ExKnow: Dynamic Example Selection and Knowledge-Enhanced Prompting for Cross-Domain Multimodal Sentiment Analysis Without Retraining

Zhenhua Feng[1,2] **Di Wu**[1,2] **Yan Xiang**[1,2] **Wenjun Wang**[1,2,*]
[1]Kunming University of Science and Technology, China
[2]Yunnan Key Laboratory of Artificial Intelligence, China
[*]Correspondence: 20250156@kust.edu.cn

Abstract

Multimodal Sentiment Analysis (MSA) aims to recognize human sentiment states by integrating heterogeneous signals from text and image. However, existing models face significant challenges in cross-domain scenarios due to distribution shifts. While Large Language Models (LLMs) offer strong generalization, their massive computational overhead hinders practical deployment. Conversely, small-scale LLMs are efficient but often fail to handle unseen domains without costly retraining. In this paper, we propose ExKnow (Example-Knowledge Enhanced Prompting), a lightweight inference-stage framework that empowers small-scale LLMs for cross-domain MSA. ExKnow bypasses the need for parameter updates through two synergistic modules: (1) Dynamic Example Selection (DES), which retrieves semantically similar and label-balanced source-domain samples to provide instance-level reasoning patterns; and (2) Domain Knowledge Generation (DKG), which synthesizes target-specific prior knowledge to interpret domain-unique sentiment expressions. By integrating these cues into a unified, knowledge-enhanced prompt, ExKnow enables a source-tuned small LLM to generalize effectively across diverse target domains. Experimental results demonstrate that ExKnow outperforms the compared baselines in the two primary transfer tasks and improves the backbone in most evaluated transfer directions, providing a high-performance solution for scalable cross-domain MSA.

1 Introduction

Multimodal Sentiment Analysis (MSA) integrates text and images to identify sentiment. Existing fusion methods, including contrastive learning (Hu et al., 2022) and disentangled representation learning (Wang et al., 2025), perform well when training and test data follow similar distributions but degrade under domain shifts. Domain adaptation approaches based on adversarial learning (Ganin and Lempitsky, 2015), reconstruction (Liu et al., 2017), or class-centered contrastive learning (Kang et al., 2019) mitigate this problem, yet typically require retraining for each target domain. LLM-based methods instead convert visual inputs into text and exploit strong semantic priors (Chay-intr et al., 2025; Huang et al., 2025; Bhattacharyya and Wang, 2025); however, large models are costly to deploy, whereas smaller models remain sensitive to unseen-domain distributions.

We therefore study whether a small LLM trained on one source domain can adapt to multiple unseen target domains without retraining. We propose ExKnow (Example-Knowledge Enhanced Prompting), a lightweight inference-stage framework with two complementary modules. Dynamic Example Selection (DES) retrieves semantically relevant, label-balanced source examples that provide diverse reasoning patterns, while Domain Knowledge Generation (DKG) supplies target-specific knowledge for interpreting domain-dependent sentiment expressions. The small LLM is fine-tuned once on source-domain text-caption pairs; at inference, DES examples and DKG knowledge are combined in a unified prompt. This design enables efficient and scalable cross-domain MSA without target-domain parameter updates. We will release the code upon acceptance.

2 RELATED WORK

2.1 Multimodal sentiment analysis

MSA predicts sentiment polarity from complementary modalities such as text, images, and audio (Zhang et al., 2022a). Representative methods model cross-modal dependencies through attention (Poria et al., 2016; Shi and Huang, 2023) or unify sentiment analysis and reasoning as a generative task (Hu et al., 2022). These approaches generally assume matched training and test distributions. ExKnow retains the generative multimodal formu-

lation but uses knowledge-enhanced prompting to improve robustness under domain shift.

