Research on the Application of Artificial Intelligence-Driven Cross-Modal Semantic Communication System in the Tourism Industry

Dandan Lu; Xue Gong*; Liudan Qiu

School of Business Administration, Guangxi University of Finance and Economics, Nanning, Guangxi, China

*Email:2019220055@gxufe.edu.cn

Abstract. In response to the current research gaps in cross-modal semantic communication within the tourism industry, this study proposes the architecture, core concepts, key technologies, practical applications, and challenges of a cross-modal semantic communication system driven by artificial intelligence. The aim of this research is to further advance the theoretical and applied studies in this new direction within the tourism industry. It is anticipated that this work will have a positive impact on the fields of multimedia communication and information processing, particularly in the application scenarios of the tourism industry.

Keywords: Tourism industry; Cross-modal semantic communication; Artificial intelligence; Semantic association.

1 Introduction

Claude Shannon's communication theory divides systems into three levels: syntactic, semantic, and pragmatic [1]. Traditional systems focus on syntactic transmission, while semantic communication emphasizes meaning, reducing data volume and improving efficiency. The pragmatic level addresses the purpose and context of communication. With the rise of multimodal services, cross-modal communication has emerged, leveraging semantic correlations between modalities like audio, video, and tactile signals for collaborative transmission and processing. Cross-modal semantic communication combines these paradigms [2], enhancing resource use and user experience.

However, cross-modal semantic communication faces challenges in core concepts, technology, and practical applications, particularly in tourism, where efficient communication is critical. This paper explores AI-driven cross-modal semantic communication in tourism, offering theoretical and practical innovations in tourism e-commerce and multimedia communication.

Y. Guan et al. (eds.), *Proceedings of the 2024 International Conference on Digital Econimics and Marxist Economics (ICDEME 2024)*, Advances in Economics, Business and Management Research 314,
https://doi.org/10.2991/978-2-38476-344-3_5

The paper first reviews the research background on cross-modal semantic communication, then proposes an AI-based architecture addressing key concepts and technologies. Finally, it explores tourism applications, analyzing challenges and opportunities while providing research references.

2 Overview of Cross-Modal Semantic Communication Research Background

2.1 Overview of Semantic Communication

Currently, semantic communication is divided into unimodal and multimodal. Unimodal focuses on extracting and transmitting semantic information from a single modality, such as text, image, or speech, for tasks like text analysis or machine translation [3-5]. Multimodal communication deals with semantic transmission across two modalities, such as text and image [6].

Semantic communication systems face two key challenges: ambiguity and noise. Ambiguity occurs when meaning is unclear without enough context, like the phrase "burdened" could refer to financial or psychological stress. Noise involves semantic interference during transmission, leading to misinterpretation, such as mistaking "grape" for "cherry." These challenges highlight the need for further research to improve precision and reliability in semantic communication systems.

2.2 Overview of Cross-Modal Communication

To support new multimedia services such as audio, video, and tactile signals, cross-modal communication has emerged[7][8]. It aims to explore the potential correlations between different modalities and build an architecture that can cooperatively transmit and comprehensively process various signals to achieve efficient transmission and processing.

At the transmitting end, different modal signals assist each other in compression to reduce redundant transmission. At the receiving end, features from different modalities are integrated to reconstruct a complete signal, ensuring the quality of multimodal services and enhancing the user experience. This approach leverages the strengths of each modality, allowing for more robust and efficient communication systems capable of handling the complex demands of modern multimedia applications.

2.3 Initial Exploration of Cross-Modal Semantic Communication

To address the challenges of ambiguity and noise in semantic communication, experts and scholars introduced the concept of cross-modal semantic communication. This concept combines the advantages of semantic communication and cross-modal communication, aiming to meet the demands of new multimedia services for low latency, high reliability, high capacity transmission, and immersive experiences.

