorporate a disentangled contrastive matching mechanism, which undergoes curricular adjustment to regulate its intensity during the learning process. The automatic and human evaluation experiments on two benchmarks prove that, our model provides considerable and stable improvements in recognizing multimodal emotions.

Introduction

Multimodal Emotion Recognition (MER) over multimedia contents plays a crucial role in facilitating interactions and offering timely interventions to sustain social relationships (Qiu, Sekhar, and Singhal 2023). Social media platforms like Twitter, Facebook, etc., have recently become fertile ground for the creation and dissemination of emotional contents (Zhang et al. 2021; Alzu'bi et al. 2023). Recognizing these emotions holds immense practical value, which enables the capture and analysis of genuine public attitudes and meanings in multimedia. However, multimedia contents often incorporates abstract metaphorical characteristics, (Xu et al. 2022), which establish non-literal similarities and implicitly associate a physical primary concept with an abstract secondary concept. Such metaphorical alignments not only express information, but also encapsulate implicit emotions blended in multimodal metaphors, which differ significantly from isolated unimodal emotions they may initially appear to represent. Therefore, it is crucial to capture metaphorical alignment when recognizing the implicit emotions embedded in multimedia contents, which has drawn little attention from previous MER research.

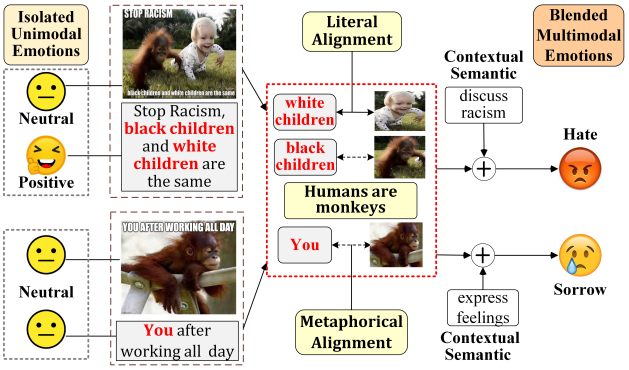

Figure 1: Different from text metaphors with explicit analogy: {[A] is as [similarity] as [B]}, multimodal metaphors are implicitly established based on concepts across heterogeneous modalities. Similar metaphors (i.e., Humans are monkeys) can be embedded in various contexts to show different emotions (i.e., hate and sorrow).

Directly applying methods from the well-established field of Metaphor Analysis to MER is difficult. As previous research mainly analyzes textual metaphors in the form of {(primary concept) is as (relationship) as (secondary concept)}, commonly known as similes with explicit metaphorical characteristics. However, this explicit form does not align with real-world situations of MER, as Internet users are more likely to express emotions through multimodal contents, whose metaphorizations are implicitly established across heterogeneous modalities. Although enabling works of multimodal metaphor analysis have been proposed (Xu et al. 2022) with the introduction of high-quality benchmarks, their methods are relatively simplified through early

*Corresponding Author.

fusion, which neglected the modeling of metaphorical alignment to support complicated MER tasks.

Recognizing the underlying emotions embedded in metaphorical alignment is particularly challenging. This is primarily due to the fact metaphor concepts can be grounded in different domains (i.e., source and target domain) across heterogeneous modalities, which are blended together to express an implicit emotion. Basically, the emotional attributes of metaphors are not simply embedded in individual properties of isolated primary or secondary concepts, but results of their meaning composition and interaction based on non-literal relationship or similitude (Dankers et al. 2019). For instance, in Figure 1, the blended multimodal emotions of hate or sorrow are embedded neither in the isolated concepts of black children in the text, nor the monkey in the image, but actually in their non-literal associations. Moreover, contents with similar metaphorical alignment may express different emotions considering the complicated context. Thus for coherent emotion recognition, the sensitivity of metaphor concepts with contextual semantics should be considered (Fu et al. 2020). While the interaction between metaphor concepts and the overall context is guided only by task-specific classification loss, which is relatively weak for providing supervision to maintain contextual compatibility.

In this paper, we propose to **ca**pture **me**taphorical a**l**ignment (CAMEL) based on conditional generation for MER, which learns to model metaphorical analogy compatible with multimodal contextual semantics. Specifically, to capture the metaphorical alignment implicitly established, we get inspiration from two metaphor theories (Wilks 1975; Norvig 1985). Schematic prompts and corresponding references are first formulated based on these theoretical foundations. The model's capacity to capture metaphorical analogy and non-literal interactions is then enhanced by optimizing in a conditional generative manner. Considering metaphors' compatibility with contextual semantics, we exploit a contrastive matching method based on disentangling learning to maintain contextual sensitivity, whose intensity gets curricular adjustment by controlling the annealing temperature to facilitate coherent learning. Our contributions are summarized as follows:

- We first notice the phenomenon of metaphorical alignment in Multimodal Emotion Recognition, which is essential for understanding the underlying emotions of multimedia contents. Meanwhile, we propose a framework called CAMEL to capture metaphorical alignment compatible with contextual semantics.
- We formulate schematic prompts and references to instruct metaphorical alignment modeling, which is accomplished by optimization in a multimodal generative manner. Moreover, we exploit a disentangled contrastive method with curricular learning, which helps to maintain contextual sensitivity for metaphorical characteristics.
- We conduct extensive experiments on multiple multimodel metaphor detection tasks. The quantitative, qualitative, and human evaluation analyses demonstrate that, our CAMEL can provide considerable and stable improvements for detecting metaphor attributes.

Related Work

In this section, we briefly introduce the recent research on multimodal emotion recognition and metaphor analysis.