2.2 Domain adaptation

Unsupervised domain adaptation aligns labeled source and unlabeled target distributions. Prior methods learn domain-invariant representations through gradient reversal (Ganin and Lempitsky, 2015), auxiliary label-distribution alignment (Tang et al., 2020), or information-bottleneck representations with label-distribution matching (Zhang et al., 2024). Because these methods typically retrain for every new target domain, their scalability is limited. ExKnow instead combines DES and DKG for plug-and-play inference-time adaptation.

2.3 Large language model fine-tuning

Full-parameter instruction tuning is computationally expensive (Wei et al., 2021). LoRA reduces trainable parameters through low-rank updates (Hu et al., 2021), but target-domain tuning still requires labeled target data. ExKnow fine-tunes a small LLM only on the source domain, then uses DKG to inject target-specific prior knowledge at inference time without further parameter updates.

3 METHODOLOGY

The overall architecture of our proposed framework is illustrated in Fig 1. We consider the source domain dataset $\mathcal{D}_s = \{x_i^s, c_i^s, y_i^s\}_{i=1}^{N_s}$, where x_i^s denotes textual input, c_i^s represents the image caption, and y_i^s is the corresponding sentiment label. The unlabeled target domain dataset is $\mathcal{D}_t = \left\{x_j^t, c_j^t\right\}_{j=1}^{N_t}$, with N_s and N_t indicating the number of samples in each domain. Our approach operates in two main phases: (1) source domain fine-tuning, which establishes base sentiment reasoning capabilities, and (2) target domain inference, where we introduce novel adaptation mechanisms to enable cross-domain generalization without retraining. The prompt design for target inference is detailed in Fig. 2.

3.1 Source Domain Fine-tuning

Although LLMs demonstrate strong textual semantic understanding, they lack the capability to directly process visual inputs. Small-scale multimodal models, on the other hand, often exhibit limited textual comprehension, which restricts their effectiveness in sentiment reasoning. Following

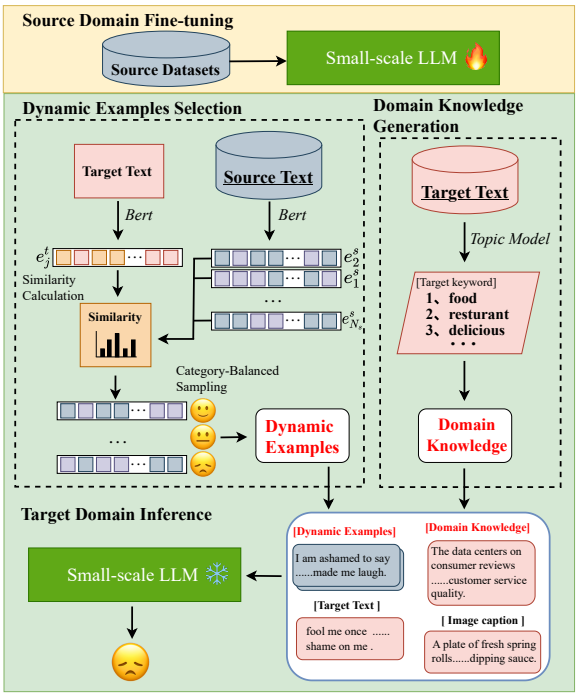


Figure 1: The overall framework of ExKnow.

prior studies (Guo et al., 2025), we adopt a caption-based strategy to bridge this gap: each image is transformed into a textual description c_i^s by a multimodal LLM (MLLM). The generated caption is then concatenated with the corresponding text to construct a unified multimodal input. To equip the model with preliminary sentiment reasoning ability, we fine-tune the small-scale LLM on the source-domain dataset $\mathcal{D}_s$, which contains paired text–caption inputs with sentiment labels. Specifically, we adopt supervised fine-tuning, where the model learns to predict the correct sentiment label given the joint text–caption representation. For parameter-efficient training, we employ LoRA ensuring that only a small subset of parameters is updated while keeping the backbone model frozen.

