

多模态意图感知Token通信方法

于 静^{1,2,3*}, 茹志鹏^{1,2,3}, 王路娟^{1,2,3}, 徐 浩^{1,2,3}, 张翰林^{1,2,3}

(1. 中央民族大学人工智能研究院, 北京 100081; 2. 中央民族大学信息工程学院, 北京 100081;

3. 民族语言智能分析与安全治理教育部重点实验室, 北京 100081)

摘 要: 面向第六代移动通信(sixth-generation mobile communication, 6G)人工智能原生网络中可理解、可推理和可协同的多模态传输需求, 传统语义通信方法多以文本、图像或多模态特征为传输对象, 通常关注源内容压缩和恢复, 难以在通信前显式理解并压缩用户由多源信息表达的任务意图。为验证多模态意图在Token通信中的传输价值, 本文采用组合图像检索作为下游任务, 并提出一种多模态意图感知Token通信方法(Multimodal Intent-Aware Token Communication, MIA-TC)。该方法面向多模态输入中蕴含的任务意图如何在通信前被理解、压缩并传输这一问题, 将发送端多模态意图抽取、目标描述生成、文本Token化传输和接收端任务检索组织为统一流程。在具体实现中, 发送端不直接传输原始图像、操作文本或高维多模态特征, 而是利用多模态大语言模型(Multimodal Large Language Models, MLLM)从参考图像与操作文本中理解用户检索意图, 生成面向任务的目标意图描述, 并将该描述离散化为少量文本Token后进入基础Token通信链路。接收端恢复目标描述后, 利用视觉语言模型完成文本到图像检索。针对发送端意图压缩, 本文设计了包含视觉定位、意图推断、反思过滤和压缩输出的提示词, 使模型在传输前完成任务相关信息选择, 并把中间推理信息压缩为可通信的目标描述。本文进一步提出Token级通信感知检索得分(Token-level Communication-Aware Retrieval Score, Token-CARS)与信道使用级通信感知检索得分(Channel-use-level Communication-Aware Retrieval Score, Channel-CARS), 将下游任务性能分别与Token数量和信道资源消耗进行联合评价; 通信链路采用基础Token分组、校验、信道编码、调制以及单次传输和可靠重传流程。组合图像检索(Composed Image Retrieval on Common Objects in context, CIRCO)验证集上的实验结果表明, MIA-TC与原始图文Token传输后推理方案在下游任务指标上保持相近水平, 并在中高信噪比(Signal-to-Noise Ratio, SNR)的可靠传输设置下取得更好结果; 在SNR = 5 dB的可靠重传模式下, MIA-TC在下游任务性能指标前5位召回率(Recall at 5, Recall@5)取得了16.82, Token-CARS@5达到0.506, Channel-CARS@5达到7.124, 均高于传输后MLLM推理检索和两阶段语义压缩检索; 相比原始图文Token方案, 平均Token数量减少93.58%, 信道使用次数减少87.34%, 平均端到端推理时延减少11.89 ms。相比两阶段语义压缩方法, MIA-TC在下游任务检索性能以及Token-CARS和Channel-CARS综合指标上均表现更好。尤其在Token-CARS和Channel-CARS指标上, MIA-TC均高于原始图文Token方案和两阶段语义压缩方法, 说明发送端先进行多模态意图理解与语义压缩, 能够在保持任务有效性的同时降低通信负载, 为6G AI原生语义通信中的多模态意图传输提供了一种可复现的研究范式。

关键词: Token Communication; 多模态大语言模型; 组合图像检索; 6G智能通信网络; 多模态意图理解

基金项目: 新一代人工智能国家科技重大专项(No.2025ZD0123303); 北京市高层次创新创业人才支持计划科技新星计划资助(No.20250484921)

中图分类号: TN911.2; TP181

文献标识码: A

文章编号: 0372-2112(XXXX)XX-0001-19

电子学报URL: <http://www.ejournal.org.cn>

DOI: 10.12263/DZXB.20260703

Multimodal Intent-Aware Token Communication

YU Jing^{1,2,3*}, RU Zhipeng^{1,2,3}, WANG Lujuan^{1,2,3}, XU Hao^{1,2,3}, ZHANG Hanlin^{1,2,3}

(1. Institute of Artificial Intelligence, Minzu University of China, Beijing 100081, China;

2. School of Information Engineering, Minzu University of China, Beijing 100081, China;

3. Key Laboratory of Ethnic Language Intelligent Analysis and Security Management of MOE, Minzu University of China, Beijing 100081, China)

Abstract: To meet the demand for multimodal transmission with understanding, reasoning, and collaboration capabilities in sixth-generation artificial intelligence (AI)-native networks, conventional semantic communication methods generally take text, images, or multimodal features as transmission objects and mainly focus on source-content compression and reconstruction. Consequently, they have difficulty explicitly understanding and compressing task intent expressed by users through multiple information sources before transmission. To investigate the value of transmitting multimodal intent

in Token communication, this paper adopts composed image retrieval as the downstream task and proposes a multimodal intent-aware token communication (MIA-TC) method. MIA-TC addresses how task intent embedded in multimodal inputs can be understood, compressed, and transmitted before communication, and organizes transmitter-side multimodal intent extraction, target-description generation, text Token transmission, and receiver-side task retrieval into a unified pipeline. Specifically, rather than directly transmitting raw images, manipulation text, or high-dimensional multimodal features, the transmitter employs a multimodal large language model (MLLM) to understand the user's retrieval intent jointly expressed by a reference image and manipulation text, generate a task-oriented target intent description, and discretize the description into a small number of text Tokens for transmission over a basic Token communication link. After recovering the target description, the receiver performs text-to-image retrieval using a vision-language model. For transmitter-side intent compression, we design a prompt consisting of visual grounding, intent inference, reflective filtering, and compressed output, enabling the model to select task-relevant information before transmission and compress intermediate reasoning information into a communicable target description. We further propose the Token-level communication-aware retrieval score (Token-CARS) and the channel-use-level communication-aware retrieval score (Channel-CARS), which jointly evaluate downstream task performance with the number of Tokens and channel resource consumption, respectively. The communication link employs basic Token grouping, error checking, channel coding, modulation, single-shot transmission, and reliable retransmission. Preliminary experiments on the composed Image retrieval on common objects in context (CIRCO) validation set show that MIA-TC achieves downstream task performance comparable to that of the raw image-text Token scheme with post-transmission reasoning and obtains better results under reliable transmission settings at medium-to-high signal-to-noise ratios (SNRs). In reliable retransmission mode at an SNR of 5 dB, MIA-TC achieves a Recall at 5 (Recall@5) value of 16.82, a Token-CARS@5 of 0.506, and a Channel-CARS@5 value of 7.124, all of which outperform post-transmission MLLM reasoning retrieval and two-stage semantic compression retrieval. Compared with the raw image-text Token scheme, MIA-TC reduces the average number of Tokens by 93.58%, the number of channel uses by 87.34%, and the average end-to-end inference latency by 11.89 ms. MIA-TC also outperforms the two-stage semantic compression method in downstream retrieval performance as well as the Token-CARS and Channel-CARS metrics. In particular, MIA-TC achieves higher Token-CARS and Channel-CARS than both the raw image-text Token scheme and the two-stage semantic compression method. These results demonstrate that performing multimodal intent understanding and semantic compression at the transmitter can reduce communication overhead while preserving task effectiveness, providing a reproducible research paradigm for multimodal intent transmission in sixth-generation AI-native semantic communication systems.

Keywords: token communication; multimodal large language model; composed image retrieval; sixth-generation intelligent communication networks; multimodal intention understanding

Foundation Item(s): New Generation Artificial Intelligence-National Science and Technology Major Project (No.2025ZD0123303); Beijing Nova Program (No.20250484921)

0 引言

香农信息论以比特可靠传输为核心,为现代通信系统建立了统一的数学基础^[1]。随着第六代移动通信(sixth-generation mobile communication, 6G)智能网络、人工智能(Artificial Intelligence, AI)原生空口、智能终端、具身机器人、低空无人系统和沉浸式交互的发展,通信系统承载的对象正在由单一信号扩展为图像、文本、语音、视频和传感信息共同组成的多模态内容,智能体可协同性^[2-4]。在这类应用中,接收端往往并不只是需要完整复原源数据,而是需要据此完成检索、决策、规划、生成或协同控制等任务。因此,通信系统的评价目标也正在从比特级准确传输,逐步转向语义是否被正确理解、任务是否被有效完成以及通信资源是否被合理使用。语义通信试图突破传统通信中源信号保真恢复的限制,将消息含义、共享知

识、通信上下文和接收端任务纳入系统设计^[2-4]。已有深度学习语义通信方法在文本、语音和图像传输中展示了较好的压缩能力和抗噪能力^[5-6],说明面向任务的语义保真可在一定程度上替代逐比特恢复。现有多数方法仍以文本、图像或多模态特征本身作为编码对象,重点解决源内容如何压缩、传输和恢复的问题^[5-7]。对于用户意图由多源信息共同表达的场景,系统若仅压缩原始内容,仍可能把大量与下游任务无关的背景、纹理、局部细节或字面信息带入通信链路,从而造成带宽浪费、误差传播和接收端任务理解负担。

Token通信为上述问题提供了新的实现途径^[7]。TokenCom等工作提出将Token作为语义内容的基本传输单元,利用生成式基础模型和多模态大语言模型中的离散Token码本、上下文建模能力和跨模态关联

知识,在发送端和接收端进行Transformer式Token处理^[7-10]。与连续潜变量或端到端黑盒特征不同,Token可表示为离散标识符(Identifier, ID),并进一步转换为比特流,可以转化为比特流并接入分组、校验、信道编码、调制和重传等成熟数字通信模块。同时,Token又保留了与大模型语义空间相连的优势,使通信系统有机会在较低载荷下表达更高层次的语义内容。Transformer^[11]、来自Transformer的双向编码器表征(Bidirectional Encoder Representations from Transformers, BERT)^[12]、生成式预训练Transformer(Generative Pre-trained Transformer, GPT)^[13]、对比语言图像预训练模型(Contrastive Language-Image Pre-training, CLIP)^[14]等模型的发展说明,Token不仅是模型输入输出的基本单元,也是语言、图像和跨模态语义关系的承载形式^[11-15]。不过,现有Token通信更多关注Token级压缩、掩码预测、丢失Token补全和接收端上下文恢复^[7-10]。现有方法默认待传对象已经确定,然后研究如何更可靠或更高效地传输这些Token。这类Token级压缩主要回答的是源内容已经被表示为Token后如何减少传输量、提高传输可靠性或恢复受损Token。本文关注的问题发生在Token化之前,即在源内容进入通信链路之前,先判断哪些内容才是下游检索任务真正需要的通信对象。对于多模态意图通信而言,这个更靠近信源端的问题仍未被充分讨论,即发送端能否先理解用户真实任务意图,再只传输对下游任务有用的紧凑意图Token。若用户输入包含图像和文本,实际需要传输的并非图像的全部像素、视觉Token或文本字面内容,而是由不同模态共同约束的目标状态、对象关系等。从标准化进展看,第三代合作伙伴计划(3rd Generation Partnership Project, 3GPP)已在第五代移动通信(fifth-generation mobile communication, 5G) Advanced和6G早期研究中关注AI Traffic, AI Agent通信以及意图驱动管理等方面。这说明未来网络正在从传统连接能力向业务感知、AI服务承载和意图驱动管理演进。本文的MIA-TC在此背景下,研究多模态任务意图如何在发送端被理解,压缩并以少量文本Token进入通信链路。

