Key Points of Multimodal Communication Design Driven by AI: Exploration of Visual, Auditory, and Textual Information Integration

Tianyang Chen

University of Arizona, Tucson, Arizona, USA
Email: tianyangchen@arizona.edu (T.Y.C.)
Manuscript received December 20, 2024; accepted March 17, 2025; published July 1, 2025.

***Abstract*— In order to solve the problem of low efficiency in multimodal information processing, this article analyzes the design of multimodal communication driven by AI, and explores the implementation points of integrating visual, auditory, and textual information. By introducing the idea of multimodal AI information integration, clarify some datasets that can be used. Subsequently, AI technologies are introduced in three dimensions: visual, auditory, and textual information, such as image recognition technology, video analysis technology, speech recognition technology, audio analysis technology, natural language processing technology, and sentiment analysis technology. And in the end, choose Taobao Intelligent Customer Service as a case study to illustrate the application value of multimodal AI technology. Through analysis, it is concluded that multimodal AI communication design has application value in many fields such as high-quality services, improving user experience, and launching personalized services. However, in the process of application, it is also necessary to pay attention to issues such as cost and technical loopholes, and maximize the advantages of multimodal AI communication design. It is hoped that this can provide some reference for future research by relevant personnel.**

***Keywords*—AI, multimodal, visual, auditory perception, text information integration**

I. Introduction

The application of Artificial Intelligence (AI) in multimodal fields is becoming increasingly profound, providing advanced technological support for multimodal information integration. Many specialized information processing tasks cannot rely solely on a single modality to obtain accurate information, such as sentiment analysis. If only textual information is available, diverse and vague language expressions will inevitably limit sentiment analysis and fail to capture delicate emotions. In addition to the basic information in this article, integrating visual and auditory information can collect more accurate sentiment analysis results. From this, it can be seen that in the process of designing multimodal communication driven by AI, the integration of visual, auditory, and textual information can achieve a more ideal interactive experience, thereby improving the efficiency of multimodal information transmission.

A. AI Driven Multimodal Information Integration Approach

After integrating multimodal information into an AI dataset, applying artificial intelligence algorithms can improve the accuracy of multimodal information integration and prediction. Driven by AI technology, the multimodal communication design process integrates visual, auditory, and textual information, among which the processing of visual textual modal information can be applied to datasets such as Activi tyNet, MSR-VTT, and HowTo100M [1]. Taking the ActiveNet dataset as an example, it can be used for video understanding tasks, including 7 basic activities, with over 100 types of videos in each category, and an average of 1.41 actions per video [2]. For example, the MSR-VTT dataset can be applied in video description generation, including over 20 types of visual content, such as music and games.

For audiovisual fusion tasks, it is recommended to use datasets such as AMI Corpus and Kinetics400. Taking AMI Corpus as an example, this dataset can be used for automatic speech recognition and understanding of spoken language, including recorded conference videos with a duration of up to 100 hours. For example, Kinetics400 is a dataset that can recognize video actions. The dataset includes 650000 video and audio clips and stores 400 categories of human actions, including some basic actions such as shaking hands and hugging. There are 400 video clips of one action, all of which are manually annotated and last for about 10 seconds.

For processing audio text information, it is recommended to use LibriSpeech ASR corpus, IVD, etc. For example, the LibriSpeech ASR dataset has significant application effects in automated speech recognition, storing a large amount of English speech data for 1000 hours, provided by LibriVox audiobooks. The stored English speech data is segmented and reassembled into audio files with a duration of 10 seconds, and all of them are annotated with text. The IVD dataset is provided with resources by Sogou voice assistant, which stores a large number of users' real conversation logs, with a total of about 6890000 entries and an average length of 3–4 seconds. Each voice has a corresponding text description.

For the integration of audio-visual text modalities, choices can be made in datasets such as IEMOCAP and CMU-MOSEI. For example, IEMOCAP is commonly used in multimodal emotion recognition to obtain audio-visual data from emotional interactions of 10 actors, with a duration of 12 hours. This includes capturing facial movements and audiovisual videos, which are stored in the dataset and manually labeled with emotional tags. CMU-MOSEI is an improvement based on CMU-MOSI, which is a key dataset for multimodal sentiment analysis and integrates the opinions of commentators on different topics. After improvement, capturing 5000 videos resulted in 23453 video clips with labels.

