Multimodal Emotion Understanding with Vision-Language Models: A Cross-Modal Attention Framework for Sentiment Analysis

Leo L. Hayes
School of Computing, Clemson University, Clemson, SC, USA.
leo.work@clemson.edu

Zixan Heng
Department of Computer Science, University of Houston, Houston, TX, USA.
zixuan778@uh.edu

Tejas Saini
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
saini2000@binghamton.edu

Abstract

The emergence of vision-language models has enabled substantial progress in multimodal sentiment analysis, where emotional cues arise from the interplay between visual expressions and textual content. This paper presents a systems-oriented investigation of a cross-modal attention framework for emotion understanding using large-scale vision-language models. We examine the architectural design space, discussing the trade-offs inherent in dual-encoder versus fusion-based attention mechanisms, and the challenges of aligning heterogeneous modalities for reliable sentiment classification. Beyond algorithmic performance, we focus on infrastructure requirements for training and deploying these models, including distributed computing, model serving, and edge-optimized variants. Robustness to noisy or adversarial multimodal inputs, fairness across demographic groups in facial affect recognition, and privacy concerns in real-world deployment are analyzed in depth. We further articulate governance frameworks and policy guidelines necessary for responsible fielding of emotion AI systems. The paper contributes a comprehensive perspective that bridges technical architecture, operational sustainability, and socio-technical considerations, illustrating how cross-modal attention frameworks can be designed not only for accuracy but also for equity, transparency, and deployability at scale.

Keywords

multimodal sentiment analysis; vision-language models; cross-modal attention; emotion understanding; system architecture; fairness; deployment infrastructure.

1. Introduction

Multimodal sentiment analysis has evolved from unimodal text or image classification into a rich interdisciplinary field that integrates visual, acoustic, and linguistic signals [1]. The proliferation of vision-language models pretrained on vast internet-scale image-text pairs, such as CLIP [2], has opened new frontiers for understanding nuanced emotional expressions that transcend the boundaries of a single modality. By jointly encoding visual and textual information, these models can capture subtle interplay—a sarcastic text paired with a smiling

face, or a joyful caption alongside a somber photograph—that unimodal systems inevitably miss. However, the effective fusion of modalities in sentiment analysis demands more than simple concatenation; it requires carefully designed cross-modal attention mechanisms that dynamically weight the relevance of each modality for the affective judgment.

A significant body of research has demonstrated that cross-modal attention, as realized in multimodal transformers, substantially improves sentiment and emotion recognition by allowing each modality to condition the representation of the other [3]. Yet, the majority of these efforts have focused narrowly on maximizing accuracy on curated benchmark datasets, often neglecting the broader system-level challenges that determine whether such models can be responsibly and reliably deployed. These challenges include the scalability of training pipelines for billion-parameter vision-language models, the computational and energy costs of real-time inference, the robustness of emotion predictions under distribution shift, and the fairness implications of applying facial expression analysis across diverse populations [4]. Moreover, the deployment of emotion AI in sensitive domains such as education, mental health, and public surveillance raises profound governance and policy questions that remain underexplored in the technical literature [5].

This paper provides an interdisciplinary examination of a cross-modal attention framework built on vision-language models for multimodal sentiment analysis, emphasizing structural trade-offs, infrastructure design, and socio-technical implications. Rather than presenting a singular algorithmic contribution, we adopt a systems lens to dissect how architectural choices—such as dual-encoder versus single-stream fusion, static versus adaptive attention, and centralized cloud versus federated edge deployment—interact with considerations of fairness, robustness, sustainability, and governance. Through this analysis, we aim to bridge the gap between high-performing emotion understanding models and the practical, ethical, and infrastructural realities of their deployment. The remainder of the paper is organized as follows. Section 2 reviews related work on multimodal sentiment analysis, attention mechanisms, and vision-language pretraining. Section 3 details the system architecture of our cross-modal attention framework. Section 4 discusses infrastructure and deployment strategies. Section 5 addresses robustness, fairness, and governance. Section 6 concludes with implications for future research and policy.

