XXXX

边缘智能计算的协同推理关键技术研究

张智涵[1]，张天魁[1]，党梅梅[2]，程伟强[3]

（1. 北京邮电大学信息与通信工程学院，北京 100876；
2. 中国信息通信研究院技术与标准研究所，北京 100191；
3. 中国移动通信有限公司研究院，北京 100053）

摘　要：本文剖析了边缘智能计算技术在边缘计算网络中的实现逻辑，并围绕边缘协同推理任务的推理时延与推理质量需求，分别从判别式与生成式人工智能模型的特性出发，系统分析了边缘协同推理关键技术的实现方法。在此基础上，提出了一种面向生成式人工智能模型的边端协同推理系统设计实例，验证了其在推理时延与推理质量方面的优势，为边缘智能计算在边缘计算网络中高效部署与应用实现提供了可行的技术方案。

关键词：人工智能；智能计算；边缘计算网络；协同推理

中图分类号：TP393

文献标志码：A

doi: 10.11959/j.issn.1000−0801.

Key Technologies for Collaborative Inference in Edge Intelligence Computing

ZHANG Zhihan[1], ZHANG Tiankui[1], DANG Meimei[2], CHENG Weiqiang[3]

1. School of Information and Communication Engineering, Beijing University of Posts and Telecommunication, Beijing 100876, China
2. China Technology and Standards Research Institute, China Academy of Information and Communications Technology, Beijing 100191, China
3. The Research Institution of China Mobile, Beijing 100053, China

Abstract: This paper provided an in-depth analysis of the implementation logic of edge collaborative intelligent computing technologies in edge computing networks. Focusing on the latency and inference quality requirements of edge collaborative inference tasks, it systematically analyzed the characteristics of discriminative and generative artificial intelligence models, respectively, and summarized the core implementation approaches of key technologies for edge collaborative inference. On this basis, a device‐edge collaborative inference system design tailored for generative artificial intelligence models was proposed, and its advantages in terms of inference latency and inference quality were

收稿日期：2026−02−XX；**修回日期：**XXXX−XX−XX

通信作者：张天魁，zhangtiankui@bupt.edu.cn

基金项目：国家自然科学基金资助项目（No.62371068），鹏城实验室科教基金会-中国移动科创基金项目（SQ2510300009）

Foundation Items: The National Natural Science Foundation of China (No.62371068), Education and Research Foundation of PCL - Technological Innovation Foundation of China Mobile (SQ2510300009)

validated. The proposed approach provided a technical reference for the efficient deployment and practical implementation of edge collaborative intelligent computing technologies.

Key words: artificial intelligence, intelligent computing, edge computing networks, collaborative inference

1 引言

当前，在国家战略的持续引领下，我国在人工智能（AI，Artificial Intelligence）领域已跻身国际领先行列，有力推动了制造业、金融、农业和物流等行业的智能化转型升级[1]。“‘十五五’规划建议”明确提出要抢占AI产业应用制高点，推动人工智能技术与实体经济深度融合。实现上述目标的关键在于充分释放AI技术的能力优势，并构建与之相匹配的智能计算技术支撑体系。智能计算技术是融合AI算法、高性能计算与大数据处理，实现自主学习、自适应推理与智能决策的新型计算技术。显然，AI技术赋能千行百业高度依赖智能计算技术的服务水平，既需要大规模数据的积累也需要高性能计算资源的支撑，因此AI技术落地与规模化应用需要高效、可靠的算力基础设施。智能计算的性能表现不仅取决于算法设计本身，还受到计算资源供给、数据可用性、时延约束以及能效需求等多重因素的共同制约[2]。传统云中心化架构依赖将大量原始数据远程传输至云端进行集中处理，容易引发高时延和网络拥塞等问题，从而在实时性要求高、部署规模大的AI技术应用场景中逐渐暴露出性能瓶颈，限制了智能计算的服务质量和用户满意度。

为突破上述制约，边缘计算网络作为承载智能计算的关键基础设施，成为推动AI技术应用落地的关键路径。边缘计算网络通过在网络边缘侧提供计算、存储与数据处理能力，使智能计算能够在靠近数据源的位置本地执行；同时，边缘计算网络可支撑多节点协同与实时数据交互，实现AI模型在网络边缘的分布式部署[3]。通过终端与边缘节点之间的协同计算，能够进一步释放智能计算的潜力，推动内生智能与通算智能的一体化发展。

2 相关工作

近年来，围绕边缘智能与协同推理已有一定研究。一类工作从智能计算演进与边缘网络优化层面讨论边缘智能的部署与推理。文献[1]总结了感知、理解与生成等任务在算力组织和系统支撑方面的演进趋势。文献[4]等针对大规模多接入边缘计算场景，提出了结合部分可观测建模、长短期记忆网络与平均场近似理论的任务卸载算法；文献[5]面向无人机辅助边缘计算系统，将计算卸载与轨迹规划纳入多指标意图驱动框架进行联合优化。这些研究表明，资源分配与卸载决策是边缘智能的重要基础，但其关注对象多为任务或业务流层面的调度问题，对于模型不同推理阶段如何匹配异构节点资源，仍缺乏更细致的分析。

另一类工作关注边缘智能计算中的协同机制与体系重构。文献[6]针对工业智能体互联网提出“通信—控制”协同原子化重构机制，从系统架构层面讨论了动态协同关系与拓扑重构问题。文献[7]从云边端协同计算的总体架构与编排机制角度对协同关系进行了系统归纳。这类研究对于理解边缘环境中的多主体协作、任务编排和跨层协同具有启发意义，但其研究场景主要面向工业网络控制，尚未进一步区分判别式人工智能（DAI，Discriminative AI）与生成式人工智能（GAI，Generative AI）在推理目标、时延约束、计算方式和协同粒度上的差异。

综上，现有研究已分别从智能计算演进、资源优化和协同机制等角度为边缘智能研究提供了基础，但尚缺少一个能够在统一框架下同时讨论

DAI 模内协同与 GAI 模间协同的分析视角。基于此，本文将 DAI 与 GAI 纳入同一边缘协同推理框架进行梳理，并结合“边缘前序推理+终端解码推理”的 GAI 边端协同设计实例，分析不同推理范式在时延、推理质量与协同粒度上的权衡关系。

3　边缘智能计算技术概述

为了让边缘计算网络有效支撑智能计算，边缘协同智能计算技术从 AI 模型结构优化至网络协同层面展开了系统性的探索。在 AI 模型结构优化层面，为应对边缘设备资源受限和异构计算环境的挑战，不仅需要通过模型压缩降低参数量，还应对模型结构进行适配和调整，相关技术包括模型剪枝、量化、结构重参数化以及模块化设计等[8]。在网络协同层面，如图 1 所示，边缘计算网络由终端节点与边缘节点构成，通过节点协同将计算任务拆解分配至不同计算单元。根据协同主体与层次，边缘协同分为三种方式：边侧协同（边缘节点间）、边端协同（终端与边缘节点间）及端侧协同（终端设备间）。

在上述协同机制支撑下，边缘协同技术包括边缘协同训练与边缘协同推理两类。其中，边缘协同训练通过跨节点的参数、特征或预测信息协同更新，在异构数据与设备环境中实现模型联合优化；边缘协同推理技术则通过任务拆解与节点协同，实现边缘资源受限环境下的高效推理。

4　边缘协同推理关键技术

根据人工智能模型再推理阶段的任务目标和输出形式差异，推理任务通常被划分为 DAI 与 GAI 两类。DAI 侧重于学习输入与输出之间的判别关系，用于完成分类、回归等预测任务[9]；GAI 则通过建模数据的潜在分布，实现对新样本或内容的生成[10]。边缘协同推理技术突破了传统推理任务“单点部署、独立执行”的范式，依据不同 AI 模型特性实现推理过程的分布式协同执行。针对 DAI 模型的结构特性，边缘协同推理通过模型模块化拆分实现模内协同，单个 AI 模型的推理任务可以分为多个模块在多个节点间的高效分担；而针对 GAI 模型，边缘协同推理则采用模间协同的方式，在不同节点间实现多个 AI 模型分

