International Journal of Innovative Computer Science and IT Research

Volume 01, Issue 03 – May-Jun 2025 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

Multimodal Artificial Intelligence for Integrated Text, Image, and Speech Analysis

Dr. Christopher Pal

Professor, University de Montréal / Mila, Canada

Abstract

The rapid growth of digital information has produced increasingly heterogeneous data environments in which text, images, and speech collectively describe events, entities, interactions, and user experiences. Conventional artificial intelligence systems typically process individual modalities independently, limiting their ability to exploit complementary information distributed across different forms of data. Multimodal Artificial Intelligence (AI) addresses this limitation by integrating heterogeneous modalities into unified representations that can support more comprehensive analysis, reasoning, prediction, and decision-making.

This paper investigates multimodal AI for integrated text, image, and speech analysis, with particular emphasis on representation learning, cross-modal alignment, feature fusion, attention mechanisms, transformer architectures, multimodal large language models, and evaluation strategies. The study develops a conceptual framework in which textual, visual, and speech inputs are independently encoded and subsequently aligned through shared representation and multimodal fusion mechanisms. The resulting representation is used for downstream tasks such as classification, sentiment analysis, visual question answering, speech understanding, event interpretation, and multimodal reasoning. Existing literature indicates that multimodal learning can exploit complementary information that may not be available through individual modalities alone, while recent research has increasingly focused on data-level, feature-level, and output-level fusion as well as contrastive, attention-based, generative, and large-language-model-based approaches.

The analysis compares representative multimodal approaches according to representation capability, alignment requirements, computational complexity, robustness, and application suitability. Major challenges include heterogeneous data structures, temporal misalignment, noisy speech, missing modalities, modality imbalance, computational requirements, hallucination, privacy, and interpretability. The paper concludes that multimodal AI can provide a more contextually complete understanding of complex information when modality-specific representations are effectively aligned and fused. Future research should emphasize efficient multimodal architectures, robust learning under missing or noisy modalities, speech-aware multimodal language models, privacy-preserving training, explainable fusion, and resource-efficient deployment.

Keywords: Multimodal Artificial Intelligence, Multimodal Machine Learning, Text Analysis, Image Analysis, Speech Analysis, Multimodal Fusion, Cross-Modal Learning, Transformer, Artificial Intelligence, Multimodal Large Language Models.

International Journal of Innovative Computer Science and IT Research

Volume 01, Issue 03 – May-Jun 2025 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

1. Introduction

Artificial Intelligence has traditionally developed through specialized models designed to process individual forms of information. Natural language processing systems analyze textual content, computer vision models interpret images, and speech-processing systems recognize and analyze spoken language. Although these approaches have achieved considerable progress, real-world information is rarely restricted to a single modality. A social media post may contain written text, photographs, and recorded speech; a customer interaction may involve spoken language, visual information, and textual records; and an intelligent assistant may need to understand written instructions, images, and voice commands simultaneously. Multimodal AI seeks to address this complexity by enabling computational systems to process and integrate multiple forms of information.

Multimodal machine learning combines heterogeneous data sources to improve representation, prediction, and reasoning. A recent systematic review describes major multimodal challenges in terms of representation, translation, alignment, fusion, and co-learning, demonstrating that multimodal intelligence requires substantially more than simply combining independent models. The central objective is to identify relationships between modalities and develop representations that preserve modality-specific information while also capturing shared semantic characteristics. Text provides linguistic and contextual information, images provide spatial and visual information, and speech contributes linguistic, acoustic, prosodic, and temporal characteristics.

The development of transformer-based architectures and large pretrained models has accelerated multimodal AI research. Multimodal large language models increasingly connect language models with visual and auditory encoders, allowing systems to reason over heterogeneous inputs. Recent surveys identify multimodal alignment, instruction tuning, cross-modal reasoning, and hallucination reduction as important areas of research. Such architectures enable applications ranging from image-based question answering and speech understanding to multimodal content generation and intelligent interaction.

