Multimodal Large Language Models Driven Privacy-Preserving Wireless Semantic Communication in 6G

Daipeng Cao[†], Jun Wu[†*] and Ali Kashif Bashir[‡]
[†]Graduate School of Information, Production and Systems, Waseda University, Japan
[‡]Department of Computing and Mathematics Manchester Metropolitan, University Manchester, U.K.
[*]Corresponding Author (Email: junwu@aoni.waseda.jp)

2024 IEEE International Conference on Communications Workshops (ICC Workshops) | 979-8-3503-0405-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCWORKSHOPS59551.2024.10615340

***Abstract*—In the context of 6G and artificial intelligence, the amount of data on the internet is rapidly increasing. However, traditional syntactic communication encoding techniques are approaching the Shannon limit and have reached a bottleneck. Semantic communication can significantly reduce data traffic and enhance communication efficiency by converting multimodal data, such as images, audio, and video, into semantic information for transmission. However, current semantic communication technology is still in its early stages and faces many challenges. The main challenges in semantic communication currently include data heterogeneity, semantic consistency, and privacy issues. Multimodal large language models (MLLM) could perform semantic understanding of multimodal data and semantic-based multimodal data generation, which offer a unified and consistent multimodal semantic conversion. Therefore, we propose a Privacy-Preserving Semantic Communication scheme based on MLLM (MLLM-PSC). The semantic interpretation ability of MLLM-PSC is cultivated based on the pre-training and fine-tuning of MLLM, without the need for additional knowledge base (KB) alignment. We take textual semantics as a medium for consistently conversion of multimodal data, while safeguarding sensitive semantic information based on few-shot learning. Our simulation results demonstrate the superior performance of MLLM-PSC.**

***Index Terms*—6G, Semantic Communication, Motimodal Large Language Model, Privacy Protection.**

I. Introduction

As 6G, big data and artificial intelligence (AI) continue to evolve, the volume of traffic data on the internet is experiencing an exponential increase. According to a report by the International Telecommunication Union (ITU), the overall mobile data traffic is projected to reach 5 zettabytes (ZB) per month by 2030 [1]. Concurrently, the emergence of technologies such as the metaverse, virtual reality (VR), and digital twins is intensifying the demand for the transmission of various heterogeneous data, posing significant challenges to traditional communication technologies.

According to the communication model proposed by Shannon, communication levels, in ascending order of complexity, are syntactic, semantic, and pragmatic communication. Traditional communication technologies focus on studying source encoding (such as Huffman encoding, algorithmic coding, etc.) and channel encoding (such as LDPC codes, polar codes, etc.) based on syntactic communication. Traditional communication encoding techniques are approaching the Shannon limit, making it challenging to further improve data transmission efficiency on this basis. Therefore, to meet the growing demand for data transmission in 6G, wireless semantic communication is emerging as a new direction to address the issue of exploding data traffic [2].

Traditional syntactic communication focuses on compressing and transmitting information through encoding, striving to avoid loss during transmission and pursuing the consistency of information between the sender and receiver. In contrast, semantic communication seeks to maintain the intrinsic semantic consistency of information during transmission, without requiring the source of information to be identical between sender and receiver. This characteristic of semantic communication allows for the conversion of large amounts of data (such as images, video data, etc.) into condensed semantic information for transmission, significantly reducing the overhead associated with direct data transmission [3].

Mainstream semantic communication systems consist of a semantic encoder, a channel encoder, a channel decoder, a semantic decoder, and a KB. Semantic encoders and decoders typically employ methods such as deep learning models to extract and summarize information, converting it into semantic information [4]. Channel encoders and decoders are responsible for transmitting data transformed by semantic encoders and decoders over the physical channel. The KB ensures alignment during the semantic interpretation.

Current semantic communication systems face several major challenges [5]. First, the issue of data heterogeneity. In real-world semantic communication applications (such as the metaverse, VR, etc.), the data to be transmitted is often heterogeneous multimedia data, including text, images, audio, and video. Traditional semantic communication approaches rely on training different semantic encoder-decoder models for each type of data, which leads to high training costs and coordination problems between different semantic encoders and decoders. Second, there's a challenge of consistency in semantic understanding. Adapting deep learning models to individual user semantics is challenging because deep learning

models require a significant amount of relevant data to perform well. KB requires extensive data input from users to improve semantic understanding, which raises new privacy concerns. Finally, the use of KB means that user privacy is exposed to components outside of the semantic encoders and decoders, posing a risk of privacy leakage during the semantic transfer process or KB updates.