2. Related Work

2.1 Unimodal and Early Multimodal Sentiment Analysis

The sentiment analysis literature initially progressed along modality-specific trajectories. In text, the introduction of deep bidirectional transformers such as BERT [6] enabled models to capture complex semantic and syntactic dependencies, substantially outperforming earlier recurrent and convolutional architectures on benchmark sentiment datasets. In parallel, visual sentiment analysis leveraged deep convolutional networks, with architectures like ResNet [7] serving as backbones for transfer learning to emotion-related image classification tasks including facial expression recognition and art sentiment. However, unimodal analysis inherently misses the contextual richness provided by co-occurring signals, as human emotion is rarely expressed through a single channel. Early multimodal fusion approaches attempted to address this by combining features from different modalities either at the input level or at the decision level. The Tensor Fusion Network [8] introduced a principled tensor-based approach to model multimodal interactions, explicitly capturing the multiplicative relationships among text, video, and audio features. The Memory Fusion Network [9] extended this line by incorporating an attention-based memory module that tracks cross-modal interactions over

time, improving performance on sequential sentiment tasks. Despite their innovations, these methods relied on modality-specific feature extractors that were not jointly pretrained across modalities, limiting their ability to exploit shared semantic structures.

2.2 Vision-Language Pretraining and Cross-Modal Attention

The advent of vision-language pretraining marked a paradigm shift. Models such as ViLBERT [10] and VL-BERT [11] extended the BERT architecture to jointly process images and text, pretraining on large-scale image–caption datasets with tasks like masked language modeling and image-text matching. These architectures demonstrated that a unified representation space could be learned, where visual and linguistic tokens attend to each other through cross-modal attention layers, leading to strong performance on downstream vision-language tasks. For sentiment analysis, such pretrained models offer the potential to capture nuanced affective cues, such as irony or subtle facial micro-expressions, that depend on deep alignment between modalities. Yet most vision-language models have been applied to factual reasoning tasks rather than affective computing. Meanwhile, in text-only sentimentanalysis, the integration of cross-modal attention has remained a frontier where emotion recognition demands more than factoid alignment. The fine-grained coupling between visual expressions and emotionally charged language requires architectures that can interpret sarcasm, hyperbolic facial displays, and culture-specific gestures. Prior work on multimodal sentiment analysis with vision-language backbones has explored early fusion of CLIP embeddings [2] with lightweight attention bridges, but these efforts often treat the cross-modal attention as a static merging mechanism. A more promising avenue lies in adaptive attention frameworks that learn to weight modalities dynamically based on input characteristics, a principle that has shown considerable success in text-only sentiment classification through dynamic attention and contrastive objectives [12]. Extending such dynamic cross-modal attention to vision-language sentiment models introduces new system-level complexities, including the need to manage asymmetric information content—where text may carry the primary affective load and vision serves a disambiguating role—and the requirement to maintain robustness when one modality is missing or corrupted.

2.3 Cross-Modal Attention Mechanisms in Emotion AI

The architecture of cross-modal attention for emotion understanding can be situated within a broader landscape of attention-based fusion techniques. Transformer-based multimodal models such as MulT [3] defined pairwise cross-modal attention streams that allow each modality to refine its representation by attending to the other, establishing a framework later adopted for vision-language pretraining. For sentiment analysis, the VisualBERT architecture [13] extended single-stream transformers by treating visual region features as token inputs alongside text, enabling full self-attention across modalities from the early layers. This design fosters rich interaction, but can suffer from computational overhead and a tendency to overfit spurious visual-textual correlations. Co-attention mechanisms, where separate encoders produce modality-specific representations that are then aligned through a bidirectional cross-attention layer, offer a more parameter-efficient alternative that parallels the query-key-value attention paradigm used in retrieval-augmented generation.