The integration of text, image, and speech introduces important technical considerations. Each modality has different structural characteristics. Text consists primarily of discrete linguistic tokens, images are represented through spatially organized visual information, and speech is a continuous temporal signal containing linguistic and acoustic information. Consequently, direct concatenation of raw inputs is generally ineffective. Multimodal systems require modality-specific encoders followed by alignment and fusion mechanisms capable of transforming heterogeneous inputs into compatible representations. Recent research categorizes fusion according to data-level, feature-level, and output-level strategies, while more advanced methods employ contrastive learning, cross-attention, generative learning, and multimodal language models.

Another important development is the increasing use of pretrained foundation models. Vision-language models such as CLIP demonstrated that image and text representations can be aligned within a shared semantic space through large-scale contrastive learning. Subsequent multimodal systems extended these principles to conversational reasoning and increasingly diverse modalities. Recent surveys describe the transition from conventional multimodal models toward multimodal large language models capable of processing visual, textual, and increasingly audio-based inputs.

The purpose of this study is to examine the role of multimodal AI in integrated text, image, and speech analysis. The study reviews major architectural approaches, analyzes multimodal representation and fusion strategies, identifies practical applications, and discusses challenges associated with building

reliable multimodal systems. A conceptual framework is also presented for integrating text, image, and speech information through modality-specific encoding, cross-modal alignment, multimodal fusion, and task-specific prediction. The analysis aims to establish a technical foundation for researchers developing intelligent systems capable of understanding complex information distributed across multiple modalities.

2. Objectives

2.1 Primary Objective

The primary objective of this study is to examine how multimodal artificial intelligence can integrate text, image, and speech information to improve intelligent analysis and decision-making. The study investigates representation learning, cross-modal alignment, fusion strategies, transformer-based architectures, and multimodal language models while identifying the conditions required to achieve accurate, robust, scalable, and context-aware multimodal analysis.

2.2 Analysis of Multimodal Representation

The first objective is to examine how text, image, and speech information can be represented using modality-specific and shared feature spaces. The study focuses on language encoders, vision encoders, speech encoders, embeddings, temporal representations, and cross-modal alignment mechanisms. It evaluates how these representations can preserve unique modality characteristics while enabling computational models to identify semantic relationships across different information sources.

2.3 Evaluation of Fusion and Alignment Strategies

The second objective is to analyze early, intermediate, late, hybrid, attention-based, contrastive, and transformer-based fusion approaches. The study examines how different strategies combine heterogeneous representations and how cross-modal attention can identify relationships between textual tokens, image regions, and speech segments. The objective is to identify suitable strategies for integrated analysis under different computational and application requirements.

2.4 Identification of Technical Challenges

The third objective is to identify major challenges affecting multimodal AI systems, including data heterogeneity, modality misalignment, missing information, noisy inputs, computational requirements, hallucination, interpretability, privacy, and model robustness. The study further identifies future research opportunities for efficient, explainable, privacy-preserving, and resource-aware multimodal architectures.

3. Review of Literature

3.1 Multimodal Machine Learning

Multimodal machine learning has emerged as an important research field because real-world information frequently exists in multiple complementary forms. Traditional unimodal models may fail when critical information is distributed across modalities. Multimodal learning attempts to overcome this limitation by jointly modeling heterogeneous information. A systematic review of multimodal machine learning identifies representation, translation, alignment, fusion, and co-learning as central research challenges.

Multimodal integration can occur at different stages of the processing pipeline. Early fusion combines information before or during representation learning, intermediate fusion integrates modality-specific features within a shared model, and late fusion combines predictions from independently processed modalities.

More recent approaches use cross-attention and transformer architectures to dynamically determine relationships between modalities. Research on multimodal fusion emphasizes that the appropriate fusion strategy depends on data quality, temporal relationships, computational constraints, and the target task.

3.2 Text-Image Integration

Text-image integration has become one of the most mature areas of multimodal AI. Text supplies semantic descriptions while images provide visual evidence. Models such as CLIP demonstrated that large-scale contrastive learning can align visual and linguistic representations within a shared embedding space. Subsequent vision-language models expanded these capabilities toward image captioning, visual question answering, image retrieval, visual grounding, and multimodal reasoning.

Recent surveys of multimodal large language models report substantial progress in connecting visual encoders with language-model backbones. Models such as Flamingo and related architectures use cross-attention mechanisms to integrate visual representations with language processing. This development has established an important foundation for broader multimodal systems that incorporate additional modalities such as speech and audio.