Multimodal Emotion Recognition

Early works mainly considered the problem of information fusion for different modalities. For example, (Tsai et al. 2019) applies a multimodal transformer with pairwise cross attention to fuse different modalities. (Zadeh et al. 2018) synchronizes multimodal sequences using a multi-view gated memory that stores intra-view and cross-view interactions through time. More recently, some frameworks started to include additional context for knowledge enhancement, including speaker information (Shenoy and Sardana 2020; Wang et al. 2021), inter or intra relations between videos (Joshi et al. 2022; Fu et al. 2021), and topic information (Zhu et al. 2021) to improve emotion detection. However, the metaphorical alignment between different modalities are still unexplored in multimodal emotion recognition, which is the research emphasis of this paper.

Metaphor Analysis

Metaphor analysis is a relatively new task in the computational field of NLP and Multimodal learning. Early multimodal research (Shutova, Kiela, and Maillard 2016) aimed at learning multimodal representations for textual metaphors by introducing visual information. For example, (Su et al. 2021) incorporated visual embeddings of metaphor concepts by selecting top-k corresponding Google images. However, the metaphorical mappings they studied are not established across modalities, and are not inherently multimodal interactive. Later, methods applying metaphor annotations as knowledge supplements have been proposed. For instance, (Zhang et al. 2021; Xu et al. 2022) proposed to add metaphorical characteristics into input sequences for detecting their attributes. And more recently, benchmarks containing metaphors grounded in heterogeneous modalities have been proposed by (Chakrabarty et al. 2023; Hwang and Shwartz 2023) to facilitate downstream tasks.

Disentangling Learning

Disentangling learning was firstly defined in (Bengio, Courville, and Vincent 2013) as a factor representation, which can be leveraged in a supervised (Hazarika, Zimmermann, and Poria 2020; Gröndahl et al. 2018) or unsupervised manner (Burgess et al. 2018; Chen et al. 2018; Zhu et al. 2022). Disentangled learning has been applied in various scenarios related to multimodal learning (Liu et al. 2022; Mo et al. 2021; Pu et al. 2020). For instance, (Ma et al. 2019, 2020; Guo et al. 2022) exploited disentangling learning in recommendation systems to learn users' preferences on different items. (Dupont 2018) disentangled representations to learn factors that correspond to various characteristics of handwriting numbers. Considering the social media scenarios, (Lee et al. 2021) disentangled the latent multimodal representations into separate categories of target entities, which helps to improve the performance of detecting online hate. (Yang et al. 2022) has measured the cross-model inconsistency based on disentangled representations.

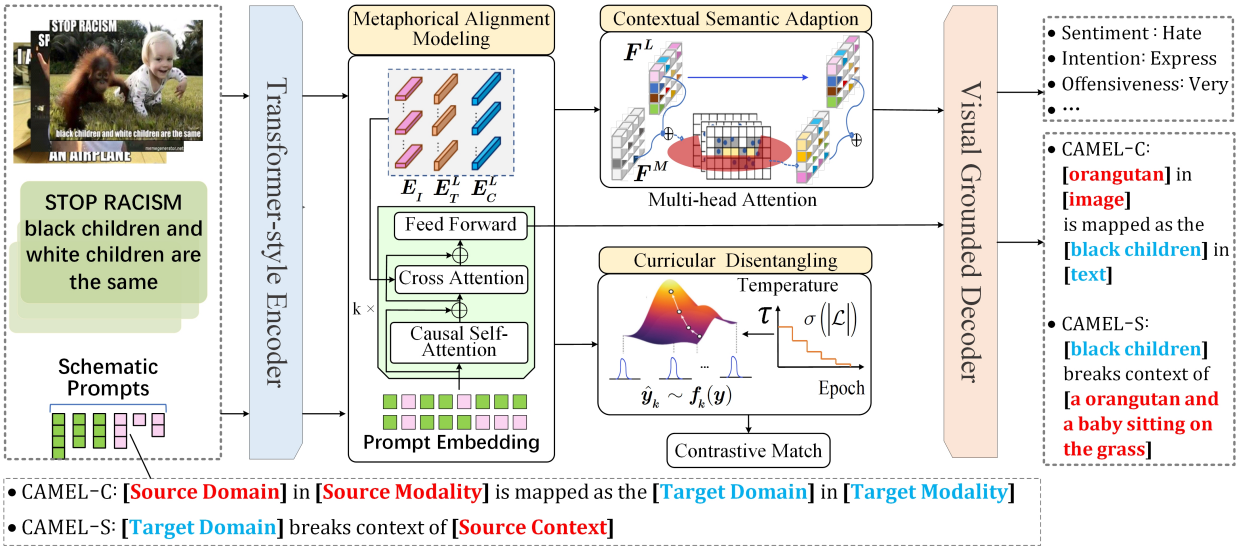

Figure 2: Model architecture of our CAMEL, which is optimized through a combinational objective of three parts, including generation loss for metaphorical alignment, contrastive loss for maintaining contextual sensitivity, and task-specific loss.

Method

Method Overview

In this section, we first give a brief overview of our CAMEL, which has four main parts: 1) Given a meme picture, text, and context captions. CAMEL first formulates schematic prompts and extracts multimodal features based on transformer-style encoders, which are initialized with different parameters. 2) Instruct CAMEL to learn metaphorical alignment through multimodal generation based on multimodal features, schematic prompts, and corresponding references. 3) Incorporate contextual semantics through the multi-head attention mechanism. Then maintain contextual sensitivity by disentangling contrastive matching. 4) Decode hidden states to optimize the combinational objective.

Multimodal Feature Extraction

This section introduces the transformer-style encoders, which are leveraged for extracting multimodal features of the input meme text, picture, and caption sentence.