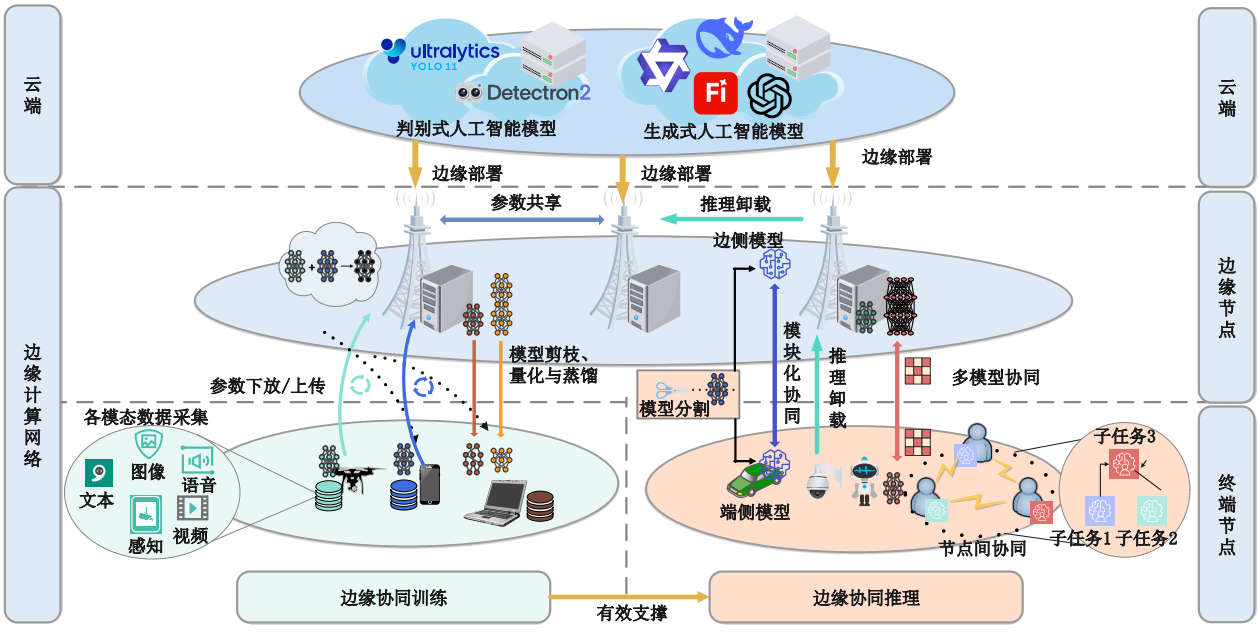

图 1　面向智能计算的边缘计算网络架构

阶段的推理任务处理。

4.1 面向判别式人工智能的协同推理技术

DAI推理任务对推理完成的端到端时延有着严格要求，例如工业质检中的缺陷识别任务、自动驾驶任务与实时视频处理均需要毫秒级的快速响应。因此，面向判别式人工智能的协同推理通过将深度神经网络拆分为可调度的独立模块，实现模型内部的模内协同推理，并在终端与边缘节点分布式部署，充分利用各节点的通算资源。如图2所示，利用拆分后模型的推理特性，通过模型分割、信息压缩与推理提前退出等技术，面向DAI推理的模间协同技术能够降低计算负载和通信开销，同时保障判别精度与整体响应速度。

模型分割技术通过将完整深度神经网络按层间或层内划分为多个子模块或计算块，并在不同终端与边缘节点上分段执行前向传播[11]。层间分割沿网络层级划分模型，每个节点仅承担部分前向计算，可根据节点算力、负载和通信条件动态调整分割位置。例如，细粒度层间分割通过多个分割点将网络拆分为连续子模型片段，例如将编码器、解码器或注意力模块分布在不同节点执行，中间特征通过网络传输到下一节点，充分利用分布式算力，降低端到端延迟和能耗。层内分割则针对单层计算结构拆分任务，如将卷积层输入特征图按通道或空间块划分，在多个节点并行完成卷积运算，再通过聚合或拼接生成该层输出，能够显著提升计算并行度以实现推理加速。

信息压缩技术通过缩减节点间传输的特征信息量来降低通信开销。对于模内协同而言，跨节点传输对象通常并非原始输入，而是中间特征表示，特征张量的冗余度成为通信开销的关键。信息压缩技术可分为语义压缩和特征维度压缩两类。语义压缩通过筛选并优先传输关键特征通道，利用注意力机制或互信息度量识别高信息量特征，在尽量保持判别精度的前提下降低传输数据量[12]；特征维度压缩则借助主成分分析、自动编码器降维或标量量化等手段减少特征数目，实现中间表示的轻量化传输[13]。

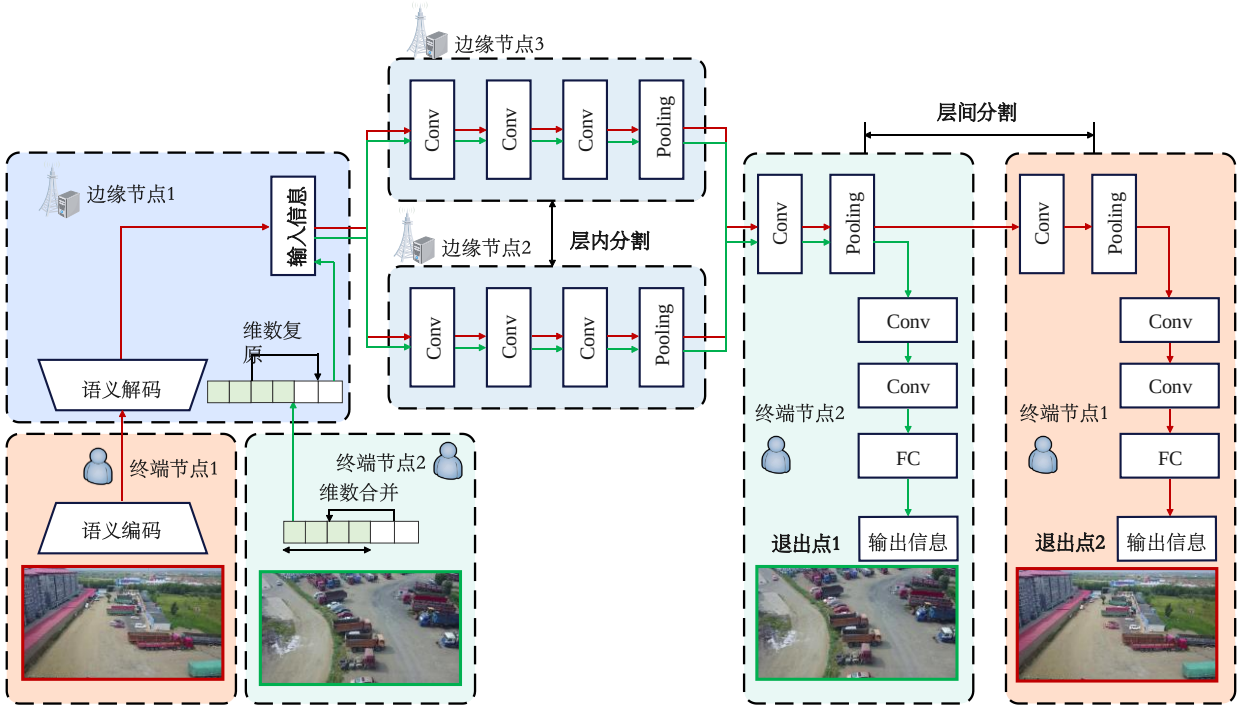

图2 面向判别式人工智能的协同推理技术（模内推理）

提前退出技术通过在深度神经网络的中间层设置可选输出节点，使输入样本在达到特定置信度或判定条件时即可完成推理，从而减少端到端延迟[14]。该机制允许边缘节点根据输入特征自适应决定推理深度，实现计算资源的高效利用，同时在保证判别准确性的前提下提升系统整体响应速度。该机制尤其适用于存在明确时效窗口的判别式任务，因为其可在易样本上提前终止后续计算，将有限算力优先分配给难样本或后续到达任务，从而降低排队积压。