多模态大语言模型(Multimodal Large Language Models, MLLM)为发送端意图理解和语义压缩提供了可行基础^[16-18]。MLLM能够同时处理视觉内容和自然语言指令,并基于预训练知识进行对象定位、属性识别、关系推理和意图解释。与传统图像描述模型相比,MLLM不仅能够生成图像描述,而是可以根据问题或操作文本选择性关注与任务有关的视觉证据。链式思维、零样本推理和反思式反馈等提示方法进一步表明,大模型可以通过显式中间推理提升复杂意图

理解的稳定性^[19-21]。本文关注的压缩过程并不是普通文本压缩,也不是机械缩短句子,而是先在发送端完成多模态意图理解,再根据下游任务进行语义选择和表达压缩。由此,意图理解决定压缩内容的语义方向,语义压缩决定通信载荷的表达形式,二者共同构成本文发送端语义信源编码的核心。

组合图像检索(Composed Image Retrieval, CIR)为验证这一思想提供了合适的任务载体^[22-23]。该任务输入参考图像与操作文本,目标是检索符合修改意图的目标图像。用户真正希望传递的不是参考图像中的所有像素,也不是操作文本的孤立字面含义,而是一个目标图像应当如何变化的跨模态任务意图。已有组合图像检索研究表明,用户意图同时分布在视觉和文本模态中,并需要在共同语义空间中进行推理和检索^[22-29]。因此,CIR既具有清晰的多模态意图输入,又具有可量化的检索输出,能够连接发送端意图压缩、通信链路失真和接收端任务性能。从组合图像检索角度看,免训练方法通常可分为两阶段和单阶段思路。两阶段方法先生成参考图像描述,再结合操作文本得到目标图像描述,但由于初始描述生成阶段未引入操作文本,容易遗漏后续修改所需的关键视觉细节^[29]。单阶段思路直接将参考图像和操作文本输入MLLM,使模型在同一推理过程中联合观察视觉证据和文本修改,从而减少中间表示带来的信息损失^[23]。本文借鉴这一任务特性,但研究目标并不是单纯提升CIR检索精度,而是把CIR作为多模态意图通信的可复现实验闭环,用来考察发送端先理解再传输是否能够降低通信负载并保持下游任务有效性。

基于上述动机,本文提出多模态意图感知Token通信方法(Multimodal Intent-Aware Token Communication, MIA-TC)。本文定义的通信对象并不是原始参考图像、操作文本或高维多模态特征本身,而是发送端利用MLLM从多模态输入中抽取出的、能够驱动下游任务的目标意图描述。具体的,MIA-TC将原始多模态内容先转化为面向任务的文本Token载荷,再进入Token通信链路。接收端恢复目标描述后,使用视觉语言模型完成文本到图像检索。与常规语义通信相比,MIA-TC不追求重建输入源,而追求保留与任务目标相关的意图约束。与常规组合图像检索相比,本文进一步把无线信道失真、Token预算、重传开销和任务收益纳入同一分析框架,并通过Token级通信感知检索得分(Token-level Communication-Aware Retrieval Score, Token-CARS)与信道使用级通信感知检索得分(Channel-use-level Communication-Aware Retrieval Score, Channel-CARS)刻画单位Token预算或单位信道资源所获得的检索收益。这样,多模态意图理解、意图驱

动语义压缩和通信资源消耗可以在统一端到端流程中进行评价。图1给出了本文问题的动机。

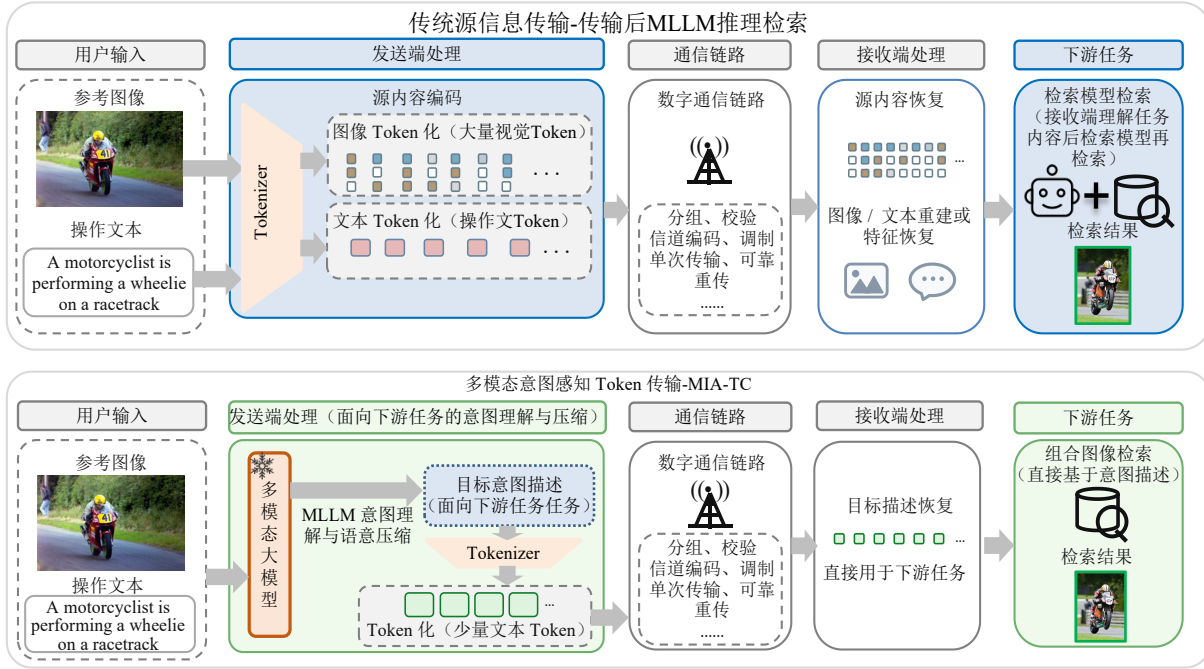


图1 源信息传输与多模态意图感知Token传输的动机对比图

Figure 1 Motivational comparison between source information transmission and multimodal intent-aware Token transmission

综上,本文的主要贡献如下。

(1) 提出多模态意图感知Token通信任务。本文的核心是面向多模态输入中的用户任务意图,研究其如何在发送端被理解、压缩并以文本Token形式进入Token通信链路。不同于直接传输原始图像、操作文本或高维多模态特征,本文将多模态意图理解与基于意图的语义压缩引入通信前端,只传输面向下游任务的目标意图描述。组合图像检索在本文中作为验证载体,用于提供清晰的多模态意图输入和可量化的下游任务输出,从而为6G AI原生语义通信中的多模态意图传输提供可复现的研究范式。

(2) 设计基于MLLM的多模态意图理解与意图驱动语义压缩机制。该机制利用MLLM的多模态理解和推理能力,在发送端完成视觉定位、意图推断、反思过滤和压缩输出,将多模态输入中的任务相关信息凝练为目标意图描述Token。该过程不把中间推理链作为通信载荷,而只传输接收端检索所需的目标描述,从而在保留意图约束的同时提升Token载荷的任务表达效率。

(3) 提出Token-CARS与Channel-CARS通信感知检索指标。上述指标分别从Token预算和信道资源消耗角度统一衡量下游检索收益与通信代价,避免只看检索精度或只看通信开销造成的片面评价。基于组合图像检索(Composed Image Retrieval on Common

Objects in context, CIRCO)验证集的实验表明,MIA-TC在保持下游检索有效性的同时显著压缩平均Token预算,并在Token-CARS和Channel-CARS上优于原始图文Token方案和两阶段语义压缩方法。

1 相关工作

1.1 Token Communication

语义通信的基本目标是使接收端获得与任务相关的含义,而不是逐比特恢复源消息。早期研究从通信含义、知识共享和解释一致性角度讨论语义通信的理论基础^[2],近年来研究进一步将上下文、任务目标和智能体行为纳入通信设计^[3-4]。深度语义通信(Deep Semantic Communication, DeepSC)^[5]等方法利用Transformer直接学习句子级语义编码,在低信噪比条件下相较传统分离式通信具有更好的语义鲁棒性^[5]。图像通信领域的深度联合信源信道编码通过端到端神经网络避免显式压缩和纠错分离,在加性高斯白噪声(Additive White Gaussian Noise, AWGN)^[6]和Rayleigh(Rayleigh fading)^[6]信道中缓解了传统数字链路可能出现的悬崖效应^[6]。这些工作为Token Communication的出现提供了语义通信基础,即通信系统可以围绕语义内容、上下文知识和接收端任务共同设计,而不必始终以源信号级重建作为唯一目标。

Token Communication进一步把通信单位从连续

语义特征转向离散Token。大模型中的Token既是计算单元,也是知识表示单元。文本Token可由字节对编码(Byte Pair Encoding, BPE)、WordPiece或Unigram等子词建模方法产生^[30-32],图像Token可来自视觉Transformer的Patch表示、离散视觉码本或矢量量化生成对抗网络(Vector Quantized Generative Adversarial Network, VQGAN)类视觉Tokenizer^[15, 33-34],音频和视频也可通过频谱切片或时空Patch表示为Token序列^[7]。由于Token ID可以转化为比特流,Token通信能够与分组、校验、信道编码、调制和重传等传统数字通信模块自然衔接^[7]。TokenCom框架进一步指出,预训练Token码本可作为发送端和接收端共享的知识基础,并可在语义源压缩、语义信道编码、语义多址和网络协议等层面利用生成式基础模型(Generative Foundation Model, GFM)或MLLM的上下文建模能力^[7]。这类方法的优势在于既保留数字通信可部署、可校验和可重传的工程特性,又能够借助Transformer模型对Token之间依赖关系的学习实现跨模态语义表达^[7, 11-15]。

在标准方面,3GPP也已从意图驱动管理、空口智能、6G业务需求、系统架构和媒体处理等层面开展相关研究。业务与系统第五工作组(Service and System Aspects working group 5, SA5)已围绕移动网络意图驱动管理服务开展标准化工作。无线接入网第一工作组(Radio Access Network working group 1, RAN1)已研究AI和机器学习辅助的新空口(New Radio, NR)技术,无线接入网第二工作组(Radio Access Network working group 2, RAN2)的调度、重传和资源管理机制可为未来Token化AI流量承载提供协议基础。业务与系统第一工作组(Service and System Aspects working group 1, SA1)已开展6G用例与业务需求研究,业务与系统第二工作组(Service and System Aspects working group 2, SA2)进一步面向6G系统架构开展Stage 2研究。业务与系统第四工作组(Service and System Aspects working group 4, SA4)已启动6G媒体系统研究。这些进展可为Token通信中的业务感知、资源调度和差异化传输提供标准化基础。

现有Token通信研究通常沿着两类思路提升传输效率。第一类思路在发送端进行Token筛选、掩码、重要性排序或语义源压缩,利用语言模型的可预测性和视觉离散码本减少待发送载荷^[7, 10, 33-34]。第二类思路在接收端利用自回归预测或掩码预测恢复受损Token,通过上下文补全降低误包或重传带来的语义损失^[7, 12-13]。这两类方法的共同点是,待传对象通常已经被表示为文本Token、视觉Token序列,研究重点是对已有Token进行筛选、压缩、保护或恢复。

