专题：AI赋能通信网络

面向实时通信的边缘智能关键技术研究

王辰[1]，白雪茜[1]，魏彬[1]，宋月[1]，张强[2]
（1. 中国移动通信有限公司研究院，北京 100053；
2. 中兴通讯股份有限公司，广东 深圳 518063）

摘　要：针对当前实时通信网络架构复杂、新业务引入缓慢等问题，结合未来实时通信智能化的演进趋势，提出了智能内生实时通信网络边缘智能架构，包含统一控制面、统一智能面、边缘敏捷媒体面等，以支撑传统实时通信网络从“通话入口”向“应用入口”“超级入口”的演进；研究了网络人工智能（artificial intelligence，AI）内生体系，提出边缘智能基本策略、单智能体通信架构；分析了实时通信边云协同推理关键技术，提出了分布式AI模型管理、边云协同推理机制和构建用户知识图谱。

关键词：实时通信；边缘智能；智能体；边云协同推理；多智能体交互
中图分类号：TN915.81
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025194

Research on the key technologies of edge intelligent for real-time communication

WANG Chen[1], BAI Xueqian[1], WEI Bin[1], SONG Yue[1], ZHANG Qiang[2]
1. China Mobile Research Institute, Beijing 100053, China
2. ZTE Corporation, Shenzhen 518063, China

Abstract: In view of the pain points faced by the complex architecture and slow introduction of new services of the current real-time communication network , and the future evolution trend of intelligent real-time communication, an edge intelligent architecture of intelligent agile real-time communication network was proposed, including an unified control plane, a unified intelligent plane and an edge agile media plane, so that the intelligent communication network can evolve from “call entrance” to “application entrance” and “super entrance”. The intrinsic network artificial intelligence(AI) system was studied, and the basic strategy of edge intelligence and the communication architecture of a single AI agent were proposed. The key technologies of collaborative inference of edge-cloud for real-time communication were analyzed. The construction of distributed AI model management, collaborative inference mechanisms of end-edge-cloud and user’s knowledge maps were proposed.

Key words: real-time communication, edge intelligent, AI agent, edge-cloud collaborative inference, multi-agent interaction

收稿日期：2025−06−13；修回日期：2025−08−08
通信作者：魏彬，weibin@chinamobile.com

0　引言

目前，全球移动通信技术正在经历从5G网络向5G-Advanced（5G-A）、6G时代逐步迈进的重要时期，网络架构及业务功能都在悄然发生变化[1]。传统IP多媒体子系统（IP multimedia subsystem，IMS）网络是实现网络互通的关键。IMS作为提供多媒体服务的核心，支持着服务质量的保障、用户授权等多种功能[2]。近些年，由于新业务模式的冲击，IMS网络面对新需求已经逐渐力不从心，扩展现实（extended reality，XR）、全息交互、多模态交互等新型业务模式也逐步进入用户视野，IMS网络面临人工智能（artificial intelligence，AI）、沉浸式交互等新兴媒体形态的冲击与挑战[3]。

实时通信网络层级结构复杂，网元种类众多。由于行业需求变化，功能性网元数量逐步累积，给网络的管理和运维等带来了压力。另外，网络层级较多、网元功能分散等特点对新业务的开发和新功能的快速上线造成较大阻碍，为网络带来升级困难、业务研发周期延长等问题，导致传统电信网络业务发展严重滞后于互联网应用[4]。

1　未来实时通信网络的演进趋势

结合未来业务演进趋势，AI技术与实时通信技术的深度结合已成为未来业务发展的基础，基于大语言模型（large language model，LLM）的智能体（AI agent）技术，逐渐成为未来媒体交互的核心能力。边缘网络的AI功能可为用户提供原子化、轻量化的网络AI能力[5]。

未来通信演进趋势如图1所示，未来实时通信网络的变化趋势主要根据业务类型的变化可以大致分为以下3个阶段。

阶段1：DC+实时通信。以VoNR+网络为代表，基于IMS网络实现功能深度拓展，引入数据通道（data channel，DC）技术，实现趣味通信、同声翻译、背景替换、人像风格替换等增值功能，为用户带来全新的交互体验。

阶段2：AI+实时通信。融合AI技术升级用户体验，形成个人信息的服务入口，初步具备智能通信能力，由智能体处理用户的基本需求，安全助理、智能助聊、智能代聊等成为该阶段的主流业务。