Since 2023, the development of MLLMs has offered potential solutions for semantic communication. MLLMs are pre-trained on massive datasets, which allows them to gain robust semantic understanding capabilities. MLLMs such as GPT-4 can process and generate outputs for multimodal data through a unified interface, demonstrating their ability to handle different types of data.These features enable MLLMs to accurately perform semantic understanding and interpretation tasks. In specific scenarios, they can also be fine-tuned efficiently through prompt fine-tuning [6], avoiding the additional overhead associated with transfer learning via KB.

Based on the analysis presented, we introduce MLLM-PSC, a novel Privacy-Preserving Semantic Communication scheme leveraging the multimodal semantic capabilities of MLLM. Capitalizing on the adeptness of MLLM in multimodal semantic interpretation, MLLM-PSC could accurately process semantic interpretation tasks for multimodal data. Furthermore, we have devised a privacy protection mechanism grounded in user profiles, ensuring the confidentiality of personal semantic content related to privacy during communication. Our main contributions are as follows:

- We propose a multimodal data semantic communication framework based on MLLM, which employs a unified semantic interpretation method. This strategy mitigates the overhead associated with training separate semantic encoders and decoders for different data formats and ensures the accuracy of semantic interpretation across various modalities.
- We design a privacy protection mechanism based on profiles. This mechanism, which involves prompt fine-tuning for semantic interpretation and protection, enhances the security of user privacy during semantic transmission.
- We conduct human evaluations of MLLM-PSC in interpreting the semantics on images data, demonstrating the superior performance of our approach.

The remainder of this paper is organized as follows: First, we present the background of semantic communication and multimodal large language model (MLLM), analyzing the potential of MLLM in addressing challenges in semantic communication. Next, we introduce the architecture and principles of our designed Privacy-Preserving Semantic Communication scheme based on MLLM (MLLM-PSC), including mathematical foundations and formulations. Thereafter, we demonstrate the superiority of MLLM-PSC through simulation experiments. Finally, we conclude this paper.

II. Background

A. Semantic Communication

In recent years, a widely recognized semantic communication scheme is the deep learning-based source-channel joint coding scheme (DJSCC). This scheme has the potential to achieve efficient and reliable communication by simultaneously accomplishing semantic compression and recognizing and protecting key semantic tasks through a source-channel joint encoder constructed based on deep neural networks [7]. Fig 1 illustrates the architecture of semantic communication. The data is encoded and decoded at the semantic level through a channel encoder and decoder, respectively. The encoded semantic data is then transmitted over a conventional physical channel.

DJSCC-based methods require extensive training to learn encoding and decoding strategies that meet the requirements. This training also needs to be supported by a large amount of data. However, in the field of wireless communication, both training time and training data are very scarce resources. To solve this problem, DJSCC schemes that introduce semantic KBs have been proposed and received much attention in recent years. In this scheme, a semantic KB is a structured, memory-enabled knowledge network model that can provide relevant semantic knowledge descriptions for data information, which can speed up the training of the source-channel joint coder and decoder and reduce the dependence of training on data.

In addition, the above solutions currently face the following problems. First, for data heterogeneity, there is currently no unified solution for DJSCC systems. For various multimodal data, traditional deep learning models cannot process them semantically at the same time. As a result, different models need to be trained for different data formats, which leads to additional training costs and the work of using a KB for multimodal semantic alignment. Second, deep learning-based systems are prone to the problem of semantic ambiguity. Especially when the relevant training data is not sufficient, the deep learning model may be prone to ambiguity when semantically transforming the data. Meanwhile, different communicators have different emphasis on semantic concerns, which poses a challenge for semantic KBs to support user-definable semantic content. Finally, privacy and security issues in semantic communication are also a concern. The process of interaction between codecs and KBs and the transfer of semantics may risk the leakage of users' private data, especially for user profile-related semantic content, which requires additional protection.