3.2 Target Domain Inference with Example-Knowledge Enhancement

During target domain inference, we enhance the source-fine-tuned model's cross-domain generalization through two novel lightweight modules: Dynamic Examples Selection and Domain Knowledge Generation.

3.2.1 Dynamic Examples Selection (DES)

To mitigate domain shift, we propose a dynamic example selection mechanism that retrieves semantically relevant and sentiment-balanced examples from the source domain as contextual guidance. The process consists of two steps:

User *Source Domain Fine-tuning*

You are a sentiment analysis expert. You need to read the input image caption and the given text and then determine its sentiment label. Sentiment labels include positive, negative, and neutral. Make sure to analyze both the image caption and the text to determine the sentiment.
Image caption: {Image caption c_i^s }, Text: {text x_i^s}.

Assistant
{sentiment label y_i^s}

User *Domain Knowledge Generation*

Please provide a general summary for a dataset. The dataset contains multiple topics, each with a set of keywords. Please generate a general summary based on the following topic keywords, briefly describing the main content and trends of the dataset. 10-50 words are required.
Keyword: {Image caption W_n }.

Assistant
{Target Domain Knowledge K}

User *Target Domain Inference*

You are a sentiment analysis expert.You need to read the input image caption and the given text and then determine its sentiment label. Sentiment labels include positive, negative, and neutral. Make sure to analyze both the image caption and the text to determine the sentiment.
Below are some examples:
Example 1: Image caption:{image caption c_1^s}, Text:{text x_1^s}.
.....
Example k: Image caption:{image caption}, Text:{text x_k^s}.
Now, please analyze the following example.
{Target Domain Knowledge K}
Image caption:{image caption c_j^t}, text:{text x_j^t}.

Assistant
{Sentiment label y_j^t}

Figure 2: The constructed prompt.

Semantic Retrieval: We encode the target text x_j^t and each source-domain text $\{x_i^s\}_{i=1}^{N_s}$ into feature vectors e_j^t and $\{e_i^s\}_{i=1}^{N_s}$using a BERT-based text encoder, and compute their cosine similarity:

$$\mathrm{sim}(x_j^t, x_i^s) = \frac{\langle e_j^t, e_i^s \rangle}{\|e_j^t\|_2 \|e_i^s\|_2}, \quad i = 1, \ldots, N_s \quad (1)$$

Label-Balanced Selection: Let $\mathcal{C}$ denote the set of sentiment labels. For each label $c \in \mathcal{C}$, we rank the source-domain examples with label c by their cosine similarity to the target text and select the top-k' examples. The resulting demonstration set contains $k = |\mathcal{C}|k'$ examples in total. In our three-class setting, $k = 3$ corresponds to one positive, one neutral, and one negative example. Thus, retrieval preserves semantic relevance while providing equal representation of the three sentiment labels.

3.2.2 Domain Knowledge Generation (DKG)

Small-scale LLMs often lack sufficient domain-specific knowledge, limiting their ability to interpret specialized terminology and contextual sentiment expressions in cross-domain scenarios. To address this, we introduce a knowledge generation module that enhances model adaptation through structured external knowledge.

We collect unlabeled text from the target domain and preprocess it by removing invalid characters, filtering stop words, and tokenizing the text. BERTopic then extracts representative target-domain keywords $W_n = \{w_1, \ldots, w_n\}$. We provide these keywords to GPT-4o through the prompt in Fig. 2 to generate a short target-domain description K. The description is included in the inference prompt without updating the parameters of the source-fine-tuned model.

3.2.3 Target Domain Inference

We concatenate the multimodal joint input $\left\{x_j^t, c_j^t\right\}$ from the target domain, the dynamic example set E_k, and the target domain knowledge K to form a target domain inference prompt (Fig. 2) . This composite input is then fed into the source-fine-tuned LLM, to perform sentiment prediction in the target domain.