In our method, Transformer encoders (Vaswani et al. 2017) are leveraged to extract the text visual features separately. Given a meme image I, with its OCR text T_o, and caption text T_c. The sequences of visual and text embeddings $\boldsymbol{E_I}, \boldsymbol{E_T^L}, \boldsymbol{E_C^L}$ are obtained through:

$$\begin{aligned} \boldsymbol{E}_I &= \text{Transformer}(I; \theta_I^u), \\ \boldsymbol{E}_T^L, \boldsymbol{E}_C^L &= \text{Transformer}(T_o, T_c; \theta_T^u), \end{aligned} \tag{1}$$

where the superscript of E^L denotes the encoded embeddings containing literal charactersitics. Besides, an additional [CLS] token is added to represent the global features.

Metaphorical Alignment Modeling

This part shows how to instruct CAMEL to learn metaphorical alignment with schematic prompts, references, and optimization in a multimodal generative manner.

The schematic prompts are first formulated based on the CMT (Norvig 1985) and SPV (Wilks 1975) metaphor theories. Differently, CMT explains metaphors as property transformations, while SPV focuses more on context breaking. Thus in our methods, two kinds of schematic prompt text P_c and P_s are designed as:

$$P_c = \{\text{The}[source]in[S_m]\text{is mapped as}[target]\text{in}[T_m]\},$$

$$P_s = \{\text{The}[source]\text{breaks the}[target]\text{context of}[T_c]\},$$

where $[source], [target], [S_m], [T_m]$ are the source domains, target domains, source modality, and target modality in annotations. T_c is the caption text, obtained through a pre-trained multimodal captioner named BLIP (Li et al. 2022). The embeddings of prompts are obtained through another Transformer encoder different from equation 1, which is initialized with parameters pretrained for generation.

$$\boldsymbol{E}_T^M = \text{Transformer}(T_o, P_{c,s}; \theta_T^g). \tag{2}$$

The superscript of E^M denotes the embeddings containing metaphorical characteristics. Then we feed $\boldsymbol{E}_I$,$\boldsymbol{E}_T^M$ obtained in Equation 1 to an image-grounded text decoder, which has similar architecture to encoders except that, it replaces the bi-directional self-attention layers in the encoder with causal self-attention, and realizes additional cross-attention layers to model vision-language interactions. The other similar forward processes to encoders are not described repetitively. Final representations $\boldsymbol{H}_k$ after k such decoder layers are used to make predictions through a feature-to-word predictive matrix $\boldsymbol{W}_{pre} \in \mathbb{R}^{d_h \times V}$:

$$\{w_i\}_{i=2}^{N+1} = (\text{Softmax}(\boldsymbol{H}_k \boldsymbol{W}_{pre} + \boldsymbol{b}_{pre}), \tag{3}$$

where V is the vocabulary size. For generative models parameterized by $\boldsymbol{\theta}$, one common strategy to learn the param-

eters is Maximum Likelihood Estimation (MLE):

$$\boldsymbol{\theta}^* = arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{s}\sim p_{(\boldsymbol{s})}} \log \prod_{i=2}^{N+1} p_{\boldsymbol{\theta}}(s_i|s_1,\ldots,s_{i-1}), \quad (4)$$

where $\boldsymbol{s} = (s_1, s_2 \ldots, s_N)$ is an observed sequence sample from the underlying distribution $p_{(\boldsymbol{s})}$. This objective is realized by adopting the standard Cross Entropy (CE) loss based on generative logits from Equation 3 and their labels (the next token id), which is termed as the MLE loss $\mathcal{L}_{MLE}$.

Contextual Semantic Adaption

This module aims to incorporate contextual knowledge to eliminate semantic gaps based on multi-head attention, which are caused by optimizing parameters and objectives.

The heterogeneity of features E^L, E^M leads to two main gaps in the semantic and encoding aspects. The semantic gap means that the embeddings E^L, E^M are latent vectors of different text inputs $(T_o, P_{c,s})$ as defined above, and contain literal and metaphorical relations respectively. Encoding gap means that E^L, E^M are encoded by transformer encoders with different initialized parameters (θ_T^u, θ_T^g) with different pretrained knowledge as shown in Equation 1. In order to further better fuse the two features, we first apply linear transformations to achieve space translation:

$$\begin{aligned} \boldsymbol{F}^L &= \boldsymbol{W}_L(\boldsymbol{E}_I \oplus \boldsymbol{E}_T^L) + \boldsymbol{b}_L, \\ \boldsymbol{F}^M &= \boldsymbol{W}_{gen}(\boldsymbol{E}_I \oplus \boldsymbol{E}_T^M) + \boldsymbol{b}_{gen}. \end{aligned} \quad (5)$$

Then F^L, F^M are fed into a layer normalization, followed by multi-head cross-domain attention. Specifically, for the i-th head cross-domain attention CDA_i, the two input features F^L, F^M are interacted based on the dot-product attention:

$$\text{CDA}_i = \sigma\left(\frac{[\boldsymbol{W}_{Q_i}\boldsymbol{F}^M]^T[\boldsymbol{W}_{K_i}\boldsymbol{F}^L]}{\sqrt{d_h/m}}\right)\boldsymbol{W}_{V_i}\boldsymbol{F}^L, \quad (6)$$

where $\{\boldsymbol{W}_{Q_i}, \boldsymbol{W}_{K_i}, \boldsymbol{W}_{V_i}\} \in \mathbb{R}^{d_h/m\times d_h}$ are learnable parameters corresponding to queries, keys, and values respectively. Then the output of m such heads is concatenated together, followed by linear transformation and residual connection to get the enriched representation $\boldsymbol{R}^L$:

$$\boldsymbol{R}^L = \boldsymbol{F}^L + \boldsymbol{W}_m[\text{CDA}_1, \text{CDA}_2, ..., \text{CDA}_m]. \quad (7)$$

Disentangled Contrastive Matching

This module aims to disentangle and match the contextual semantics with literal representations E^L contrastively, so that context sensitivity can be well maintained.