B. Moltimodal Large Language Model

Multimodal Large Language Model (MLLM) is a recently emerged research hotspot that accomplishes multimodal tasks by using a large language model (LLM) as a brain, and is the latest achievement of LLM that has gained the ability to process multimodal data [8]. MLLM builds on the powerful knowledge capability of LLM and has new capabilities that are not available in traditional deep learning models, such as

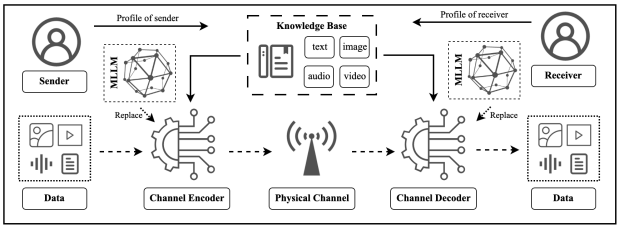

Fig. 1: Architecture for the semantic communication in 6G

summarizing a story based on a picture or video, or generating pictures based on text, etc. The emergence of MLLM (such as GPT-4V [9], Gemini [10], etc.) provides a new solution for unified processing of multimodal data semantics.

Compared to LLMs that focus solely on solving Natural Language Processing (NLP) tasks, MLLMs support a broader range of tasks, including image and audio processing. Models like GPT-4V, as an MLLM, facilitate multimodal input and output through text semantics as an intermediary. MLLMs undergo extensive pre-training across diverse datasets, equipping them with robust semantic understanding capabilities. This allows them to accurately interpret even semantically ambiguous sentences through inference. MLLMs inherit the substantial knowledge capacity of LLMs, ensuring high accuracy in semantic transformation tasks across various scenarios. Additionally, MLLMs also support technologies like prompt fine-tuning, enabling performance enhancement in specific scenarios through tailored adjustments [11].

The utilization of MLLMs in semantic communication tasks provides several advantages. Firstly, MLLMs can effectively replace semantic encoders and decoders in such tasks, as illustrated in Fig 1, and are capable of uniformly processing multimodal data. Semantic communication schemes based on the MLLMs eliminate the need for designing multiple encoders and decoders to accommodate different data formats. The use of MLLMs in semantic communication can simplify architectural complexity and improve accuracy. Additionally, MLLMs reduce the need for intricate KBs and semantic constraint rules, resulting in more efficient performance enhancement through prompt fine-tuning.

III. Design

This section discusses the principles and architecture of MLLM-PSC. It covers the design of using MLLM as a replacement for traditional semantic encoders and decoders in semantic communication, adaptation to semantic contexts based on user-defined profiles and analysis of sensitive information, and implementation of privacy-preserving encryption transmission based on the results of the sensitivity analysis.

As illustrated in Fig 2, within the architecture of MLLM-PSC, we eschew the traditional semantic encoder-decoder for semantic transformation, instead leveraging the MLLM as a proxy for semantic understanding and conversion. The impressive semantic capabilities of the MLLM enable it to effectively accomplish the task of semantic conversion. In this process, natural language text data serves as the medium for semantic transmission. The rationale for employing text data to preserve semantics is twofold: Firstly, the MLLM is built upon the foundation of LLM, which excels in processing natural language text, thus utilizing natural language text data to retain semantics maximizes the potential of MLLM, while obviating the need for additional syntactic constraints or definitions. Secondly, text data, compared to other data formats, occupies less space; therefore, using text data as the medium for semantic transmission minimizes the physical layer transmission costs in semantic communication, thereby enhancing the efficiency of resource utilization inherent in semantic communication.

Algorithm 1 MLLM-PSC for Multimodal Data Transmission

Input: Multimodal Data ($Data$), User Profile ($Profile$)
Output: Recovered Data
procedure TRANSMISSION($Data, Profile$)
  $Prompt \leftarrow$ INITIALIZEPROMPT($Profile$)
  $Encoder \leftarrow$ FEWSHOTLEARNING($Profile$)
  $SemanticContent \leftarrow$ ENCODE($Data, Encoder, Prompt$)
  $PrivacyFilter \leftarrow$ INITIALIZEMLLM($Profile$)
  $PrivateContent \leftarrow$ empty list
  for each $item$ in $SemanticContent$ **do**
    if ISSENSITIVE(item, PrivacyFilter) **then**
      $PrivateContent$.ADD($item$)
      Remove $item$ from $SemanticContent$
    end if
  end for
  $EncryptedData \leftarrow$ ENCRYPT($PrivateContent, K_{pub}$)
  $TransmissionData \leftarrow$ CONCATENATE($EncryptedData, SemanticContent$)
  TRANSMIT(TransmissionData)
  $Decoder \leftarrow$ INITIALIZEDECODER()
  $DecryptedData \leftarrow$ DECRYPT($TransmissionData, K_{priv}$)
  $RecoveredData \leftarrow$ DECODE($DecryptedData, Decoder$)
  return $RecoveredData$
end procedure

