European Journal of AI, Computing & Informatics

Vol. 1 No. 4 2025

Review **Open Access**

Cross Modal Data Understanding Based on Visual Language Model

Bukun Ren [1,*]

[1] College of Engineering, University of California Berkeley, Berkeley, 94720, USA
* Correspondence: Bukun Ren, College of Engineering, University of California Berkeley, Berkeley, 94720, USA

Received: 24 October 2025
Revised: 15 November 2025
Accepted: 11 December 2025
Published: 13 December 2025

Abstract: With the widespread adoption of multimodal data in artificial intelligence, visual language models that integrate cross-modal information have emerged as a prominent research hotspot. These models are capable of jointly processing and interpreting both image and text information, enabling a range of complex multimodal tasks such as image captioning, visual question answering, cross-modal retrieval, and content summarization. By effectively bridging visual and linguistic modalities, visual language models facilitate more intelligent and context-aware systems that enhance human-computer interaction and decision-making processes. This article provides a comprehensive introduction to visual language models, covering their definitions, fundamental operations, and core methodologies. Key techniques analyzed include visual-language joint embedding, attention mechanisms, graph convolutional networks, and generative adversarial networks, all of which play critical roles in enabling accurate cross-modal understanding and representation. The paper further examines the practical applications of these models in multiple domains, including product labeling and categorization on e-commerce platforms, intelligent home control systems, social media sentiment analysis, and personalized recommendation systems. Through this research, it is evident that the integration of cross-modal data understanding technologies can substantially improve the operational performance and intelligence of systems in complex, real-world scenarios. The ability to accurately interpret and fuse visual and textual information not only enhances system efficiency but also expands the potential for innovative applications. These findings underscore the promising application prospects of visual language models, highlighting their significance for future developments in AI-driven multimodal understanding and intelligent system design.

Keywords: visual language model; cross modal data understanding; image processing

1. Introduction

Cross-modal data understanding refers to the process of acquiring, integrating, and analyzing information from multiple modalities-such as images, text, audio, and video-to enable computers to perform more intelligent and context-aware decision-making. As the volume and diversity of multimodal data continue to grow rapidly in various forms, including digital images, textual documents, videos, and social media content, efficiently summarizing and comprehensively interpreting the information from these distributed sources has become a critical challenge in the field of artificial intelligence.

Visual Language Models (VLMs) have emerged as a pivotal deep learning framework that bridges and integrates visual and textual information, significantly

advancing the development and practical applications of cross-modal understanding. By projecting image features and textual features into a shared semantic representation space, VLMs enable effective alignment and interaction between different modalities. This capability facilitates a wide range of multimodal tasks, including image annotation, visual question answering, cross-modal retrieval, caption generation, and content recommendation, enhancing both system intelligence and user interaction.

In addition to their functional advantages, VLMs leverage core techniques such as joint embedding of visual and textual features, attention mechanisms, transformer architectures, and graph-based reasoning networks to capture complex relationships across modalities. These approaches not only improve the accuracy and efficiency of multimodal understanding but also expand the applicability of AI systems to real-world scenarios, including e-commerce product labeling, intelligent recommendation systems, smart home control, and social media sentiment analysis.

This article systematically summarizes the fundamental concepts, key technologies, and application examples of visual language models, aiming to provide both theoretical references and practical technical guidance for cross-modal data understanding. By exploring these models and their implementations, the study seeks to promote the sustainable development of multimodal AI technologies and to support the design of more intelligent, efficient, and adaptable systems capable of operating in complex and dynamic environments.

2. Basic Concepts of Visual Language Model

2.1. Definition of Visual Language Model

Visual language model is a deep learning model that can simultaneously study visual images and text information. It is an effective understanding mode for the integration of visual language and text language, and can improve the performance of various tasks. Due to the traditional isolation of graphic linguistics and linguistics, visual language models can integrate different elements to achieve a common goal, such as computers being able to understand both images and human speech, and improve the performance of various tasks. Visual language models can mine semantic representations of images and text to support various tasks, such as image classification and annotation, answering visual questions, cross modal retrieval, etc. To achieve this goal, visual language models typically share embedded spatial regions, transforming images and text into a fused semantic space to compare, blend, and generate information from different domains. In recent years, with the development of deep learning, visual language models have been able to make breakthroughs in image text interaction tasks and have become a multimodal learning tool for advancing AI [1].