3.3 Speech and Audio Integration

Speech provides information that is complementary to written language and visual content. Automatic speech recognition can convert spoken language into textual representations, while acoustic analysis can capture characteristics such as tone, intensity, rhythm, pauses, and speaker-related information. Consequently, speech-aware multimodal systems can exploit both linguistic and paralinguistic information.

The development of audio-visual and audio-language models has expanded multimodal AI beyond conventional text-image combinations. Recent research specifically examining audio-visual large language models highlights the importance of temporally dynamic audio-visual information and cross-modal alignment. Integrating speech with text and images can therefore improve applications involving meetings, educational content, human-computer interaction, media analysis, accessibility, and intelligent assistants.

3.4 Multimodal Alignment and Fusion

Alignment is a central component of multimodal AI because corresponding information may appear at different temporal or semantic positions. For example, a spoken sentence may correspond to a particular image region or visual event. Effective alignment mechanisms must determine which elements across modalities represent related concepts.

Recent multimodal research categorizes alignment and fusion into structural approaches involving data-level, feature-level, and output-level integration, as well as methodological approaches involving contrastive, attention-based, generative, and large-language-model-based techniques. The

increasing complexity of multimodal systems has therefore shifted research toward adaptive fusion mechanisms capable of dynamically weighting modalities according to their relevance and reliability.

4. Methodology

4.1 Research Design and Conceptual Framework

This study adopts a conceptual and analytical research design for examining integrated multimodal AI. The methodology is based on systematic examination of established multimodal learning principles, contemporary multimodal architectures, fusion strategies, and application requirements. The analytical framework considers text, image, and speech as complementary information channels that are independently processed before being aligned within a shared multimodal representation.

The proposed framework contains four major stages: modality-specific encoding, cross-modal alignment, multimodal fusion, and downstream analysis. Textual information is converted into contextual embeddings, images are represented through visual features, and speech is transformed into acoustic or linguistic representations. The resulting representations are aligned through shared embeddings or cross-attention mechanisms before being combined for classification, retrieval, reasoning, prediction, or generation.

4.2 Modality-Specific Representation

The first methodological stage involves transforming each input modality into a computational representation. Text can be processed using transformer-based language encoders, images through convolutional or vision-transformer architectures, and speech through acoustic encoders or speech-language models. Each encoder is designed to preserve information that is particularly important for its modality.

The representation process can be summarized through the following components:

- **Text encoder:** Converts textual tokens into contextual semantic embeddings.
- **Image encoder:** Extracts spatial and semantic visual representations.
- **Speech encoder:** Captures acoustic, phonetic, linguistic, and temporal information.
- **Normalization layer:** Reduces representation-scale differences.
- **Projection layer:** Maps modality-specific features into a compatible representation space.
- **Temporal alignment:** Connects speech segments with corresponding textual or visual events.

The resulting representations can be expressed conceptually as:

$T = f_t(X_t)$

$I = f_i(X_i)$

$S = f_s(X_s)$

where X_t, X_i, and X_s represent text, image, and speech inputs, while f_t, f_i, and f_s represent modality-specific encoders.

4.3 Cross-Modal Alignment and Fusion

The second stage establishes relationships among modality-specific representations. Alignment can be performed using shared embedding spaces, contrastive objectives, cross-attention, co-attention, or transformer-based interaction mechanisms. The goal is to determine which textual, visual, and speech elements correspond to one another.

Volume 01, Issue 03 – May-Jun 2025 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

Fusion can then be represented as:

$F = \Phi(T, I, S)$

where Φ represents the multimodal fusion function. A simple fusion method may concatenate representations, whereas more advanced systems use attention mechanisms to assign dynamic importance to individual modality features. Recent surveys classify these approaches across data-level, feature-level, and output-level fusion and emphasize the importance of alignment for reliable multimodal learning.

4.4 Training and Evaluation Framework

The proposed analytical framework considers supervised, self-supervised, contrastive, and instruction-based learning. Training may involve modality-specific objectives combined with multimodal objectives. Contrastive learning can encourage corresponding text-image or speech-image representations to become closer in shared embedding spaces, while instruction tuning can teach multimodal systems to respond appropriately to complex user requests.