A. Workflow

Fig 2 illustrates the workflow of MLLM-PSC in image data transmission. The process begins with the sender transferring the original data to the MLLM Encoder, which can be a locally deployed model. Under the guidance of a profile defined by the sender, the MLLM Encoder performs semantic extraction on the data. This involves leveraging a prompt provided by the sender, which outlines the desired semantic focus, enabling the MLLM Encoder to engage in few-shot learning to meet and adapt to the semantic extraction requirements of the sender.

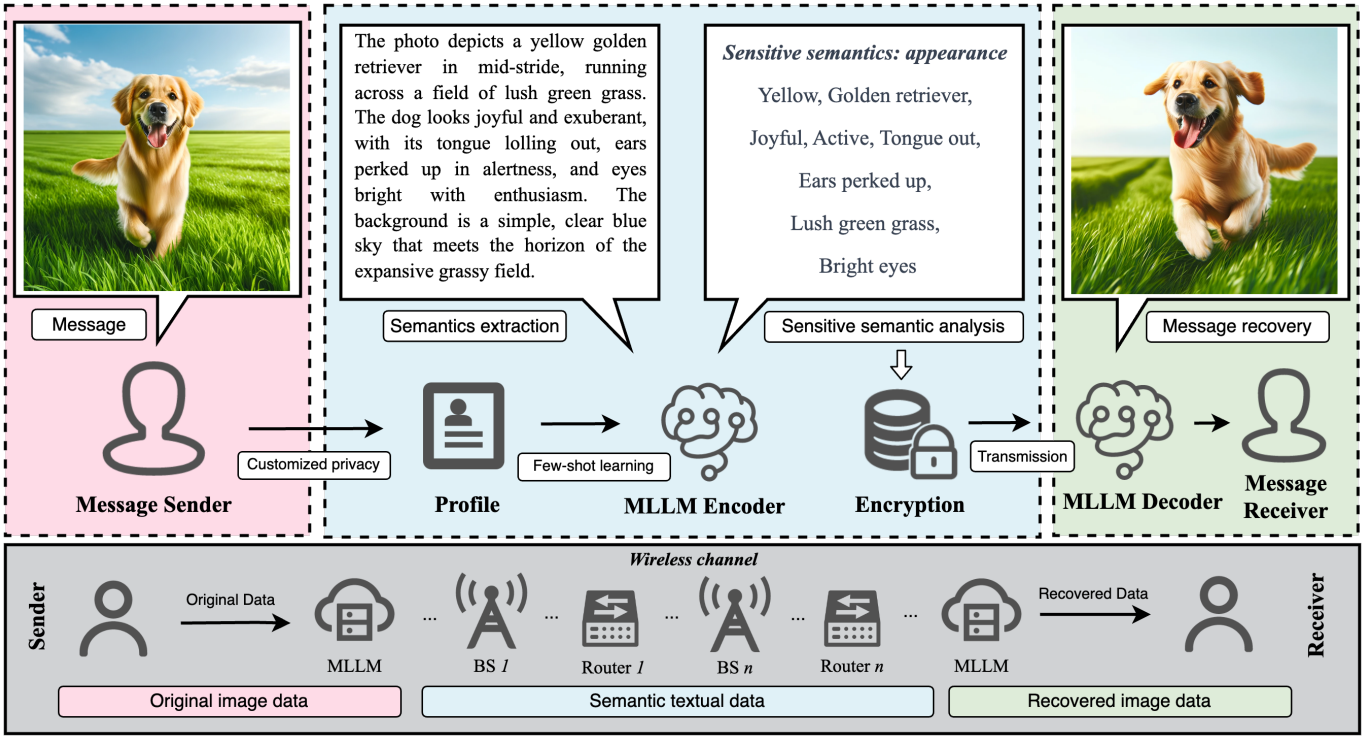

Fig. 2: A demonstration of semantic communication transfer via MLLM-PSC.