阶段3：智能内生实时通信。实现内置AI功能的新型实时通信网络，完整的智能体将改变现

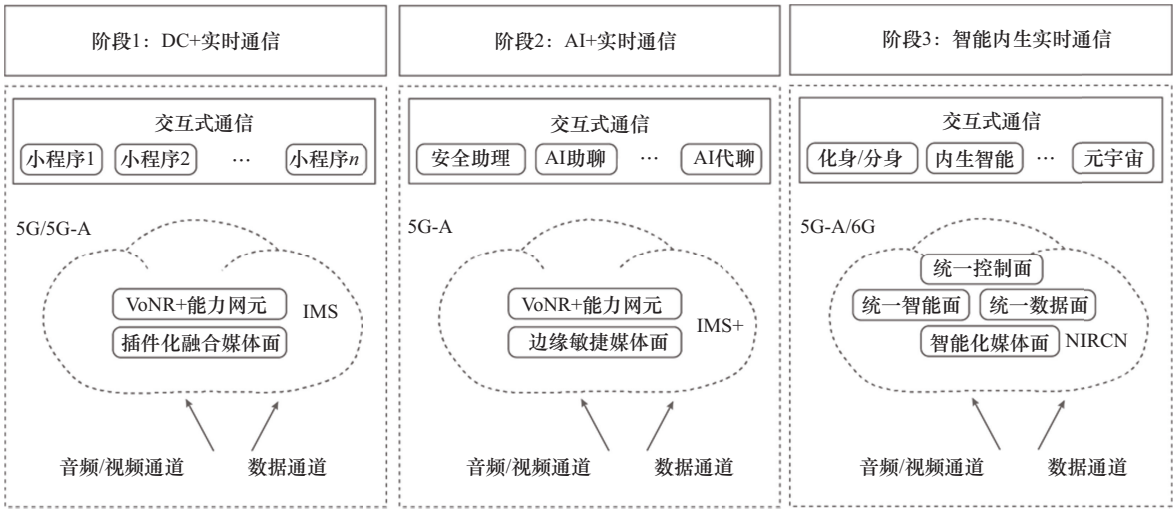


图1　未来通信演进趋势

有业务模式，为用户提供全天候的实时服务，如订票、购物，AI办公、自动化收发邮件、虚拟会议等。

2 智能内生实时通信网络

AI技术和信息通信技术（information and communications technology，ICT）的深度融合，是实现IMS网络破茧重生的基础。智能内生实时通信网络（native intelligent real-time communication network，NIRCN）架构主要包含统一应用面、统一控制面、统一智能面、统一数据面和边缘敏捷媒体面（edge agile media plane，EAMP）分别承担不同职责的网元功能[6]，图2展示了NIRCN基本网络架构。实时通信网络AI能力集成于统一智能面，向整个网络提供基础AI能力，统一控制面、EAMP通过协作配合，按需向用户提供AI能力，减少网络资源浪费，该架构将有效助力实时通信网络的数智化转型，为用户提供更加智能、敏捷的运营商网络服务[7]。

（1）统一控制面

统一控制面是NIRCN提供基础通信和实现智能决策的重要网元，承担策略生成、资源调度等核心功能，通过信令路由、媒体资源的控制实现网络资源的全局协调。结合网络服务质量（quality of service，QoS），从宏观角度针对多个用户的不同需求（如XR低时延、8K超高清）实时调整资源分配，合理化媒体优先级控制，保障用户体验。

（2）EAMP

EAMP采用插件化的微服务架构设计，实现了功能的动态扩展与卸载，以用户需求为核心按需装配。结合未来实时通信业务的高算力、低时延诉求，EAMP配合统一智能面、统一数据面形成组合AI系统，完善推理时延与精度间的平衡。

（3）统一智能面

统一智能面集成了AI模型的管理及推理等关键功能，是实现基础AI模型管理的核心网元。统一智能面既具备完整的AI模型，又可以推送定制化的AI模型。统一智能面的引入大大提升了网络资源的智能化效率。

（4）统一数据面

统一数据面的主要功能覆盖数据采集、存储、管理、过滤及数据资产的认证，为各网元按需提供数据调度和保障，同时统一数据面也是用户隐私安全的核心屏障，通过动态知识库及智能加密能力，为用户信息安全提供有效保障。

（5）统一应用面

统一应用面基于其他层所开放的基础能力提供运营商扩展应用或第三方定制化应用能力。

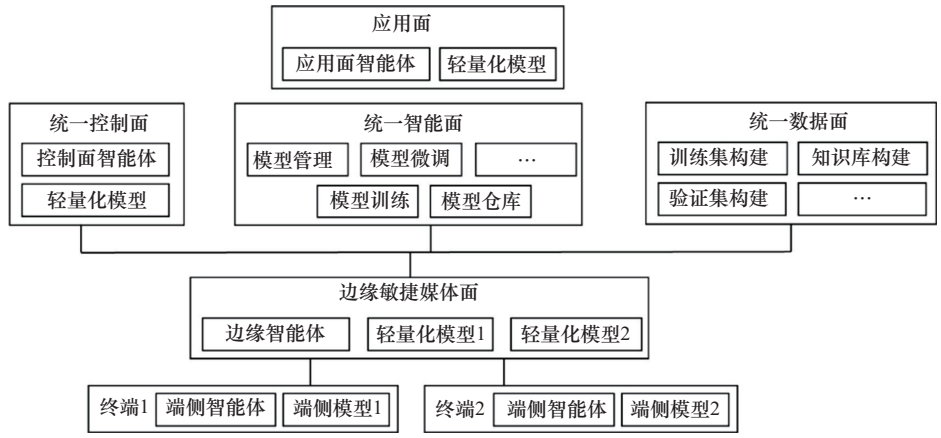


图2 NIRCN基本网络架构

3　边缘智能基本策略

通过部署多组客制化的轻量化大语言模型（lightweight large language model，L-LLM）及其他AI模型组成的模型群组，避免了单一模型在特定领域泛化能力差、资源消耗高等缺点，边缘网络由于其响应迅速、带宽消耗低及可再生等特点，成为网络提供AI服务的核心。