2.2. Working Principle of Visual Language Model

The key working process of visual language models consists of two parts: feature extraction and modal fusion. Image features are extracted based on convolutional neural networks (CNN) or visual transformers to extract visual features, while text features are understood and extracted for text semantics through natural language modeling (NLP) models (BERT, GPT) [2]. In addition, for image and text matching, it is necessary to process to ensure the coordination of the text and image, and finally perform modal matching, that is, the method of fusing image and text features to generate multimodal single view expressions, and then perform cross modal alignment to map them to a common space. The role of attention mechanism is crucial, as it can dynamically adjust the fusion weights of the characteristics of each modality, allowing the model to dynamically focus on the modality information most relevant to the target and achieve higher accuracy in multimodal understanding. (See Figure 1).

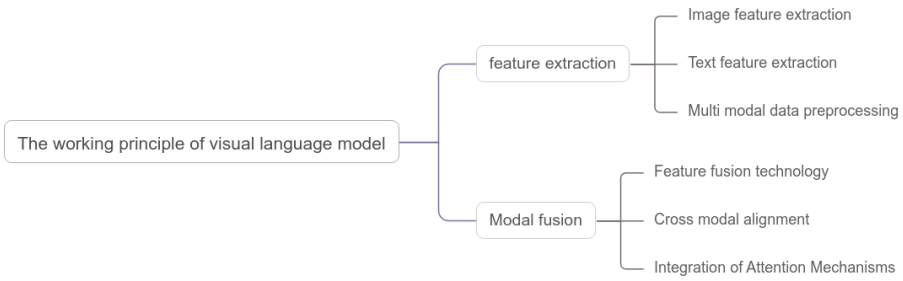

Figure 1. Working principle of visual language model.

3. Cross modal Data Understanding Techniques

3.1. Visual Language Joint Embedding and Shared Feature Learning

Visual language joint embedding refers to mapping image and text data to a unified feature representation space, and based on this, completing accurate comparison, retrieval, and prediction, so that image and text data can be mapped to each other in the same space of meaning, thereby improving the performance or efficiency of cross modal tasks. In practice, deep learning models such as Convolutional Neural Networks (CNN) are commonly used to process visual characteristics of images, and pre trained language models (such as BERT, GPT) are used for text semantic feature parsing to obtain higher dimensional feature representations, thereby extracting more diverse visual and linguistic information. During the process of visual language joint embedding, contrastive learning or multi task joint training methods are usually used to adjust the feature space of images and texts to ensure that they can match in a unified feature space and be closer to semantics. At the same time, similarity measurement methods such as cosine similarity or Euclidean distance are used to measure the similarity between images and text, thereby completing tasks such as image text matching or cross modal retrieval [3].

3.2. Cross Modal Information Weighted Fusion Based on Attention Mechanism

Cross modal information weighted fusion achieves weighted and selective attention to different modal information through attention mechanisms [4]. Due to the differences in feature expression between image and text information, directly concatenating or simply adding the features of the two modalities often fails to capture the deep semantic associations between them. Weighted fusion based on attention mechanism assigns different weights to features of different modalities, enabling the model to automatically focus on key information and improve the performance of cross modal tasks. Attention weights can be calculated in the following way:

$$\alpha_I = \frac{exp(score(f_I, f_T))}{\sum_{i=1}^{N} exp(score(f_I, f_T))} \quad (1)$$

$$\alpha_T = \frac{exp(score(f_T, f_I))}{\sum_{i=1}^{N} exp(score(f_T, f_I))} \quad (2)$$