As shown in Fig 2, the sender transmits an image depicting a golden dog running on grass.

The image data is initially processed by the MLLM Encoder for semantic extraction, where the MLLM Encoder comprehensively and meticulously extracts the descriptive information contained within the image. Notably, the MLLM Encoder synthesizes a detailed description based on the entirety of the image's content, encompassing aspects such as characters, location, and appearance. Subsequently, based on the sender's privacy preference settings, the MLLM Encoder analyzes and extracts private information from the semantic content, a task also executed by MLLM itself. In the illustration, the sender considers information related to appearance as sensitive; therefore, descriptions pertaining to appearance are extracted by MLLM. For private content identified by the MLLM Encoder, encryption safeguards are applied prior to transmission. MLLM-PSC encrypts the private information and transmits it via the physical layer channel. The MLLM Decoder then reconstructs the original data based on the textual semantic information, thereby achieving semantic-consistent recovery of the source data.

As shown in algorithm 1, The transmission of MLLM-PSC begins by generating a tailored prompt based on the user's profile, which guides the Few-Shot Learning-based MLLM encoder in transforming the data into semantic content. A crucial aspect of MLLM-PSC is its emphasis on privacy: a MLLM filter, initialized according to the user's privacy preferences, identifies and segregates sensitive information within the semantic content. This sensitive data is then encrypted, ensuring its security during transmission. The encrypted private content is concatenated with the non-sensitive semantic content for transmission. At the receiving end, a decoder reconstructs the original data from the decrypted content. This process not only ensures the integrity and confidentiality of sensitive information but also allows for a flexible and adaptive handling of various types of multimodal data.

B. Methodology

MLLM-PSC offers a promising solution to the challenges of data heterogeneity, semantic consistency, and privacy concerns. This section delineates the core mechanisms underpinning the MLLM-PSC.

1) MLLM for Semantic Interpretation: The transformer-based MLLM, which is pivotal in interpreting and transforming multimodal data into consistent textual semantics. The self-attention mechanism of the transformer model processes the input data X through a series of transformations:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{1}$$

Where Q, K, V are the Query, Key, and Value matrices, respectively.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2}$$

d_k denotes the dimensionality of the key vectors. This self-attention facilitates the dynamic focusing on various aspects

of the input data, essential for achieving semantic consistency in the heterogeneous data landscape.

2) Semantic Extraction and Privacy Preservation: In addressing privacy concerns, MLLM-PSC integrates user profile-based semantic processing. The methodology involves two key processes:

1) Semantic information extraction:

$$S_{info} = MLLM_{encode}(I, P; \theta_{semantic}) \quad (3)$$

where I is the input, P is the user profile influencing the semantic extraction function $MLLM_{encode}$, $\theta_{semantic}$ is the fine-tuned parameters of the MLLM.

2) Privacy-sensitive content extraction:

$$P_{content} = MLLM_{encode}(S_{info}, P; \theta_{privacy}) \quad (4)$$

capturing sensitive content through privacy analysis function $F_{privacy}$.

3) Secure Transmission: Ensuring the secure transmission of the processed data, the MLLM-PSC employs encryption for the privacy-sensitive content:

$$E_{data} = Encrypt(P_{content}; K_{pub}) \quad (5)$$

and utilizes the physical protocol for data transmission. Where O represents the output of recovered data using the $MLLM_{decode}$ function:

$$O = MLLM_{decode}(Concat(S_{info}, Decrypt(E_{data}))) \quad (6)$$

This methodology ensures that MLLM-PSC adeptly handles the semantic interpretation and transformation of multimodal data into textual semantics. Meanwhile, it safeguards sensitive semantic information, aligning with the needs of secure and privacy-preserving semantic communication.

IV. Simulation Results

This section presents the evaluation of MLLM-PSC performance through a series of experiments. The investigation primarily focuses on the accuracy of semantic transformation in MLLM-PSC, the impact of signal attenuation during transmission on semantic restoration, and the performance enhancement afforded by selective encryption and decryption of private content as opposed to full encryption.