The cross-modal semantic communication concept leverages the strengths of both semantic and cross-modal communication to enhance the efficiency and effectiveness of data transmission and processing. However, research in this area still has significant gaps. The core concepts of cross-modal semantic communication remain unclear, the system architecture and key technologies are not yet established, and there are no reported practical implementations or application scenarios. These gaps limit the theoretical development and practical application of cross-modal semantic communication.

3 Artificial Intelligence-Driven Cross-Modal Semantic Communication System

3.1 System Architecture

This paper proposes an AI-driven cross-modal semantic communication architecture, based on the framework in literature [9]. The architecture comprises five modules: intra-modal and inter-modal semantic encoders and decoders, and a semantic knowledge base (Figure 1).

This approach separates encoding and decoding into intra-modal and inter-modal levels to enhance semantic processing. Intra-modal encoders extract and compress data within a single modality, while inter-modal encoders capture cross-modal semantic correlations. At the receiving end, intra-modal decoders reconstruct original semantics, and inter-modal decoders integrate cross-modal data. The semantic knowledge base aids in resolving ambiguities and reducing noise.

Semantic features are extracted and compressed at the transmitter via intra-modal and inter-modal encoders, producing residual semantic features and correlations. At the receiver, inter-modal decoders reconstruct these features, followed by intra-modal decoders restoring the signals. The semantic knowledge base ensures precise extraction, transmission, and reconstruction, improving communication efficiency and reducing ambiguity and noise for an immersive experience.

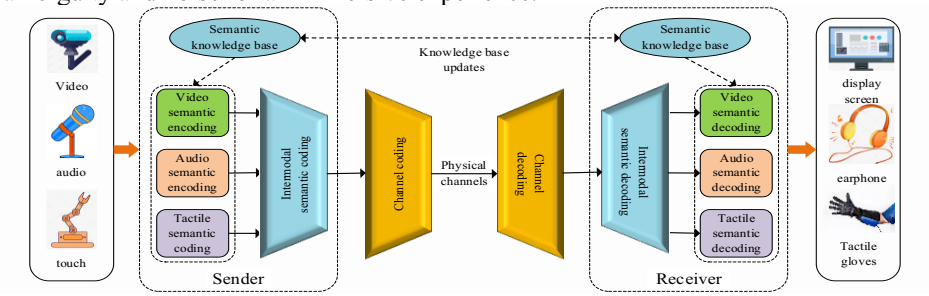

Fig. 1. AI-Driven Cross-Modal Semantic Communication Framework

The innovation of this paper lies in subdividing the cross-modal encoding and decoding process into two sub-processes: intra-modal and inter-modal. This approach allows for more effective compression of transmitted data and the integration of semantic features across different modalities. The ultimate goal is to ensure that the receiving

end accurately understands the semantic information conveyed by the transmitting end and restores the source signal as precisely as possible.

3.2 Core Concept

In semantic communication, the main task is to extract and convey "intra-modal semantics," or the meaning within each signal. Cross-modal communication, on the other hand, focuses on "inter-modal semantics," uncovering correlations between modalities like audio, video, and tactile signals to enhance multimodal information processing. Cross-modal semantic communication integrates both intra-modal and inter-modal semantics for more efficient transmission and reception.

Traditional research has developed separate systems for semantic and cross-modal communication. For example, literature [10] introduces key concepts for unimodal semantic communication, such as semantic channel, noise, and entropy, while literature [11] defines semantic entropy and rate-distortion theory for cross-modal communication. However, a theoretical framework for cross-modal semantic communication is still lacking. This paper builds on these theories and the framework in Figure 1 to redefine the objective function at the transmitter, optimizing efficiency in cross-modal semantic communication.

The objective function can be defined as:

$$F_{encode} = I(S_v ; W_{\Delta v}, W_{vah}) + I(S_a ; W_{\Delta a}, W_{vah}) + I(S_h ; W_{\Delta h}, W_{vah}) + \mu\Psi(I_c ; W_{vah}, W_{\Delta v}, W_{\Delta a}, W_{\Delta h}, \delta)$$

where Sv, Sa, Sh represent the video, audio, and haptic semantics obtained after intra-modal semantic encoding, respectively. WΔv, WΔa , WΔh represent the residual semantics of each modality obtained after inter-modal semantic encoding, respectively.