DOI: 10.7763/IJET.2025.V17.1318

II. Implementation of Multi Modal Communication Design Driven by AI

A. Image and Video Information Design

Image recognition technology and video analysis technology are two effective techniques for designing and processing image and video information [3]. Applying image recognition technology to accurately identify important information such as objects and faces in images. For example, facial recognition mainly involves four steps: collecting and detecting facial images, preprocessing the recognized facial images, extracting facial image features, and matching information and features for recognition, in order to obtain accurate and valuable key elements. At present, facial recognition technology has been widely applied in the field of smart homes, such as Huawei's 3D facial recognition smart door lock, which uses facial recognition to unlock.

The application of video analysis technology can perform real-time analysis and processing of video data, and extract valuable data from the processed information. For example, video analysis technology has been widely applied in the research of autonomous driving. Cameras are installed in vehicles to capture videos, helping vehicles quickly identify pedestrians, obstacles, etc. on surrounding roads, and assisting autonomous driving in measuring and making intelligent decisions. For example, in the design of intelligent buildings, security monitoring systems will be installed, which will be applied to video analysis technology. After real-time analysis of building monitoring videos, if any abnormalities are found, timely warnings can be issued.

B. Auditory Information Design

For the integration and processing of auditory information, it is recommended to apply speech recognition technology and audio analysis technology to efficiently process auditory information. Speech recognition technology can accurately recognize the voice commands issued by users and quickly convert them into text information [4]. For example, the widely used intelligent customer service system is supported by voice recognition technology, which recognizes user feedback voice data and converts it into text information, so that staff can handle and solve problems in a timely manner. For example, Xiaomi's smart home also applies voice recognition technology. When the user shouts "Xiao Ai classmate", the smart system will automatically respond and use voice control to meet the user's needs [5].

Audio analysis technology can extract key data such as sound characteristics and emotions from a large amount of audio data information. For example, navigation software introduces different character voices as voice navigation, and uses audio analysis technology to collect characteristic information of character voices, including four dimensions: expressiveness, sound quality, complexity, and naturalness. The collected audio information is used to create a speech library, which can convert text into audio.

To understand the characteristics of acoustic systems, audio generation and testing are crucial. There are three main types of test audio: pure frequency, sweep frequency, and slide frequency. The generation formula for each type is:

Pure frequency generation formula:

$$y = \sin(2\pi ft + \varphi) \tag{1}$$

In Eq. (1), f represents frequency, t represents time, and φ represents phase.

Sweep frequency generation formula:

$$f(t) = f_1 + (f_2 - f_1) \times t / T \tag{2}$$

In Eq. (2), f1 represents the starting frequency, *f2* represents the ending frequency, and T refers to the length of time.

Sliding frequency generation formula:

$$F(t) = F_r \cdot (2^{1/3})^{T_l + t} \tag{3}$$

In Eq. (3), $F(t)$ represents sliding frequency, F_r represents reference frequency, i represents time slice index, and T_l is transmission loss.

Taking sliding frequency as an example, sliding frequency changes with time, covering the starting frequency and ending frequency. According to formula (3) for calculation, considering that there may be variable factors in the calculation, the angular position θ should be calculated using integration.

If $T_l = -17$, then $F(0) = 1000 \cdot (2^{1/3})^{-17} \approx 19.7$, $F(29) = 1000 \cdot (2^{1/3})^{12} = 16000$, which can basically cover the sliding frequency range, that is, the sound can be recognized.

C. Text Information Design

The text information in multimodal information integration can be applied with natural language processing technology, sentiment analysis technology, etc. driven by AI to achieve multimodal communication design. By applying natural language processing technology, it is possible to automatically recognize and understand the meaning of human language, summarize writing relationships, and achieve automatic translation and intelligent answers after text recognition. For example, the photo recognition function in Taobao software has applied natural language processing technology. The software identifies key information from images uploaded or taken by users and recommends potentially relevant products to them. For example, many translation software also use this function. Users can directly upload images, and the software recognizes the information in the images for automatic translation [6].

Compared with natural language technology, sentiment analysis technology can identify the emotional tendencies contained in text, such as positive, negative, or neutral emotions. Taking WeChat's voice to text function as an example, users use voice input to convey information, and the other party can directly convert it into text. The software applies sentiment analysis technology to automatically recognize the user's input emotions and automatically equip corresponding emoticons to enhance the tone. Of course, this technology is also widely used in mental health assessment to help users manage their emotions [7].

III. Case Study on Multimodal AI Information Integration of Visual, Auditory, and Textual Modalities

A. Application of Multimodal AI Technology

The integration of multimodal information driven by AI integrates visual, auditory, and textual information, and after analysis, forms a comprehensive understanding of the integrated information. For example, applying deep learning and neural network technologies in AI can integrate information from multiple sources and improve information processing efficiency.