4.2　面向生成式人工智能的协同推理技术

GAI 模型强大的内容生成与推理能力依赖于长上下文建模与自回归生成，在边缘场景下面临着算力受限的问题。以 Stable Diffusion 为例，其完成一次高质量图像生成推理通常需要约 10 TFLOPs 量级的浮点运算量，并产生约 10 GB 的显存开销。为支撑不同 GAI 推理任务在边缘侧的高效执行，面向 GAI 的协同推理技术通过在边缘计算网络中部署多个生成式模型，实现模间协同推理，使推理任务能够在模型层面协同完成。根据任务目标与协同方式的差异，该类协同推理技术可进一步划分为单主体模间推理与多主体模间推理。

单主体模间推理技术通过在异构算力设备上部署多个 GAI 模型协同推理，实现对同一任务的多阶段处理。该技术通过模型间的低维中间信息传输，替代完整输入或输出的直接传输，从而有效降低通信开销[15]。图 3 以边端协同方式为例，展示了单主体模间协同推理技术在该节点协同方式下的三种推理策略，即端侧推理策略、边侧推理策略和协同推理策略。

端侧推理策略中，终端节点负责请求解析并生成任务种子，边缘节点根据种子完成内容生成，最终结果再下发至终端节点。由于上行链路仅需传输种子信息，该策略能够显著降低上行带宽消耗，适用于对上传负载敏感的生成任务场景。与其相反，边侧推理策略则将推理过程部署于边缘节点，由边缘节点完成任务解析与种子生成，并将种子返回终端节点以执行后续生成过程。该策略能够有效减少下行链路中生成内容的传输开销，适用于下行数据量较大的生成任务。协同推理策略进一步强化端边协同，仅在端边之间传输种子信息。终端节点生成任务种子，边缘节点完成从任务种子到生成种子的映射，终端节点依据返回的生成种子完成最终内容生成。该策略在最大程度节省通信资源的同时，利用种子信息的鲁棒性，有利于保障生成质量的稳定性。

与单主体模间推理侧重于同一目标下的模间协同推理不同，多主体推理强调模型间的功能分工，各模型作为独立主体承担不同角色职责，并

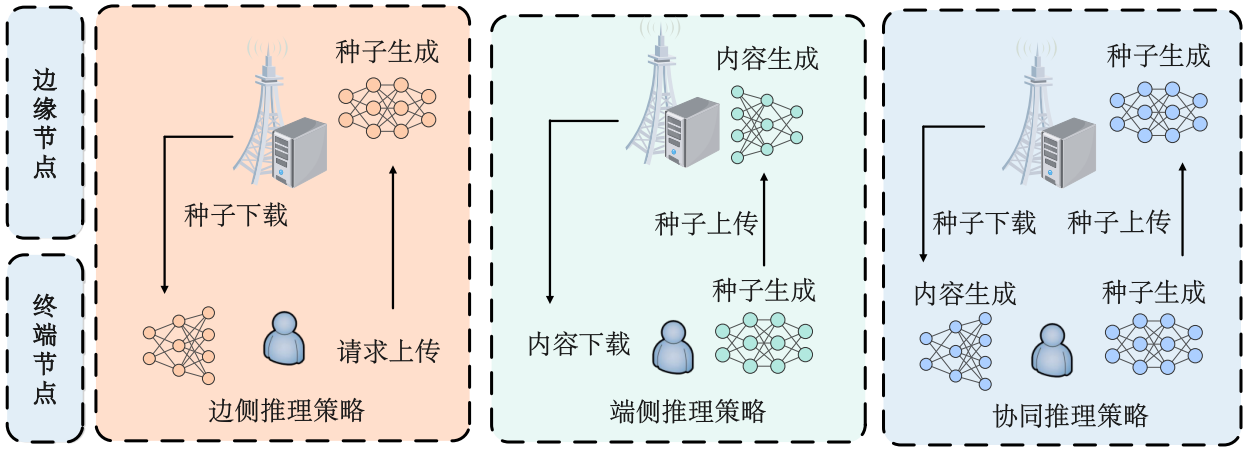

图3　面向生成式人工智能推理的单主体模间推理技术

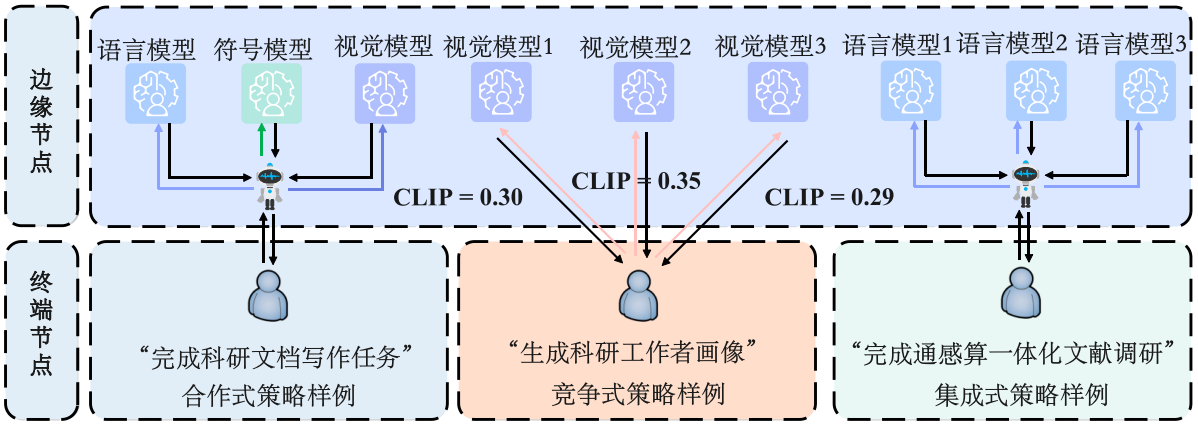

图4　面向生成式人工智能推理的多主体模间推理技术

通过显式或隐式的信息交互实现整体推理能力的提升。根据协同方式的差异，多主体模间推理技术可分为合作式、竞争式与集成式多主体策略。图4以边端协同方式为例，展示了这三类策略的推理任务样例。

在合作式多主体策略中，多个GAI模型围绕统一目标开展协同推理，系统通过任务拆分或角色分配，将复杂问题划分为若干子任务，由不同主体分别负责规划、分析或评估，并通过迭代反馈不断细化推理结果，适用于结构复杂的推理场景。竞争式多主体策略引入模型间的对抗或博弈机制，不同主体基于各自假设独立生成推理结果，系统通过比较或裁决筛选最优输出，以竞争驱动模型发挥更高推理性能，适合对结果严谨性要求较高的任务。集成式多主体策略则采用并行推理，各GAI模型独立完成推理，由系统依据预设规则对输出进行综合决策，从而降低单模型误差对最终结果的影响[16]。

在边缘网络资源受限与链路时变条件下，多主体模间推理的核心挑战在于协同过程的可收敛性、可控开销与可验证性：过多交互轮次会使通信与重复推理开销抵消质量增益；消息传递需在语义充分性与带宽占用间取得平衡，过度摘要或冗余均会损害推理质量；此外，缺乏显式真值参照时，多主体输出的一致性评估与裁决本身也是难点。为此，工程实现通常限定共享上下文、消息格式与轮次上限，引入结构化中间表示与一致性验证机制对多路输出进行筛选，并仅在不确定或冲突时触发复核，以在保证协同收益的前提下将交互开销与时延控制在可接受范围内。

5　边端协同推理系统设计实例

5.1　面向生成式人工智能推理的模间协同技术

为克服文本-图像生成式推理在边缘侧响应延迟长、生成质量受限的问题，如图5所示，本文构建了一种面向GAI模型边缘推理的边端协同推理系统，采用“边缘前序推理+终端解码推理”协同范式，依据计算复杂度与节点资源的适配原则进行任务划分：前序扩散阶段计算密集、显存需求高，部署于边缘节点；解码阶段复杂度较低，适于在终端侧完成局部特征重建。该范式适用于移动边缘计算（MEC，Mobile Edge Computing）网络中的实时图像生成、6G超密集网络中的分布式AI推理及边缘辅助渲染等场景，能够在保障生成质量的同时有效降低无线信道传输开销与端到端时延。具体推理流程如下：