Wvah represents the inter-modal semantic correlation. I denotes the mutual information among the intra-modal semantics, residual semantics, and inter-modal semantic correlations of the three modalities. Ic represents the channel capacity, δ represents the range of inter-modal semantic correlation representation, Ψ imposes constraints on the channel capacity, inter-modal semantic correlations, and residual semantics, and μ represents the control coefficient. During encoding, a higher value of mutual information indicates a greater degree of semantic correlation, which means that the amount of data to be transmitted can be compressed more significantly.

The Ψ term represents the adjustment of data transmission rates according to channel capacity. When channel resources are abundant, the semantic compression rate is reduced to increase the transmission rate; when resources are limited, the compression rate is increased to lower the data rate. This ensures maximum semantic information transmission without exceeding channel capacity. By optimizing the objective function Fencode, the design and optimization of intra-modal and inter-modal semantic encoding at the transmitter can be effectively guided.

The overall objective function at the receiver end can be defined as:

$$F_{\text{decode}} = H(\hat{W}_{vah}, \hat{W}_{\Delta v}) - H(\hat{W}_{v}) + H(\hat{W}_{vah}, \hat{W}_{\Delta a}) - H(\hat{W}_{a}) + H(\hat{W}_{vah}, \hat{W}_{\Delta h}) - H(\hat{W}_{h}) + \lambda \cdot d(\hat{W}_{v}, \hat{W}_{a}, \hat{W}_{h}; l)$$

where $H(\hat{W}vah, \hat{W}\Delta v)$, $H(\hat{W}vah, \hat{W}\Delta a)$, $H(\hat{W}vah, \hat{W}\Delta h)$ represent the joint semantic entropy of the received semantic correlation and the residual semantics of the three modalities, respectively. $\hat{W}v$, $\hat{W}a$, $\hat{W}h$ are the video, audio, and haptic modality semantic features obtained after inter-modal semantic decoding, respectively. $H(\hat{W}v)$, $H(\hat{W}a)$, $H(\hat{W}h)$ denote the semantic entropy of each modality after decoding. l represents the common semantic label, d denotes the semantic discriminator, and λ is the control coefficient. During inter-modal decoding, the goal is to minimize the difference between the joint semantic entropy of each modality and the intra-modal semantic entropy to achieve intra-modal semantic recovery. The d term is used to determine whether the semantics of the three modalities are consistent, thereby enhancing the quality of semantic recovery. Ultimately, by minimizing the objective function Fdecode, the design and optimization of the intra-modal semantic decoder and the inter-modal semantic decoder at the receiver end can be guided.

3.3 Key Technologies

(1)Modality-Specific Semantic Encoding Technology: This technology creates dedicated input pathways for each modality to extract unique semantic features. Due to the varying characteristics of different signals, diverse encoders are required. For example, CNNs can extract video features, while RNNs capture semantic information from sequential tactile signals [11][12]. Recently, large AI models have shown great success in areas like computer vision and NLP. This study suggests these models as efficient modality-specific encoders, such as the ViT-e model for video tasks [13] and the LLaMA model for language and time-series processing [14]. Their attention mechanisms enhance precise semantic information extraction, as shown in Figure 2.