4 EXPERIMENTS

4.1 Experiment Settings

Dataset. We conduct experiments on three benchmark datasets for multimodal sentiment analysis: MVSA(Xu and Mao, 2017), Yelp(Asghar, 2016), and Twitter2017(Rosenthal et al., 2017), all of which consist of paired text–image inputs annotated with three sentiment categories (positive, neutral, and negative). To rigorously assess cross-domain generalization, we adopt two experimental configurations. First, we follow the conventional two-domain transfer setting used in prior research, evaluating performance on MVSA→Yelp and Yelp→MVSA. Furthermore, we design a more rigorous multi-target adaptation setting to evaluate the model's scalability. Specifically, ExKnow is fine-tuned on a single source dataset and subsequently evaluated in a zero-shot manner across multiple remaining domains (e.g., MVSA → {Yelp, Twitter}).

Baselines. We compare ExKnow against a diverse set of competitive baselines, categorized into three major technical paradigms: (1) Conventional Multimodal Networks: We include TFN(Zadeh et al., 2017) and VistaNet(Truong and Lauw, 2019), which primarily leverage tensor fusion and hierarchical attention mechanisms to capture inter-modal dynamics. (2) Domain Adaptation (DA) Frameworks: We incorporate established DA approaches such as MMAN(Ma et al., 2019), MDANN(Qi et al., 2018), DiSRAN(Zhang et al., 2022b), and

Table 1: Cross-Domain Sentiment Analysis Results on MVSA → Yelp and Yelp → MVSA

Paradigm	Model	MVSA → Yelp			Yelp → MVSA		
		Prec.	Rec.	Macro-F1	Prec.	Rec.	Macro-F1
Traditional Multimodal Methods	TFN (EMNLP 2017)	0.330	0.331	0.331	0.322	0.316	0.164
	VistaNet (AAAI 2019)	0.403	0.413	0.331	0.404	0.403	0.351
DA Methods	MMAN (TMM 2019)	0.429	0.453	0.401	0.431	0.414	0.360
	MDANN (MM 2018)	0.469	0.463	0.439	0.401	0.403	0.369
	DiSRAN (TNNLS 2022)	0.538	0.510	0.492	0.467	0.456	0.424
	CASP (AAAI 2025)	–	–	0.457	–	–	0.440
LLMs-based Methods	Baichuan2-7B	0.481	0.478	0.478	0.403	0.437	0.403
	Llama3-8B	0.479	0.480	0.479	0.413	0.454	0.416
	ExKnow (Baichuan2-7B)	0.545	0.531	0.536	0.458	0.472	0.459
	ExKnow (Llama3-8B)	**0.555**	**0.560**	**0.556**	**0.487**	**0.518**	**0.498**

the state-of-the-art CASP(Guo et al., 2025). These methods emphasize bridging modality and domain gaps through adversarial learning or alignment, but typically require parameter updates for target-domain adaptation. (3) LLMs-based Methods: We fine-tune Baichuan2-7B (Yang et al., 2023) and Llama3-8B (Grattafiori et al., 2024) on the source domain using Low-Rank Adaptation (LoRA) and subsequently perform inference on the target domain.

Implementation Details. Experiments are conducted with LLaMA-Factory on two NVIDIA RTX 4090 GPUs (24GB each). We use Baichuan2-7B and Llama3-8B as the base LLMs. The maximum input sequence length is 2048 tokens. Source-domain training runs for 3 epochs with a batch size of 1, using AdamW and an initial learning rate of 5×10^{-5}. We apply LoRA with rank 8 and scaling factor 16. Qwen2-VL-7B generates image captions, and GPT-4o generates target-domain descriptions from the extracted keywords.

4.2 Main Results

Table 1 presents the performance comparison of ExKnow against various baselines in two primary transfer directions. The analysis leads to the following observations:

ExKnow (Llama3-8B) obtains the highest Macro-F1 among the compared methods in Table 1 on both transfer directions. Compared with CASP, its Macro-F1 is higher by 0.099 absolute points on MVSA→Yelp (0.556 vs. 0.457) and by 0.058 points on Yelp→MVSA (0.498 vs. 0.440). These results demonstrate the effectiveness of the complete framework in the two evaluated settings.