（1）前序GAI模型推理流程

给定输入条件y（如文本、语义图），前序GAI模型首先通过特征编码器$\mathcal{E}$将其转化为条件上下文向量$\mathbf{c}=\mathcal{E}_{\phi^{\text{enc}}}(y)$，其中$\phi^{\text{enc}}$为特征编码器模

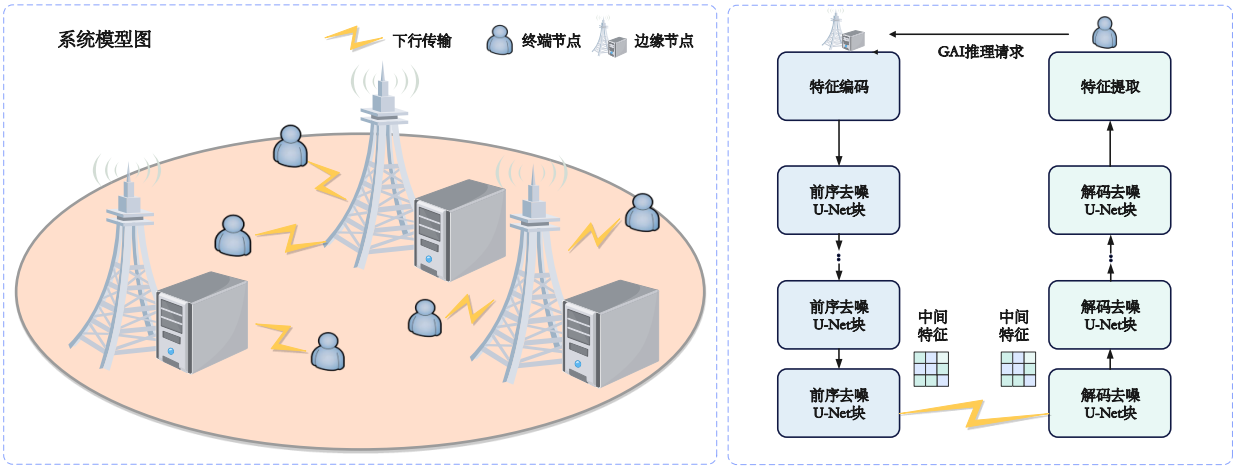

图5　所提出的面向文本-图像GAI模型边缘推理的单主体多模型推理系统

型参数，$\mathbf{c}\in R^{L\times d_p}$，$L$表示序列长度，$d_p$表示每一个分词被映射到的高维特征向量长度。在前序GAI模型的隐空间扩散中，设时间步的下标为i，记$\mathbf{z}_i$为第i个时间步的潜空间图像，$\boldsymbol{\epsilon}_i^{\text{prior}}$为第$i$个时间步对潜空间图像噪声的估计值，其反向去噪过程可表示为$\boldsymbol{\epsilon}_i^{\text{prior}}=f_{\phi^{\text{UNet}}}(\mathbf{z}_i,i,\mathbf{c})$，其中$\phi^{\text{UNet}}$表示前序去噪U-Net块的模型参数。以去噪扩散概率模型采样算法为例[17]，潜空间图像的步进更新公式满足

$$\mathbf{z}_{i-1}=\frac{1}{\sqrt{\alpha_i}}\left(\mathbf{z}_i-\frac{1-\alpha_i}{\sqrt{1-\bar{\alpha}_i}}\boldsymbol{\epsilon}_i^{\text{prior}}\right)+\gamma_i\mathbf{r},\quad \mathbf{r}\sim\mathcal{N}(0,\mathbf{I}) \tag{1}$$

其中系数α_i随步数i变化，表示在第i步推理中，保留原始信号的比例。$\bar{\alpha}_i=\prod_{k=1}^{I}\alpha_k$为系数$\alpha_k$从第1步到第$i$步的连乘积。$\gamma_i\mathbf{r}$为随机噪声扰动，$\gamma_i$为扰动的标准差。前序GAI模型的前序去噪U-Net块在$L_{\max}^{\text{prior}}$步内生成中间语义特征$\mathbf{z}^{\text{prior}}$，下行传输至终端节点的解码GAI模型。

（2）下行传输阶段

记J^{dl}为边缘节点传输至终端节点的中间信息比特数，边缘节点到终端节点在时隙t的传输速率记为$R(t)=B^{\text{dl}}\log_2\left(1+\frac{P^{\text{dl}}|h(t)|^2}{\sigma^2}\right)$，其中，$B^{\text{dl}}$为下行带宽，$P^{\text{dl}}$为下行传输功率，$h(t)=\sqrt{10^{-\frac{\text{PL}}{10}}}\hat{h}(t)$为信道增益，其中PL表示大尺度衰落，$\hat{h}(t)$表示时隙$t$的小尺度衰落，服从复高斯分布，$\sigma^2$为接收机噪声功率。可以得到中间信息的期望下行传输时延满足$\mathbb{E}[T^{\text{dl}}]=\frac{J^{\text{dl}}}{\mathbb{E}[R(t)]}$。

（3）解码GAI模型推理阶段

记解码GAI模型隐空间扩散中的时间步的下标为j。不同于前序GAI模型的前序去噪U-Net块，解码GAI模型的解码去噪U-Net块同时接受文本嵌入向量$\mathbf{c}$与中间语义特征$\mathbf{z}^{\text{prior}}$的双重引导，每一步$j$的预测噪声的表达式为

$$\boldsymbol{\epsilon}_j^{\text{dec}}=g_{\varphi^{\text{UNet}}}(\mathbf{z}_j,j,\mathbf{0})+\text{GS}\cdot(g_{\varphi^{\text{UNet}}}(\mathbf{z}_j,j,\{\mathbf{c},\mathbf{z}^{\text{prior}}\})-g_{\varphi^{\text{UNet}}}(\mathbf{z}_j,j,\mathbf{0})) \tag{2}$$

其中φ^{UNet}为解码去噪U-Net块的模型参数，GS为引导系数，其数值越大，对模型生成泛化内容的倾向惩罚性越强。$\{\mathbf{c},\mathbf{z}^{\text{prior}}\}$表示通过注意力机制或拼接进行融合文本嵌入向量与中间语义特征的联合条件。在$L_{\max}^{\text{dec}}$步内完成解码去噪推理后，解码器$\mathcal{D}$将最终的潜空间向量$\mathbf{z}^{\text{final}}$映射至全尺寸图像$\mathbf{x}^{\text{final}}=\mathcal{D}_{\varphi^{\text{dec}}}(z^{\text{final}})$，其中$\varphi^{\text{dec}}$为解码器的模型参数，$\mathbf{x}^{\text{final}}\in R^{H\times W\times 3}$，$H$与$W$分别为图像的高度和宽度。

（4）边端协同信令开销分析

边端协同推理架构涉及两类信令交互。系统原有信令包括接入信令、调度请求、功率控制与切换等，是移动通信系统的固有开销[3]。协同推理信令为本文架构额外引入，包括：任务请求（TR, Task Request）由终端发往边缘节点；模型步数分配由边缘节点下发，告知前序与解码阶段的步数分配 L^{prior} 与 L^{dec}，这是本架构的内禀属性；资源状态反馈（RS，Resource Status）由终端上报本地算力与内存状态。由于RS与步数分配信令均为轻量字段（队列与算力状态、两个步数参数），其额外控制开销相对数据载荷可忽略。

为便于说明本文架构中多主体的信令交互关系，图6给出了典型同址部署，即边缘用户面功能（UPF，User Plane Function）与MEC服务器平台同址、本地分流的信令示意图。

典型配置下，协同信令开销可近似估计为1-2 KB，远小于数十KB至数MB的数据载荷。高并发时可预留约10%–15%的控制信道资源以保证信令传输可靠性。若进一步将控制开销折算为控制资源占用系数 ξ_{ctrl}，则系统可用于数据承载的有效通信能力可近似写为 $R_{\text{eff}}(t)=(1-\xi_{\text{ctrl}})R(t)$，从而更加清晰地体现额外控制开销对有效通信容量的压缩作用。