In the emotion domain, localizing cross-modal attention to affective cues becomes critical. A facial expression may span only a few visual tokens, while sentiment-laden words like “devastatingly beautiful” carry disproportionate semantic weight. Thus, fine-grained token-level cross-attention, rather than global pooling, is essential. Additionally, hierarchical attention schemes that first attend within a modality to isolate salient affective regions—such

as eyes and mouth in faces, or emotionally charged n-grams in text—and then perform cross-modal alignment have shown improved resilience against noisy backgrounds or neutral filler text [14]. The system design must also account for the temporal dimension when video is involved, where cross-modal attention straddles not only modality but also time steps, creating a three-dimensional alignment problem that strains memory and compute budgets. Balancing this complexity against the latency requirements of real-time applications such as mental health chatbots or customer experience analytics is a central architectural trade-off that we address in the following section.

3. System Architecture of a Cross-Modal Attention Framework

3.1 Dual-Encoder versus Single-Stream Fusion Architectures

The foundational decision in designing a cross-modal attention framework for sentiment analysis is whether to adopt a dual-encoder architecture with late cross-attention or a single-stream architecture where visual and textual tokens are concatenated and jointly processed through unified transformer layers. Dual-encoder designs, exemplified by the original CLIP model, encode images and text with separate transformers and then compute a similarity score in a shared embedding space through a dot product or a shallow projection head. This approach is highly modular: the image encoder can be independently optimized, updated, or swapped without affecting the text encoder, and it naturally supports pre-computed indexing of image or text representations for efficient retrieval and zero-shot classification. For sentiment analysis, a dual-encoder with an added cross-attention module that fuses the two representations only at the final layers achieves a favorable trade-off between computational efficiency and multimodal alignment accuracy. The downside is that fine-grained, token-level interactions between modalities are delayed until the late fusion stage, potentially missing subtle dependencies such as a specific word (“irony”) influencing the interpretation of a specific facial region.

In contrast, single-stream architectures, in which visual patch embeddings and word token embeddings are interleaved and fed into a shared transformer encoder from the start, allow attention heads to freely attend across modalities at every layer. This design maximizes the cross-modal binding capacity and has demonstrated state-of-the-art performance on tasks that require tight visual grounding, including visual question answering and image captioning. For emotion understanding, early and dense cross-modal attention can capture how text sentiment colors the perception of a neutral face, or how a micro-expression recontextualizes an ambiguous phrase. However, the single-stream approach incurs a quadratic increase in the number of token interactions in the self-attention layers, which becomes particularly burdensome when high-resolution images or long text sequences are involved. For deployment in environments with constrained GPU memory or strict latency bounds, such as mobile emotion coaching apps, this cost can be prohibitive. Therefore, our cross-modal framework adopts a hybrid strategy: a dual-encoder backbone pretrained on large-scale vision-language data, followed by a lightweight but expressive cross-attention fusion network that can be finetuned on domain-specific sentiment data. This design preserves modularity and allows the bulk of the heavy lifting to be performed once during offline pretraining, while the fusion module can be adapted rapidly to new sentiment domains.

3.2 Dynamic Attention Allocation and Modality Gating

Static cross-attention, where every visual token attends to every text token with equal potential, introduces noise when one modality carries little affective signal. In many real-

world scenarios, the text “Great.” paired with a scowling face requires the model to heavily down-weight the linguistic signal in favor of visual evidence, while a detailed emotional description of a scene with a generic stock image demands the opposite. Our framework incorporates a dynamic modality gating mechanism that learns a scalar weight conditioned on the input representations, modulating the contribution of each modality before the cross-attention stage. This gate is a small multi-layer perceptron that takes as input summary vectors from both modalities—such as the mean-pooled representations of the textual and visual encoders—and outputs a soft weight that multiplies the value vectors in the cross-attention computation. The gating function can be viewed as a learned analog of modality reliability estimation, which has been shown to improve robustness to missing or degraded modalities [15].