ExKnow yields remarkable performance gains over its backbone models (Baichuan2-7B and Llama3-8B). For instance, on the Yelp→MVSA task, ExKnow improves the macro-F1 of Llama3-8B from 0.416 to 0.498. This indicates that source-domain fine-tuning alone is insufficient for zero-shot domain transfer, underscoring the critical roles of our proposed DES and DKG modules in bridging distribution shifts.

ExKnow achieves these results without updating the source-fine-tuned model's parameters for a new target domain. Adapting to that domain still requires unlabeled target-domain text for keyword extraction, image captions from Qwen2-VL-7B, and a domain description generated by GPT-4o. We therefore distinguish the absence of target-domain parameter updates from the computational cost of preparing these inputs.

4.3 Generalization across Multiple Target Domains

Table 2 reports Macro-F1 across six transfer directions. ExKnow improves over the source-fine-tuned Baichuan2-7B backbone in all six directions and over Llama3-8B in five directions. The exception is Twitter→MVSA with Llama3-8B, where Macro-F1 decreases from 0.445 to 0.433. Thus, the method benefits most of the evaluated transfers, while its effect depends on the backbone and transfer direction.

4.4 Ablation Study

To evaluate the contribution of each component within the ExKnow framework, we conduct extensive ablation studies using Llama3-8B as the backbone, as summarized in Table 3. The results validate our design from the following aspects:

4.4.1 Component Ablation

The full model obtains the highest Macro-F1 in both evaluated directions: 0.556 on MVSA→Yelp

Table 2: Cross-domain performance (Macro-F1) comparison between the baseline and ExKnow across multiple target domains.

Setting	Twitter→MVSA	Twitter→Yelp	MVSA→Yelp	MVSA→Twitter	Yelp→MVSA	Yelp→Twitter	Avg.
Backbone: Baichuan2-7B							
Base Model	0.403	0.493	0.478	0.436	0.403	0.393	0.434
ExKnow (Baichuan2-7B)	**0.438**	**0.530**	**0.536**	**0.467**	**0.459**	**0.471**	**0.484**
Improvement	*+8.7%*	*+7.5%*	*+12.1%*	*+7.1%*	*+13.9%*	*+19.8%*	*+11.5%*
Backbone: Llama3-8B							
Base Model	0.445	0.459	0.479	0.467	0.416	0.423	0.448
ExKnow (Llama3-8B)	0.433	**0.496**	**0.556**	**0.494**	**0.498**	**0.488**	**0.494**
Improvement	*-2.7%*	*+8.1%*	*+16.1%*	*+5.8%*	*+19.7%*	*+15.4%*	*+10.3%*

Table 3: Ablation Study Results on MVSA → Yelp and Yelp → MVSA.

Ablation Group	Setting	MVSA → Yelp			Yelp → MVSA		
		Prec.	Rec.	Macro-F1	Prec.	Rec.	Macro-F1
Full Model	ExKnow (Llama3-8B)	0.555	0.560	0.556	0.487	0.518	0.498
Component Ablation	w/o DES & DKG	0.479	0.480	0.479	0.413	0.454	0.416
	w/o DES	0.557	0.555	0.549	0.485	0.505	0.493
	w/o DKG	0.527	0.534	0.529	0.468	0.521	0.483
Knowledge Form in DKG	Key-DKG	0.527	0.528	0.525	0.471	0.501	0.482
Exemplar Selection in DES	Rand-DES	0.522	0.529	0.524	0.479	0.507	0.489
	Sim-DES	0.545	0.549	0.547	0.490	0.507	0.496
Modality Contribution	Text Only	0.546	0.544	0.545	0.473	0.550	0.465

and 0.498 on Yelp→MVSA. Some ablations have higher precision or recall, so the full model is not best on every metric. Removing DES reduces Macro-F1 by 0.007 and 0.005, whereas removing DKG reduces it by 0.027 and 0.015, respectively. DKG therefore has the larger observed contribution in these two settings.