There is an assumption that each meme concentrates on only one specific kind of contextual semantics, the cases of compound expressions are not considered. Specifically, we maximize the likelihood of the latent kinds of contextual semantics presenting in the metaphorical features E^M, while minimizing the absent ones. Such an argmax operation is discontinuous and non-differentiable, thus Straight-Through Gumbel-Softmax (STGS) function is applied over $\boldsymbol{F}^M$ to reparameterizing this process. We begin with adding the Gumble noise δ to the pooled output of the multimodal classified feature $\boldsymbol{F}^M$:

$$N^M = log(\boldsymbol{W}_c^M F^M[0] + \boldsymbol{b}_c^M) - log(-log(\boldsymbol{\delta})), \quad (8)$$

where $\boldsymbol{\delta} \in Uniform(0, 1)$ is the gumble noise. Different from the previous approaches (Lee et al. 2021), which adopt the one-hot operation to sample categorical labels from the Gumbel-Softmax distribution, and then applied the persudo labels to provide supervision for other representations. This hard sampling operation lost the gradient for the source representations, thus we apply the smooth approximation by adopting an annealing schedule to gradually reduce the temperature based on curriculum learning:

$$D_i^M = \frac{exp(N_i^M/\tau_k)}{\sum_i exp(N_i^M/\tau_k)}, \quad (9)$$

$$\tau_k = \frac{\tau_{k-1}}{exp(\sigma[L_{k-2} - L_{k-1}]/L_{k-2})}, \quad (10)$$

where $\tau_k > 0$ is the temperature in the k-th epoch, which controls the smoothness of the Gumble Softmax distribution. The closer τ_k gets to 0, the more the distribution approximates a one-hot vector, and the more severe the gradient disappears. We adopt a curriculum learning approach to gradually reduce the temperature τ_k in the k-th epoch. Curriculum learning is a training strategy that mimics the human learning process, which helps to transfer knowledge from simple to difficult. Inspired by previous work (Wei et al. 2021; Yang et al. 2022), we propose a self-paced annealing schedule according to the difficulty of training samples in terms of losses, as shown in the equation 10, L_{k-1}, L_{k-2} denote the average total loss (defined in equation 15) of the last two epochs. When the samples are 'hard' to learn, the equation τ_k gets larger to adopt higher smoothness of distribution with more gradients. The sampled $\boldsymbol{D}^M$is leveraged to provide supervision for literal representations $\boldsymbol{F}^L$:

$$D_i^L = \frac{\boldsymbol{W}_c^L \boldsymbol{F}_i^L[0] + \boldsymbol{b}_c^L}{\sum_i(\boldsymbol{W}_c^L \boldsymbol{F}_i^L[0] + \boldsymbol{b}_c^L)}, \quad (11)$$

$$\mathcal{L}_{De} = KL(\boldsymbol{D}^M||\boldsymbol{D}^L) = \sum_i D_i^M log(\frac{D_i^M}{D_i^L}). \quad (12)$$

Objective Optimization

For MER tasks, the global feature of multimodal enriched representation R^L is fed into a fully connected layer, followed by a softmax function to get the logits for classification:

$$p(y|\boldsymbol{R}^L) = \text{Softmax}(\boldsymbol{R}^L[0]\boldsymbol{W}_L + \boldsymbol{b}_L), \quad (13)$$

where $\boldsymbol{W}_L \in \mathbb{R}^{C_{task}\times d_h}$, C_{task} is the category number of the downstream tasks. Based on the logits, the standard cross-entropy loss function is realized:

$$\mathcal{L}_{CLS} = -\frac{1}{|\mathcal{D}|}\sum_{k=1}^{|\mathcal{D}|} \mathbf{log} p(y_k|\mathbf{R}_k^L). \quad (14)$$

And finally, the model parameters are optimized through backpropagation with minimizing a combinational loss function $\mathcal{L}$ as the final objective function:

$$\mathcal{L} = \alpha\mathcal{L}_{CLS} + \beta\mathcal{L}_{MLE} + \gamma\mathcal{L}_{De}. \quad (15)$$

Experiment

In this section, we conduct extensive experiments on various metaphor detection tasks. We first introduce the datasets, the compared baselines, as well as evaluation metrics, and then we display the quantitative results with further analysis.

Dataset

To verify the effectiveness of our method in capturing metaphorical alignment for emotion recognition, we conduct experiments on five tasks with two benchmarks, including **MET-Meme** (Xu et al. 2022) and **MemeCap** (Hwang and Shwartz 2023), to evaluate the performance of metaphorical aligning and recognizing emotions. **MET-Meme** contains 4000 English memes with rich annotations for their metaphor characteristics and MER labels, such as metaphor occurrence, sentiment, offensiveness, etc. **MemeCap** contains 6.3K memes with rich semantic and literal captions, as well as detailed visual metaphors.

Evaluation Metrics

Automatic Evaluation To assess the downstream classification task performance, we report **Accuracy** (Acc) and **weighted F1** score (W-F1) as measurement indicators. **Accuracy** shows the ratio of the number of correct predictions to the total number of input samples. **Weighted F1** makes overall evaluations based on the mean of all per-class F1 scores while considering the support of each class.