Furthermore, within the cross-attention layer, attention heads can be specialized to attend to specific emotional dimensions. Attention head pruning and specialization strategies, akin to those used in multi-task learning for vision-language models, allow us to dedicate subsets of heads to discrete emotion categories such as valence, arousal, and dominance, or to sentiment polarity dimensions. This architectural choice not only improves interpretability—a crucial requirement for high-stakes applications of emotion AI—but also reduces interference between conflicting affective signals that may co-occur in an input, such as a text expressing sadness paired with an image depicting hope. The resulting attention maps can be visually inspected by human reviewers to verify that the model is focusing on relevant facial regions and emotionally salient words, thereby supporting human-in-the-loop governance processes.

3.3 Training Objectives and Multimodal Alignment

Training the cross-modal attention framework involves a combination of objectives designed to align visual and textual representations in a sentiment-aware joint space. The primary loss function is a contrastive objective that pulls together positive pairs—images and corresponding emotionally annotated texts—while pushing apart negative pairs in the embedding space. This is enhanced by a supervised contrastive loss that leverages sentiment labels to structure the representation space such that samples with the same sentiment polarity cluster together regardless of modality [12]. Such an objective encourages the model to learn sentiment-invariant cross-modal representations, which improves generalization to unseen affective expressions.

In addition to contrastive alignment, we apply a masked multimodal modeling objective during finetuning, where visual patches or text tokens are randomly masked and the model must predict them using cross-modal context. This forces the model to rely on cross-modal cues to reconstruct missing information, directly training the attention mechanism to bridge modalities in a semantically meaningful way. A cross-modal sentiment classifier head, consisting of a small transformer or multi-layer perceptron, is trained jointly on the fused representation to predict fine-grained emotion labels. To prevent catastrophic forgetting of the pretrained vision-language knowledge, we employ elastic weight consolidation and gradual unfreezing schedules, which preserve the rich semantic knowledge in the base encoders while allowing the fusion module to specialize. This multi-objective training regime necessitates careful hyperparameter balancing and distributed data parallelism, which we address in the infrastructure section next.

4. Infrastructure and Deployment Considerations

4.1 Training Infrastructure and Scalability

Pretraining and finetuning large vision-language models for multimodal sentiment analysis pose substantial infrastructure demands. The base encoders, typically ViT-Large or ViT-Huge for images and a correspondingly deep transformer for text, can have upwards of 300 million parameters each. Full model finetuning on high-resolution multimodal affective datasets such as AffectNet or CMU-MOSEI requires distributed training across multiple GPU nodes, often using model and data parallelism strategies. We adopt a sharded data parallel training paradigm with ZeRO optimization stages [16] to handle memory constraints, partitioning model states across devices. For the cross-modal fusion network, which is purposefully kept small, gradient accumulation and mixed-precision training allow efficient gradient updates without sacrificing numerical stability.

The heterogeneous nature of multimodal datasets—images with widely varying resolutions, text of different lengths, and potentially audio or video streams that must be synchronized—complicates the construction of efficient data pipelines. We employ dynamic batching with bucketing by sequence length and image aspect ratio to minimize padding overhead, and prefetching with multi-threaded data loaders to keep the GPUs saturated. In cloud environments, provisioning spot instances for non-critical pretraining stages and reserving on-demand instances for time-sensitive finetuning yields cost savings while meeting deadlines. The total compute budget for a full training cycle, including pretraining alignment, contrastive fine-tuning, and cross-modal fusion training, can exceed thousands of GPU-hours. Consequently, careful logging of energy consumption and carbon footprint, using tools that integrate with cloud provider telemetry, is essential to align with institutional sustainability commitments and emerging regulatory requirements on AI environmental reporting.

4.2 Inference Serving and Edge Deployment

Deploying cross-modal sentiment models in production environments introduces latency and throughput constraints that differ markedly from offline training. Applications such as real-time video sentiment analysis in call centers, emotion-aware educational software, or wearable mental health monitors demand inference latencies on the order of tens of milliseconds, often on resource-constrained edge devices. To meet these demands, we implement several serving optimizations. The dual-encoder design allows precomputation of text embeddings for static content, such as frequently used affective prompts or standard responses, while the image encoder can be replaced with a MobileViT or EfficientNet backbone for edge scenarios. The cross-modal fusion module, being shallow, can be compiled with inference-focused runtimes like ONNX Runtime or TensorRT, and quantized to INT8 precision with minimal accuracy degradation through quantization-aware training.