4.4.2 Knowledge Form in DKG

Using keywords alone yields Macro-F1 scores of 0.525 and 0.482, below the full model's 0.556 and 0.498. The keyword-only results are also slightly below those obtained without DKG (0.529 and 0.483). Consequently, these results support the complete generated description within the full framework, but do not show that keywords alone improve performance.

4.4.3 Exemplar Selection in DES

Similarity-only selection obtains Macro-F1 scores of 0.547 and 0.496, compared with 0.556 and 0.498 for the full selection strategy. The observed gains from label balancing are therefore 0.009 and 0.002. Random selection scores 0.524 and 0.489, below the corresponding results without DES (0.549 and 0.493). These comparisons show that example choice matters, while the small advantage over similarity-only selection should be interpreted cautiously without repeated-run results.

4.4.4 Modality Contribution

The "Text Only" setting retains DES and DKG but removes image captions. According to Table 3, adding captions increases Macro-F1 from 0.545 to 0.556 on MVSA→Yelp and from 0.465 to 0.498 on Yelp→MVSA. The corresponding gains are 0.011 and 0.033. Thus, the contribution of captions differs across the two transfer directions and is particularly relevant to Yelp→MVSA.

4.5 Comparison with Multimodal Large Language Models

The performance comparison in Figure 3 highlights the advantages of utilizing LLMs over direct multimodal fusion via MLLMs. ExKnow consistently outperforms both Qwen2-VL-7B and LLaVA-1.5-7B across all transfer directions. We attribute this superiority to the following factors:

First, for models of a comparable parameter scale, LLMs dedicate their entire capacity to linguistic reasoning, whereas MLLMs must allocate a portion of their parameters to process visual modal-

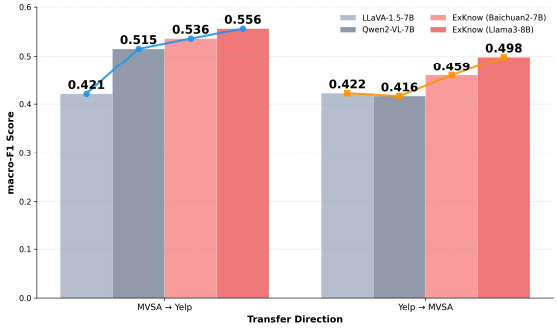


Figure 3: Macro-F1 performance comparison between MLLMs (LLaVA-1.5-7B/Qwen2-VL-7B) and ExKnow.

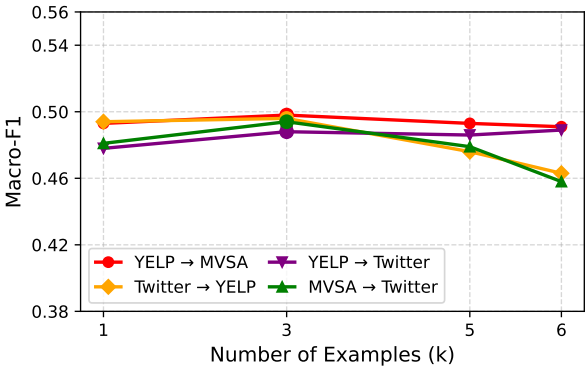


Figure 4: Ablation Study on the Number of Examples (k) in DES

ities. Given that cross-domain MSA remains an inherently text-centric task, the superior textual inference depth of LLMs provides a more robust foundation for capturing emotional nuances than MLLMs. Second, following prior studies (Guo et al., 2025), we adopt an image captioning strategy to translate visual signals into high-level linguistic descriptions. This approach allows the model to operate within a unified and stabilized semantic space, avoiding the generalization bottlenecks and model degradation that MLLMs often encounter during domain shifts.