5.2 模型推理质量评估

为避免单一指标对文生图质量的刻画偏差，本文进一步引入FID、KID、CLIP与LPIPS四类指标共同用于刻画生成质量。表2给出了上述四类指标的公式定义与侧重维度，其中 $f(\cdot)$ 表示CLIP特征提取器的输出值[18]，μ_r,Σ_r 与 μ_g,Σ_g 分别为真实与生成图像特征的均值与协方差，$\mathcal{F}_r,\mathcal{F}_g$ 表示真实/生成图像的特征集合，$\text{tr}(\cdot)$ 表示矩阵迹，$\text{MMD}^2(\cdot,\cdot)$ 表示平方最大均值差异（MMD，Maximum Mean Discrepancy），$\phi_l(\cdot)$ 为第 l 层深度特征映射，w_l 为对应层的通道权重，H_l,W_l 为该层特征图的高与宽，$\odot$ 表示Hadamard积，$\|\cdot\|_2$ 为 ℓ_2 范数，$\cos(\cdot,\cdot)$ 为余弦相似度，f_I,f_T 表示归一化的图像/文本嵌入向量。

本文采用Cascaded Diffusion模型[19]的前两个低维扩散网络模型作为前序GAI模型，并将最后一个高维扩散阶段作为解码GAI模型。图7展示了各输出指标随前序去噪步数与解码去噪步数的变化，前序去噪步数 L^{prior} 的增加对输出指标的提升显著高于解码去噪步数 L^{dec} 的增益。这种现象的物理本质在于，前序阶段负责在低维潜空间中确立图像的核心语义骨架与特征分布，一旦 $\mathbf{z}^{\text{prior}}$ 在该阶段未能实现与文本嵌入向量 $\mathbf{c}$ 的有效对齐，后续增加解码阶段的去噪步数 L^{dec}，也仅

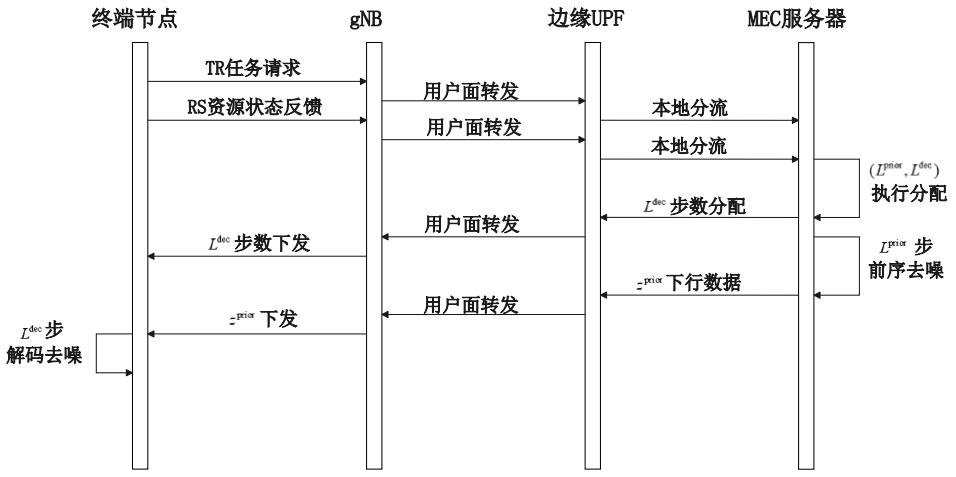

图6 典型同址部署的信令示意图

表 1　生成质量指标定义与物理意义

指标	数学定义	物理意义
FID	$\|\mu_r-\mu_g\|^2+\mathrm{tr}(\Sigma_r+\Sigma_g-2(\Sigma_r\Sigma_g)^{1/2})$	衡量真实与生成图像在特征空间分布的距离；数值越小表示生成分布越接近真实分布。
KID	$\mathrm{MMD}^2(\mathcal{F}_r,\mathcal{F}_g)$	基于平方最大均值差异的分布差异度量；数值越小表示两者分布差异越小，且具有无偏估计特性。
CLIP	$\max(100\cdot\cos(f_I,f_T),0)$	衡量文本条件与生成图像在CLIP嵌入空间特征提取器映射后的语义对齐程度；数值越大表示对齐更强。
LPIPS	$\sum_l\frac{1}{H_lW_l}\sum_{h,w}\|w_l(\phi_l(x)_{hw}-\phi_l()x'_{hw})\|_2^2$	基于深度特征的感知距离，用于有参考图像场景下刻画感知相似性；数值越小表示感知差异越小。

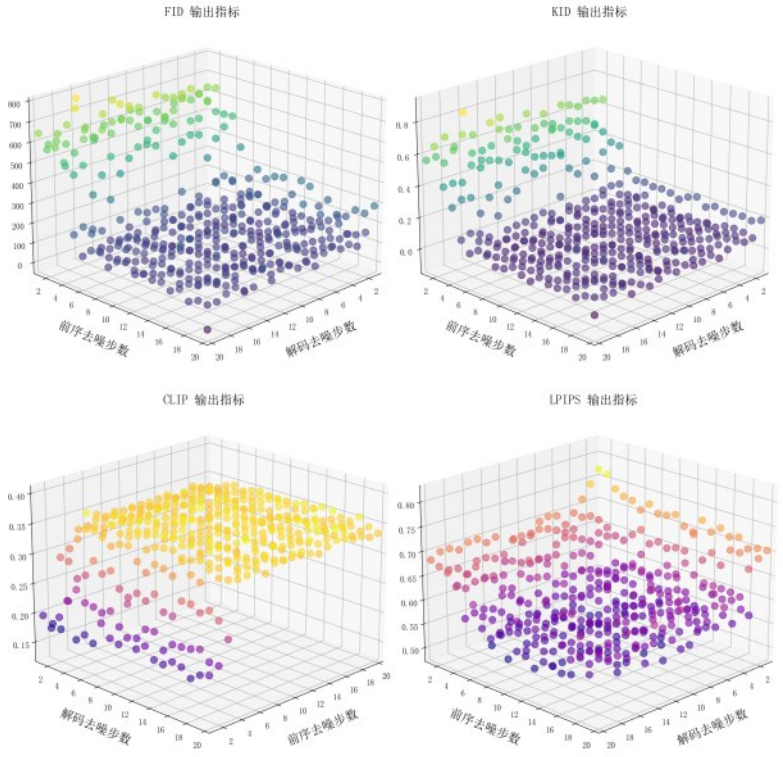

图7　平均CLIP输出指标随前序与解码去噪步数的变化

能在错误的语义基础上徒增纹理细节，而无法挽回全局语义映射的偏差[20]。因此，在所提出的推理架构中，将前序GAI模型部署在算力资源更为充足的边缘节点，能够更有效地保障生成的准确度。

5.3　系统逗留时延期望与时延违反概率上界分析

（1）　系统逗留时延期望分析

记λ为推理任务的到达速率，推理任务到达服从泊松分布。记边缘节点与终端节点的每秒浮点运算次数分别为f^{ES}与f^{UE}，特征编码与特征提取所需的浮点运算次数分别为O^{E}与O^{D}，前序去噪U-Net块与解码去噪U-Net块进行单步去噪的浮点运算次数分别为$O_{\mathrm{UNet}}^{\mathrm{prior}}$与$O_{\mathrm{UNet}}^{\mathrm{dec}}$，前序去噪步数与解码去噪步数分别为$L^{\mathrm{prior}}$与$L^{\mathrm{dec}}$，可以得到总推理时延满足

$$T^{\mathrm{infere}}=\frac{O^{\mathrm{E}}+O_{\mathrm{UNet}}^{\mathrm{prior}}L^{\mathrm{prior}}}{f^{\mathrm{ES}}}+\frac{O^{\mathrm{D}}+O_{\mathrm{UNet}}^{\mathrm{dec}}L^{\mathrm{dec}}}{f^{\mathrm{UE}}}\quad(3)$$

以上处理流程可以看作一个M/G/1排队系统，根据Pollaczek-Khintchine公式，排队等待时

延的期望$\mathbb{E}[T^{\text{wait}}]$可以表示为

$$\mathbb{E}[T^{\text{wait}}]=\frac{\rho+\lambda\mu\mathbb{D}[T^{\text{serv}}]}{2(\mu-\lambda)} \tag{4}$$