Among them, $\mathrm{score}(\mathrm{f_I}, \mathrm{f_T})$ and $\mathrm{score}(\mathrm{f_T}, \mathrm{f_I})$ is a function that calculates the similarity between images and text, and common choices include dot product or trainable similarity functions. After weighting the image and text features through attention mechanism, weighted fusion is performed to obtain a joint representation f_{fuse}:

$$f_{fuse} = \alpha_I \cdot f_I + \alpha_T \cdot f_T \quad (3)$$

here, $\mathrm{f_{fuse}}$ is a weighted cross modal joint feature representation that can integrate semantic information of images and texts. The specific methods include using self attention mechanism and cross modality attention mechanism, the latter of which can

dynamically adjust the weights between images and texts, so that the model pays more attention to the most relevant parts between texts and images during inference.

3.3. Feature Propagation and Fusion of Cross Modal Graph Convolutional Networks

Graph Convolutional Network (GCN) belongs to the method of constructing multimodal data through graphical models, and then propagating and fusing features through graph convolution operations to enhance the understanding of multimodal feature information. In the early stage, this method was mainly used for graphical analysis. Therefore, later on, this method began to be used to express the relationship between different modalities (such as images and texts), where each modality (such as images or articles) is regarded as a node in the graph, and the relationship between nodes is the edge in the graph, which is understood as the existence of a line connecting adjacent images and texts, indicating a certain similarity between images and texts. GCN uses graph convolution to facilitate information exchange and fusion between nodes, discovering deep level correlations between patterns [5]. During the graph convolution process, not only can the semantic relationship between images and text be found, but also the influence and interaction between various modal elements are considered. Therefore, cross modal GCN has good flexibility and efficiency in handling complex modal relationships, which is beneficial for cross modal data understanding of image text, video interpretation, and social media data collection. Cross modal GCN utilizes efficient graphic construction, feature fusion, and deep mining analysis to obtain detailed and comprehensive cross modal data statistics from a modal perspective, and further enhances the semantic representation ability of cross modal data. It has great advantages, especially in the application scenarios of deep mining and inference analysis of multiple modal data.

3.4. Generation and Discrimination Mechanism of Cross Modal Generative Adversarial Networks

There is a method of cross modal generative adversarial networks (GANs) that models two or more types of data (such as image data and text data), and then reads and analyzes them through a decoder. The generator attempts to generate another type of data (such as text) under one type (such as images), continuously evolving this generation process. The generator can generate text descriptions based on a given image, and due to the large deviation between the generated data and the real data, the discriminator is used to capture the gap between the generated data and the real data to distinguish whether the generated content is close enough to the real data. The generator and discriminator compete with each other to train the generator to generate more high-quality data, while making the discriminator better at distinguishing truth from falsehood. As the generator and discriminator evolve in opposition to each other, the generator generates high-quality multimodal data to meet various tasks, and finally, through deep training of the generator, various problems can be solved (see Table 1).

Table 1. Generation and Discrimination Mechanism of Cross Modal Generative Adversarial Networks.

step	describe	Method/Technology
Preparation of input data	Utilize input data from each modality separately.	Extract image and text representation features through an encoder.
Generator	Generator learning is the learning of data from one modality to another modality.	The means of reaching the target schema by transitioning from schema to text or other schemas.
Discriminator	The tester determines the difference between the	Assess the consistency between the output data and the measured data.

	generated data and the actual data.	
Adversarial training	Adversarial optimization of generators and discriminators can generate more realistic samples.	Train adversarial improvement generators and recognizers.
Optimization and convergence	Using BP algorithm to achieve the process of finding the optimal network, that is, adjusting the error of the network.	Using a loss function to adjust the generator and recognizer to ensure the quality of the generated data.
Generate data output	High precision data can be generated by training the generator.	Establish high-quality cross modal data (such as illustrations, text generation, etc.).

The table intuitively describes the specific steps and methods of constructing a cross modal adversarial network, involving input data preparation, generator and discriminator, adversarial training, optimization process, and the entire process of generating data.