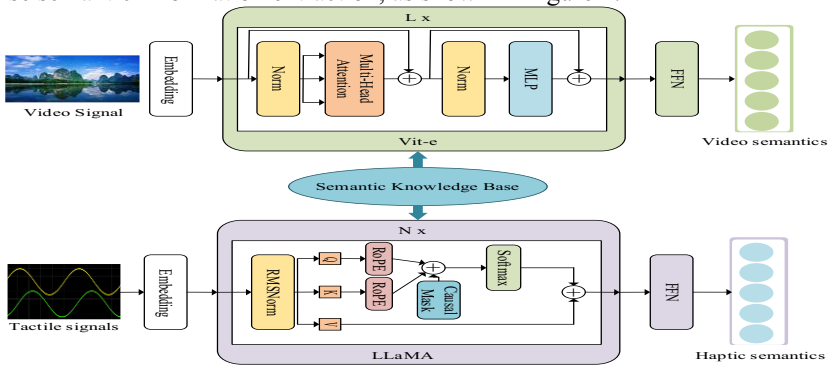

Fig. 2. Intra-Modal Semantic Encoder

(2)Cross-Modal Semantic Encoding: This study aims to use the semantic features of video and tactile signals as inputs to deeply explore and analyze the potential semantic correlations between them. The goal is to obtain the semantic correspondence between video and tactile signals, as well as the residual semantic information within each modality. In existing research, literature [11] revealed potential semantic links by manually annotating semantic relationship matrices, while literature [15] used network models based on attention mechanisms to discover semantic correlations between video and tactile modalities. Based on this analysis, this study posits that the Cross-Attention mechanism in the Transformer structure [15] and the Merged-Attention mechanism in the Transformer structure [16] can effectively extract the semantic correspondence between video and tactile signals and their respective residual semantics, as illustrated in Figure 3.Specifically, the core advantage of these two Transformer structures is their ability to distill the most critical parts from complex semantic information, thereby efficiently constructing potential links between video and tactile modalities. Furthermore, based on the semantic correspondence between video and tactile signals and fully considering the limitations of channel capacity and transmission resources, this study optimizes the objective function in Equation (1) to achieve effective extraction of video residual semantics and tactile residual semantics.

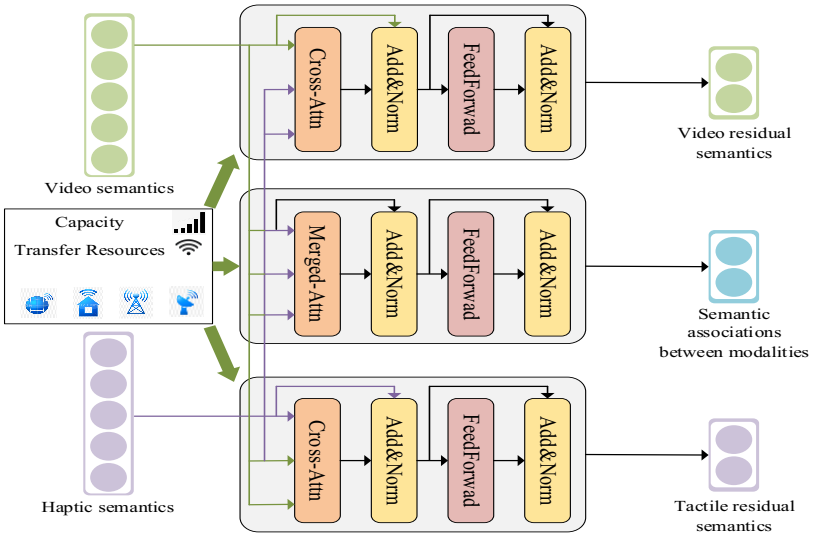

Fig. 3. Inter-Modal Semantic Encoder

(3)Cross-Modal Semantic Decoding: The core task at this stage is to decode the cross-modal semantic correlations between video and tactile signals and their respective residual semantics back into the original video and tactile semantics. Given that semantic noise during transmission can lead to semantic distortion and ambiguity, this study introduces a fusion module based on the Cross-Attention structure [15] during the decoding process. Supported by the Transformer model and combined with a self-supervised learning mechanism, this module effectively integrates video residual seman-

tics, tactile residual semantics, and cross-modal correlated semantics to ensure the complete recovery of video and tactile semantic features, as shown in Figure 4. It is important to note that the implementation of the self-supervised learning mechanism can rely on manual annotations, synchronized timestamps in haptic and video streams, or guidance and cloud-edge collaboration from cloud servers. By optimizing the objective function in Equation (2), the recovery of video and tactile semantic features is achieved.