Evaluation should consider both individual-modality performance and integrated multimodal performance. Important metrics include classification accuracy, precision, recall, F1-score, retrieval accuracy, word error rate for speech recognition, multimodal reasoning accuracy, robustness, latency, and computational cost. Evaluation should also test missing-modality conditions and noisy inputs because real-world multimodal systems rarely receive perfectly synchronized and complete data.

5. Analysis

5.1 Comparative Analysis of Multimodal Architectures

Multimodal systems differ substantially in how they represent and combine information. Traditional fusion approaches often use concatenation or weighted prediction, while contemporary systems increasingly rely on cross-attention and transformer-based architectures. Multimodal large language models further extend this approach by using language models as general reasoning backbones connected to modality-specific encoders.

Table 1: Comparative Analysis of Multimodal AI Approaches

Approach	Main Modalities	Fusion Mechanism	Major Strength	Major Limitation
Early Fusion	Text, Image, Speech	Feature concatenation	Simple integration	Sensitive to representation differences
Late Fusion	Text, Image, Speech	Prediction aggregation	Modularity	Limited cross-modal interaction
Hybrid Fusion	Multiple	Multiple fusion stages	Flexible integration	Higher complexity
Contrastive Learning	Text-Image/Audio	Shared embedding space	Strong cross-modal alignment	Requires large paired datasets

Approach	Main Modalities	Fusion Mechanism	Major Strength	Major Limitation
Cross-Attention	Multiple	Attention-based interaction	Captures detailed relationships	Computationally expensive
Multimodal Transformer	Multiple	Self/cross-attention	Strong contextual modeling	High training requirements
Multimodal LLM	Text, Image, Audio	Encoder-projector-LLM	General-purpose reasoning	Hallucination and high computation

Interpretation: Table 1 demonstrates that multimodal architectures have progressed from relatively simple feature or decision aggregation toward sophisticated attention-driven and language-model-based integration. Early and late fusion remain useful when computational simplicity and modularity are priorities, whereas cross-attention and multimodal transformers provide stronger cross-modal interaction. Multimodal LLMs provide broad reasoning capabilities but introduce substantial computational and reliability challenges. Contemporary research similarly identifies attention-based, generative, contrastive, and LLM-based methods as important directions in multimodal alignment and fusion.

5.2 Integrated Text, Image, and Speech Analysis

Integrated analysis becomes particularly valuable when individual modalities provide incomplete or ambiguous information. Text may explain an event but omit visual details; an image may provide visual evidence without describing context; and speech may communicate linguistic content while simultaneously revealing emotional or acoustic characteristics. Combining these modalities can produce a more comprehensive representation.

For example, an intelligent meeting-analysis system could combine meeting transcripts, presentation slides, and speaker audio. Text could identify topics, images could identify visual content, and speech could provide speaker-specific and acoustic information. A multimodal model could then determine relationships between spoken statements, displayed information, and conversational context.

Table 2: Applications of Integrated Text, Image, and Speech Analysis

Application	Text Information	Image Information	Speech Information	Expected Multimodal Benefit
Intelligent Assistants	User instructions	Camera input	Voice commands	Context-aware interaction
Education	Learning materials	Diagrams and slides	Lectures	Integrated learning analysis
Healthcare	Clinical notes	Medical images	Patient speech	Multisource decision support

Volume 01, Issue 03 – May-Jun 2025 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

Application	Text Information	Image Information	Speech Information	Expected Multimodal Benefit
Social Media	Posts and comments	Photos	Recorded speech	Contextual sentiment analysis
Security	Reports	Surveillance imagery	Audio events	Improved event detection
Customer Service	Chat records	Product images	Calls	Comprehensive customer profiling
Media Analysis	Captions	Video frames	Dialogue	Event and content understanding
Accessibility	Written content	Visual scenes	Spoken descriptions	Enhanced assistive interaction

Interpretation: Table 2 indicates that integrated multimodal AI can provide significant value in situations where information is distributed across multiple communication channels. The combination of modalities can improve contextual understanding, reduce dependence on a single potentially noisy information source, and support more natural human-computer interaction. However, the effectiveness of such systems depends on the quality and synchronization of the underlying modalities.

5.3 Multimodal Reasoning and Decision-Making

The emergence of multimodal large language models has expanded multimodal analysis from feature extraction toward reasoning. Instead of simply combining predictions, modern systems can interpret relationships among different modalities and generate natural-language responses. Surveys of MLLMs identify visual reasoning, visual grounding, multimodal instruction following, and cross-modal interaction as important capabilities.