边缘网络AI功能架构如图3所示，包含意图分析、提示词、检索增强生成（retrieval augmented generation，RAG）及本地知识库等核心模块，但边缘节点算力资源少，且推理保障能力较差。基于上述现实，云端统一智能面支撑边缘网络智能体便成为必然选择，该方案为网络的AI能力动态适配提供了平台基础。

3.1　实时通信单智能体交互

网络的智能体架构以大语言模型为核心，集成推理、执行与记忆等关键能力，智能体主要为满足用户多样化的新业务需求，提供即时的AI服务，因此智能体需要具备任务执行、任务状态管理及记忆管理等功能，为实现“人、物、虚”多模态通信提供基础。图4展示了实时通信单智能体交互流程，本文提出了插件化、敏捷单智能体实时通信的架构，内嵌多模态智能体，融合意图识别、背景替换等AI插件，灵活实现了实时通信业务的敏捷上线[8]。

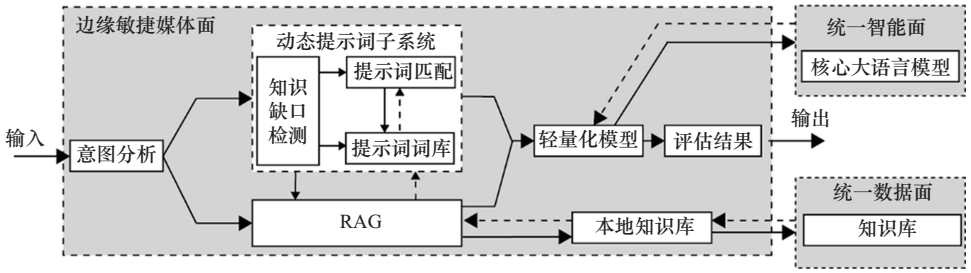


图3　边缘网络AI功能架构

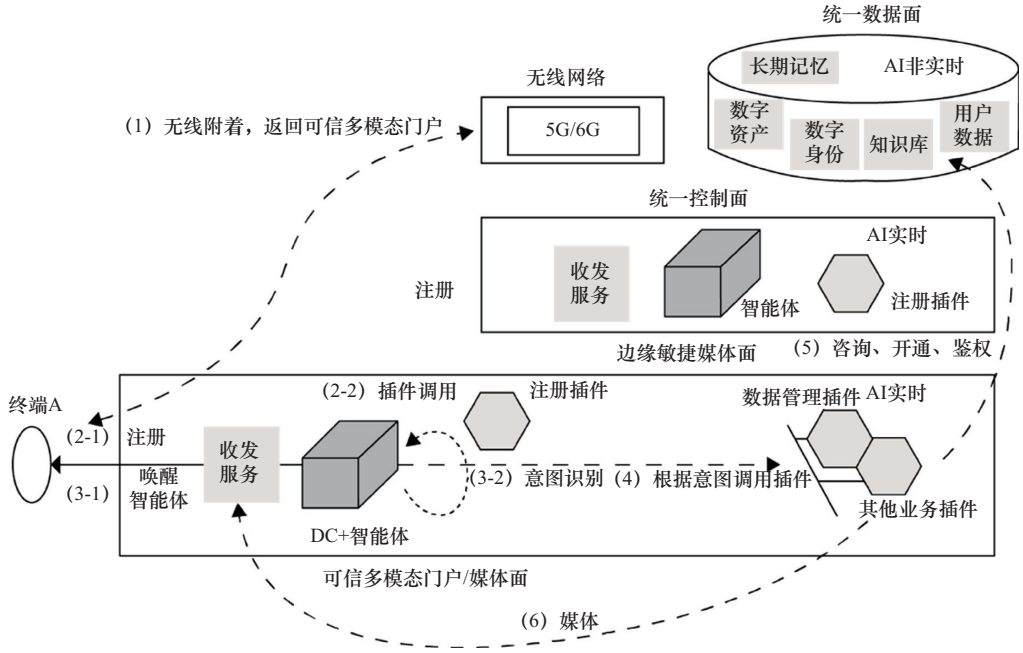


图4　实时通信单智能体交互流程

首先，针对多模态交互需求，本文提出在EAMP内嵌具备多模态信息处理能力的智能体，提升网络的业务覆盖能力。为了对接未来业务的新需求，智能体以多模态大语言模型（multi modal large language model，MLLM）为基础实现文本、图像、音频等多种数据的同步处理，完成多模态信息在不同参数空间内的信息融合，提取特征并在状态空间完成映射与特征融合，该方法可完善推理结果准确性，提高AI推理在多模态业务领域的精度。完善MLLM的跨模态理解能力，使得AI模型能够从多维角度理解用户需求，承接更多业务类型。