For applications requiring continuous video sentiment analysis, we incorporate temporal smoothing via a lightweight recurrent policy network on top of per-frame cross-modal predictions, reducing jitter while maintaining responsiveness. Model cascading is another effective strategy: a fast, deterministic rule-based system or a tiny distilled model can handle clear-cut cases, and only ambiguous inputs are escalated to the full cross-modal model, thereby reducing average serving latency and energy consumption. In settings where user privacy is paramount, such as in-car emotion monitoring for driver safety, the entire inference pipeline runs on-device, leveraging federated learning to periodically update models without raw data leaving the device. Federated averaging with differential privacy noise injection ensures that model updates do not leak sensitive emotional data, a point we elaborate on in Section 5.

4.3 Energy and Environmental Sustainability

The carbon footprint of training and deploying large-scale AI models has drawn increasing scrutiny from researchers and policymakers. While large vision-language models provide remarkable capabilities, their environmental cost must be weighed against the benefits, particularly for applications like sentiment analysis that may be deployed at a vast scale across millions of user interactions daily. Architected choices such as the dual-encoder with lightweight fusion significantly reduce the inference-time compute compared to single-stream models, as the heavy cross-modal attention is confined to a small set of layers. Further, knowledge distillation from a large teacher model into a compact student model, combined with structured pruning, can yield a model with less than ten percent of the original parameters while retaining over 95 percent of the sentiment classification accuracy, as demonstrated in distillation studies for multimodal transformers [17].

Operational sustainability also involves optimizing the serving infrastructure. Autoscaling policies that spin down GPU-backed instances during periods of low demand, coupled with the use of hardware accelerators designed for efficient matrix multiplication such as TPUs or custom ASICs, can reduce the energy per query. From a lifecycle perspective, selecting data center regions powered by renewable energy and participating in carbon offset programs are governance actions that responsible organizations can adopt. Ultimately, the system must be instrumented with real-time energy monitoring dashboards that provide visibility into the carbon emissions associated with model serving, enabling operators to make informed trade-offs between service availability, latency, and environmental impact.

5. Robustness, Fairness, and Governance

5.1 Robustness to Noisy, Partial, and Adversarial Multimodal Inputs

Real-world deployment of multimodal sentiment models exposes them to a variety of corruptions that are rarely present in curated benchmark datasets. Images may suffer from motion blur, poor lighting, or occlusions, while text may contain typos, idiomatic expressions, or slang. A core system requirement is graceful degradation: the model should rely more heavily on the cleaner modality when the other is degraded, rather than producing erratic or overconfident predictions. Our dynamic gating mechanism inherently provides a degree of robustness, as the model learns to assign lower weight to corrupted modalities during training with synthetic data augmentation. We systematically inject multimodal noise during training by applying randomized image transformations—such as Gaussian blur, pixelation, and cutout—and text corruptions—such as character-level perturbations and word dropping—and enforce consistency regularization between the corrupted and clean versions of the same sample. This trains the model to maintain stable predictions under distribution shift.

Adversarial robustness is a distinct concern, especially in high-stakes applications where malicious actors might attempt to manipulate emotion recognition systems. An adversary could alter a few pixels in an image or add imperceptible text perturbations to change the predicted sentiment, potentially leading to harms such as unjustified customer prioritization or misleading mental health assessments. Defensive strategies include adversarial training with projected gradient descent attacks during the finetuning phase, gradient masking, and input transformation ensembles that randomize preprocessing to thwart gradient-based attacks. However, these defenses must be balanced against the increased computational cost and potential loss in clean accuracy. At the system level, anomaly detection modules that flag inputs with suspiciously high reconstruction error under a variational autoencoder trained on clean multimodal data can serve as an additional safeguard, routing potentially adversarial inputs for human review.