本文与上述Token级压缩思路的区别在于,MIA-TC是在Token化之前先改变待传对象。对于多模态意图任务,原始图像中的许多背景纹理和局部细节并不服务于用户目标,操作文本中的语义目标也必须结合图像上下文才能被正确解释。若直接发送原始图文Token,再在接收端理解任务意图,通信资源仍可能消耗在大量与下游任务弱相关的信息上。因此,MIA-TC在基础Token通信链路之前加入面向任务意图的语义信源编码阶段,先由发送端MLLM将参考图像和操作文本共同约束的任务意图凝练为目标意图描述,再对该描述进行Token化和传输。该过程使待传Token具有更高的任务相关性和更紧凑的语义载荷,从而为后续Token通信提供更有效的输入。

1.2 组合图像检索

组合图像检索要求模型同时理解参考图像与文本修改意图,并检索符合用户期望变化的目标图像^[22]。早期监督式方法通常依赖参考图像、操作文本和目标图像构成的三元组标注数据^[22]。FashionIQ提供服饰图像与自然语言反馈,适合评估细粒度属性修改^[24]。CIRCO则基于COCO图像(Common Objects in Context, COCO)构建开放域组合图像检索基准,提供多个目标答案和语义类别标注,用于缓解单一目标标注带来的假负例问题^[28, 35]。这类任务具有明确的意图驱动特征。用户并不是要求系统找到与参考图像完全相似的图片,而是在保持部分视觉内容的同时施加文本中的修改约束。也就是说,CIR查询中的有效信息并不等同于参考图像和操作文本的简单拼接,而是两种模态共同指向的目标状态。

随着CLIP等视觉语言模型出现,组合图像检索从监督训练逐渐扩展到零样本和免训练场景。CLIP通过大规模图文对比学习建立共享语义空间,使文本描述可以直接用于图像检索^[14]。在此基础上,研究者提出了多种查询生成方式。基于特征融合的方法将参考图像特征与文本特征在共享空间内组合^[26]。Pic2Word把图片映射为伪词Token,使图像内容可以像文本一样参与组合查询^[27]。SEARLE等文本反演方法把图像编码成可插入文本提示的伪Token,从而利用预训练文本编码器完成检索^[28]。免训练组合检索方法进一步利用大语言模型(Large Language Model, LLM)^[13, 20]或多模态大语言模型(MLLM)^[16-18]生成目标描述^[23, 29]。典型两阶段方法通常先用图像描述模型生成参考图像说明,再根据操作文本修改说明得到目标图像描述^[29]。这种方式具有较好的可解释性,但初始图像说明是在不知道操作文本的条件下生成的,容易遗漏后续修改所依赖的关键视觉细节。为解决这一问题,本文在通信前端使用MLLM的一阶

段跨模态推理能力,将参考图像和操作文本同时输入模型,使其在同一推理过程中联合观察视觉证据和文本修改要求,直接生成面向下游检索的目标意图描述^[23]。

本文并不把组合图像检索视为唯一研究目标,而是把它作为多模态意图通信的可度量下游任务。CIR的输入包含跨模态意图,输出目标描述适合作为文本Token传输,检索结果又可以用平均精度均值(mean Average Precision, mAP)和召回率(Recall)等指标定量评估^[22-23, 28]。因此,CIR能够连接发送端意图压缩、通信链路失真和接收端任务性能。本文按照意图处理在通信系统中的位置组织对比方法,包括基础直接检索方法、传输后处理方法和发送前处理方法。据此可评估不同位置进行意图理解对通信效率和检索性能的影响,也能说明在Token通信系统中先压缩意图再传输,与先传输源内容再理解意图之间的差异。

1.3 MLLM的意图理解与语义压缩能力

大模型的语义压缩能力来源于其对Token序列的概率建模能力。Transformer通过自注意力在Token之间建立全局依赖^[11],BERT和GPT分别展示了双向掩码建模与自回归生成在上下文理解中的有效性^[12-13]。从压缩角度看,语言模型越能准确预测下一个Token或被掩码Token,说明它越能用内部知识解释外部序列的不确定^[10]。在通信系统中,这种能力可以转化为更短的语义描述和更强的错误容忍。多模态大语言模型把视觉编码器、语言模型和跨模态对齐模块结合起来,使模型能够在同一推理过程中处理图像和文本。LLaVA等视觉指令调优模型证明了图文对话和视觉推理能力可以通过指令数据进一步增强^[16],GPT-4系列展示了多模态理解和复杂推理的潜力^[17],Qwen系列视觉语言模型也提供了适合本地或边缘部署的开源选择^[18]。这些模型不只是把图像转换为说明文字,还能够根据问题或操作文本选择性关注视觉内容,因此更适合承担发送端语义信源编码器的角色。

提示词工程对MLLM的意图压缩质量有直接影响。链式思维提示能够引导模型显式生成中间推理步骤,提升复杂任务中的推理稳定性^[19]。零样本推理研究表明,恰当的提示语可以在没有任务样例的情况下激发模型的推理能力^[20]。反思式语言反馈则能够帮助模型检查自身推断,过滤不一致或无关的中间结果^[21]。不过,在通信场景中,中间推理过程本身并不适合作为待传载荷。过长的推理文本会增加Token数,且推理链中的不确定表达可能被接收端误用。

基于上述观察,本文将MLLM的作用限定为发送

端意图理解和语义压缩,并将其部署于通信链路之前的发送侧具备AI算力的终端用户设备或端侧智能体。MLLM在发送端本地侧读取参考图像和操作文本,按照本文设计的提示词完成视觉定位、意图推断和反思过滤,但最终只输出一条简短目标意图描述。这样的设计兼顾了可解释性和通信效率。提示词内部允许模型进行推理,但通信链路只承载压缩后的目标描述。由于输出是自然语言文本Token,接收端不需要部署同等规模的MLLM即可使用CLIP检索模型完成任务。

2 多模态意图感知Token通信方法

2.1 问题描述与总体框架

本文研究的输入为一个查询三元组,其中参考图像记为 I_r ,操作文本记为 T_m ,候选图像库记为 D 。传统组合图像检索直接在 I_r 、 T_m 和 D 之间建立匹配关系。本文进一步考虑通信链路, I_r 和 T_m 位于发送端的终端用户设备并处理得到目标意图描述,下游检索在接收端或边缘检索服务器执行。通信系统的目标不是恢复 I_r 和 T_m ,而是恢复足以完成检索任务的目标意图描述。

给定发送端MLLM记为 Φ_θ ,意图压缩提示词记为 p_c ,MIA-TC首先使用部署在发送端的MLLM理解参考图像 I_r 与操作文本 T_m 共同表达的用户意图,其中MLLM部署于具备AI算力的终端用户设备或端侧智能体,并生成面向下游检索任务的目标意图描述 T_t 。 T_t 是一段自然语言,用来描述用户希望检索到的目标图像。随后,文本Tokenizer把 T_t 映射为离散Token ID序列 s ,该序列经定长编码、分组、循环冗余校验(Cyclic Redundancy Check, CRC)、信道编码和调制后进入6G网络。本文中的6G网络不直接接收原始参考图像和操作文本,也不替代发送端MLLM完成原始图文意图解析,而是对目标意图Token分组进行业务感知承载。具体而言,网络侧可根据Token数量、分组长度、时延约束、可靠性需求和错误容忍度等信息进行无线资源调度、链路自适应、重传控制和差异化保护。接收端完成解调、译码和CRC检查,得到恢复Token序列 $\hat{s}$,并反解码为恢复描述 $\hat{T}_t$ 。最后,CLIP文本编码器把 $\hat{T}_t$ 映射到语义空间,与候选图库的图像特征计算相似度并排序。图2展示了完整流程。

上述流程把发送端意图理解、目标描述生成、Token化传输和接收端检索连接为一个闭环。MIA-TC的关键是在Token通信链路之前加入面向下游任务的意图语义信源编码。若目标意图描述能够准确保留检索所需的主体、属性、动作、空间关系和必要背景,则通信链路只需保护少量文本Token即可支撑后续

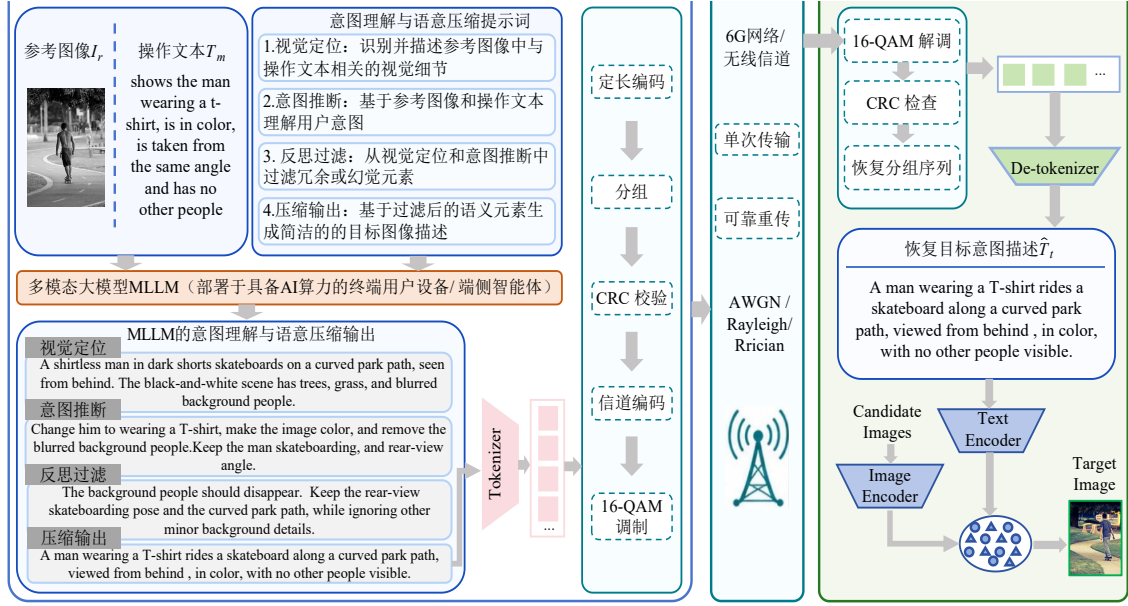


图2 MIA-TC总体框架

Figure 2 Overall framework of MIA-TC

任务执行。本文涉及的主要符号及其含义如表1所示。

表1 MIA-TC主要符号说明

Table 1 Main notations of MIA-TC

符号	含义	说明
I_r	参考图像	用户给出的视觉参照
T_m	操作文本	描述期望修改的自然语言
p_c	意图压缩提示词	控制 MLLM 输出目标意图描述
T_i	目标意图描述	真正进入通信链路的语义文本
s	Token ID 序列	由 Tokenizer 得到的离散序列
V	词表大小	决定定长编码 bit 数
G	每包 Token 数	本文取 4
U	信道使用次数	衡量通信资源
$\hat{T}_i$	恢复目标描述	由恢复 Token 反解码得到