其次，针对业务上线的差异化需求，EAMP、统一控制面具备插件化功能部署能力，为智能体业务提供即插即的能力扩展（如XR通信助手、全息会议代理、AI客服等），平台包含注册、数据管理、背景替换、意图识别等多种服务。搭配插件化的工具箱扩展智能体的能力范围[9]。

智能体根据多模态信息识别结果，调用数据管理插件，从统一数据面获取数字资产、数字身份、知识库、用户数据等，根据业务场景识别意图并调用相关能力，如在背景替换场景中，用户只需通过语音即可完成背景替换。实现场景示例：语音（用户A：“帮我换一个新通话背景、二次元风格的”）由自动语音识别（automatic speech recognition，ASR）转换成文本后交由自然语言处理（natural language processing，NLP）模型识别用户意图并返回结果，在背景替换插件的支持下完成二次元风格背景的自动替换。

3.2 边缘与分布式智能体架构

EAMP是网络面向用户提供服务的首要网元，但目前用户群体差异较大，导致任务类型既包含简单化且低时延要求的日常任务，也存在资源消耗较高的复杂任务。在此背景基础上，针对不同需求，实现分级处理的智能体便成为下一代网络的设计重点。NIRCN智能体总架构如图5所示，网络包含集成于边缘网络的边缘智能体与各组件分离部署的分布式智能体。多智能体耦合的设计可高效响应不同的用户需求，提升媒体处理效率[10]。

（1）边缘智能体

边缘智能体由轻量化模型与其他组件构成（推理、执行及短期记忆等模块），插件化架构可提供高效的任务编排能力。所有组件在边缘侧部署，包括轻量化模型及其他生成类模型，可提供丰富的能力空间。轻量化模型主要负责用户任务的理解与规划，而图形或者其他视觉类模型则完成不同模态信息的处理。

（2）分布式智能体

分布式智能体以边缘侧的核心模块为主，搭配轻量化模型及核心大模型，具备更强的推理能

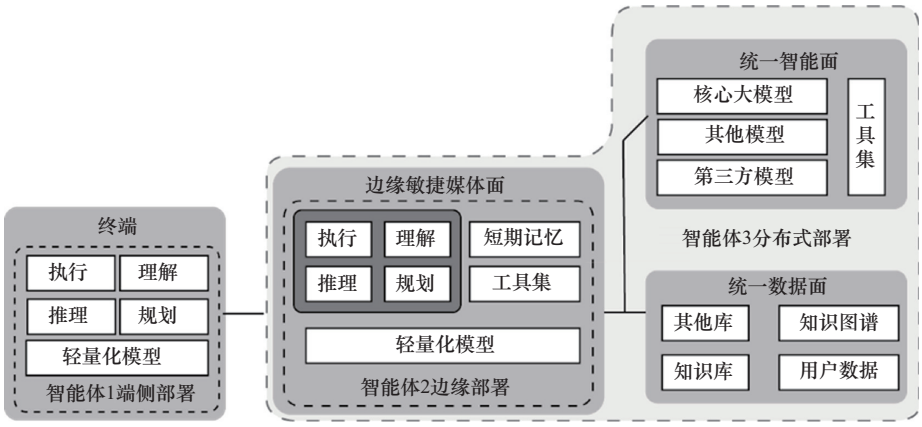


图5 NIRCN智能体总架构

力，成为未来网络的“变形金刚”。多模态感知模块完成用户输入的信息统一，通过边缘网络完成任务规划，同时积累用户数据，为训练用户画像的知识图谱提供基础。记忆功能由边缘侧的短期记忆模块与统一数据面的知识库共同实现，提供更长的上文处理能力，完成相对复杂的任务处理。

3.3　动态提示词机制

良好的提示词工程可为AI系统充分节省资源，提高推理效率与精确度，动态提示词根据不同的任务场景和用户需求，选择不同的提示词模板。动态提示词系统流程如图6所示，简单意图由提示词库匹配模板，复杂或不明确的意图，由知识缺口检测模块调用RAG，完成陌生词的检索及提示词库的更新，生成更符合上下文的提示词，引导模型生成更加精准的结果。

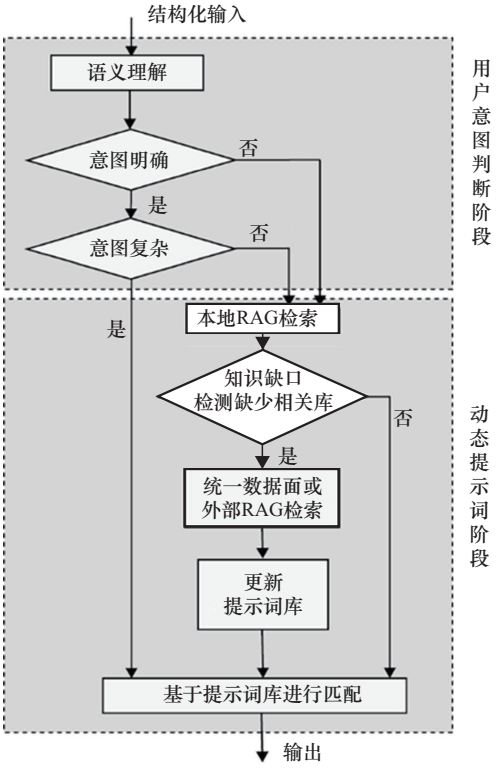


图6　动态提示词系统流程

3.4　RAG知识库更新机制

为了增强边缘智能体的推理精确度，降低幻觉概率，EAMP的RAG模块通过本地知识库支撑提示词系统。RAG知识库系统流程如图7所示，RAG模块可同时对接第三方或统一数据面的知识库更新数据，保证知识时效性，提高推理结果的准确度。