其中$\mu=\frac{1}{\mathbb{E}[T^{\text{serv}}]}$表示平均服务速率，$T^{\text{serv}}=T^{\text{infere}}+T^{\text{dl}}$表示服务时延，$\mathbb{D}[T^{\text{serv}}]$为服务时延的方差，$\rho=\frac{\lambda}{\mu}$为输入负荷。系统逗留时延$T^{\text{tot}}$包括排队等待时延$T^{\text{wait}}$和服务时延$T^{\text{serv}}$，其期望为

$$\mathbb{E}[T^{\text{tot}}]=\mathbb{E}[T^{\text{wait}}]+\mathbb{E}[T^{serv}]=\frac{\rho+\lambda\mu\mathbb{D}[T^{\text{serv}}]}{2(\mu-\lambda)}+\frac{1}{\mu} \tag{5}$$

（2） 时延违反概率上界分析

记时延违反阈值为T_0，系统在时隙t的逗留时延超过阈值的违反概率可以表示为$\Pr(T^{\text{tot}}(t)>T_0)$。为了分析该时延违反概率的理论上界，本文引入矩母函数（MGF，Moment-Generation Function）法对系统进行分析[21]。记MGF函数的形式为$\mathbb{M}_X(\theta,\tau)=\mathbb{E}[e^{\theta X(t-\tau,t)}]$，其中$X(t-\tau,t)=\sum_{u=t-\tau}^{t-1}X(u)$为时间区间$[t-\tau,t)$上随机过程$X(u)$的累积量，$\theta$为一正实数。记$A^{\text{arrive}}(t-\tau,t)$表示时间$[t-\tau,t)$的到达的累计GAI推理请求数量，可以得到其MGF函数$\mathbb{M}_{A^{\text{arrive}}}(\theta,\tau)$的表达为

$$\mathbb{M}_{A^{\text{arrive}}}(\theta,t-\tau,t)=\mathbb{E}[e^{\theta A^{\text{arrive}}(t-\tau,t)}]=\sum_{k=0}^{\infty}e^{\theta k}\frac{e^{-\lambda\tau}(\lambda\tau)^k}{k!}=e^{\lambda\tau(e^{\theta}-1)} \tag{6}$$

记时间区间$[t-\tau,t)$累计推理的请求数量为$S^{\text{infere}}(t-\tau,t)$，累计下行传输的请求数量为$S^{\text{dl}}(t-\tau,t)$，两者的MGF函数可以表示为

$$\mathbb{M}_{S^{\text{infere}}}(-\theta,\tau)=\mathbb{E}[e^{-\theta S^{\text{infere}}(t-\tau,t)}]=e^{-\theta\left\lfloor\frac{f^{\text{UE}}\tau}{O^{\text{E}}+O^{\text{prior}}_{\text{UNet}}L^{\text{prior}}}\right\rfloor}e^{-\theta\left\lfloor\frac{f^{\text{ES}}\tau}{O^{\text{D}}+O^{\text{dec}}_{\text{UNet}}L^{\text{dec}}}\right\rfloor} \tag{7}$$

$$\mathbb{M}_{S^{\text{dl}}}(-\theta,\tau)=\mathbb{E}[e^{-\theta S^{\text{dl}}(t-\tau,t)}]=\{\mathbb{E}[e^{-\theta S^{\text{dl}}(u)}]\}^{\tau}=\{\mathbb{E}[e^{-\theta\left\lfloor\frac{R(u)}{f^{\text{dl}}}\right\rfloor}]\}^{\tau} \tag{8}$$

其中$\lfloor\cdot\rfloor$表示向下取整，$u\in[t-\tau,t)$。根据服务级联定理，可以得到总服务过程的MGF函数$\mathbb{M}_{S^{\text{tot}}}(-\theta,\tau)$满足

$$\begin{aligned}\mathbb{M}_{S^{\text{serv}}}(-\theta,\tau)&=\mathbb{E}[e^{-\theta(S^{\text{infere}}\otimes S^{\text{dl}})(t-\tau,t)}]\\&=\mathbb{E}[e^{-\theta\inf\limits_{0\le u\le\tau}(S^{\text{infere}}(t-\tau,u)+S^{\text{dl}}(u,t))}]\le\sum_{u=0}^{\tau}\mathbb{M}^{\text{infere}}(-\theta,u)\mathbb{M}^{\text{dl}}(-\theta,\tau-u)\end{aligned} \tag{9}$$

系统中，到达过程和服务过程相互独立，因此可以推导出时延违反概率满足

$$\begin{aligned}&\Pr(T^{\text{tot}}(t)>T_0)=\Pr\left(A^{\text{arrive}}(0,t)>\inf_{0\le s\le t+T_0}\{A^{\text{arrive}}(0,s)+S^{\text{serv}}(s,t+T_0)\}\right)\\&=\Pr\left(\exists s\in[0,t]:A^{\text{arrive}}(s,t)>S^{\text{serv}}(s,t+T_0)\right)\\&\le\sum_{s=0}^{t}\Pr(A^{\text{arrive}}(s,t)-S^{\text{serv}}(s,t+T_0)>0)\le\sum_{s=0}^{t}\mathbb{E}\left[e^{\theta A^{\text{arrive}}(s,t)}\right]\cdot\mathbb{E}\left[e^{-\theta S^{\text{serv}}(s,t+T_0)}\right]\\&=\sum_{\tau=0}^{t}\mathbb{M}_{A^{\text{arrive}}}(\theta,\tau)\cdot\mathbb{M}_{S^{\text{serv}}}(-\theta,\tau+T_0)\le\mathbb{M}_{S^{\text{serv}}}(-\theta,T_0)\sum_{\tau=0}^{t}\mathbb{M}_{A^{\text{arrive}}}(\theta,\tau)\mathbb{M}_{S^{\text{serv}}}(-\theta,\tau)\end{aligned} \tag{10}$$

为了得到以上上界的闭式解，本文对MGF函数进一步进行参数化表达，记$\rho_A(\theta)=\lambda(e^{\theta}-1)$，则到达过程的MGF函数形式可以改写为$\mathbb{M}_{A^{\text{arrive}}}(\theta,\tau)=e^{\rho_A(\theta)\tau}$。对于总服务过程，设单位时间的有效容量$r_{\text{eff}}(\theta)$约束了MGF函数$\mathbb{M}_{S^{\text{serv}}}(-\theta,\tau)$的上界，即$\mathbb{M}_{S^{\text{serv}}}(-\theta,\tau)\le e^{-\theta r_{\text{eff}}(\theta)\tau}$，当$t\to+\infty$时，若不等式$\lambda(e^{\theta}-1)<\theta r_{\text{eff}}(\theta)$成立，即指数级数能够收敛，系统可以达到稳态，可以得到时延违反概率的上界满足

$$\Pr(T^{\text{tot}}>T_0)\le \mathbb{M}_{S^{\text{serv}}}(-\theta,T_0)\sum_{\tau=0}^{\infty}\left(\mathbb{M}_{A^{\text{arrive}}}(\theta,\tau)\cdot\mathbb{M}_{S^{\text{serv}}}(-\theta,\tau)\right)$$
$$\le \mathbb{M}_{S^{\text{serv}}}(-\theta,T_0)\sum_{\tau=0}^{\infty}\left(e^{\lambda(e^{\theta}-1)}\cdot e^{-\theta r_{\text{eff}}(\theta)}\right)^{\tau} \tag{11}$$
$$=\mathbb{M}_{S^{\text{serv}}}(-\theta,T_0)\cdot\frac{1}{1-e^{\lambda(e^{\theta}-1)-\theta r_{\text{eff}}(\theta)}}=\frac{\sum_{u=0}^{T_0}\mathbb{M}_{S^{\text{infere}}}(-\theta,u)\mathbb{M}_{S^{\text{dl}}}(-\theta,T_0-u)}{1-e^{\lambda(e^{\theta}-1)-\theta r_{\text{eff}}(\theta)}}$$

最终时延违反概率上界可以表示为

$$\overline{\Pr}(T^{\text{tot}}>T_0)=\inf_{\theta>0}\frac{\sum_{u=0}^{T_0}\mathbb{M}_{S^{\text{infere}}}(-\theta,u)\mathbb{M}_{S^{\text{dl}}}(-\theta,T_0-u)}{1-e^{\lambda(e^{\theta}-1)-\theta r_{\text{eff}}(\theta)}} \tag{12}$$