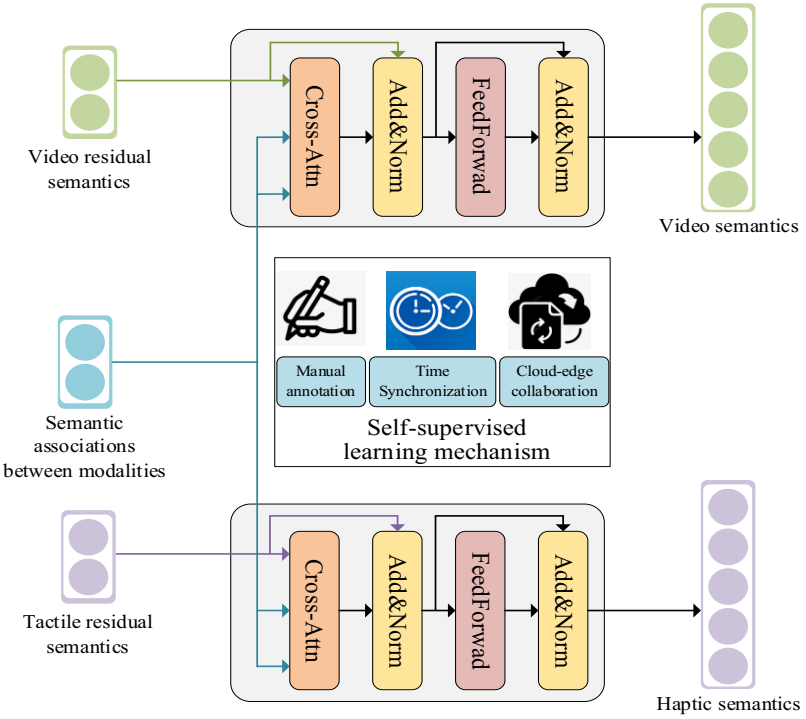

Fig. 4. Inter-Modal Semantic Decoder

(4)Intra-Modal Semantic Decoding: Guided by the background knowledge provided by the semantic library, this module is responsible for recovering video semantic features and tactile semantic features into video signals and tactile signals, respectively. In current research, Generative Adversarial Networks (GANs) are the mainstream method for achieving this process, as shown in Figure 5. Additionally, diffusion models have demonstrated significant success in the field of video generation and recovery. Based on this, the present study proposes using diffusion models to optimize the intra-modal semantic decoding process. Specifically, two intra-modal semantic decoders based on diffusion models will be constructed to handle video feature semantics and tactile feature semantics, respectively. Furthermore, techniques such as knowledge distillation and transfer learning will be employed to integrate the background knowledge from the semantic knowledge base into the diffusion models, aiming to generate high-quality video signals and tactile information.

(5)Semantic Knowledge Base: In a cross-modal semantic communication system, the semantic knowledge base is vital for supporting intra-modal encoding and decoding. It aids in extracting semantic features during encoding and compensating for semantic distortion during decoding. The knowledge base stores vast entities and their relationships. This study proposes using a large generative AI model-based knowledge base, trained on extensive corpora, to extract semantic features and implicitly store them in model parameters. Additionally, integrating this knowledge base into cloud-edge networks allows for efficient updates via localized fine-tuning, minimizing synchronization costs between the transmitter and receiver.