For integrated text, image, and speech analysis, multimodal reasoning can operate through a sequence in which each modality contributes evidence to a shared reasoning process. A system could analyze a spoken statement, inspect an associated image, compare the image with textual descriptions, and then determine whether the information is consistent. Such capabilities are particularly relevant to intelligent assistants, multimedia search, educational technologies, accessibility tools, and content analysis systems.

6. Discussion

6.1 Improved Contextual Understanding

The primary advantage of multimodal AI is its ability to integrate complementary information. A unimodal system may produce an incomplete interpretation when relevant information is distributed across different sources. Multimodal analysis can reduce this limitation by combining linguistic, visual, and acoustic evidence.

Volume 01, Issue 03 – May-Jun 2025 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

Research on multimodal learning emphasizes that different modalities can provide complementary information and improve representation and task performance. This suggests that multimodal AI should not simply be viewed as a larger version of unimodal AI but as a framework for learning relationships among heterogeneous information sources.

6.2 Cross-Modal Reasoning

Cross-modal reasoning is an important development in multimodal AI. Traditional fusion systems may combine features without explicitly reasoning about relationships between modalities. Attention-based and transformer-based architectures provide mechanisms for identifying which components of one modality are relevant to another.

For example, a model may associate a specific spoken phrase with a particular object in an image and use accompanying text to establish broader context. This form of reasoning enables more sophisticated applications than independent modality classification. Recent multimodal research increasingly focuses on alignment and cross-modal interaction as fundamental components of reliable multimodal intelligence.

6.3 Computational Efficiency and Scalability

The integration of three modalities increases computational complexity. Each modality generally requires its own encoder, while cross-modal attention and multimodal reasoning introduce additional processing requirements. Large multimodal models may require substantial memory, specialized hardware, and extensive training data.

Efficient multimodal learning therefore represents an important research direction. Parameter-efficient adaptation, knowledge distillation, sparse attention, modality routing, quantization, and smaller specialized encoders may reduce computational requirements. Efficient architectures will be especially important for edge devices and real-time applications where latency and energy consumption are constrained.

7. Challenges

A major challenge in multimodal AI is the heterogeneity of input data. Text, images, and speech differ substantially in structure, dimensionality, temporal characteristics, and noise patterns. A text encoder may produce discrete token representations, whereas speech contains continuously varying acoustic signals and images contain spatially structured information. Establishing meaningful relationships among these representations requires sophisticated alignment mechanisms. Recent multimodal surveys identify representation differences and alignment as central technical challenges.

Another major challenge is missing, incomplete, or noisy modalities. Real-world systems may receive poor-quality speech, partially obscured images, incomplete text, or asynchronous information. A multimodal model that assumes the availability of all modalities may experience significant performance degradation when one source is unavailable.

Research on robust multimodal fusion increasingly emphasizes noisy, incomplete, imbalanced, and dynamic multimodal data as important reliability problems. A third challenge involves computational requirements and reliability. Large multimodal models require substantial training data and computational resources. Furthermore, multimodal systems can produce incorrect cross-modal

Volume 01, Issue 03 – May-Jun 2025 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

associations or hallucinated information. Recent surveys explicitly identify hallucination, computational feasibility, and cross-modal correspondence as continuing challenges for multimodal large language models. Ensuring that multimodal outputs remain grounded in available evidence is therefore essential for practical deployment.

8. Future Research

8.1 Robust Multimodal Learning

Future research should develop multimodal architectures capable of maintaining reliable performance when one or more modalities are noisy, incomplete, or unavailable. Modality dropout, adaptive modality weighting, uncertainty-aware fusion, and missing-modality reconstruction can improve robustness. Rather than assuming equal reliability across modalities, future systems should dynamically determine which information sources provide the strongest evidence for a particular decision. Such approaches could significantly improve deployment in uncontrolled environments.

8.2 Efficient Multimodal Foundation Models

Research should increasingly focus on reducing the computational requirements of multimodal foundation models. Smaller modality-specific encoders, parameter-efficient fine-tuning, quantization, sparse attention, and knowledge distillation could enable multimodal models to operate on resource-constrained systems. Future research should also investigate adaptive computation in which the model activates only the modalities and processing components required for a particular task.