由上述表达式可知，系统稳定性还可由总负荷$\rho_{\text{tot}}=\lambda\mathbb{E}[T^{\text{infer}}]$刻画，且需满足$\rho_{\text{tot}}<1$，才能避免队列积压在高负载下持续放大。为找到等式右端的下界，最优参数$\theta^{\star}$可以通过二分法找到[22]。

为进一步量化所提出的协作推理范式对GAI推理阶段划分所带来的通信与计算收益，表2给出了关键公式：

表2　三类推理范式的时延构成

推理范式	推理时延	下行传输时延
协作推理范式	$\frac{\mathcal{O}^{\text{E}}+\mathcal{O}^{\text{UNet}}_{\text{prior}}L^{\text{prior}}}{f^{\text{ES}}}+\frac{\mathcal{O}^{\text{D}}+\mathcal{O}^{\text{UNet}}_{\text{dec}}L^{\text{dec}}}{f^{\text{UE}}}$	$\frac{J^{\text{dl}}}{\mathbb{E}[R(t)]}$
边缘节点推理范式	$\frac{\mathcal{O}^{\text{E}}+\mathcal{O}^{\text{D}}+\mathcal{O}^{\text{UNet}}_{\text{prior}}L^{\text{prior}}+\mathcal{O}^{\text{UNet}}_{\text{dec}}L^{\text{dec}}}{f^{\text{ES}}}$	$\frac{J^{\text{img}}}{\mathbb{E}[R(t)]}$
终端节点推理范式	$\frac{\mathcal{O}^{\text{E}}+\mathcal{O}^{\text{D}}+\mathcal{O}^{\text{UNet}}_{\text{prior}}L^{\text{prior}}+\mathcal{O}^{\text{UNet}}_{\text{dec}}L^{\text{dec}}}{f^{\text{UE}}}$	无下行图像传输

在相同模型结构与可比质量约束下，所提出的协作推理范式在时延上严格优于边缘节点推理范式与终端节点推理范式的充分必要条件可等价写为以下两条不等式同时成立：

$$\frac{J^{\text{img}}-J^{\text{dl}}}{\mathbb{E}[R(t)]}>\left(\mathcal{O}^{\text{D}}+\mathcal{O}^{\text{UNet}}_{\text{dec}}L^{\text{dec}}\right)\left(\frac{1}{f^{\text{UE}}}-\frac{1}{f^{\text{ES}}}\right) \tag{13}$$

$$\frac{J^{\text{dl}}}{\mathbb{E}[R(t)]}<\left(\mathcal{O}^{\text{E}}+\mathcal{O}^{\text{UNet}}_{\text{prior}}L^{\text{prior}}\right)\left(\frac{1}{f^{\text{UE}}}-\frac{1}{f^{\text{ES}}}\right) \tag{14}$$

进一步地，可将边端协同带来的收益概括为相对时延节省率$\eta_T=\frac{T^{\text{edge-only}}-T^{\text{split}}}{T^{\text{edge-only}}}$与相对载荷节省率$\eta_J=\frac{J^{\text{img}}-J^{\text{dl}}}{J^{\text{img}}}$，二者分别刻画计算侧与通信侧的收益幅度。其中，$T^{\text{edge-only}}$为边缘独立推理范式的端到端总时延，$T^{\text{split}}$为边端协同范式的端到端总时延。

5.4　仿真结果分析

仿真参数设置遵循典型5G/MEC部署与Cascaded Diffusion模型规模折算，参数的含义与取值见表3。

表3　仿真参数设置

仿真参数	参数数值	仿真参数	参数数值
P^{dl}	31dBm	$[\mathcal{O}^{\text{D}},\mathcal{O}^{\text{UNet}}_{\text{dec}}]$	[1.1170, 3.5492] TFLOPs
B^{dl}	10MHz	$[\mathcal{O}^{\text{E}},\mathcal{O}^{\text{UNet}}_{\text{prior}}]$	[0.1940, 6.7598] TFLOPs
σ^2	-100dBm	$[J^{\text{img}},J^{\text{dl}}]$	[5.2429, 1.0486] Mbits

图8给出了不同时延界阈值下各阶段实际时延违约概率的仿真与理论上界结果。可以看出，理论上界与实际违约概率吻合紧密，验证了所推导上界的有效性与紧致性，表明该上界能够准确表征扩散推理各阶段的端到端时延统计特性。

图9展示了三类推理范式在不同指定时延界下的时延违反概率变化情况。结果表明，在相同推理请求到达速率条件下，所提出的推理范式在时延违反概率上界上均显著优于两类基线对比范式，其主要原因在于该范式通过在边缘与终端节点之间合理划分推理阶段，在充分利用边缘节点计算资源的同时，仅需传输低维潜空间中间表示。图10给出了不同推理范式下的推理质量对比