For example, Taobao's intelligent customer service integrates text recognition, speech recognition, and image recognition technologies. If users encounter any questions while shopping, they can directly ask the intelligent customer service to answer their questions as quickly as possible, optimizing their shopping experience. For example, if a user chooses to ask customer service questions through voice, the intelligent system can quickly recognize the voice information, understand the question, and automatically retrieve the user's purchase record to provide targeted suggestions for the user; Users search for products of interest, upload images in Taobao software, and the system uses image recognition technology to identify and push products to users. Through the integration of various information processing technologies, the speed of intelligent customer service response to users has been effectively improved, and answers can be provided based on the understanding of the questions raised by users, ensuring that users receive the best service experience. The application effect comparison of multimodal AI technology on Taobao platform is shown in Table 1, and the technical principle is shown in Fig. 1.

Table 1. Comparison of application effects of multimodal AI technology in Taobao intelligent customer service field

Scene	Data Modality	Effect
Intelligent recommendation of products	Image, text, and speech recognition technology	Users can search for products more directly
Personalized recommendation of products	Retrieve browsing history, purchase history, and social data	Effectively improved information conversion rate and optimized users' shopping experience on the platform
Virtual try on and trial	Image, video, 3D modeling technology	Reduce user return rates and enhance e-commerce shopping information
Optimize inventory	Shelf images, seasonal trends, sales data	Save inventory costs and improve product turnover rate

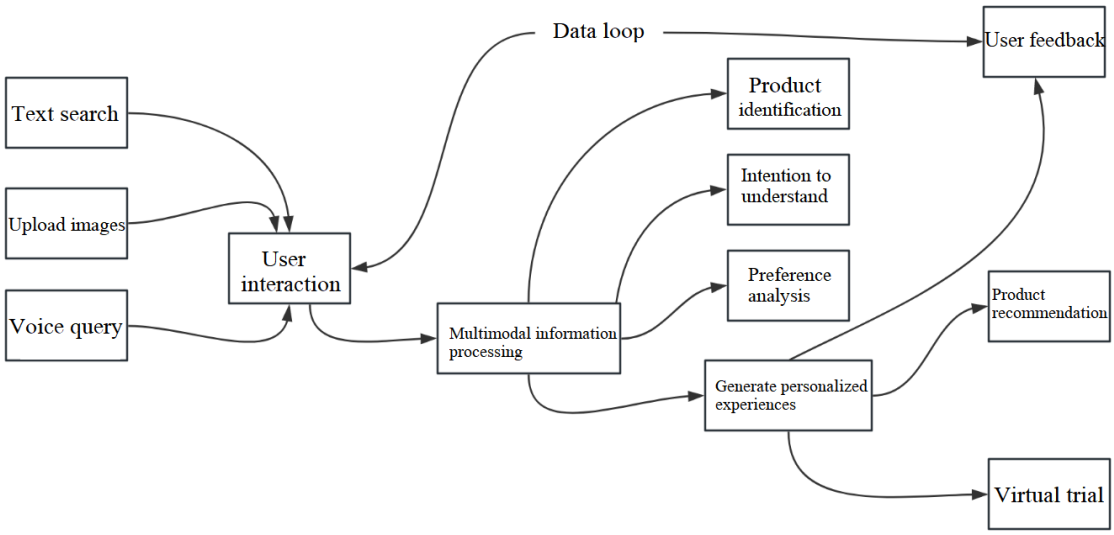

Fig. 1. Application process of multimodal AI technology on Taobao platform.

B. Application Effect

Through the analysis of the application of multimodal AI technology on the Taobao platform, given that intelligent customer service has become the main research and development goal in multiple fields, the competition in the market is relatively fierce. Therefore, the application of AI multimodal communication design in this field, with the integration of diversified AI technologies, can provide platform users with higher quality services. The AI multimodal intelligent customer service system launched on the Taobao e-commerce platform has many functions during use, mainly including: (1) accurate recognition of user questions and instructions, and deep understanding of user text information. With image recognition technology as an auxiliary tool, it helps users solve problems in product retrieval, purchase, and other aspects. (2) Users can directly propose to the intelligent customer service, and the interactive methods include voice interaction, text interaction, image interaction, etc. Users can choose the most convenient method to receive the most comprehensive answer from the system. (3) The system introduces an emotion analysis module, which can quickly capture the user's emotional state and make adjustments to the next response strategy to avoid using a harsh tone that may affect service quality. According to the data displayed on the Taobao platform, once this intelligent customer service

system was launched, users' problems were effectively solved, with a one-time problem solving rate that increased by 25% compared to the original, and user satisfaction with the platform increased by 18%. This shows the application effect of multimodal AI technology on the platform.