A. Experiments Setup

In the context of semantic communication for multimodal scenarios, there currently exists no suitable dataset. Therefore, we have created a simulated dataset based on real communication scenarios. To effectively quantify the accuracy of the MLLM in extracting semantics from multimodal data, we have defined twenty standard data sets in an image format. Images are generated as initial messages using our predefined template, formatted as "*[3 appearance descriptions] [character] [action] on/in [location].*" Key information in an image is divided into four parts: "appearance descriptions," "character," "action," and "location." Utilizing this template, we constructed twenty predetermined semantics. Subsequently, we input these defined semantics into DALL·E 3 [12] to generate corresponding images. The resultant twenty sets of data serve as the foundation for evaluating the performance of MLLM-PSC. In these experiments, GPT-4V [9] served as the foundational MLLM underpinning our experiments.

B. Semantic Extraction Accuracy

The initial information was provided to the MLLM Encoder to extract semantic information from images, with the accuracy results displayed in Fig 3. We divided the accuracy assessment into four dimensions: "*Appearance*," "*Character Identification*," "*Behavior*," and "*Environment*," corresponding to the four dimensions in our initial template setting. This division facilitates the quantification of the MLLM Encoder's accuracy in semantic extraction. The semantic information extracted by the MLLM Encoder was evaluated manually to determine if it included the semantics from the predefined template.

The MLLM Encoder demonstrates high accuracy across all four evaluation dimensions. Notably, the accuracy for character and environment recognition reached 100%, indicating the MLLM Encoder's strong capability in discerning semantics related to characters and environments in images. Additionally, for the extraction of appearance and behavior semantics, the MLLM Encoder maintained an accuracy rate above 75%, showcasing its proficiency in extracting semantics from multimodal data.

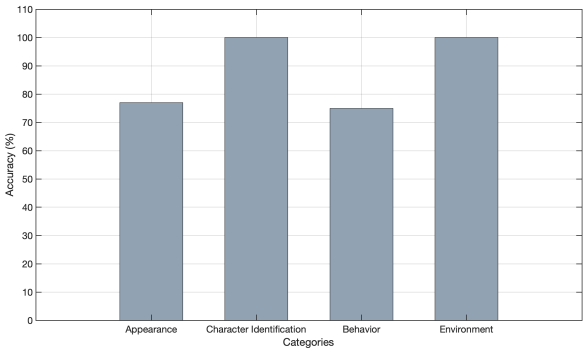

Fig. 3: Accuracy of Encoder for Semantic Extraction

C. Transmission Loss and Data Recovery

As illustrated in Fig 4, we simulated data transmission over AWGN (Additive White Gaussian Noise) and Rayleigh fading channels and calculated the error rates [4]. For the AWGN channel, Gaussian noise was added to the transmitted signal, with the noise level varying according to specified SNR (Signal-to-Noise Ratio) values ranging from 0 dB to 50 dB. In the case of the Rayleigh fading channel, the transmitted signal was multiplied by a Rayleigh distributed random variable, followed by the addition of Gaussian noise, with the noise level also varying based on the SNR values. The error rate was defined as the number of erroneous characters during transmission divided by the total number of characters. We

simulated the error rates over AWGN and Rayleigh fading channels at different SNR levels (ranging from 0 dB to 50 dB, with an increment of 5 dB). The experimental results show that the error rate for transmissions, across various message quantities and SNR levels, does not exceed 30%.

Considering the semantic information error rate range of 0-30% during wireless signal transmission. We processed the semantic information with an error rate set between 0-30%, subsequently providing it to the MLLM Decoder for image restoration. Additionally, we manually conducted semantic consistency checks on these restored images. As depicted in Fig 5, MLLM demonstrates remarkable robustness against semantic information loss, exhibiting high accuracy across four evaluation dimensions.