5.2 Fairness Across Demographics and Cultural Contexts

Facial expression analysis has consistently been shown to exhibit biases related to race, gender, and age, largely because training datasets overrepresent certain demographic groups and specific cultural display rules [4]. When integrated into a multimodal sentiment framework, visual biases can be amplified or mitigated depending on how the cross-modal attention weighs the visual signal relative to text. If the model learns to unduly trust visual cues for certain populations while relying more on text for others, it may produce systematically different sentiment inferences across groups, leading to fairness harms in applications such as automated interview analysis or content moderation. To address this, we incorporate fairness-aware training procedures, including reweighting losses based on demographic annotations where available, and adversarial debiasing where a separate fairness critic attempts to predict protected attributes from the fused representation, while the main model is trained to minimize sentiment loss while making the critic's task difficult.

Moreover, sentiment expression is culturally dependent: a thumbs-up gesture is positive in many Western cultures but offensive in others, and smiling norms differ across societies. A cross-modal system that has learned associations primarily from English-speaking, Western internet data will fail to generalize to global user bases. We advocate for geographically distributed data collection and annotation that captures local cultural norms, alongside continual learning pipelines that allow models to be updated as new cultural data becomes available. At the architectural level, culture-specific adapters—small bottleneck modules inserted into the fusion network—can be trained with limited data from each cultural context, while the shared backbone remains frozen, enabling efficient personalization without requiring full retraining.

5.3 Privacy and Data Governance in Emotion AI

Emotion AI systems operate on highly sensitive personal data. Facial expressions and textual self-reports, especially when linked, can reveal mental health states, political attitudes, and intimate preferences, raising profound privacy concerns. The traditional cloud-based inference paradigm, where user data is transmitted to remote servers, is increasingly at odds with privacy regulations such as the GDPR and the evolving landscape of biometric information privacy laws. Our framework is designed from the outset to support privacy-preserving deployment through on-device processing and federated learning. By keeping raw images and text on the user's device and only transmitting anonymized model updates, we mitigate the risk of data breaches and unauthorized surveillance. Furthermore, techniques such as differential privacy in federated training, where Gaussian noise is added to gradients before aggregation, provide formal privacy guarantees that can be communicated to regulators and users.

Data governance extends beyond technical measures to encompass data provenance, consent, and data subject rights. Any multimodal sentiment dataset used for training must be curated with explicit informed consent for emotion inference, with mechanisms for users to opt out and request deletion of their data. The system must also be auditable: logs of model decisions and the evidence on which they were based (such as attention heatmaps) should be preserved in a tamper-proof manner to support ex-post fairness and accuracy audits. Data minimization is a critical principle; the system should collect and store only the minimal information necessary for the sentiment task, ideally processing data in a transient, stateless fashion. When emotion AI is deployed in sensitive environments like schools or mental health clinics,

institutional review board oversight and adherence to ethical guidelines established by professional bodies are necessary layers of governance.

5.4 Policy Frameworks and Institutional Mechanisms

The governance of multimodal emotion understanding systems cannot be left to technical design alone; it requires a robust policy ecosystem that spans organizational guidelines, industry standards, and regulatory mandates. Organizations that deploy such systems should establish internal AI ethics boards that include diverse stakeholders—engineers, legal experts, psychologists, and civil society representatives—to review use cases and monitor ongoing impacts. Consequentialist risk assessments, such as algorithmic impact assessments, should be conducted prior to deployment and at regular intervals thereafter, evaluating the potential for discriminatory outcomes, psychological harm, and chilling effects on free expression.