8.3 Speech-Integrated Multimodal Intelligence

Although text-image multimodal systems have received substantial attention, integrating speech as a first-class modality presents additional opportunities. Future systems should jointly model linguistic, acoustic, temporal, and visual information rather than simply converting speech into text. Recent research on audio-visual large language models demonstrates growing interest in architectures capable of processing temporally dynamic audio-visual information. This direction can support more natural conversational agents, accessibility systems, intelligent education platforms, and human-computer interfaces.

9. Conclusion

Multimodal Artificial Intelligence provides an important foundation for developing computing systems capable of understanding complex information distributed across text, images, and speech. Conventional unimodal systems are effective within specialized domains but can fail to capture relationships that exist between different information sources. Multimodal AI addresses this limitation by transforming heterogeneous inputs into modality-specific representations and subsequently aligning and integrating those representations. The study demonstrates that multimodal learning has progressed from simple feature concatenation and prediction-level fusion toward contrastive learning, cross-attention, multimodal transformers, and multimodal large language models. These developments have significantly expanded the capacity of artificial intelligence systems to perform integrated analysis, reasoning, retrieval, classification, and interactive tasks.

The analysis shows that text, image, and speech provide complementary information that can improve contextual understanding when effectively combined. Text contributes linguistic and semantic information, images provide visual and spatial evidence, and speech provides linguistic, acoustic, temporal, and paralinguistic information. Applications include intelligent assistants, education, healthcare, social-media analysis, customer service, security, accessibility, and multimedia intelligence. Contemporary research confirms the importance of multimodal alignment and fusion, while recent MLLM research demonstrates the increasing ability of foundation models to integrate multiple forms of information. However, successful multimodal integration cannot be achieved simply by increasing the number of modalities. Effective systems require appropriate representation learning, cross-modal alignment, robust fusion, efficient training, and reliable evaluation.

Future multimodal AI research should therefore move toward systems that are robust, efficient, explainable, and capable of operating under imperfect real-world conditions. Missing modalities, noisy data, computational cost, privacy concerns, and multimodal hallucination remain important barriers to large-scale deployment. Research on adaptive modality selection, uncertainty-aware fusion, efficient multimodal foundation models, privacy-preserving learning, and speech-integrated reasoning can address these limitations. The long-term development of multimodal AI is likely to involve systems that dynamically determine which modalities are relevant, align information across time and semantic space, and reason over heterogeneous evidence within a unified computational framework. Such systems can provide more comprehensive and context-sensitive artificial intelligence capabilities while establishing a stronger technological foundation for next-generation intelligent information systems.

References

1. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607
2. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. *Proceedings of the 28th International Conference on Machine Learning*, 689–696.
3. Srivastava, N., & Salakhutdinov, R. (2012). Multimodal learning with deep Boltzmann machines. *Advances in Neural Information Processing Systems, 25*, 2222–2230.
4. Ngiam, J., Khosla, A., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. *Proceedings of the 28th International Conference on Machine Learning*, 689–696.
5. Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine, 34*(6), 96–108. https://doi.org/10.1109/MSP.2017.2738401
6. 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. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558–6569. https://doi.org/10.18653/v1/P19-1656
7. Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., & Gong, B. (2021). VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems, 34*, 24207–24220.

8. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning, 139*, 8748–8763.
9. Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y.-H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. *Proceedings of the 38th International Conference on Machine Learning, 139*, 4904–4916.
10. Li, J., Selvaraju, R. R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems, 34*, 9694–9705.
11. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, J., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., … Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems, 35*, 23716–23736.
12. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI.
13. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186. https://doi.org/10.18653/v1/N19-1423
14. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460.
15. Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291
16. Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., & Wu, Y. (2021). W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *IEEE Spoken Language Technology Workshop*, 244–250.
17. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Proceedings of the 39th International Conference on Machine Learning, 162*, 12888–12900.
18. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., et al. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems, 35*, 23716–23736.
19. Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Cornia, M., & Cucchiara, R. (2024). The revolution of multimodal large language models: A survey. *Findings of the Association for Computational Linguistics: ACL 2024*, 141–157.