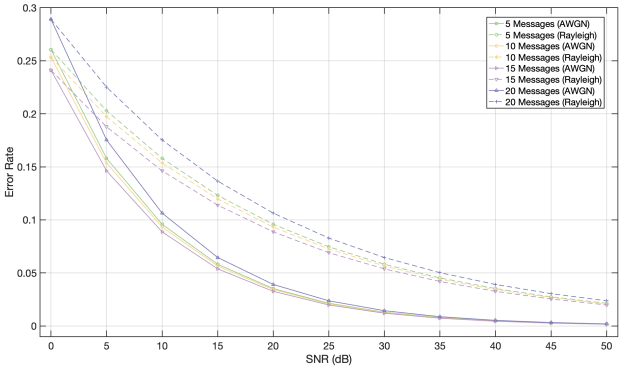

Fig. 4: Error rate under different SNR conditions

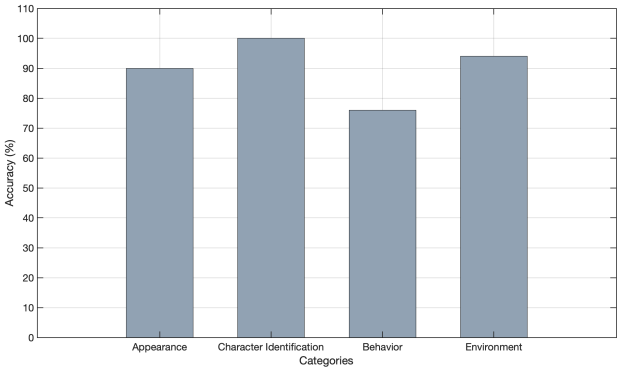

Fig. 5: Accuracy of MLLM-PSC Decoder for Data Recovery

D. Encryption Overhead

Our approach specifically encrypts the sender's private information based on the profile, ensuring privacy security while avoiding the overhead associated with full encryption. Using twenty message sets as the basis, we designated "*Appearance*" as the privacy-sensitive content, from which the MLLM Encoder extracted the private information. Subsequently, we compared the overhead of full encryption with our approach. Fig 6 shows that the average encryption overhead of our method is 30% of that incurred by full encryption.

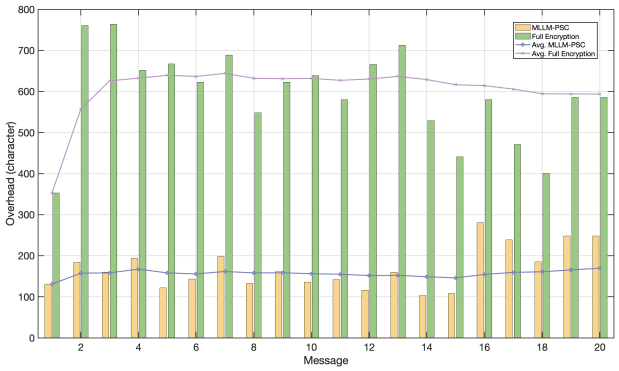

Fig. 6: Overhead of full encryption and MLLM-PSC

V. Conclusion

In conclusion, our proposed MLLM-PSC effectively handles the issues of data heterogeneity and semantic consistency. By leveraging the semantic interpretation capabilities of MLLM, the scheme ensures a unified and consistent conversion of multimodal data into textual semantics without relying on external KB alignment. Additionally, MLLM-PSC employs few-shot learning to protect sensitive semantics.

Acknowledgment

This work was supported in part by the JSPS KAKENHI under Grants 23K11072, in part by the National Natural Science Foundation of China under Grants U21B2019 and 61972255.

References

[1] "Imt traffic estimates for the years 2020 to 2030," ITU. [Online]. Available: https://www.itu.int/pub/R-REP-M.2370-2015
[2] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213–250, 2023.
[3] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210–219, 2022.
[4] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.
[5] K. Lu, Q. Zhou, R. Li, Z. Zhao, X. Chen, J. Wu, and H. Zhang, "Rethinking modern communication from semantic coding to semantic communication," *IEEE Wireless Communications*, vol. 30, no. 1, pp. 158–164, 2023.
[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
[7] J. Xu, T.-Y. Tung, B. Ai, W. Chen, Y. Sun, and D. Gündüz, "Deep joint source-channel coding for semantic communications," *IEEE Communications Magazine*, vol. 61, no. 11, pp. 42–48, 2023.
[8] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," 2023.
[9] OpenAI, "Gpt-4 technical report," 2023.
[10] G. Team, "Gemini: A family of highly capable multimodal models," 2023.
[11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.
[12] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, vol. 2, p. 3, 2023.