At the industry level, the development of benchmarks and certification programs for fair and robust emotion AI can incentivize responsible innovation. Standardized testing protocols that measure model performance across demographic subgroups, cultural contexts, and adversarial scenarios would allow purchasers of emotion AI services to make informed comparisons, much like safety ratings for automobiles. Cross-institutional collaborations, such as the creation of open, demographically balanced multimodal sentiment datasets with transparent annotation protocols, can help level the playing field and reduce the risk that proprietary black-box models entrench societal biases.

Finally, legislative frameworks must evolve to specifically address the unique risks of emotion AI. While general AI regulations such as the EU AI Act provide a foundation, emotion recognition systems warrant sector-specific rules, particularly in employment, education, and healthcare. Provisions regarding the prohibition of real-time emotion monitoring without explicit consent, the right to human review of automated affective decisions, and the requirement to disclose when one is interacting with emotion-aware AI should be codified. International coordination is essential, as emotion AI systems are often deployed across borders, and divergent regulatory standards can create both protection gaps and compliance burdens. The research community, together with policymakers, must engage in ongoing dialogue to ensure that scientific progress in multimodal emotion understanding aligns with societal values and human rights.

6. Conclusion

This paper has presented a comprehensive systems-oriented analysis of a cross-modal attention framework for multimodal emotion understanding built upon vision-language models. We examined the architectural trade-offs between dual-encoder and single-stream designs, proposing a hybrid approach that combines modular pretrained encoders with a lightweight, dynamically gated cross-attention fusion network tailored for sentiment analysis. Our discussion moved beyond accuracy to encompass the full lifecycle of these systems: distributed training infrastructure, inference optimization for cloud and edge environments, energy sustainability, and the critical dimensions of robustness, fairness, and privacy. We argued that technical choices—such as modality gating, attention head specialization, and federated on-device inference—are inseparable from socio-technical outcomes, and we outlined a policy and governance framework to guide responsible deployment.

The cross-modal attention approach offers a path toward emotionally intelligent systems that can interpret the rich tapestry of human expression, yet substantial challenges remain. Future research should develop more efficient pretraining strategies that reduce the environmental

footprint, more comprehensive fairness auditing tools that account for intersectional identities, and more robust multimodal foundation models that generalize across languages and cultural contexts. Equally important is the creation of institutional mechanisms that ensure these powerful capabilities are harnessed to augment human well-being rather than undermine it. By integrating system architecture, infrastructure, and governance into a unified research agenda, the multimodal sentiment analysis community can build technologies that are not only performant but also trustworthy, sustainable, and aligned with the public good.

References

1. Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. Image and Vision Computing, 65, 3–14.

2. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748–8763). PMLR.

3. Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 6558–6569).

4. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77–91). PMLR.

5. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18(1), 32–80.

6. Devlin, J., Chang, M.-W., Lee, K., & 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 (pp. 4171–4186).

7. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).

8. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1103–1114).

9. Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L.-P. (2018). Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

10. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in Neural Information Processing Systems (pp. 13–23).

11. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2020). VL-BERT: Pre-training of generic visual-linguistic representations. In International Conference on Learning Representations.

12. Li, Q. (2026). Dynamic Adaptive Attention and Supervised Contrastive Learning: A Novel Hybrid Framework for Text Sentiment Classification. arXiv preprint arXiv:2604.10459.

13. Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). VisualBERT: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

14. Rahman, W., Hasan, M. K., Lee, S., Zadeh, A., Mao, C., Morency, L.-P., & Hoque, E. (2020). Integrating multimodal information in large pretrained transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2359–2369).

15. Kim, W., Son, B., & Kim, I. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. In International Conference on Machine Learning (pp. 5583–5594). PMLR.

16. Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (2020). ZeRO: Memory optimizations toward training trillion parameter models. In SC20: International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1–16). IEEE.

17. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

18. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (pp. 1273–1282). PMLR.

19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684–10695).

20. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning (pp. 1597–1607). PMLR.

21. Liang, P. P., Zadeh, A., & Morency, L.-P. (2018). Multimodal local-global ranking fusion for emotion recognition. In Proceedings of the 2018 ACM International Conference on Multimodal Interaction (pp. 472–476). ACM.