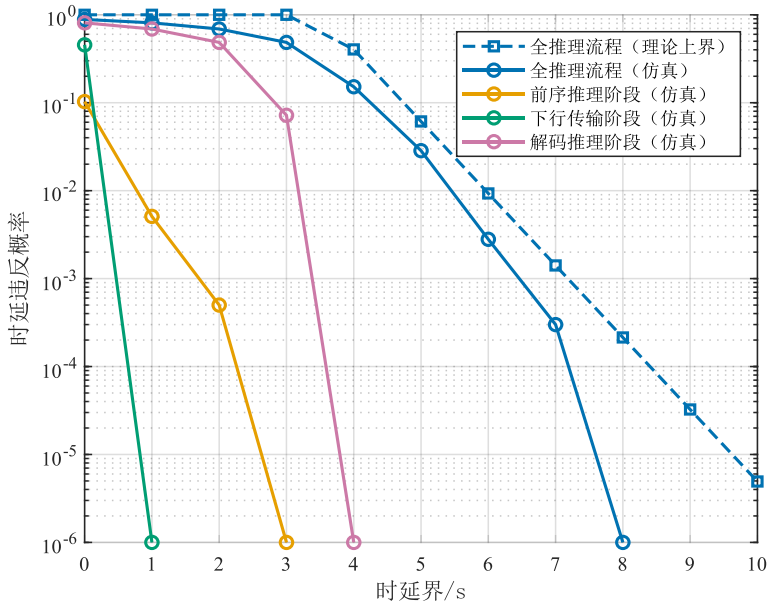

图8　各推理阶段的仿真与理论时延违反概率随时延界的变化

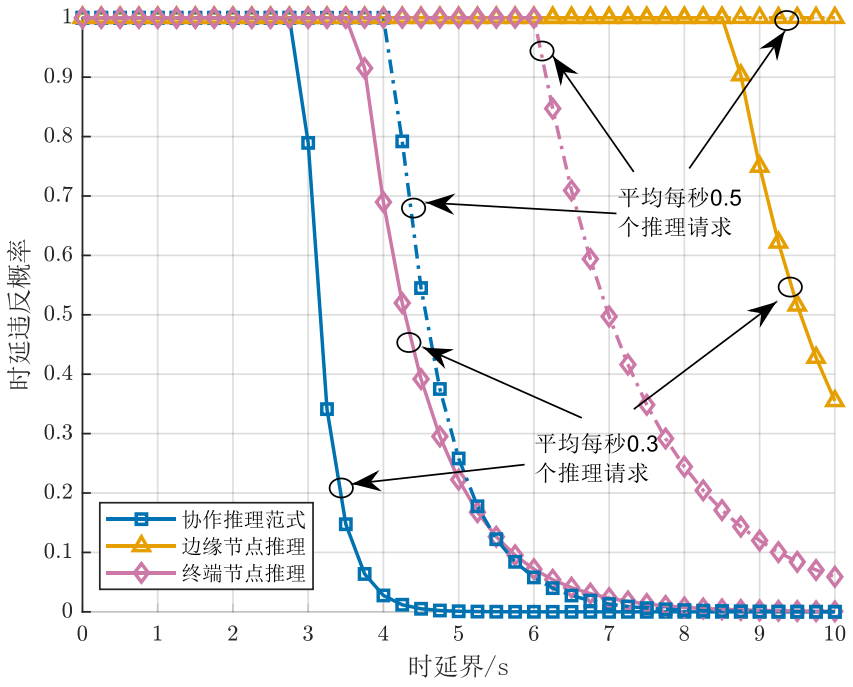

图9　各推理范式时延违反概率上界随时延界变化图

结果。可以看出，在相同生成质量指标约束条件下，所提出的推理范式在推理质量方面同样优于两类基线范式。

6　结束语

智能计算技术的快速发展正引发新一轮产业

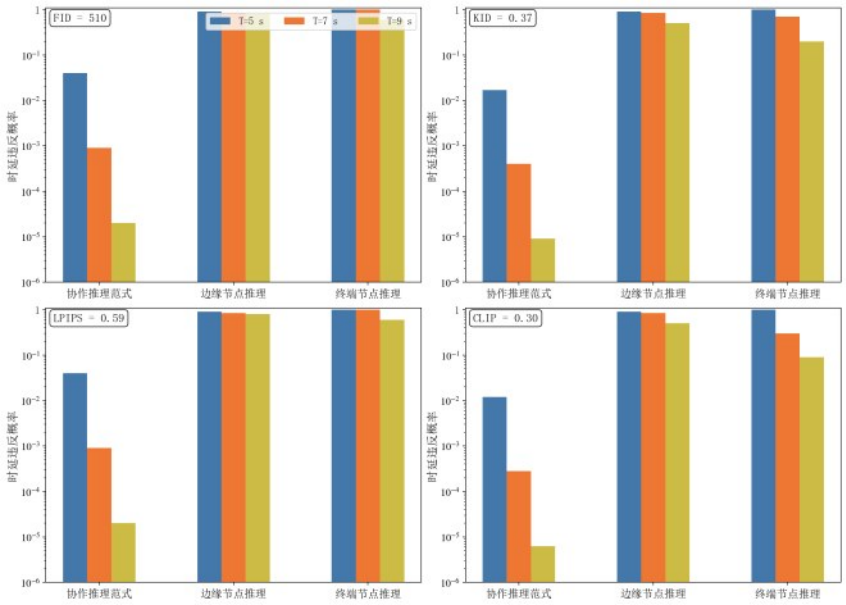

图 10　各推理范式在不同平均推理质量指标要求下的时延违反概率上界

形态与应用模式的深刻变革，而将智能计算从云端下沉至网络边缘，是充分释放其低时延、高可靠和场景自适应优势的重要途径。围绕这一发展趋势，本文系统梳理了边缘智能计算技术的实现路径与分类，进一步总结了边缘协同推理关键技术。在此基础上，本文进一步提出了一种面向生成式人工智能模型边缘推理的单主体模间推理系统示例，并通过性能对比分析表明，该系统在推理时延与推理质量方面相较传统边缘或终端独立推理方式具有明显优势。将来，随着边缘计算基础设施和网络协同机制的持续演进，边缘协同推理技术必将迎来更加广泛的应用。通过边缘计算资源分配与调度优化、AI 模型计算效率提升，边缘协同推理技术将在智能计算能力的广泛落地和规模化应用中发挥至关重要的支撑作用，为未来的智能社会铺设更加坚实的技术基础。与此同时，随着技术的不断成熟，边缘协同智能计算还将推动新型产业形态的诞生，开辟全新的应用领域，进一步推动全球数字经济的发展与进步。

参考文献：

[1]　朱文武. 多媒体智能计算若干研究进展[J]. 中国科学: 信息科学, 2025, 55(09): 2153-2164.

[2]　郑远鹏, 张天魁, 庞博, 等. 面向边缘智能的移动算力网络关键技术研究[J]. 邮电设计技术, 2023, (05): 88-92.

[3]　WANG C X, YOU X, GAO X, et al. On the road to 6G: visions, requirements, key technologies, and testbeds[J]. IEEE Communications Surveys & Tutorials, 2023, 25(2): 905-974. DOI: 10.1109/COMST.2023.3249835.

[4]　卢先领, 李德康. 面向大规模多接入边缘计算场景的任务卸载算法[J]. 电子与信息学报, 2025, 47(1): 116-127. DOI: 10.11999/JEIT240624.

[5]　林鹏, 黄新梁, 宁兆龙, 等. 多指标意图驱动的无人机计算卸载与轨迹规划自适应优化策略[J]. 通信学报, 2026, 47(3): 195-208. DOI: 10.11959/j.issn.1000-436x.2026056.

[6]　陈乐, 马彰超, 董芃, 等. 面向工业智能体互联网的“通信-控制”协同原子化重构机制[J]. 通信学报, 2026, 47(3): 42-63. DOI: 10.11959/j.issn.1000-436x.2026053.

[7]　WANG Y, YANG C, LAN S, et al. End-edge-cloud collaborative computing for deep learning: a comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2024, 26(4): 2647-2683. DOI: 10.1109/COMST.2024.3393230.

[8]　牛涛. 面向边缘智能的计算加速研究[D]. 北京: 北京邮电大学, 2024. DOI: 10.26969/d.cnki.gbydu.2024.000146.

[9]　傅文军, 谭伟, 胡露航. 从判别式人工智能到生成式人工智能的演进逻辑及场景策略研究[J]. 中国仪器仪表, 2024(10): 17-21.

[10]　BANH L, STROBEL G. Generative artificial intelligence[J]. Electronic Markets, 2023, 33(1): 63. DOI: 10.1007/s12525-023-

00680-1.

[11] DAI P, HAN B, LI K, et al. Joint optimization of device placement and model partitioning for cooperative DNN inference in heterogeneous edge computing[J]. IEEE Transactions on Mobile Computing, 2025, 24(1): 210-226. DOI: 10.1109/TMC.2024.3457793.

[12] XIAO X, ZHANG J, WANG W, et al. DNN-driven compressive offloading for edge-assisted semantic video segmentation [C]//IEEE INFOCOM 2022-IEEE Conference on Computer Communications. London, United Kingdom: IEEE, 2022: 1888-1897.

[13] HAO Z, XU G, LUO Y, et al. Multi-agent collaborative inference via DNN decoupling: intermediate feature compression and edge learning[J]. IEEE Transactions on Mobile Computing, 2023, 22(10): 6041-6055.

[14] PACHECO R G, SHIFRIN M, COUTO R S, et al. AdaEE: adaptive early-exit DNN inference through multi-armed bandits [C]//ICC 2023-IEEE International Conference on Communications. Rome, Italy: IEEE, 2023: 3726-3731.

[15] ZHONG R, MU X, ZHANG Y, et al. Mobile edge generation: a new era to 6G[J]. IEEE Network, 2024, 38(5): 47-55. DOI: 10.1109/MNET.2024.3420240.

[16] LUO H, et al. Toward edge general intelligence with multiple-large language models (Multi-LLM): architecture, trust, and orchestration[J]. IEEE Transactions on Cognitive Communications and Networking, 2025, 11(6): 3563-3585.

[17] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2020: 6840-6851.

[18] GAO S, YANG P, KONG Y, et al. Characterizing and scheduling of diffusion process for text-to-image generation in edge networks[J]. IEEE Transactions on Mobile Computing, 2025, 24(10): 11137-11150. DOI: 10.1109/TMC.2025.3574065.

[19] HO J, SAHARIA C, CHAN W, et al. Cascaded diffusion models for high fidelity image generation[J]. Journal of Machine Learning Research, 2022, 23(1): 1-33.

[20] PERNIAS P, RAMPAS D, RICHTER M L, et al. Wuerstchen: an efficient architecture for large-scale text-to-image diffusion models[C]//Proceedings of the 2024 International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.

[21] 李松, 王新荣, 王博文, 等. 基于随机网络演算的车联网边缘计算多跳任务卸载性能分析[J]. 电子与信息学报, 2023, 45(7): 2459-2466.

[22] CUI P, HAN S, LI L, et al. Roundtrip interaction delay analysis of immersive communications: a stochastic network calculus perspective[J]. IEEE Transactions on Wireless Communications, 2025, 24(3): 2188-2202. DOI: 10.1109/TWC.2024.3518589.