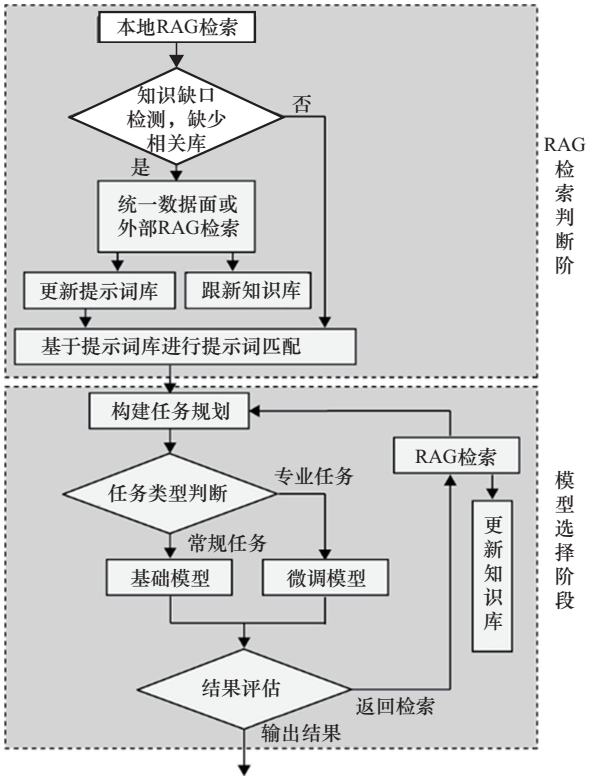


图7　RAG知识库系统流程

4　实时通信边云协同关键技术

边云协同AI处理能力是NIRCN未来实现快速、精准AI服务的基石，分布式的模型管理及用户知识图谱等功能是边云协同推理的基础[11]。

4.1　分布式AI模型管理

模型训练成本高、特定领域泛化能力不足是AI能力与NIRCN结合过程中无法避开的挑战。分布式AI模型管理流程如图8所示，AI系统中的模型管理可实现轻量化模型与核心大模型的高低搭配，降低资源开销。统一智能面具备的模型管理能力包含训练验证、蒸馏及微调等，为轻量化模型的统一管理提供了基础[12]。

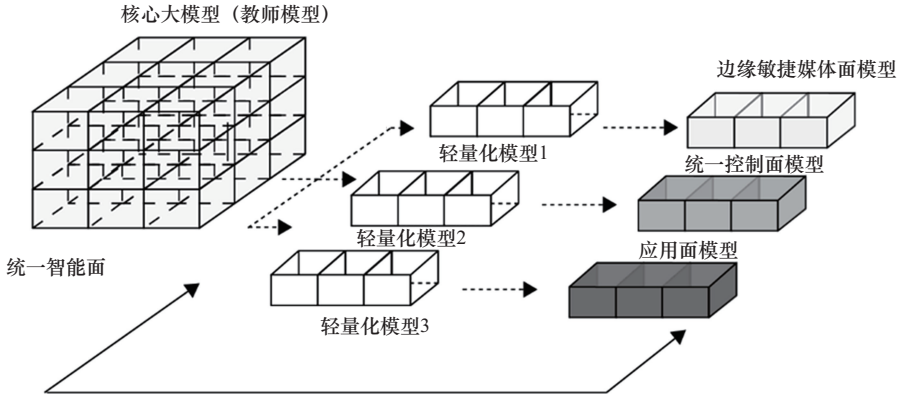


图8　分布式AI模型管理流程

（1）模型训练验证

统一智能面通过统一数据面提供的标准数据完成教师模型（teacher model，TM）的预训练，或直接部署第三方模型。自有模型的可靠性及安全性由统一智能面负责评估。

（2）模型蒸馏

边缘侧轻量化模型的划分机制如图9所示，部署于统一智能面的核心大模型首先经过知识蒸馏、量化等步骤完成模型的压缩，基于核心大模型分离出适合部署在EAMP的轻量化模型，即学生模型（student model，SM），依靠灰度发布策略推送至各网元[13]。

协同AI推理需要在轻量化模型的中间层找到合适的模型划分点，直接影响AI推理的效率，推理的时延（包括边缘模型的计算时延、数据传输时延及云端推理时延）和最终推理精度存在相互制约的矛盾。