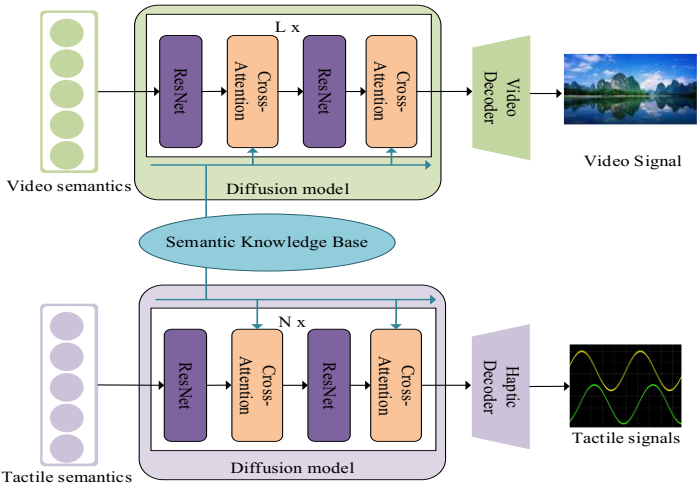

Fig. 5. Intra-Modal Semantic Decoder

3.4 Practical Implementation

Additionally, this paper introduces several existing successful platforms for semantic communication and cross-modal communication, as shown in Table 1.

Table 1. Successful Platforms for Semantic Communication and Cross-Modal Communication

Platform Name	Features	Advantages	Disadvantages
Semantic Communication Prototype for Detection Tasks [3], Such as ecological environment detection platform.	Integrated with cameras, edge servers, USRP radio devices, and antennas, and transmits data based on the UDP protocol.	Applied in tourism ecological environment monitoring, replacing traditional manual detection, enhancing efficiency and objectivity.	Limited to video modality signal detection, with room for accuracy improvement. Consider combining with field surveys and tactile modality.
Task-Oriented Real-Time Mobile Semantic Communication System Prototype [3], Such as smart tourism public service platform.	Utilizes Raspberry Pi, WiFi modules, and display screens to achieve semantic encoding/decoding and feature selection, transmitted via WiFi.	Enhances robustness to semantic ambiguities, selecting task-relevant semantic information transmission, reducing communication costs.	Focused on video modality single-task communication, not involving general tasks and data transmission security.

Multi-User Semantic Communication System for Text and Image Query [6], Such as scenic spot tourism identification system.	Equipped with two single-antenna transmitters and one multi-antenna receiver, converting semantic features into complex values for transmission.	For image transmission, greatly reduces the number of transmission symbols and computational complexity, saving processing time.	Text transmission requires more symbols, slightly increasing text transmission time.
Visual-Tactile Human-Machine Interaction System [8], such as wisdom health service platform.	Combines robotic arms, linear servo-driven tactile sensing gloves, and Kinect cameras.	Uses video signals to compensate for the loss of tactile signals, enhancing cross-modal signal reconstruction reliability.	Symbol-level transmission and cross-modal reconstruction may introduce delays, challenging to meet ultra-low latency requirements.

4 Applications and Challenges

4.1 Application Scenarios

Based on the above analysis, this paper suggests that the application of cross-modal semantic communication systems in the tourism e-commerce scene includes the following aspects:

(1) Remote Tourism Experience: With technological advances, "cloud tourism" has gained popularity among those unable to visit destinations in person. Cross-modal semantic communication enhances remote tourism by using AI for motion capture and object recognition. Tourists can wear VR headsets and haptic suits to experience scenic spots, feeling sensations like the sea breeze or walking on sand, providing an immersive experience.

(2) Remote Intelligent Visual Search: Cross-modal semantic communication can enhance visual search in tourism. Tourists can instantly identify landmarks, artworks, or natural attractions using visual algorithms on e-commerce platforms. This feature offers detailed information, such as historical background and personalized recommendations, making tourism more interactive and educational.

(3) Remote Heritage Protection: Heritage deterioration is a complex challenge. Applying cross-modal semantic communication allows automated deterioration identification through remote data analysis and environmental adjustments. Tactile perception monitors environmental conditions, while visual perception detects early signs of aging, enabling timely interventions and efficient heritage protection.