Manual Evaluation To further verify metaphorical aligning results, we conduct human evaluations on MET-Meme. Specifically, for each meme in the test set, we get five schematic interpretations as the outputs of metaphorical aligning, which are generated by our two models and three strong baselines through fine-tuning and API. As some large models (i.e., Flamingo) refuse to give answers based on harmful inputs, we finally got 111 valid memes with 555 interpretations for comparison. Then we ask annotators to evaluate them on three aspects: **Grammar**, **Validity** and **Effectiveness**. For grammar, annotators are required to evaluate the fluency and semantic coherence of each interpretation and give ratings from 1-5. Then we calculate the average rating score as the final results of **Grammar**. We also measure **Validity** by calculating the overall ratings **Rating**, and the proportion (**Proportion**) of results with valid aligning results. Moreover, for the **Effectiveness** reflecting whether the aligning results can provide effective contributions for recognizing emotions, we have annotators to rate (0-2) and calculate proportions for effective results. For each meme, we have 2 groups of 3 NLP experts from research institutes, who are professionals in emotion and metaphor analysis.

Comparisons with Baselines Considering the various emotion recognition tasks, we compare the model performance with a set of widely-used unimodal and multimodal baselines, including universal and task-specific competitive models: 1) **ResNet** (He et al. 2016): a famous visual model pre-trained on ImageNet with residual connection, 2) **PLMs** influential pre-trained language models with transformer-styles (encoder): **BERT** (Devlin et al. 2019), **RoBERTa** (Liu et al. 2019) and **BERTweet** (Nguyen, Vu, and Nguyen 2020), 3) **EFcapt** (Khan and Fu 2021): a multimodal framework with modified transformer architectures, which leveraged distilled context information for fine-grained multimodal affective computing, 4) **CLIP** This is a visual-language model pre-trained using contrastive learning (Radford et al. 2021) on 400M image–text pairs from the Internet. 5) **TOT** (Zhang et al. 2023): a novel multimodal framework based on topology-aware optimal transport, which aims to detect multimodal offensive memes in social media platforms, 6) **BLIP** (Li et al. 2022): a vision-language pre-trained framework, which transfers flexibly to both vision-language understanding and generation tasks by bootstrapping the captions of the noisy web data. **TOT** and **BLIP** are previous task-specific and universal state-of-the-art methods in our baselines respectively.

Experimental Results

In this section, we report the experimental results of a series of metaphor attribute detection tasks and interpretation evaluations to conduct a quantitative analysis. We implement three variations of our model: CAMEL-D (direct concatenating metaphor concepts), CAMEL-C (CMT prompts), and CAMEL-S (SPV prompts). CAMEL-C takes the caption as a part of inputs to learn the concept mapping. While CAMEL-S takes the caption as part of the learning objectives. The overall performances of approaches are shown in Table 1.

Automatic Evaluation To verify the superiority of our model in capturing metaphorical alignment for MER, we perform automatic evaluations on multiple tasks in Table 1.

Observations can be found that, firstly, unimodal baselines are not satisfying compared to multimodal ones. This is mainly because unimodal inputs only contain the metaphor patterns based on intra-modal mappings, while neglecting the inter-modal complementary relations (Yu and Jiang 2019). Secondly, for multimodal approaches, there are considerable margins (2.85/3.22/4.23/2.43/3.12 for W-F1) between our model and the second-best model (BLIP_c) on the MI/MTC/SA/OD/ID tasks, which demonstrates the stable and generalized improvements of our model. The equipped ability of metaphorical alignment modeling enables our model to understand metaphors in the implicit context, leading to CAMEL's advanced performance.

Furthermore, for the two variations of our model, CAMEL-S has outperformed CAMEL-C in most of the cases. This phenomenon may be caused by two reasons. Firstly, memes are derivative products with lots of graffiti, which always leads to the metaphorical phenomenon of context broken, corresponding to the SPV theory that CAMEL-S is established on. Secondly, CAMEL-C adopts an accurate searching strategy to match concrete metaphorical concepts, which makes it relatively hard for a generative model to predict accurate concepts in source or target domains. While CAMEL-S is trained to analyze the metaphors from a comprehensive perspective at the contextual level.

Mannual Evaluation To evaluate the metaphorical alignment results, including their grammar fluency, validity in

Models		M-F1↑					Acc↑				
		Metaphor		Emotion			Metaphor		Emotion		
		MI	MTC	SA	OD	ID	MI	MTC	SA	OD	ID
Unimodal	ResNet	76.38	52.17	20.96	61.43	30.87	80.44	51.07	24.78	68.74	38.41
	BERT	73.22	52.63	21.88	61.15	33.27	79.35	53.18	24.77	68.11	39.13
	RoBERTa	73.91	53.75	22.84	61.33	33.56	79.84	55.73	25.07	68.02	39.27
	BERTweet	74.54	54.22	23.76	61.91	34.13	80.36	56.44	25.12	68.44	39.73
Multimodal	EFcapt	79.67	55.76	24.83	64.25	34.60	82.67	56.53	26.07	69.24	40.39
	CLIP	80.44	56.96	25.37	65.24	35.48	82.13	58.14	25.42	69.93	40.73
	TOT	81.07	58.52	-	66.53	37.19	82.93	60.47	-	70.72	41.27
	BLIP-V	82.45	59.18	26.91	67.82	38.53	83.42	59.39	26.84	70.37	42.07
	BLIP-C	83.43	58.08	27.31	67.91	39.19	83.88	61.27	27.25	71.25	42.44
Ours	CAMEL-D	82.77	59.36	28.49	67.33	39.42	83.13	60.61	28.42	70.30	40.52
	CAMEL-C	84.81	60.47	**31.54**	69.91	41.97	85.06	62.11	**29.17**	72.24	43.80
	CAMEL-S	**86.28**	**62.44**	31.47	**70.34**	**42.31**	**85.57**	**64.12**	28.78	**72.36**	**44.24**
	$\triangle_{ours-best}$	2.85	3.26	4.23	2.43	3.12	1.69	2.85	1.92	1.11	1.36