（3）模型微调

EAMP具有多组平行部署的轻量化模型，由统一智能面进行管理，根据任务及业务类型进行区分，边缘轻量化模型微调后的输出精度（即模型输出置信度）由统一智能面进行调控。

（4）异构模型管理

系统包括Transformer架构模型及普通深度神经网络（deep neural network，DNN）模型，DNN模型中包含大量卷积神经网络（convolutional neural network，CNN）模型。这两类模型在性能和机制上存在差异，在进行任务推理过程中，两者需要搭配使用，因此统一智能面的异构模型管理便成为模型管理的重点能力之一[14]。

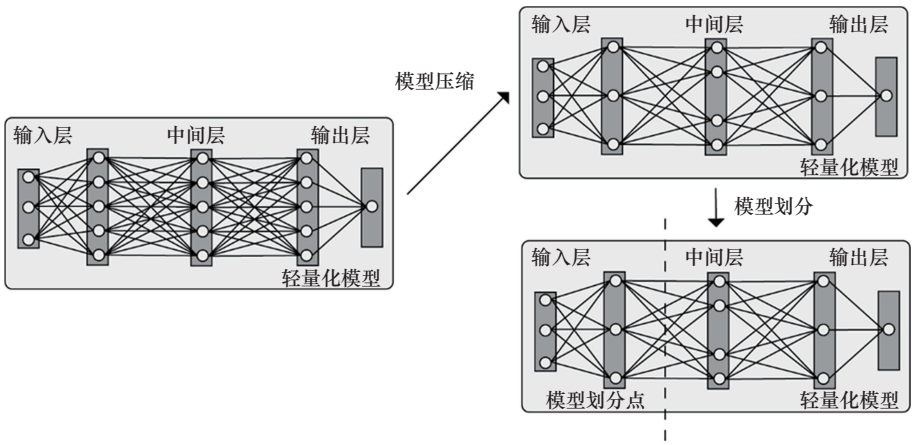


图9　边缘侧轻量化模型的划分机制

4.2　边云协同AI推理

现有AI技术关注模型的泛化能力与结果精度，对推理时延关注度不高，边缘节点的有限资源无法完成复杂推理。硬件及软件资源丰富的云端可完成核心大模型的调用，但推理时延较大，因此需要设计一种边缘和云端协同完成AI推理的方案[15]。

NIRCN边云协同推理机制如图10所示，边云协同系统由边缘推理节点及云推理节点共同完成。EAMP部署的轻量化模型由统一智能面获得，边缘网络的轻量化模型在训练过程中包含适合于云边协同推理的模型划分点设计，在模型主干增加分类头，模型划分点的选择直接决定边缘端和云端的计算负载分配，影响系统的推理延迟、通信开销和模型精度[16]。

边云协同推理的模型划分机制如图11所示，EAMP模型的前几层分析用户的任务需求，完成输入数据（图像帧、音频）的分类推理，得到多个中间特征结果及对应的置信度。

轻量化模型S与核心模型T均为L层模型，任意两层间均可划定模型划分点r，其中$r \in \{1, 2, \cdots, L-1\}$。增加模型划分机制的Transformer架构模型如图12所示，其中包含数个中间评分窗口，