4.2 Challenges

Cross-modal semantic communication holds great potential for tourism e-commerce but faces several challenges. First, it is not merely a combination of semantic and cross-modal communication. Thus, developing an information entropy theory specifically for cross-modal semantic communication is essential for better processing tourism information and enhancing user experience.

Second, while the proposed architecture offers interpretability, efficiency improvements are needed. Optimizing intra-modal and inter-modal encoding/decoding and the transmission process is an area for further research.

Lastly, security concerns, including attacks during encoding/decoding and potential privacy risks with the semantic knowledge base, require addressing to protect tourists' private information and ensure secure data transmission.

5 Conclusion

This study delves into AI-driven cross-modal semantic communication systems and provides an overview of the relevant background of cross-modal semantic communication. Based on this, the study constructs the architecture of cross-modal semantic communication and clarifies its core concepts, key technologies, and factors to be considered in practical applications. Finally, the study focuses on the application scenarios and challenges of cross-modal semantic communication systems in the field of tourism e-commerce, aiming to provide theoretical support and practical guidance for further development in this area.

References

1. Shannon C E, And Weaver W. The mathematical theory of communication [M].Urbana: University of Illinois Press, 1949:1-125.
2. Li A, Wei X, Wu D, et al. Cross-modal semantic communications [J]. IEEE Wireless Communications, 2022, 29(6): 144-151.
3. Buhalis, D. Technology in tourism-from information communication technologies to eTourism and smart tourism towards ambient intelligence tourism[J]. A perspective article. Tourism Review, 2020,75(1): 267-272.
4. Qingqing Yan, Shu Li, Zongtao He, Xun Zhou, Mengxian Hu, Chengju Liu, Qijun Chen. Decoupling semantic and localization for semantic segmentation via magnitude-aware and phase-sensitive learning[J].Information Fusion,2024,3(107):1-16.
5. LI Ji. The System Construction and Research of the Information Art Design on the Digital Service Platform in Tourism Attractions[D]. Shanghai: Shanghai University,2018,(3):9-12.
6. Gretzel U.Intelligent systems in tourism: a social science perspective[J].Annals of Tourism Research,2011,38(3):757-779.
7. Weng Z Z, Qin Z J, Tao X M, et al. Deep learning enabled semantic communications with speech recognition and synthesis[J]. IEEE Transactions on Wireless Communications, 2022: doi. 10.1109/ TWC.2023. 3240969.
8. Zhang Y C, Zhang P, Wei J B, et al. Semantic communication for intelligent devices: architectures and a paradigm[J]. Scientia Sinica (Informationis), 2022, 52(5): 907-921.
9. Wentao Huang, Maosong Yin, Jie Xia, Xiaoshuan Zhang.A review of cross-scale and cross-modal intelligent sensing and detection technology for food quality: Mechanism analysis, decoupling strategy and integrated applications[J].Trends in Food Science & Technology,2024,(151):68-77.
10. Bao J, Basu P, Dean M K, et al. Towards a theory of semantic communication[C]//2011 IEEE Network Science Workshop. IEEE, 2011: 110-117. DOI:10.1109/NSW.2011.6004632
11. YUAN Z, KANG B, WEI X, et al. Exploring the benefits of cross-modal coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(12):8781-8794. DOI:10.1109/TCSVT.2022.3196586

12. Afonso Castro, Joel Baptista, Filipe Silva, Vitor Santos.Classification of handover interaction primitives in a COBOT–human context with a deep neural network[J].Journal of Manufacturing Systems,2023,(68):289-302.
13. Chen X, Wang X, Changpinyo S, et al. PaLI: A Jointly-scaled multilingual language-image model [EB/OL]. (2022-09-14) [2023-06-05].
14. Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundationlanguage models [EB/OL]. [2023-03-27].
15. Tan H, And Bansal M. Lxmert: Learning cross modality encoder representations from transformers[EB/OL]. (2019-08-20) [2019-10-03].
16. Ludan Ruan, Qin Jin.Survey: Transformer based video-language pre-training[J].AI Open, 2022, (3):1-13.