4.6 Effect of the Number of Examples in DES

We vary the total number of selected demonstrations $k \in \{1, 3, 5, 6\}$, as shown in Fig. 4. The setting $k = 3$ uses one example from each sentiment class and achieves the best or near-best Macro-F1 on most evaluated transfers. Increasing the number to $k = 5$ or $k = 6$ does not produce a consistent further improvement. These results suggest that more demonstrations do not necessarily provide more useful context.

5 Conclusion

We presented ExKnow, which combines source-domain demonstrations and target-domain descriptions to adapt a source-fine-tuned LLM without target-domain parameter updates. The method improves Macro-F1 over its backbone in most evaluated transfer directions, although one Llama3-8B direction shows a decrease. In the two reported ablation settings, removing DKG produces larger decreases than removing DES. These findings show the potential of combining instance-level and domain-level information while defining the limits of the observed gains.

6 Limitations

ExKnow requires unlabeled target-domain text for keyword extraction and uses Qwen2-VL-7B and GPT-4o to prepare captions and domain descriptions. End-to-end latency, memory use, and API cost have not been quantified in this study. The gains from label-balanced retrieval are small in some settings, and the method does not improve every transfer direction.

Acknowledgments

This work is supported by Key basic research project of Yunnan Province (Grant NO. 202501AS070147), National Natural Science Foundation of China (Grant NO. 62266027, U21B2027, 62266028), Yunnan provincial major science and technology special plan projects (Grant NO. 202302AD080003), Yunnan Province Young and Middle-aged Academic and Technical Leaders Reserve Talent Program(Grant NO: 202305AC160063).

References

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*.

Sree Bhattacharyya and James Z Wang. 2025. Evaluating vision-language models for emotion recognition. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1798–1820.

Thodsaporn Chay-intr, Yujun Chen, Kobkrit Viriyayudhakorn, and Thanaruk Theeramunkong. 2025. Llavac: Fine-tuning llava as a multimodal sentiment classifier. *arXiv preprint arXiv:2502.02938*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zirun Guo, Tao Jin, Wenlong Xu, Wang Lin, and Yangyang Wu. 2025. Bridging the gap for test-time multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16987–16995.

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

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 7837–7851.

Dawei Huang, Qing Li, Chuan Yan, Zebang Cheng, Yurong Huang, Xiang Li, Bin Li, Xiaohui Wang, Zheng Lian, and Xiaojiang Peng. 2025. Emotion-qwen: Training hybrid experts for unified emotion and general vision-language understanding. *arXiv preprint arXiv:2505.06685*.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4893–4902.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30.

Xinhong Ma, Tianzhu Zhang, and Changsheng Xu. 2019. Deep multi-modality adversarial networks for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 21(9):2419–2431.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE.

Fan Qi, Xiaoshan Yang, and Changsheng Xu. 2018. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 429–437.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Tao Shi and Shao-Lun Huang. 2023. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766.

Hui Tang, Ke Chen, and Kui Jia. 2020. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8725–8735.

Quoc-Tuan Truong and Hady W Lauw. 2019. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 305–312.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. 2025. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21180–21188.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2399–2402.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Xinxin Zhang, Jun Sun, Simin Hong, and Taihao Li. 2024. Amanda: Adaptively modality-balanced domain adaptation for multimodal emotion recognition. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14448–14458.

Yawei Zhang, Liangqing Wu, Jingjing Wang, and Shoushan Li. 2022a. Multimodal emotion recognition based on multilayer LSTM fusion. *Journal of Chinese Information Processing*, 36(5):145–152. In Chinese.

Yuhao Zhang, Ying Zhang, Wenya Guo, Xiangrui Cai, and Xiaojie Yuan. 2022b. Learning disentangled representation for multimodal cross-domain sentiment analysis. *IEEE transactions on neural networks and learning systems*, 34(10):7956–7966.