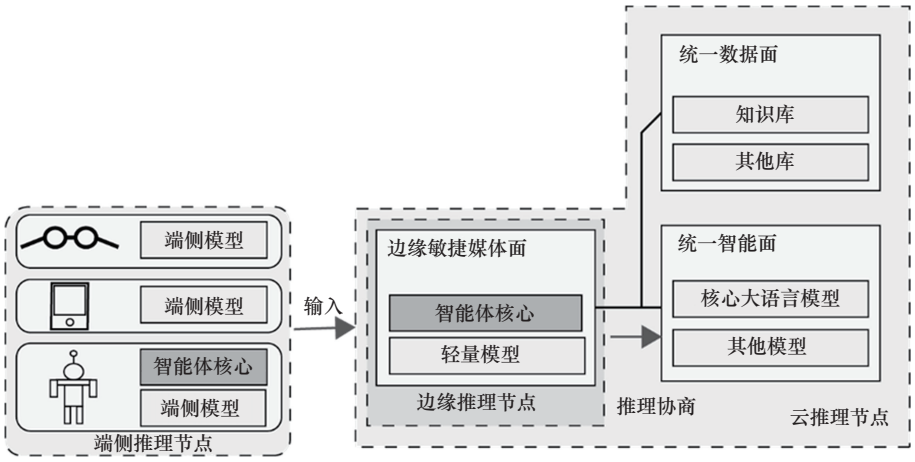


图10　NIRCN边云协同推理机制

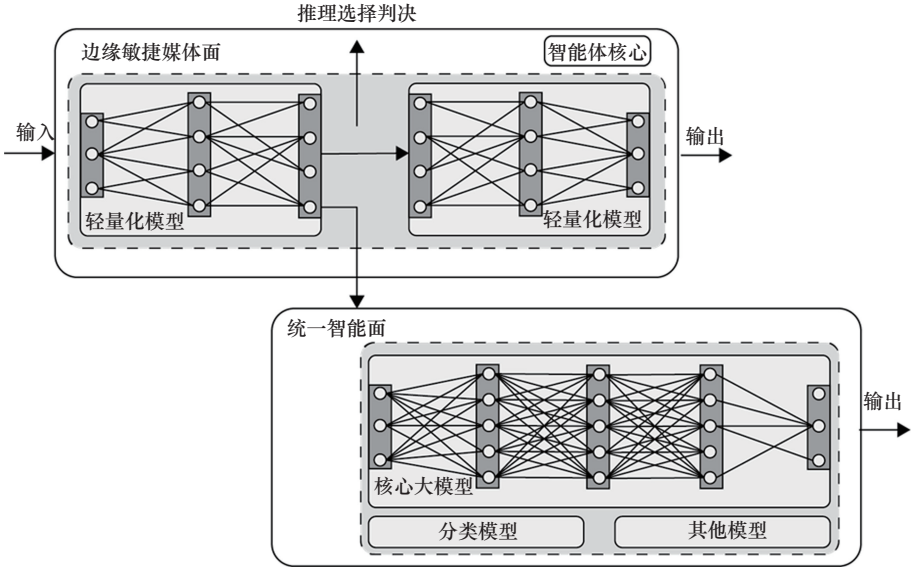


图11　边云协同推理的模型划分机制

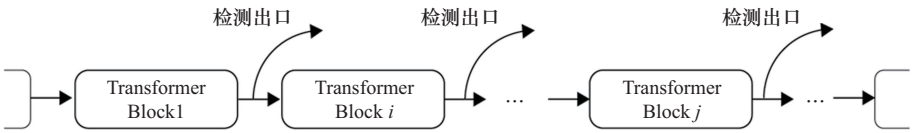


图12　增加模型划分机制的Transformer架构模型

Transformer Block为Transformer架构模型的基本单元。

具备模型划分功能的轻量化模型S训练完成后，推理过程中针对输入样本x不断进行前向传播，第i层遇到旁路出口[17]，系统由softmax函数将第i层处出口输出的评分z_i转化为概率分布向量$\boldsymbol{p}_i$：

$$\boldsymbol{p}_i=\mathrm{softmax}\left(z_i\right) \tag{1}$$

进而获得置信度分数u_i：

$$u_i=\max\left[\mathrm{softmax}\left(z_i\right)\right] \tag{2}$$

与预先设定的置信度阈值u_0进行比较，如果置信度超过阈值，即$u_i>u_0$则认为足够可信，触发早退机制，由EAMP完成推理并返回结果；如果$u_i<u_0$，则继续进行前向传播，在第j层再次遇到旁路出口时继续计算置信度u_j并继续判断。如果从模型的j_0层一直到j_r层$(r\in\mathbf{R})$，其所包含的旁路出口的置信度u_j较低，则将j_r作为模型划分点，将结果通过压缩数据传输至云端模型触发二次推理（核心大模型的完整推理），其中旁路出口r在模型中的位置在轻量化模型的训练阶段决定，以推理时延和推理精度作为优化项进行计算得到最佳平衡。由此边缘的轻量化模型与云端的核心大模型可实现高效、节能运行，找到推理精度与时延的最佳平衡点[18-19]。

NIRCN中存在DNN模型及Transformer架构模型，两种模型结构存在本质区别，Transformer架构模型的每一层都包含全局信息，并且Transformer架构模型输入信息长度的不同直接影响输出结果的大小，因此控制模型的输入信息长度与模型划分点的位置，便成为边云协同推理的必要考虑。

4.3　用户知识图谱构建

知识图谱即部署于统一数据面便于AI系统为用户精准推送个性化服务，范围涵盖用户的行为习惯、业余爱好、设备信息等。生成用户画像时，采用联邦学习等方法对用户的敏感信息，如号码、位置等进行脱密，识别出某用户的分组信息后，分组或画像信息可在数据面采用加密存储。良好的数据源是构建用户知识图谱的基础，因此统一数据面需要具备包括但不限于向量的嵌入生成、数据过滤等基本功能[20]。

用户知识图谱可以看作一个有向多重图，可定义如下：

$$G=\left(E,R,T\right) \tag{3}$$

其中，$E=\{e_1,e_2,\cdots,e_n\}$是属性实体集合，数量为$|E|$，代表用户地理位置、用户行为、使用业务类型等。$R=\{r_1,r_2,\cdots,r_n\}$是各实体间的关系，数量为$|R|$，代表如用户地理位置与用户行为的关系、用户行为与使用业务类型的关系等。$T=\{t_1,t_2,\cdots,t_n\}$是事实集合，数量为$|T|$。

评分函数（scoring function，SF）被用于评价事实三元向量组$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$的真实性，是知识图谱嵌入的灵魂，其中$\boldsymbol{h}$代表头实体，$\boldsymbol{t}$代表尾实体，$\boldsymbol{r}$代表相互关系。良好的数据源可有效减少网络资源浪费并增强方案的可靠性，因此统一数据面需要具备包括但不限于嵌入生成、数据过滤等基本功能。