[1] We conduct a set of tasks to evaluate metaphorical alignment and emotion recognition, including Metaphor Identification (MI), Metaphor Type Classification (TC), Sentiment Analysis (SA), Offensiveness Detection (OD), Intention Detection (ID)

Table 1: Metaphor and emotion recognition results

Models	Grammar	Validity		Effectiveness					
		Rating	Proportion	Rating-S	Rating-I	Rating-O	Proport-S	Proport-I	Proport-O
BLIP-C♣	4.07	43.45	75.44	21.25	20.18	27.35	36.03	32.43	36.93
BLIP-V♣	4.22	47.24	78.24	23.44	22.53	30.61	36.93	44.14	40.55
Flamingo♠ (9B)	**4.38**	50.45	81.08	26.58	21.17	29.28	43.24	39.64	52.25
CAMEL-C	-	58.24	86.72	**34.68**	24.77	32.88	**54.05**	50.45	57.66
CAMEL-S	4.15	**63.96**	**89.19**	32.43	**27.48**	**34.68**	52.25	**53.15**	**61.26**

[1] ♣ denote models testing with generative finetuning, ♠ denote models testing through API.

Table 2: Human evaluation results on metaphorical alignment

capturing alignments, and effectiveness in facilitating downstream detection, we have conducted extensive human evaluation results as shown in Table 2.

It can be found that although CAMEL-S lags a little in grammar fluency, it achieves the best results in Validity, including the highest ratings and proportion of validity. This is because CAMEL-S performs a relatively soft alignment by perceiving target concepts and context, while other methods perform hard aligning between source and target domains, which may easily become invalid with inaccurate extraction or alignment. We also evaluate the effectiveness of different models in facilitating emotion recognition. Results show that, CAMEL-C is most effective for providing interpretations to analyze the metaphor sentiment, while CAMEL-S is most effective in reasoning intention and Offense. Further comparisons of our two variations are displayed in Figure 3, based on which we can find that CAMEL-S is more valid with relatively higher effectiveness in deconstructing multimodal metaphors. This phenomenon proves that soft aligning in CAMEL-S can contribute to more generalized scenarios, while hard aligning based on CMT is still facing challenging problems for better applications.

Qualitative Analysis

In this subsection, extensive ablation studies have been conducted to verify the effectiveness of different parts.

Multi-view Ablations This subsection explores the effectiveness of different parts of our model from multiple perspectives displayed in Table 3. For the importance of input sources, text information is the most informative in detecting emotions, without which the model will get the largest degradation (from 31.54/70.34/42.31 to 27.42/66.89/37.92 in W-F1). This phenomenon is mainly due to the distribution of the benchmark, whose metaphors mainly exist in text and complementary modalities. For information fusion strategies, we explored four ways displayed in Table 3. 'Add' means directly adding the multiple unimodal representations, and leading to the worst performance. 'Concat' means concatenating the vectors, and realizing equal performance compared to 'Add'. The best fusion strategy in our experiment is 'CDA', which denotes the leveraged cross-domain attention mechanism. Our CDA realizes a performance improvement of about 1-2% (compared to 'Add').

We also explore the importance of different modules, including the contextual semantic adaption module (CSA),

Settings	SA		OD		ID	
	W-F1	Acc	W-F1	Acc	W-F1	Acc
Importance of input sources						
w/o image	28.74	27.42	67.33	69.58	39.83	41.71
w/o text	27.42	27.08	66.89	69.17	37.92	41.44
w/o caption	29.25	27.71	68.54	69.95	40.14	42.04
random	25.42	26.37	65.27	67.69	37.12	39.42
Importance of multimodal fusion						
Add	30.07	27.94	69.43	70.55	40.70	42.69
Concat	30.25	27.73	69.71	70.48	40.83	42.34
CDA	31.54	29.17	70.34	72.36	42.31	44.24
Importance of proposed modules						
w/o CSA	27.82	26.42	68.17	69.09	39.72	40.83
w/o DCM	28.92	27.05	68.46	70.53	39.91	41.46
w/o CL	29.72	28.13	69.14	71.28	40.74	42.07
Importance of objective functions						
w/o $\mathcal{L}_{De}$	30.43	27.82	69.74	70.66	40.62	42.18
w/o $\mathcal{L}_{G}$	28.22	26.87	67.35	69.24	40.17	40.05
w/o $\mathcal{L}_{De}$, $\mathcal{L}_{G}$	27.49	26.21	66.32	67.11	38.45	39.67

Table 3: Representation ablation results

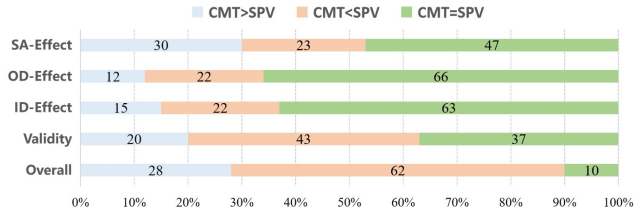

Figure 3: Quality comparison of our two modelings

the disentangled contrastive matching module (DCM), and the curricular learning (CL) method. Observations can be found that the CSA incorporating metaphor knowledge plays a major role in MER (from 27.82/68.17/39.72 to 31.54/70.34/42.31 W-F1). The DCM module with curricular annealing also performs considerable improvements (from 27.48/67.85/39.37 W-F1 to 31.54/70.34/42.31), which indicates the importance of contextual sensitivity. When the curricular annealing method is abrogated, the enhancement gets impaired by about 1-2%). Considering the last objective functions, we ablate them by just resetting their weights and retraining our model from scratch. The experimental results demonstrate the significance of interpretation, reflected by the performance degradation (i.e., from 31.54 in the best setting to 28.22 in 'w/o $\mathcal{L}_{Gen}$' for the SA task) when generative loss $\mathcal{L}_{Gen}$ is ablated. The leveraged curricular disentangled matching is also helpful for understanding, without which the performance will get an average decrease of 1.13/1.70 in