2.2 基于 MLLM 的发送端意图语义压缩

发送端意图语义压缩的核心并不是把参考图像或操作文本逐字逐像素压缩,而是先理解用户真正希望目标图像发生什么变化的意图,再把这种跨模态意图表达为可传输的目标描述。对于组合图像检索输入,操作文本中的指令通常不能脱离参考图像解释,参考图像中的视觉细节也并非全部与任务有关。因此,本文将发送端需要识别的操作意图概括为四类。第一类是保留意图,即目标图像中仍需继承参考图像的主体、姿态、类别、动作、视角和关键场景关系。第二类是修改意图,即根据操作文本改变对象属性、颜色、状态、数量、背景或观察角度。第三类是去除或弱化意图,即删除不应继续影响检索的干扰对象、背景人物或冗余纹理。第四类是补充或重构意图,即在

文本没有完全显式说明时,结合参考图像上下文推断目标图像应出现的必要环境和视觉线索。MLLM 的作用是在同一次多模态推理中同时判断这些操作意图,并把它们整理为接收端可直接使用的文本语义载荷,公式化该过程如下:

$$T_i = \phi_\theta(I_r, T_m, p_c) \quad (1)$$

其中, Φ_θ 表示发送端采用的 MLLM, p_c 表示本文设计的意图理解与语义压缩提示词, T_i 表示最终进入 Token 通信链路之前的目标意图描述。

与先生成参考图像 caption 再修改 caption 的两阶段处理方式不同, MIA-TC 在视觉理解阶段即引入操作文本,使模型围绕用户修改意图选择视觉证据,而不是生成与任务无关的完整图像说明。如图2发送端示例所示,操作文本要求同角度、彩色、穿T恤且没有其他人。因此,发送端需要保留滑板动作、背向视角和弯曲路径,修改人物着装与颜色风格,并去除背景中其他人物。

为了使 MLLM 稳定完成上述意图识别与表达压缩,本文设计了意图理解与语义压缩的思维链提示词,如图3所示。该提示词在发送端内部依次引导模型完成视觉定位、意图推断、反思过滤和压缩输出。四个阶段均服务于同一个目标,即把多模态输入中的任务相关语义凝练为一条短目标描述,而不是把完整推理链、完整图像说明或全部视觉细节送入信道。图2中的中间输出展示了这一过程。视觉定位先识别与操作文本相关的视觉细节,意图推断把用户指令转化为目标状态,反思过滤确定需要特别关注的视觉信息

以及删除与检索无关或可能误导检索的元素,压缩输出最终得到“A man wearing a T-shirt rides a skateboard

along a curved park path, viewed from behind, in color, with no other people visible.”这样的目标图像描述。

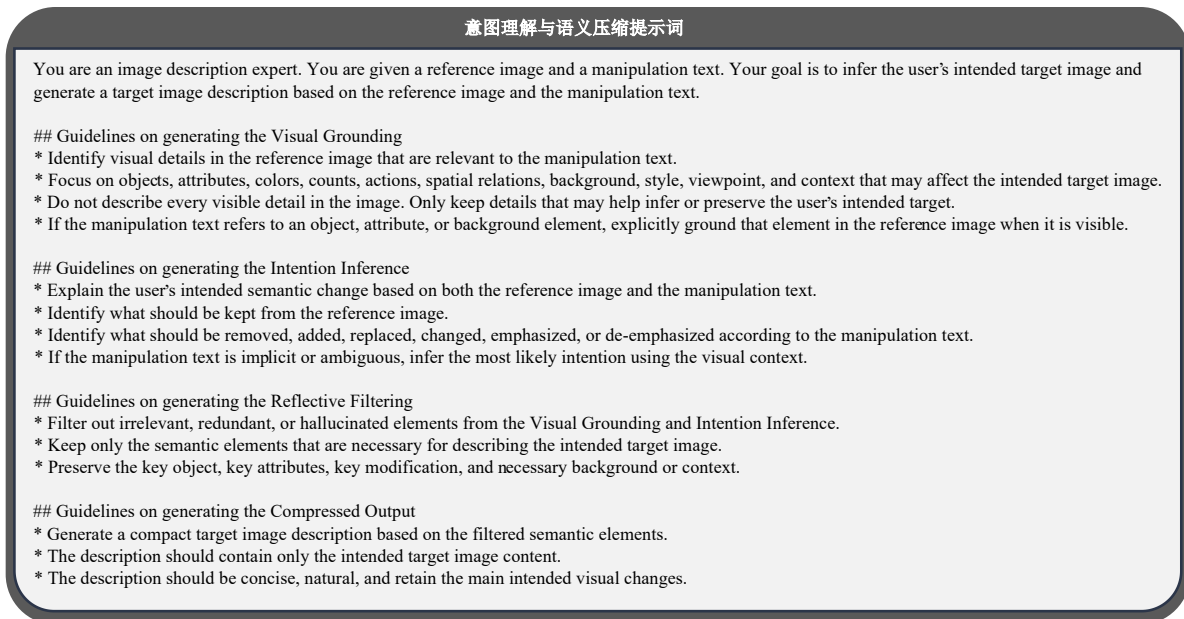


图3 意图理解与语义压缩提示词

Figure 3 Prompt design for intent understanding and semantic compression

视觉定位阶段要求模型只描述与操作文本有关的视觉证据。例如若操作文本涉及颜色变化,则优先定位对象颜色、材质和图像风格;若涉及背景变化,则优先定位背景类型、人物或环境元素;若涉及物体增删,则优先定位物体之间的空间关系和可替换区域。该阶段的重点不是生成完整图像描述,而是为后续意图推断保留必要视觉依据。例如图2中“same angle”“has no other people”等指令会使模型关注拍摄视角、公园道路和背景行人。

意图推断阶段要求模型把操作文本解释为目标图像应满足的语义状态,而不是机械拼接参考图像描述与操作文本。对于“shows the man wearing a t-shirt, is in color, is taken from the same angle and has no other people”这类指令,模型需要推断保留主体滑板动作和背向视角,修改衣着和图像颜色,并去除其他人物。该阶段的中间推理不直接进入通信链路,只用于确定最终目标描述应包含哪些语义元素。

反思过滤阶段用于检查视觉定位和意图推断中是否存在遗漏、冗余或幻觉。首先,模型需要检查关键主体、关键属性、修改动作是否被保留。例如图2中反思过滤阶段的“the curved park path”补充了在意图理解中遗漏的需要关注的视觉信息。其次,模型需要过滤不会改变检索结果但会增加Token预算的细节,例如无关背景纹理、拍摄噪声、无关小物体或过

度具体的场景描述。

压缩输出阶段要求模型只输出目标图像描述,不输出推理过程、解释文字。输出内容包含主体、关键属性、修改后的状态,同时避免过长从而引入额外Token开销和检索噪声。由于通信链路只承载压缩后的目标描述,提示词内部产生的视觉定位、意图推断和反思过滤内容不会被发送。这样既利用了MLLM的多模态推理能力,又避免把冗长中间推理链作为通信负担。

图4给出了不同类型查询下四阶段提示词的处理示例,分别覆盖对象修改与背景修改、属性修改,以及拍摄视角、空间关系和属性联合修改等场景。该流程在不同查询中均保持四阶段的结构:先围绕操作文本筛选参考图像中的任务相关视觉证据,再判断目标图像中需要保留、修改或删除的语义约束,最后过滤冗余信息并生成面向检索任务的目标意图描述。该提示词并不依赖某一特定MLLM,而是以自然语言提示词的方式规定多模态意图压缩的功能步骤,因此可迁移到其他具备图文理解能力的MLLM中。迁移时可保持四阶段结构不变,并根据不同模型的指令遵循能力适当调整输出格式、示例数量和最大生成长度。

本文的目标描述不是参考图像caption,也不是操作文本复述,而是由参考图像和操作文本共同决定的



图4 不同类型查询下四阶段意图理解与语义压缩流程示例

Figure 4 Examples of the four-stage intention understanding and semantic compression workflow for different query types

多模态意图Token载荷。若描述过短,关键主体或修改目标可能丢失;若描述过长,无关词会增加信道负载并干扰CLIP文本编码器。因此,MIA-TC先理解意图,再压缩表达,这是发送端语义信源编码的核心原则,使最终传输的文本Token既能表达下游任务所需语义,又能保持较高的通信效率。图3中的四个阶段均在发送端内部完成。最终进入通信链路的只有压缩输出阶段的目标意图描述。这样既保留了MLLM的意图理解和语义压缩能力,也避免把冗长推理链传入信道。

2.3 Token化与Token通信链路

目标意图描述 T_i 生成后,发送端首先使用与下游CLIP文本编码器一致的Tokenizer进行Token化。该步骤的作用是把可读的自然语言目标描述转化为可编码、可校验和可调制的离散Token ID序列,从而使意图描述能够进入数字通信链路。Token化过程表示为

$$s = \text{Tok}(T_i) = [s_1, s_2, \dots, s_N], s_i \in \{0, 1, \dots, V-1\} \quad (2)$$

其中, $\text{Tok}(\cdot)$ 表示文本Tokenizer, T_i 表示发送端生成的目标意图描述, s 表示由 N 个Token ID组成的离散序列, s_i 表示第 i 个Token的整数索引, V 表示Tokenizer词表大小。采用CLIP tokenizer时, V 约为49 408,因此每

个Token ID采用 $b = \lceil \log_2 V \rceil = 16$ bit定长二进制表示。

完成Token化后,发送端将Token ID转换为比特流并进行分组保护。本文使用基础Token通信链路,目的在于清晰评估发送端意图理解与语义压缩对通信负载的影响。具体而言,发送端每 G 个Token组成一个分组,本文取 $G = 4$;每个分组附加CRC-16进行错误检测,再进入码率为1/2的卷积码;调制方式采用16阶正交幅度调制(16-ary Quadrature Amplitude modulation, 16-QAM),每个符号承载4 bit^[36]。对于长度不足 G 的最后一个分组,使用填充Token补齐,并在接收端恢复后删除。有效载荷比特数公式表示如下

$$B_{\text{payload}} = N \cdot b, \quad b = \lceil \log_2 V \rceil \quad (3)$$

其中, B_{payload} 表示未加入CRC和信道编码前的有效载荷比特数, b 表示每个Token ID定长编码所需的比特数。式(4)说明,发送端意图压缩越充分, N 越小,后续进入通信链路的基础载荷也越低。

从发送端输入网络后的传输需求看,目标意图Token流与传统连续图文数据流不同,其载荷长度较短,通常表现为小分组和突发式到达,因此需要网络支持低开销的小包调度和快速接入。同时,目标意图Token服务于接收端检索任务,既需要较低时延,也

需要足够的文本恢复可靠性,因此6G网络需要支持面向AI Token业务的低时延调度、可靠传输、链路自适应、重传控制和差异化保护。本文采用的CRC检测、卷积码、16-QAM、单次传输和可靠重传构成上述网络需求的基础链路实现,用于分析不同可靠性策略对目标描述恢复质量和下游检索性能的影响。

在单次传输模式下,每个分组只发送一次。设总分组数为 J ,CRC长度为 c ,信道编码码率为 R_c ,调制阶数为 M ,则单次传输所需信道使用次数近似为

$$U_{\text{single}} = \sum_{j=1}^J \left\lceil \frac{Gb+c}{R_c \log_2 M} \right\rceil \quad (4)$$

其中, U_{single} 表示单次传输模式下的总信道使用次数, J 表示Token分组数量, G 表示每个分组包含的Token数, c 表示CRC校验位长度, R_c 表示信道编码码率, M 表示调制阶数。分母 $R_c \log_2 M$ 反映每次信道使用能够承载的有效编码比特数,向上取整表示一个分组需要整数个信道符号。