4. Cross Modal Data Application Based on Visual Language Model

4.1. Automatic Annotation of Product Images on E-commerce Platforms

The product image description in e-commerce platforms can greatly improve customer satisfaction and query performance. Through visual language models, cross modal learning mappings can be formed between images and language to achieve automatic description of product images. The specific process is as follows: Firstly, the visual features of the image are extracted through a convolutional neural network (CNN), and the product image is extracted. The visual features of the extracted image are used to describe the product, including its appearance, color, trademark, type, and other information; Then use pre trained language processing models (such as BERT, GPT) to generate relevant image description text. These image descriptors generally contain information about the type, purpose, function, and nature of the product, making it easier for customers to search for the product. A visual language model that can achieve good cross category generalization performance in a training set containing a large number of product images and their descriptive labels. Visual language model automatic description technology can effectively enrich the content of product pages, improve search engine performance, and make it more convenient for customers to find their shopping needs. In addition, this technology can also be applied to issues such as image text matching and personalized recommendations, improving the intelligence of e-commerce websites and enhancing consumer shopping experience.

4.2. Voice and Text Interaction of Smart Home Control System

With the development of smart home system technology, smart home systems can enhance the comfort of user operation in the form of voice and text. Through visual language models, they can respond to user voice commands, text input commands or information, and achieve a higher level of intelligence and humanized management through the view information in household appliances. In short, users can control household appliances such as lights, temperature, players, etc. through voice commands or text information. The system will use natural language processing technology to analyze the user's goals and connect the corresponding household appliances to complete the relevant actions. For example, if the user's command to the system is to "raise the living room temperature" or "turn on the bedroom light", the system will understand the command based on language or text, and make decisions and execute actions through the

product sensor data of household appliances (such as existing temperature or brightness). Due to the visual language model of the system, indoor images or data from product sensors enable the system to understand more environmental conditions around it. Therefore, appropriate changes can be made to make electrical products more intelligent, the system more interactive with users, and respond to instructions faster, making users live more comfortably and conveniently.

4.3. Social Media Sentiment Analysis and Sentiment Monitoring

Sentiment analysis for data processing in social media mainly includes three aspects: first, data collection and pre-processing; The second is emotion analysis and emotion recognition; The third is to monitor and provide real-time feedback on human emotions by drawing conclusions from data. In the data collection and preprocessing stage, it is necessary to use multimodal data sources such as text, images, and videos that are heterogeneous and obtained from different channels for cleaning, filtering, labeling, and other preprocessing. Emotion analysis and emotion recognition use natural language processing technology and computer vision technology to analyze the data published by users, and ultimately summarize the emotional attributes. Afterwards, multimodal modes such as text, images, and videos will be integrated to further improve accuracy. Finally, emotional monitoring and feedback are conducted, and detailed analysis is carried out based on changes in emotional trends. The resulting emotional analysis results are then used to provide recommendations for brand supervision, crisis management, or psychological counseling services. (See Figure 2).

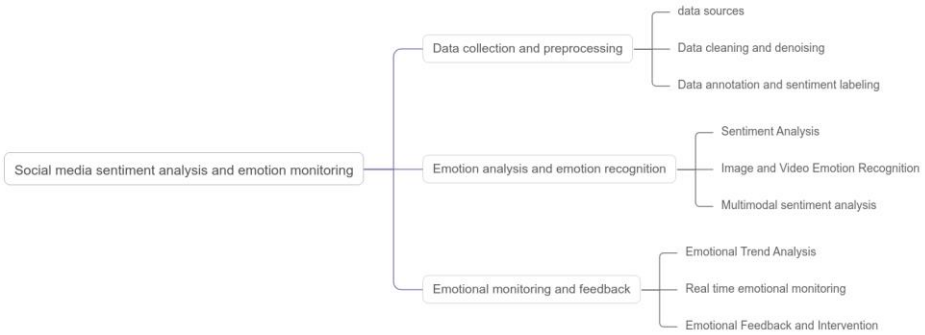

Figure 2. Social media sentiment analysis and sentiment monitoring.

4.4. Cross Modal Data Matching and Intelligent Recommendation System

On the basis of visual language models, intelligent recommendation systems can provide personalized content selection to users through matching cross modal data, thereby providing users with more accurate visual language data information. It mainly covers 5 steps of work content (see Table 2).