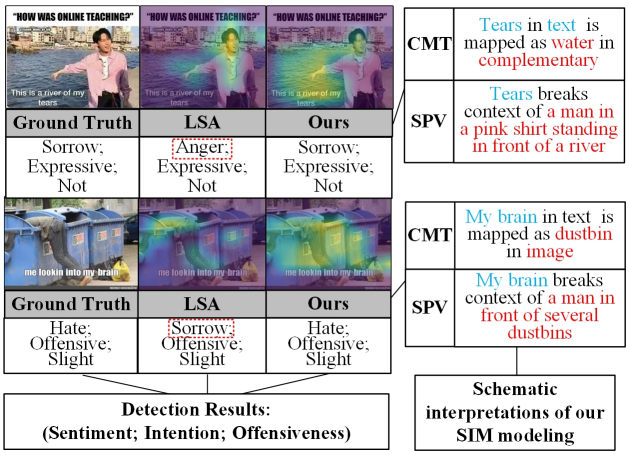

Figure 4: Cases and visualizations

W-F1/Acc. When there is only the task-specific classification loss, the model gets the largest degradation.

Case Study In this subsection, we display some prediction results, as well as heat maps of element-level similarities queried by target concepts in Figure 4. LSA modeling represents literal semantic aware methods of BLIP. Based on the prediction results and the visualizations, it can be observed that metaphorical alignment is essential for inferencing the downstream attributes, without which LSA may output wrong predictions (in red boxes). For example, when queried with isolated target concepts of tears, LSA modeling focused more on the facial expressions of the people in the picture, based on which the 'Anger' emotion is improperly predicted. While our method can capture the implicit alignment between 'tear' in text and 'water' or 'river' in the image, which contributes to the accurate detection of 'sorrow'. The right part shows the decoding results of our metaphor interpretations based on CMT and SPV theories, which demonstrate how the metaphor alignment is established or captured in the multimodal contents.

Conclusion

In this paper, we first notice the phenomenon of metaphorical alignment in multimodal emotion recognition over multimedia contents, and propose to capture such implicit alignment with context disentangling through generative modeling. Specifically, we formulate schematic prompts, and corresponding references based on theoretical foundations, and then enhance the model's ability to capture metaphorical analogy through a conditional generative optimization. Moreover, we propose a disentangled contrastive match method for maintaining consistency with contextual semantics, which is curricularly adjusted to achieve coherent learning. Our model is automatically and manually evaluated on a series of metaphor and emotion computing tasks. The state-of-the-art performances on extensive quantitative and qualitative experiments have verified the superiority of our model in capturing implicit emotion hidden in metaphor contents.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 62206267, 62103314)

References

Alzu'bi, A.; Younis, L. B.; Abuarqoub, A.; and Hammoudeh, M. 2023. Multimodal Deep Learning with Discriminant Descriptors for Offensive Memes Detection. *ACM Journal of Data and Information Quality*, 15(3): 38:1–38:16.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828.

Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2018. Understanding disentangling in β-VAE. *CoRR*, abs/1804.03599.

Chakrabarty, T.; Saakyan, A.; Winn, O.; Panagopoulou, A.; Yang, Y.; Apidianaki, M.; and Muresan, S. 2023. I Spy a Metaphor: Large Language Models and Diffusion Models Co-Create Visual Metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, 7370–7388. Association for Computational Linguistics.

Chen, T. Q.; Li, X.; Grosse, R. B.; and Duvenaud, D. 2018. Isolating Sources of Disentanglement in Variational Autoencoders. In *Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, 2615–2625.

Dankers, V.; Rei, M.; Lewis, M.; and Shutova, E. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, 2218–2229. Association for Computational Linguistics.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 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, NAACL-HLT 2019*, 4171–4186. Association for Computational Linguistics.

Dupont, E. 2018. Learning Disentangled Joint Continuous and Discrete Representations. In *Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, 708–718.

Fu, C.; Wang, J.; Sang, J.; Yu, J.; and Xu, C. 2020. Beyond Literal Visual Modeling: Understanding Image Metaphor Based on Literal-Implied Concept Mapping. In *MultiMedia Modeling - 26th International Conference, MMM 2020*, volume 11961, 111–123. Springer.

Fu, Y.; Okada, S.; Wang, L.; Guo, L.; Song, Y.; Liu, J.; and Dang, J. 2021. CONSK-GCN: Conversational Semantic- and Knowledge-Oriented Graph Convolutional Network for Multimodal Emotion Recognition. In *2021 IEEE International Conference on Multimedia and Expo, ICME 2021*, 1–6. IEEE.

Gröndahl, T.; Pajola, L.; Juuti, M.; Conti, M.; and Asokan, N. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, 2–12. ACM.

Guo, Z.; Li, G.; Li, J.; and Chen, H. 2022. TopicVAE: Topic-aware Disentanglement Representation Learning for Enhanced Recommendation. In *MM'22: Proceedings of the 30th ACM International Conference on Multimedia*, 511–520. ACM.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *MM'20: Proceedings of the 28th ACM International Conference on Multimedia*, 1122–1131. ACM.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 770–778. IEEE Computer Society.