C. Inspiration and Suggestions

Combining the case of applying multimodal AI technology on the Taobao platform, it can be clearly seen that AI driven multimodal communication design is beneficial for improving users' shopping experience on e-commerce platforms and enjoying higher quality services. The Taobao intelligent customer service system simultaneously introduces speech recognition technology, natural language processing technology, sentiment analysis technology, image recognition technology, etc., integrating multimodal information and targeting all users of the platform. On the basis of existing services and functions, it adds sentiment analysis services, personalized push services, etc., significantly improving the satisfaction of user groups. AI technology can personalize and push products of interest to users based on purchase and search records, thus completing multimodal interaction and enhancing the fun of users using the Taobao platform. At the same time, through the integration and innovative application of multimodal technology, we can also clearly recognize the important role of multimodal AI technology in optimizing user experience and strengthening data-driven capabilities [8].

Although the application of AI in multimodal communication design brings many conveniences, we also have to pay attention to some security risks that come with it. For example, user privacy protection, increased costs of technological updates, and user adaptability to new features. To address the above issues, the following three suggestions are proposed: Firstly, the platform must strictly comply with current laws and regulations, protect users' personal privacy and security, regularly check system vulnerabilities, and prevent the leakage of users' personal information. Secondly, regularly upgrade the technical architecture and evaluate the compatibility of the system, adopting a phased upgrade model to replace the original old system with modular transformation, in order to avoid cost increases caused by comprehensive upgrades. Alternatively, the system can introduce a large language model and utilize its pre training capabilities to reduce dependence on annotated data. Taobao currently uses unsupervised learning to save on long tail problem processing costs, and its dependence on traditional NLP models has also been reduced by 20%. Thirdly, every time a new feature is launched, the system can push the guidance function to users as soon as possible. Users can choose whether they need to watch the guidance introduction of the new feature, helping them quickly adapt to the platform's new features.

In summary, multimodal AI communication design has application value in many fields such as high-quality services, improving user experience, and launching personalized services. However, in the process of application, it is also necessary to pay attention to issues such as cost and technical loopholes, and maximize the advantages of multimodal AI communication design.

IV. Conclusion

In summary, multimodal AI communication design focuses on the integration of visual, auditory, and textual information. Through the rational application of technologies such as speech recognition, natural language processing, sentiment analysis, and image recognition, it further improves the efficiency of information integration and processing, brings more convenient service experience, and provides new technological research and application ideas for multimodal communication design in the era of artificial intelligence.

Conflict of Interest

The author declares no conflict of interest.

References

[1] J. Zhang, J. Shi, L. L. Qian *et al.*, "Research on complexity optimization of multimodal information interface integrating visual clues," *Packaging Engineering*, vol. 46, no. 04, pp. 36–48+95, 2025.

[2] M. M. Lu, "When creativity meets AI: The potential and limitations of AI empowering multimodal narrative in online literature," *Friends of the Editor*, vol. 01, pp. 35–40, 2025.

[3] H. L. Gui, K. Yue, and L. Duan, "Multi modal knowledge graph link prediction method integrating image and text information," *Computer Applications*, pp. 1–8.

[4] Q. Q. Li, "Research on the transformation mode of scientific and technological achievements based on multimodal information technology," *Yunnan Science and Technology Management*, vol. 37, no. 05, pp. 11–14, 2024.

[5] X. F. Liu and Z. Lu, "Progress of bioinformatics studies for multi-omics and multi-modal data in complex diseases," *Chinese Science Bulletin*, vol. 69, no. 30, pp. 4432–4446, 2024.

[6] X. Wang, J. G. Wang, Y. F. Wang *et al.*, "False information detection method based on multimodal dual collaborative gather transformer network," *Computer Science*, vol. 51, no. 12, pp. 242–249, 2024.

[7] C. Zhan and H. Zhang, "Oral images in multimodal interpreting teaching: research status and prospects," *Foreign Languages*, vol. 39, no. 06, pp. 151–158, 2023.

[8] Y. Xu, J. Mao, and G. Li, "Research on multimodal information analysis of social media for emergency management of sudden events," *Journal of Intelligence*, vol. 40, no. 11, pp. 1150–1163, 2021.