在可靠重传模式下,CRC校验失败的分组会再次发送,直到校验通过或达到最大重传次数。设第 j 个分组的平均传输次数为 τ_j ,则可靠重传模式的信道使用次数为

$$U_{\text{reliable}} = \sum_{j=1}^J \tau_j \cdot \left\lceil \frac{Gb+c}{R_c \log_2 M} \right\rceil \quad (5)$$

其中, U_{reliable} 表示可靠重传模式下的总信道使用次数, τ_j 表示第 j 个分组的平均发送次数。

单次传输模式中,若CRC失败,则该分组对应的Token直接删除,不使用接收端MLLM进行预测补全。该设置可以观察信噪比(Signal-to-Noise Ratio, SNR)下降导致的Token丢失、文本破损和检索性能下降。可靠重传模式中,失败分组触发重传,检索性能通常更接近无信道失真条件,但信道使用次数会随SNR降低而增加。两种模式分别对应低时延和高可靠两类应用需求,并且只改变链路层错误处理方式,不改变发送端意图压缩结果。基础Token通信链路参数如表2所示。

2.4 接收端语义恢复与检索

接收端完成解调、Viterbi译码^[37]和CRC校验后,

表2 基础Token通信链路参数

Table 2 Parameters of the basic Token communication link

模块	参数	本文设置	作用
Tokenizer	词表规模	约49 408	确定Token ID范围
定长编码	b	16 bit	把Token ID转为比特
分组	G	4 Token/组	控制packet粒度
校验	CRC	16 bit	检测packet错误
信道编码	卷积码	$R_c = 1/2$	提高抗噪能力
调制	16-QAM	4 bit/符号	映射为信道符号

得到恢复Token序列 $\hat{s}$ 。若采用单次传输模式,CRC失败的分组直接删除,其中包含的4个Token不进入反解码过程。本文不把失败Token替换为未知Token,因为该符号会被Tokenizer视为普通输入符号,可能给CLIP文本编码器引入额外噪声。若采用可靠重传模式,只有CRC通过的分组才会进入后续恢复,因此文本精确匹配率通常较高,但通信开销也更高。

恢复描述 $\hat{T}_i$ 由同一Tokenizer反解码得到。接收端随后执行文本到图像检索。具体而言,CLIP文本编码器 Ψ_T 用于提取恢复描述特征,CLIP图像编码器 Ψ_I 用于提取候选图像 I_c 的图像特征,二者通过归一化余弦相似度计算匹配分数,得分越高表示候选图像越符合目标意图。检索分数定义为

$$\text{Score}(I_c, T_i) = \cos(\Psi_I(I_c), \Psi_T(T_i)) \quad (6)$$

其中, $\text{Score}(I_c, \hat{T}_i)$ 表示候选图像 I_c 与恢复目标描述 $\hat{T}_i$ 之间的匹配分数, $\cos(\cdot, \cdot)$ 表示归一化余弦相似度。接收端按照该分数对候选图库排序,接收端按照匹配分数对候选图像进行降序排列,并返回排序结果。

这种接收端设计的优点在于部署简单。大模型推理被放在发送端或边缘编码侧,接收端只需要Tokenizer和CLIP检索模型。若未来系统需要更强的鲁棒性,可以在不改变MIA-TC主流程的情况下加入非均等保护、Token重要性编码或检索重排序模块。

2.5 通信感知评价指标

传统组合图像检索通常使用mAP@K和Recall@K指标,语义通信通常报告误码率、误Token率、信道使用次数或比特数。单独使用任一类指标都无法评价意图通信的整体效果。高mAP但消耗大量信道资源的方法未必适合6G低时延应用,Token数极少但任务性能很差的方法也没有实际价值。因此,本文提出通信感知检索指标,即任务性能与通信资源消耗的比值。

设下游检索性能指标为 Q_K ,本文默认取mAP@5,也可替换为mAP@10或Recall@K。为评价单位Token预算能够带来的任务收益,本文定义Token-CARS指标为

$$\text{Token-CARS}@K = \frac{Q_K}{N_{\text{avg}}} \quad (7)$$

其中,Token-CARS@K表示在评价深度K下的Token级通信感知检索得分, Q_K 表示下游检索指标, N_{avg} 表示平均目标描述Token数。该指标用于比较不同语义压缩策略的任务表达效率,数值越高表示单位Token预算获得的检索收益越大。

为进一步考虑物理信道资源消耗,本文定义Channel-CARS指标,将下游任务收益与平均信道使用次数联系起来,形式为

$$\text{Channel-CARS}@K = \frac{Q_K}{U_{\text{avg}}} \quad (8)$$

其中, Channel-CARS@K 表示在评价深度 K 下的信道使用级通信感知检索得分, U_{avg} 表示平均信道使用次数。该指标适合比较不同 SNR、不同重传策略和不同 Token 载荷长度下的端到端通信效率。

3 实验验证与结果分析

3.1 数据集、对比方法与评测指标

本文实验使用 CIRCO 数据集进行评测。CIRCO 基于 COCO 图像构建开放域组合图像检索基准, 每个查询由参考图像、操作文本和候选图库组成, 并为同一查询提供多个正样本目标和语义类别标注, 可缓解单一目标标注带来的假负例问题, 更适合使用 mAP 评价多目标检索性能^[28, 35]。本文选择 CIRCO 的验证集作为实验评测集, 其查询同时包含对象、属性、背景和关系修改, 能够覆盖多模态意图理解与语义压缩中的主要操作类型。

对比方法按通信系统中的处理位置分为三类。(1) 基础直接检索方法, 包括文本检索、图像检索和图文融合检索。文本检索只传输操作文本 Token, 接收端直接用恢复文本检索。图像检索只传输 VQGAN 图像 Token^[34], 接收端基于重建图像检索。图文融合检索同时传输图像 Token 和操作文本 Token, 并在接收端分别计算 CLIP 图像特征和文本特征, 再通过平均融合得到查询向量。(2) 发送后处理方法, 主要对比传输后 MLLM 推理检索方法。该方法在发送端分别将参考图像编码为 VQGAN 图像 Token, 并将操作文本编码为文本 Token, 经信道传输后, 接收端根据恢复的图像 Token 和文本 Token 重建参考图像与操作文本, 再将二者共同输入 MLLM 生成目标描述。为控制模型与提示词因素, 该方法与 MIA-TC 采用相同的 MLLM 和图 3 所示的四阶段意图理解与语义压缩提示词。二者的主要区别在于意图推理所处的位置及 MLLM 的输入质量, MIA-TC 在发送端基于原始参考图像和操作文本完成推理, 而传输后 MLLM 推理检索在接收端基于信道传输后重建的图文内容完成推理。(3) 发送前处理方法, 包括两阶段语义压缩检索方法和本文的 MIA-TC 方法。两阶段语义压缩检索首先使用 MLLM 和图像描述提示词生成参考图像说明, 随后将生成的图像描述与原始操作文本共同输入 LLM, 并使用目标描述生成提示词得到最终目标图像描述。由于两个阶段分别承担图像内容描述和目标意图改写功能, 其提示词设置与 MIA-TC 的一阶段多模态意图理解提示词不同。MIA-TC 则在发送端同时输入原始参考图像和操作文本, 通过一次多模态推理直接生

成目标意图描述。(4) 另设无信道上界, 用于表示目标描述不经过信道失真时的检索性能, 其中目标描述由 MLLM 使用意图理解思维链条一阶段生成。

评测指标包括 mAP@5、mAP@10、Recall@5、Recall@10、平均 Token 数、平均信道使用次数、Token-CARS 和 Channel-CARS。mAP 和 Recall 用于评价下游检索性能。平均 Token 数用于评价语义压缩强度。平均信道使用次数用于评价通信资源消耗。Token-CARS 和 Channel-CARS 用于综合评价任务收益与通信成本。

3.2 实现细节

MLLM 采用 Qwen3-VL-8B-Instruct, 推理温度设为 0 以降低随机性。每个查询输入参考图像、操作文本和本文设计的意图理解与语义压缩思维链提示词。输出经过简单清洗, 只保留最终目标意图描述, 不保留模型内部推理文本。Tokenizer 采用与 CLIP ViT-L/14 文本编码器一致的 CLIP tokenizer, 便于后续检索与 Token 统计。

通信链路采用基础 Token 传输流程, 每 4 个 Token 构成一个 packet, 每个 Token ID 按 16 bit 定长编码, packet 载荷为 64 bit, 附加 CRC-16 后为 80 bit。信道编码使用码率 1/2 卷积码, 调制方式为 16-QAM。主实验信道为 AWGN, SNR 范围为 0 到 20 dB, 步长为 1 dB。本文报告两种错误处理模式。单次传输模式中, 每个 packet 只发送一次, CRC 失败后删除对应 Token。可靠重传模式中, CRC 失败的 packet 重传, 并统计平均重传次数和总信道使用次数。

3.3 主结果分析

考虑到单次传输模式在低 SNR 下容易出现连续 packet 丢失, 前期结果波动较大, 单次传输主结果点调整为 SNR 为 8 dB, 以便在多数方法已经恢复出一定检索能力的条件下比较任务性能与通信效率。可靠重传模式用于观察在较差信道条件下通过重传换取文本恢复质量的效果, 因此仍选取 SNR 为 5 dB 作为代表点。表 3 和表 4 分别给出可靠重传和单次传输结果。

表 3 显示, 可靠重传相比于单次传输显著提升了各文本类方法的恢复质量。本文 MIA-TC 在 SNR 为 5 dB 时达到 8.67 的 mAP@5, 以及 16.82 的 Recall@5, Token-CARS 为 0.506, Channel-CARS 为 7.124, 均高于两阶段语义压缩检索方法和传输后 MLLM 推理检索方法。与原始图文 Token 传输后推理相比, MIA-TC 的信道使用次数从 961 184 次降低到 121 640 次, 说明发送端先理解意图可以在保持下游任务性能不变的同时避免把大量无关图像 Token 传过信道。

表 4 显示, 在 SNR 为 8 dB 的单次传输模式下, 本文 MIA-TC 的平均 Token 数为 17.12, 显著少于图像检索和图

表3 SNR为5 dB时可靠重传模式结果

Table 3 Results in reliable retransmission mode at an SNR of 5 dB

类别	方法	mAP@5 ↑	R@5 ↑	Token数 ↓	信道次数 ↓	Token-CARS@5 ↑	Channel-CARS@5 ↑
基础直接检索	文本检索	2.15	3.64	10.63	80 080	0.202	2.684
	图像检索	1.04	2.27	256.00	881 104	0.004	0.118
	图文融合检索	3.28	6.36	266.63	961 184	0.012	0.342
发送后处理	传输后MLLM推理检索	6.97	13.64	266.63	961 184	0.026	0.725
发送前处理	两阶段语义压缩检索	7.46	14.55	36.39	248 000	0.205	3.008
	MIA-TC检索(ours)	8.67	16.82	17.12	121 640	0.506	7.124
上限	无信道性能上界	10.02	18.64	—	—	—	—

注: Channel-CARS列数值已放大 10^5 倍显示。

文融合检索的256.00与266.63,也少于两阶段语义压缩检索的36.39。此时MIA-TC的mAP@5为5.28, Recall@5为9.55。虽然在不重传条件下其mAP略低于两阶段语义压缩检索的6.45,但MIA-TC的Token-CARS达到0.308, Channel-CARS达到12.808,均高于两阶段方法的