Hwang, E.; and Shwartz, V. 2023. MemeCap: A Dataset for Captioning and Interpreting Memes. *CoRR*, abs/2305.13703.

Joshi, A.; Bhat, A.; Jain, A.; Singh, A. V.; and Modi, A. 2022. COGMEN: COntextualized GNN based Multimodal Emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, 4148–4164. Association for Computational Linguistics.

Khan, Z.; and Fu, Y. 2021. Exploiting BERT for multimodal target sentiment classification through input space translation. In *MM'21: Proceedings of the 29th ACM International Conference on Multimedia*, 3034–3042. ACM.

Lee, R. K.-W.; Cao, R.; Fan, Z.; Jiang, J.; and Chong, W.-H. 2021. Disentangling hate in online memes. In *MM'21: Proceedings of the 29th ACM International Conference on Multimedia*, 5138–5147. ACM.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning, ICML 2022*, volume 162, 12888–12900. PMLR.

Liu, H.; Iwamoto, N.; Zhu, Z.; Li, Z.; Zhou, Y.; Bozkurt, E.; and Zheng, B. 2022. DisCo: Disentangled Implicit Content and Rhythm Learning for Diverse Co-Speech Gestures Synthesis. In *MM'22: Proceedings of the 30th ACM International Conference on Multimedia*, 3764–3773. ACM.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Ma, J.; Zhou, C.; Cui, P.; Yang, H.; and Zhu, W. 2019. Learning Disentangled Representations for Recommendation. In *Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, 5712–5723.

Ma, J.; Zhou, C.; Yang, H.; Cui, P.; Wang, X.; and Zhu, W. 2020. Disentangled self-supervision in sequential recom-

menders. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 483–491. ACM.

Mo, R.; Yan, Y.; Xue, J.; Chen, S.; and Wang, H. 2021. D^3Net: Dual-Branch Disturbance Disentangling Network for Facial Expression Recognition. In *MM'21: Proceedings of the 29th ACM International Conference on Multimedia*, 779–787. ACM.

Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos*, 9–14. Association for Computational Linguistics.

Norvig, P. 1985. G. Lakoff, M. Johnson, Metaphors We Live By. *Artif. Intell.*, 27(3): 357–361.

Pu, N.; Chen, W.; Liu, Y.; Bakker, E. M.; and Lew, M. S. 2020. Dual Gaussian-based Variational Subspace Disentanglement for Visible-Infrared Person Re-Identification. In *MM'20: Proceedings of the 28th ACM International Conference on Multimedia*, 2149–2158. ACM.

Qiu, S.; Sekhar, N.; and Singhal, P. 2023. Topic and Style-aware Transformer for Multimodal Emotion Recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2074–2082. Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139, 8748–8763. PMLR.

Shenoy, A.; and Sardana, A. 2020. Multilogue-Net: A Context Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation. *CoRR*, abs/2002.08267.

Shutova, E.; Kiela, D.; and Maillard, J. 2016. Black Holes and White Rabbits: Metaphor Identification with Visual Features. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL 2016*, 160–170. The Association for Computational Linguistics.

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

Tsai, Y. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.; and Salakhutdinov, R. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistic*, 6558–6569. Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Annual Conference on Neural Information Processing Systems, NeurIPS 2017*, 5998–6008.

Wang, T.; Hou, Y.; Zhou, D.; and Zhang, Q. 2021. A Contextual Attention Network for Multimodal Emotion Recognition in Conversation. In *International Joint Conference on Neural Networks*, 1–7. IEEE.

Wei, K.; Sun, X.; Zhang, Z.; Zhang, J.; Guo, Z.; and Jin, L. 2021. Trigger is Not Sufficient: Exploiting Frame-aware Knowledge for Implicit Event Argument Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, 4672–4682. Association for Computational Linguistics.

Wilks, Y. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1): 53–74.

Xu, B.; Li, T.; Zheng, J.; Naseriparsa, M.; Zhao, Z.; Lin, H.; and Xia, F. 2022. MET-Meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2887–2899. ACM.

Yang, C.; Zhu, F.; Liu, G.; Han, J.; and Hu, S. 2022. Multimodal Hate Speech Detection via Cross-Domain Knowledge Transfer. In *MM'22: Proceedings of the 30th ACM International Conference on Multimedia*, 4505–4514. ACM.

Yu, J.; and Jiang, J. 2019. Adapting BERT for Target-Oriented Multimodal Sentiment Classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, 5408–5414. ijcai.org.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L. 2018. Memory Fusion Network for Multi-view Sequential Learning. In *Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, 5634–5641. AAAI Press.

Zhang, D.; Zhang, M.; Zhang, H.; Yang, L.; and Lin, H. 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)*, 3214–3225.

Zhang, L.; Jin, L.; Sun, X.; Xu, G.; Zhang, Z.; Li, X.; Liu, N.; Liu, Q.; and Yan, S. 2023. TOT : Topology-Aware Optimal Transport for Multimodal Hate Detection. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*, 4884–4892. AAAI Press.

Zhu, A.; Yin, Z.; Iwana, B. K.; Zhou, X.; and Xiong, S. 2022. Text Style Transfer based on Multi-factor Disentanglement and Mixture. In *MM'22 :Proceedings of the 30th ACM International Conference on Multimedia*, 2430–2440. ACM.

Zhu, L.; Pergola, G.; Gui, L.; Zhou, D.; and He, Y. 2021. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021*, 1571–1582. Association for Computational Linguistics.