Table 2. Cross modal Data Matching and Intelligent Recommendation System.

step	describe	Method/Technology
Data feature extraction	Extract modal features such as images and text.	Image features are extracted using CNN, while text features are extracted using BERT.
Cross modal feature alignment	Embed image and text attributes into the same space.	Use comparative learning or GAN to align features across modalities.
Similarity calculation	Calculate the similarity between different modes.	Use cosine similarity and Euclidean distance to determine similarity.

recommendation generation	Recommend similar users.	Using collaborative filtering or matrix factorization to calculate personalized recommendation results.
Feedback and Optimization	Improve the recommendation system by collecting user feedback.	Adjust and improve the recommendation algorithm based on user activity records.

This table represents a workflow step that includes processing steps for cross modal data matching and intelligent recommendation systems. It includes feature extraction, cross modal alignment, similarity calculation, and the final recommendation generation and optimization process. Taking e-commerce platforms as an example, when users upload photos, the system can extract image or text features, map the image and text to the same semantic domain through cross modal correspondence algorithms, and then recommend corresponding products, movies, articles, and other content to users based on the similarity between the image and text, thereby improving recommendation accuracy, personalization, and user experience. Using a visual language model for recommendation allows the system to integrate users' historical information and make predictions based on content such as images and movies to achieve recommendations. With the development of technology, cross modal data matching and intelligent recommendation will play an increasingly important role in social networks, entertainment information recommendation, shopping websites, and other fields.

5. Conclusion

The research on cross modal data understanding based on visual language models has become a hot topic in the field of artificial intelligence, achieving the integration of information between multiple modalities (such as images, language, acoustics), and contributing to the efficient solution of multimodal problems, image description generation, visual question answering, sentiment analysis and other applications. Detailed introduction of cross modal data processing techniques based on visual and language models, such as visual language joint embedding, attention mechanism, graph convolutional network, and generative adversarial network, will greatly improve the reliability and processing speed of information in multimodality, and increase the application scope of multimodality. With the enhancement of deep learning and computational processing capabilities, cross modal data understanding based on visual language models will be an inevitable trend in various industries, such as smart homes, social networks, personalized recommendations, etc. Some of the problems faced in cross modal data understanding will also accelerate the development of models, such as data scarcity and modal fusion, which will promote more effective and intelligent model development.

References

1. T. Watanabe, A. Baba, T. Fukuda, K. Watanabe, J. Woo, and H. Ojiri, "Role of visual information in multimodal large language model performance: an evaluation using the Japanese nuclear medicine board examination," *Annals of Nuclear Medicine*, vol. 39, no. 2, pp. 217-224, 2025. doi: 10.1007/s12149-024-01992-8
2. Q. Xi, F. Wang, L. Tao, H. Zhang, X. Jiang, and J. Wu, "CM-AVAE: Cross-modal adversarial variational autoencoder for visual-to-tactile data generation," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5214-5221, 2024. doi: 10.1109/lra.2024.3387146
3. W. Wu, S. Wang, Y. Zhang, W. Yin, Y. Zhao, and S. Pang, "MOSGAT: Uniting Specificity-Aware GATs and Cross Modal-Attention to Integrate Multi-Omics Data for Disease Diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 9, pp. 5624-5637, 2024. doi: 10.1109/jbhi.2024.3415641
4. S. S. O'Neil, E. L. Pendl-Robinson, E. A. Carosella, B. D. Sullivan, and A. Sivasankaran, "The importance of community-specific survey data in understanding behavioral and social drivers of COVID-19 vaccination: Lessons learned from urban neighborhoods in four United States cities," *Vaccine*, vol. 42, no. 2, pp. 194-205, 2024. doi: 10.1016/j.vaccine.2023.12.016
5. M. Hashemi-Namin, M. R. Jahed-Motlagh, and A. T. Rahmani, "Recognition of visual scene elements from a story text in Persian natural language," *Natural Language Engineering*, vol. 29, no. 3, pp. 693-719, 2023.