0.177和7.761,说明目标意图Token在有限信道资源下具有更高的任务表达效率。在此下游任务指标上不及两阶段语义压缩检索方法的主要原因是,两阶段语义压缩检索方法没有使用压缩输出,所以目标图像描述更详细,导致在单次传输下保留更多的目标意图信息。

表4 SNR为8 dB时单次传输模式结果

Table 4 Results in single-shot transmission mode at an SNR of 8 dB

类别	方法	mAP@5 ↑	R@5 ↑	Token数 ↓	信道次数 ↓	Token-CARS@5 ↑	Channel-CARS@5 ↑
基础直接检索	文本检索	1.39	4.09	10.63	26 720	0.131	5.201
	图像检索	0.30	0.46	256.00	394 240	0.001	0.077
	图文融合检索	1.29	2.27	266.63	420 960	0.005	0.305
发送后处理	传输后MLLM推理检索	5.41	9.55	266.63	420 960	0.020	1.284
发送前处理	两阶段语义压缩检索	6.45	12.27	36.39	83 120	0.177	7.761
	MIA-TC检索(ours)	5.28	9.55	17.12	41 240	0.308	12.808
上限	无信道性能上界	10.02	18.64	—	—	—	—

注: Channel-CARS列数值已放大 10^5 倍显示。

3.4 不同SNR下的性能变化

为分析在不同信道质量下组合图像检索性能和通信感知评价结果,图5展示CIR任务mAP@5随SNR

变化的结果,左图为单次传输,右图为可靠重传。横轴为SNR,纵轴为mAP@5,曲线包含基础直接检索方法、发送后处理方法、发送前处理方法和无信道上界。

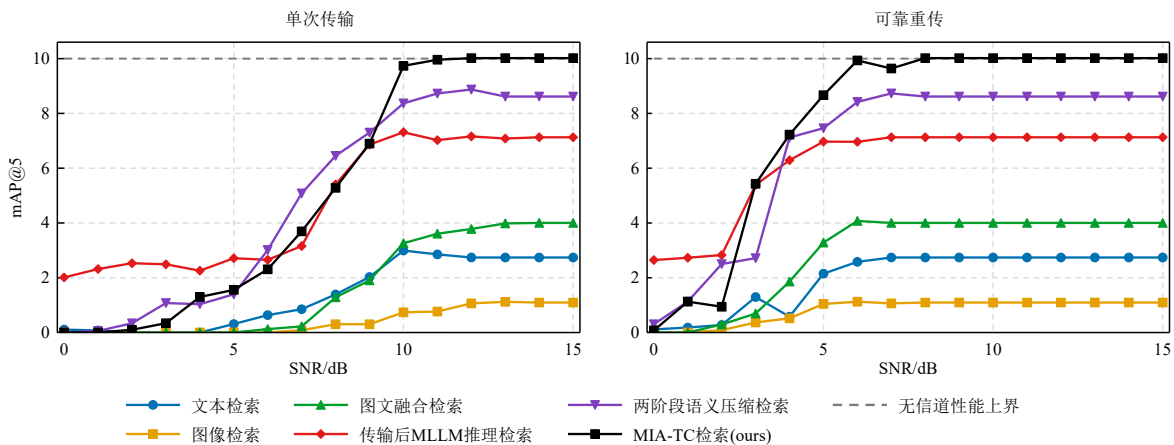


图5 不同SNR下CIR任务mAP@5变化曲线

Figure 5 mAP@5 curves of the CIR task under different SNRs

图5表明,在低SNR条件下,MIA-TC的mAP@5与两阶段语义压缩检索和传输后MLLM推理检索处于相近水平,说明即使只传输压缩后的目标意图描述,发送端MLLM完成的多模态意图理解和语义压缩仍能够保留下游检索所需的主要语义。在中高SNR条件下,MIA-TC逐渐接近无信道性能上界,并在性能上限上超过两阶段语义压缩检索和传输后MLLM推理检索。其原因在于,两阶段语义压缩检索先生成参考图像描述,再根据操作文本进行修改,第一阶段容易遗漏后续修改所依赖的关键视觉信息;对于传输后MLLM推理检索,接收端MLLM的具体输入为根据恢复图像Token重建的参考图像,以及根据恢复文本Token得到的操作文本。在单次传输模式下,CRC校验失败的分组会被删除,可能造成图像局部内容缺失或生成质量下降。这些误差发生在MLLM推理之前,会进一步影响视觉定位、修改约束识别和目标描述生成。随着SNR提高,输入Token的恢复质量改善,该方法的检索性能也随之提升,说明信道失真带来的图片质量下降是其低SNR阶段的重要影响因素。在可靠重传模式下,CRC校验与重传能

够降低Token错误,信道影响主要转化为重传次数和信道使用开销,但VQGAN图像编码与重建本身仍会损失细粒度视觉信息。相比之下,MIA-TC在原始图文上完成意图推理后才传输目标描述,信道失真主要作用于已生成的目标描述Token,不会重新改变此前完成的视觉定位和意图推断过程。因此,MIA-TC在中高SNR条件下表现出更加稳定的语义完整性和检索可用性。

图6进一步从通信感知角度验证了MIA-TC的核心优势。左图给出可靠重传模式下的Token-CARS@5,右图给出Channel-CARS@5。Token-CARS@5反映单位Token预算带来的检索收益,Channel-CARS@5反映单位信道资源带来的检索收益。可以看到,MIA-TC在两个指标上均显著高于基础直接检索、传输后MLLM推理检索和两阶段语义压缩检索。这说明本文方法不是单纯减少Token数量,而是在发送端先完成多模态意图理解,再对任务相关语义进行压缩,使传输的文本Token具有更高的任务表达密度。MIA-TC在保留目标主体、属性、背景和否定约束等意图信息的同时,提升了Token载荷的任务表达效率。

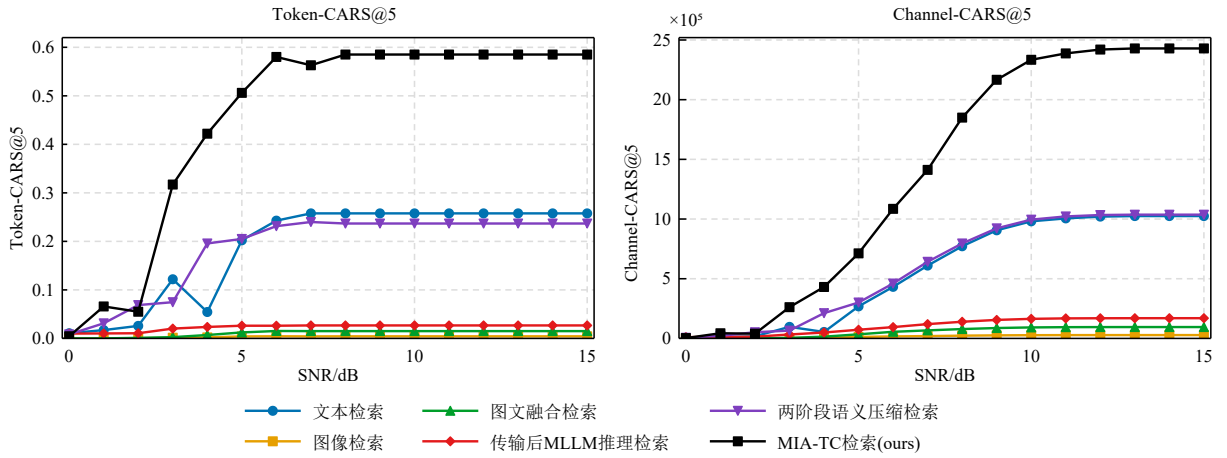


图6 不同SNR下Token-CARS@5和Channel-CARS@5变化曲线

Figure 6 Token-CARS@5 and Channel-CARS@5 curves under different SNRs

从Token-CARS@5指标中可以得出,传输后MLLM推理检索需要先传输原始图文Token,再在接收端生成目标描述,通信开销较大。两阶段语义压缩检索能够减少Token数量,但其目标描述依赖初始图像说明,容易受到视觉信息遗漏影响。MIA-TC通过一阶段跨模态意图理解直接生成目标意图描述,在语义保真和载荷紧凑性之间取得更好的平衡。从Channel-CARS@5指标中可以得出,随着SNR升高,MIA-TC的检索性能接近无信道情况,而信道资源消耗保持较低,因此Channel-CARS@5提升最快并保持最高水平。该结果说明,发送端意图压缩不仅降低了

Token预算,也降低了可靠传输中的重传代价,从而在端到端通信效率上体现出更明显优势。

Token-CARS和Channel-CARS并不用于替代mAP或Recall等下游任务指标,而是用于补充衡量任务收益与通信代价之间的关系。当SNR较低时,任务性能主要受文本恢复错误和分组丢失影响,CARS指标随恢复质量提升而明显变化。当SNR继续升高并使检索性能逐渐接近无信道上界时,CARS的增长会趋于饱和,此时指标差异主要由平均Token数和信道使用次数决定。因此,若某一方法通过极端压缩得到很小的Token数,但导致检索性能明显下降,则较高的

CARS 并不能单独说明方法更优,仍需结合 mAP、Recall、恢复文本质量和信道使用次数共同判断。本文因此将 CARS 作为通信感知补充指标,用于分析任务收益与通信资源之间的折中,而不将其作为唯一评价标准。

图 7 展示不同方法的 Token 长度分布以及可靠重传模式下的平均信道资源开销。上图给出各方法的平均 Token 长度,下图给出 SNR 为 10 dB 时可靠重传模式下的平均信道使用次数,纵轴单位为千次信道使用。图 7 进一步说明, MIA-TC 在保持较短 Token 载荷的同时,能够显著降低可靠重传模式下的信道资源消耗。由上图所示, MIA-TC 平均 Token 长度为 17.12,明显低于两阶段语义压缩检索的 36.39,也显著低于图像检索、图文融合检索和传输后 MLLM 推理检索。由下图所示,在 SNR 为 10 dB 的可靠重传模式下, MIA-TC 平均信道使用次数为 42.92,而两阶段语义压缩检索为 86.60,图文融合检索和传输后 MLLM 推理检索均为 435.26。该结果说明,发送端意图压缩不仅减少了待传 Token 数量,也进一步降低了可靠传输中的信道资源开销。

3.5 不同信道下的鲁棒性实验分析

为进一步考察 MIA-TC 在不同无线传播环境下的稳定性,我们在相同 CIRCO 验证集、相同 Token 分组、CRC 校验、卷积码和 16-QAM 调制设置下,补充 Rayleigh 衰落信道和 Rician 衰落信道实验。Rayleigh 信道用于模拟无明显直射径的多径衰落场景, Rician 信道用于模拟存在主径和散射分量共同作用的场景。为便于与前文主实验对应,展示了可靠重传模式的数据,并选取 SNR 为 5 dB。表 5 汇总了 MIA-TC 在 Rayleigh 和 Rician 两种信道下的关键鲁棒性结果。

表 5 表明,在 Rician 信道中由于存在直达径分量,衰落影响相对较弱, MIA-TC 检索(ours)仍取得 8.16 的 mAP@5 和 6.41 的 Channel-CARS@5,保持了较好的任务性能和单位信道资源收益。在 Rayleigh 衰落信道下,由于深衰落更频繁,所有方法的文本恢复质量和检索性能均受到更明显影响。传输后 MLLM 推理检

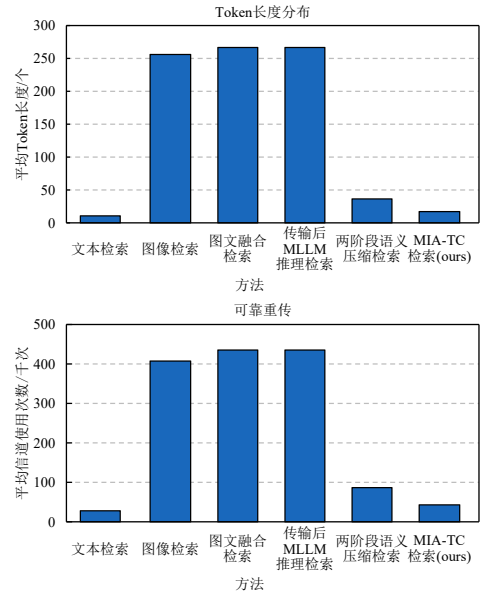


图 7 SNR 为 10 dB 时不同方法的平均信道使用次数及 Token 长度
Figure 7 Average number of channel uses and Token length of different methods at an SNR of 10 dB

索和两阶段语义压缩检索在 mAP@5 上分别达到 7.28 和 6.52,高于 MIA-TC 检索(ours)的 5.76,说明在强衰落环境中,更长的目标描述或重建后图文信息有时能够保留更多局部线索。但从通信感知指标看, MIA-TC 检索(ours)的 Channel-CARS@5 为 3.45,仍高于两阶段语义压缩检索和传输后 MLLM 推理检索。该结果说明, MIA-TC 的优势不只体现在单一检索精度上,更体现在不同信道下均能以更短 Token 载荷和更低信道资源消耗保持有效的下游任务收益,从而体现多模态意图理解与基于意图的语义压缩对可靠 Token 通信的鲁棒性价值。

3.6 推理时延分析

为分析不同方法的在线推理开销,本实验在 SNR = 5 dB 可靠重传模式下统计端到端时延,并将整体流程划分为发送端处理、通信链路、接收端处理及下游任务三个阶段,结果如表 6 所示。其中,发送端处理主要包括发送端 MLLM 或 LLM 推理、图像 Token

表 5 不同信道下 MIA-TC 的鲁棒性结果

Table 5 Robustness results of MIA-TC under different channels

方法	Rayleigh		Rician	
	mAP@5	Channel-CARS@5	mAP@5	Channel-CARS@5
文本检索	1.48	1.41	2.34	2.85
图像检索	0.47	0.04	0.91	0.10
图文融合检索	1.30	0.10	2.39	0.24
传输后 MLLM 推理检索	7.28	0.58	6.81	0.69
两阶段语义压缩检索	6.52	1.96	7.55	2.94
MIA-TC 检索(ours)	5.76	3.45	8.16	6.41

化和文本Token化;通信链路主要包括信道编码、调制、信道与自动重传请求(Automatic Repeat Request, ARQ)仿真;接收端处理及下游任务主要包括图文恢复、接收端MLLM推理、CLIP查询编码和检索。由于平均总时延还包含统一计算设备架构(Compute Unified Device Architecture, CUDA)同步、框架调度、数据搬运和函数调用等未单独列出的运行开销,且各阶段统计经过均值处理,因此三个阶段时延之和与平均总时延不完全相等。

表6 SNR为5 dB时可靠重传模式下不同方法的端到端时延

Table 6 End-to-end latency of different methods in reliable retransmission mode at an SNR of 5 dB

方法	发送端	通信链路	接收端	总时延
文本检索	0.22	0.04	4.66	5.01
图像检索	12.18	0.11	20.57	32.97
图文融合检索	12.19	0.11	24.23	36.67
传输后MLLM推理检索	11.92	0.11	567.95	580.14
两阶段语义压缩检索	1 238.34	0.06	5.03	1 243.56
MIA-TC检索(ours)	563.14	0.05	4.95	568.25

从基础直接检索方法来看,文本检索、图像检索和图文融合检索均不调用LLM或MLLM,因此平均总时延较低。然而,这类方法主要依赖原始文本、图像或简单图文组合进行检索,缺少对参考图像与操作文本之间跨模态修改意图的显式推理,因此性能较低。

在使用LLM或MLLM的方法中,两阶段语义压缩检索的主要开销集中在发送端处理阶段,其发送端处理时延达到1 238.34 ms,平均总时延为1 243.56 ms。这是因为该方法需要先生成参考图像描述,再结合操作文本生成目标描述,存在两阶段语言推理开销。相比之下,MIA-TC检索(ours)在发送端直接联合参考图像和操作文本进行一阶段多模态意图理解与压缩,发送端处理时延为563.14 ms,平均总时延为568.25 ms,较两阶段语义压缩检索降低54.30%。这说明MIA-TC能够避免两阶段描述生成带来的额外在线推理开销。

与传输后MLLM推理检索相比,MIA-TC检索的平均总时延相近但略低。二者都包含一次MLLM推理,因此总时延主要由大模型的推理时间决定;但传输后MLLM推理检索需要先对原始图像和文本进行Token化传输,再在接收端进行图文重构和MLLM推理,其额外开销主要来自图文Token化、通信链路处理以及图文重构。MIA-TC检索则将MLLM推理前移到发送端,只传输压缩后的目标意图描述,接收端仅需完成文本恢复、CLIP编码和检索。该结果表明,MIA-TC能够显著减轻接收端计算负担,并在保持多模态意图推理能力的同时降低图文传输与重构带来的额外在线开销。

3.7 消融实验

为分析意图理解与语义压缩提示词、不同多模态大语言模型对发送端目标意图描述质量的影响,选择可靠重传模式下SNR为5 dB的结果进行比较。该设置能够在信道存在明显误码风险的条件下,同时观察目标描述质量、Token载荷效率和信道资源收益的变化。

MIA-TC-Full:采用本文完整框架,默认发送端MLLM为Qwen3-VL-8B-Instruct,并使用图3所示的意图理解与语义压缩提示词。

MIA-TC-SP:去除本文设计的四阶段意图理解与语义压缩提示词,仅使用基础提示词直接要求MLLM根据参考图像和操作文本生成目标图像描述。该组用于检验结构化提示词对意图选择、冗余过滤和语义压缩的作用。

MIA-TC-LLaVA:保持通信链路与提示词不变,将发送端MLLM替换为LLaVA-1.5-7B-hf,用于评估较早视觉指令模型在多模态意图压缩任务中的适应性。

MIA-TC-InternVL:保持通信链路与提示词不变,将发送端MLLM替换为InternVL2.5-8B,用于检验不同视觉语言模型在视觉细节定位和目标描述生成上的差异。

MIA-TC-Qwen2VL:保持通信链路与提示词不变,将发送端MLLM替换为Qwen2-VL-2B-Instruct,用于分析小参数量模型在端侧或轻量部署条件下的性能和通信效率。

表7表明,完整的MIA-TC-Full相比MIA-TC-SP在mAP@5、Token-CARS@5和Channel-CARS@5上均有提升。在不使用结构化思维链提示词时,模型虽然仍能生成目标图像描述,但缺少视觉定位、意图推断和反思过滤的约束,更容易把参考图像中的无关背景或操作文本中的表面词汇带入输出,导致目标描述与真实检索意图之间存在偏差。这说明本文设计的提示词并非简单增加文本长度,而是通过显式分解多模态意图理解过程,提高了发送端语义选择的稳定性。

不同MLLM的对比进一步说明,发送端意图压缩器的多模态理解能力会直接影响下游检索效果和通

表7 消融实验结果

Table 7 Results of the ablation study

消融设置	mAP@5	Token-CARS@5	Channel-CARS@5
MIA-TC-Full	8.67	0.506	7.124
MIA-TC-SP	6.93	0.370	5.249
MIA-TC-LLaVA	3.16	0.193	2.702
MIA-TC-InternVL	8.11	0.469	6.10
MIA-TC-Qwen2VL	5.36	0.216	3.076

注:Channel-CARS列数值已放大 10^5 倍显示。

信感知效率。MIA-TC-InternVL保持了较高的检索性能和通信效率,验证了在本文框架下的有效性。相比之下,MIA-TC-LLaVA的mAP@5明显降低,表明较早的视觉指令模型在开放域CIRCO场景中容易出现细粒度视觉遗漏或修改意图理解不足。如图8所示,MIA-TC-LLaVA在第一个案例中遗漏了与people相关的视觉信息,或在第二个案例中未能保留stop sign这一关键目标约束,导致生成的目标描述偏离真实检索意图。MIA-TC-Qwen2VL的Token-CARS@5和Channel-CARS@5低于完整方法,主要与小模型的视觉理解能力、意图凝练能力较弱有关;图8中的案例也显示,Qwen2VL有时未能正确保留参考图像中的四个人物,或未能理解only one指向图像中的

自行车数量约束。上述结果说明,发送端MLLM不仅需要具备基本图像描述能力,还需要具备意图理解能力以完成稳定保留主体、属性和修改等约束,才能有效支撑后续Token通信和下游检索任务。



图8 不同发送端MLLM生成目标意图描述的典型失败案例对比

Figure 8 Typical failure case comparison of target intent descriptions generated by different transmitter-side MLLMs

4 讨论

4.1 面向6G AI原生网络的意图通信应用探讨

本文提出的MIA-TC方法虽以组合图像检索作为实验验证载体,但其核心思想并不局限于图像检索任务,而是具备扩展至6G AI原生网络中多类任务导向多模态通信场景的潜在价值。在未来智能网络中,通信对象将不再只是传统意义上的比特流或源信号,而可能进一步转向由用户目标、环境上下文和下游任务共同决定的语义意图。例如,在具身机器人协作、无

人机巡检、车路协同、智能安防和沉浸式交互等场景中,接收端往往并不需要完整恢复全部原始图像、视频或传感数据,而是需要理解“目标对象是什么”“任务应如何执行”以及“哪些状态变化与决策相关”。因此,将多模态输入压缩为面向任务的意图Token,有助于减少无关背景、纹理和冗余视觉细节的传输,从而提高有限无线资源下的任务执行效率。

从通信系统设计角度看,MIA-TC将多模态大语言模型置于发送端或边缘侧,使其在通信前完成视觉定位、意图推断和语义筛选,再将目标意图描述作为紧凑Token载荷进入传统数字通信链路。这种设计使语义信源编码不再只是对源内容进行压缩,而是根据下游任务主动选择需要传输的信息。实验结果表明,发送端意图理解能够在保持组合图像检索有效性的同时显著降低Token数量和信道使用次数,说明任务意图具有较高的通信表达密度。对于未来6G AI原生网络而言,这一特点可进一步服务于按任务、按信道和按算力联合调度的智能通信机制:当信道质量较差或终端资源受限时,系统可优先传输更短、更关键的意图Token;当任务对细粒度视觉细节要求更高时,则可适当增加描述长度或补充局部视觉Token,从而在任务性能与通信开销之间实现动态平衡。

此外,Token-CARS和Channel-CARS指标也为未来任务导向语义通信系统的资源管理提供了评价思路。传统通信指标通常关注误码率、吞吐量或重传次数,而任务导向语义通信更需要回答单位Token或单位信道资源能带来多少任务收益。本文提出的通信感知检索评价方式能够把下游任务性能与通信代价放在统一框架中分析,为多模态语义通信协议设计、边缘智能部署和语义级资源分配提供参考。未来若进一步结合多用户网络、动态信道估计和端边云协同推理,MIA-TC的意图传输思想有望扩展为一种面向AI原生网络的轻量化多模态语义交互机制。

4.2 研究局限性与未来展望

尽管本文提出的MIA-TC方法在CIRCO数据集上验证了多模态意图感知Token通信的可行性,并在Token数量、信道资源消耗和通信感知检索指标上表现出一定优势,但该方法仍存在以下局限性与待改进之处。

(1)任务与数据集的局限性:本文主要以组合图像检索作为下游任务,并基于CIRCO验证集开展实验。该任务能够较好刻画参考图像与操作文本共同表达的跨模态意图,但仍不足以覆盖6G AI原生网络中可能出现的全部多模态通信需求。例如,视频理解、视觉问答、机器人导航、协同决策和多轮交互任务中,用户意图可能具有更强的时序性、开放性和动

态变化特征。未来可进一步引入更多类型的多模态任务和数据集,验证MIA-TC在不同任务目标、不同模态组合和不同语义粒度下的通用性。

(2)通信链路建模的局限性:本文采用基础Token分组、CRC校验、信道编码、调制以及单次传输和可靠重传流程,并在AWGN、Rayleigh和Rician等信道条件下进行了初步分析。然而,真实6G网络环境中还可能多用户干扰、移动性变化、链路切换、队列时延、能耗约束和端边云协同调度等复杂因素。现有实验尚未充分考虑这些系统级因素对意图Token传输的影响。未来可进一步构建更接近真实网络的仿真平台,将语义压缩、信道编码、重传策略、资源块分配和边缘计算开销进行联合优化,以验证MIA-TC在复杂无线网络中的实际部署潜力。

总体而言,本文工作主要验证了先理解多模态意图,再传输紧凑Token这一思路在任务导向语义通信中的有效性。当前原型主要聚焦发送端意图语义信源编码与基础数字通信链路,尚未将网络侧AI Agent作为额外的意图解析或推理模块。面向未来6G AI原生网络,网络侧AI Agent可进一步结合目标意图Token及其任务类型、时延要求、可靠性需求和错误容忍度等信息,参与业务识别、任务编排、传输策略配置和边缘服务选择,从而实现更加深入的端网协同。后续还可围绕更复杂任务、更真实信道、更轻量化模型和更全面评价指标继续拓展,使多模态意图感知Token通信从初步实验范式进一步发展为可部署、可扩展、可优化的6G AI原生语义通信机制。

5 结束语

本文围绕6G AI原生语义通信中多模态信息,提出了多模态意图感知Token通信方法MIA-TC。首先,将通信对象由原始图像、操作文本或高维多模态特征进一步抽象为由多模态输入共同表达的任务意图,明确了多模态意图作为语义通信对象的研究价值;其次,将发送端意图理解与语义压缩引入Token通信前端,使通信系统能够优先传输与下游任务直接相关的目标意图Token,而不是传输大量与任务弱相关的源内容;最后,提出Token-CARS和Channel-CARS两类通信感知评价指标,从Token预算和信道资源消耗两个角度统一衡量任务性能与通信代价,为任务导向语义通信提供了更全面的评价方式。基于CIRCO验证集的实验结果表明,MIA-TC在显著减少平均Token数量和信道使用次数的同时,仍能保持有效的组合图像检索性能,并在Token-CARS和Channel-CARS指标上优于原始图文Token传输方案和两阶段语义压缩方法。上述结果说明,在发送端进行多模态意图理解与压缩

能够有效提升单位通信资源的任务收益,也验证了多模态意图传输在6G AI原生语义通信中的可行性。未来将进一步面向更多下游任务、复杂信道环境和端边协同部署场景,探索结构化意图Token、重要性保护和低成本意图压缩模型等方向。

参考文献

- [1] Shannon C E. A mathematical theory of communication[J]. Bell System Technical Journal, 1948, 27(4): 623-656.
- [2] Bao Jie, Basu P, Dean M, et al. Towards a theory of semantic communication[C]//2011 IEEE Network Science Workshop. Piscataway: IEEE, 2011: 110-117.
- [3] Gündüz D, Qin Zhijin, Aguerri I E, et al. Beyond transmitting bits: Context, semantics, and task-oriented communications[J]. IEEE Journal on Selected Areas in Communications, 2023, 41(1): 5-41.
- [4] Yang Wanting, Du Hongyang, Zi Qin Liew, et al. Semantic communications for future Internet: Fundamentals, applications, and challenges[J]. IEEE Communications Surveys & Tutorials, 2023, 25(1): 213-250.
- [5] Xie Huiqiang, Qin Zhijin, Li Geoffrey Ye, et al. Deep learning enabled semantic communication systems[J]. IEEE Transactions on Signal Processing, 2021, 69: 2663-2675.
- [6] Boursoulatz E, Burth Kurka D, Gündüz D. Deep joint source-channel coding for wireless image transmission[J]. IEEE Transactions on Cognitive Communications and Networking, 2019, 5(3): 567-579.
- [7] Qiao Li, Mashhadi M B, Gao Zhen, et al. Token communications: A large model-driven framework for cross-modal context-aware semantic communications[J]. IEEE Wireless Communications, 2025, 32(5): 80-88.
- [8] Liang Chengsi, Du Hongyang, Sun Yao, et al. Generative AI-driven semantic communication networks: Architecture, technologies, and applications[J]. IEEE Transactions on Cognitive Communications and Networking, 2025, 11(1): 27-47.
- [9] Qiao Li, Mashhadi M B, Gao Zhen, et al. Latency-aware generative semantic communications with pre-trained diffusion models[J]. IEEE Wireless Communications Letters, 2024, 13(10): 2652-2656.
- [10] Delétang G, Ruoss A, Duquenne P A, et al. Language modeling is compression[C]//Proceedings of the 12th International Conference on Learning Representations. 2024. https://proceedings.iclr.cc/paper_files/paper/2024/

- hash/3cbf627fa24fb6cb576e04e689b9428b-Abstract-Conference.html.
- [11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 2017: 5998-6008.
 - [12] Devlin J, Chang Ming-Wei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Stroudsburg: ACL, 2019: 4171-4186.
 - [13] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems 33. Curran Associates, Inc., 2020: 1877-1901. <https://proceedings.mlr.press/v139/radford21a>.
 - [14] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021, 139: 8748-8763. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
 - [15] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]//Proceedings of the 9th International Conference on Learning Representations. 2021. <https://openreview.net/forum?id=YicbFdNTTy>.
 - [16] Liu Haotian, Li Chunyuan, Wu Qingyang, et al. Visual instruction tuning[C]//Advances in Neural Information Processing Systems 36. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 34892-34916.
 - [17] OpenAI, Achiam J, Adler S, et al. GPT-4 technical report[PP/OL]. V6. arXiv (2024-03-04) [2026-07-13]. <https://doi.org/10.48550/arXiv.2303.08774>.
 - [18] Bai Shuai, Cai Yuxuan, Chen Ruizhe, et al. Qwen3-VL technical report[PP/OL]. V2. arXiv (2025-11-27) [2026-07-13]. <https://doi.org/10.48550/arXiv.2511.21631>.
 - [19] Wei Jason, Wang Xuezhi, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Advances in Neural Information Processing Systems 35. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 24824-24837.
 - [20] Kojima T, Gu Shixiang, Reid M, et al. Large language models are zero-shot reasoners[C]//Advances in Neural Information Processing Systems 35. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 22199-22213.
 - [21] Shinn N, Cassano F, Gopinath A, et al. Reflexion: Language agents with verbal reinforcement learning[C]//Advances in Neural Information Processing Systems 36. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 8634-8652.
 - [22] Vo N, Jiang Lu, Sun Chen, et al. Composing text and image for image retrieval - an empirical odyssey[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 6432-6441.
 - [23] Tang Yuanmin, Zhang Jue, Qin Xiaoting, et al. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 14400-14410.
 - [24] Wu Hui, Gao Yupeng, Guo Xiaoxiao, et al. Fashion IQ: A new dataset towards retrieving images by natural language feedback[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 11302-11312.
 - [25] Liu Zheyuan, Rodriguez-Opazo C, Teney D, et al. Image retrieval on real-life images with pre-trained vision-and-language models[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 2105-2114.
 - [26] Baldrati A, Bertini M, Uricchio T, et al. Effective conditioned and composed image retrieval combining CLIP-based features[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 21434-21442.
 - [27] Saito K, Sohn K, Zhang Xiang, et al. Pic2Word: Mapping pictures to words for zero-shot composed image retrieval[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 19305-19314.
 - [28] Baldrati A, Agnolucci L, Bertini M, et al. Zero-shot composed image retrieval with textual inversion[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 15292-15301.
 - [29] Karthik S, Roth K, Mancini M, et al. Vision-by-language for training-free compositional image retrieval[C]//Proceedings of the 12th International Conference on Learning Representations. 2024.
 - [30] Sennrich R, Haddow B, Birch A. Neural machine

- translation of rare words with subword units[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2016: 1715-1725.
- [31] Schuster M, Nakajima K. Japanese and Korean voice search[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2012: 5149-5152.
- [32] Kudo T. Subword regularization: Improving neural network translation models with multiple subword candidates[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 66-75.
- [33] Van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6309-6318.
- [34] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 12868-12878.
- [35] Lin Tsung-Yi, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[M]//Computer Vision - ECCV 2014. ChamSpringer International Publishing2014: 740-755.
- [36] Proakis J G, Salehi M. Digital communications[M]. 5th ed. Boston: McGraw-Hill, 2008.
- [37] Forney G D. The viterbi algorithm[J]. Proceedings of the IEEE, 1973, 61(3): 268-278.

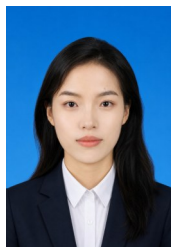
作者简介



于 静 女,1989年3月出生于河北省承德市。现为中央民族大学人工智能研究院、信息工程学院教授,博士生导师。长期从事内容安全治理、人工智能安全等方向的研究。发表学术论文100余篇,获得国际会议最佳论文奖5项。
E-mail: jing.yu@muc.edu.cn



茹志鹏 男,2003年6月出生于浙江省绍兴市。现为中央民族大学信息工程学院硕士研究生。主要研究方向为组合图像检索、多模态大模型等。
E-mail: ruzhipeng@muc.edu.cn



王路娟 女,2000年6月出生于河北省邢台市。现为中央民族大学信息工程学院硕士研究生。主要研究方向为大语言模型水印、AIGC内容安全等。
E-mail: lujuan_wang@muc.edu.cn



徐 浩 男,2001年1月出生于云南省曲靖市。现为中央民族大学信息工程学院硕士研究生。主要研究方向为多模态大模型微调、知识编辑等。
E-mail: hao_xu@muc.edu.cn



张翰林 男,2002年11月出生于北京市。现为中央民族大学信息工程学院硕士研究生。主要研究方向为大语言模型后训练、指令数据优化、大模型评测与安全对齐。
E-mail: hanlin_zhang@muc.edu.cn