4.3.1　向量嵌入与图神经网络

向量化嵌入是知识图谱的核心，通过嵌入生成可将各实体映射到低维向量空间中，将高维的复杂结构转换为低维关系，使知识变为可处理的对象。向量化嵌入过程由统一数据面部署的嵌入模型完成，嵌入模型基于向量关系计算并理解实

体之间的联系。嵌入模型包括翻译距离模型、语义理解模型、图神经网络模型等，每种模型适用于不同的领域，单一模型难以应对复杂场景，因此多模型混用逐渐成为实际部署的主流。

（1）平移嵌入模型

平移嵌入（translation embbeding，TransE）模型是知识图谱嵌入的经典模型算法。图13展示了基于平移距离的评分方式，向量空间中将头实体向量与关系向量的和与尾实体向量的距离作为评分标准，该模型无法处理对称关系，满足对称的事实关系，即事实向量组可交换，$\boldsymbol{h}+\boldsymbol{r}=\boldsymbol{t}$且$\boldsymbol{t}+\boldsymbol{r}=\boldsymbol{h}$，导致$\boldsymbol{r}$退化为零向量，因此对称关系在此类模型中较为乏力。

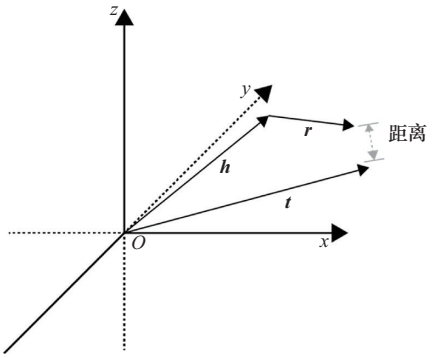


图13　基于平移距离的评分方式

在实数空间内的评分函数如下：

$$f(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})=-\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_{\mathrm{L1/L2}} \tag{4}$$

式（4）通过事实向量组在实值空间内的L1或L2范数描述评分大小。

（2）复数嵌入模型

由于NIRCN中的用户关系多数为单向关系，不满足对称性，因此针对非对称关系的复数嵌入（complex embedding，ComplEx）模型便成为必要选择，该模型是基于分布乘法（distributional multiplication，DisMult）模型在复数域上的拓展得到，图14展示了DisMult模型评价函数与节点关系，描述事实向关系的变量由矩阵$\boldsymbol{M}(\boldsymbol{r})$表示，评价函数如下：

$$f_{\mathrm{RESCAL}}(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})=\boldsymbol{h}^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{r})\boldsymbol{t} \tag{5}$$

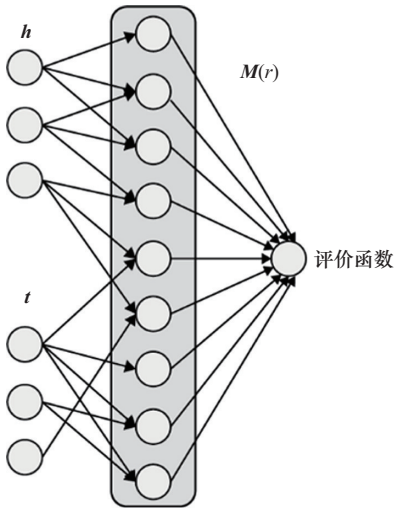


图14　DisMult模型评价函数与节点关系

关系矩阵$\boldsymbol{M}(\boldsymbol{r})$可简化为特征对角矩阵，通过向量组的点积，计算得出两个向量的相似度，以此作为评价函数的依据，判断结果可靠性，评价函数f_{DisMult}如下：

$$f_{\mathrm{DisMult}}(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})=\boldsymbol{h}^{\mathrm{T}}\mathrm{diag}\left[\boldsymbol{M}(\boldsymbol{r})\right]\boldsymbol{t} \tag{6}$$

评价函数f_{DisMult}在DisMult模型中为对角化矩阵，因此可以简化为：

$$f_{\mathrm{DisMult}}(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})=\sum_{i=1}^{k}h_i r_i t_i \tag{7}$$

其中，h_i、r_i、t_i分别为头实体向量$\boldsymbol{h}$的第i个元素、关系向量$\boldsymbol{r}$的第i个元素、尾实体向量$\boldsymbol{t}$的第i个元素，k为向量的维度。

点积的交换性导致$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$与$(\boldsymbol{r},\boldsymbol{h},\boldsymbol{t})$的评分没有任何区别。因此ComplEx模型将嵌入向量从实数域扩展到复数域，可以表示为：

$$f_{\mathrm{complEx}}(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})=\mathrm{Re}\{\boldsymbol{h}^{\mathrm{T}}\mathrm{diag}\left[\boldsymbol{M}(\boldsymbol{r})\right]\bar{\boldsymbol{t}}\} \tag{8}$$

其中，$\bar{\boldsymbol{t}}$为向量$\boldsymbol{t}$的共轭。

基于语义逻辑角度进行分析，三元组$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$可由等价三元组$(s,p,o)$表示，其中$s$、$p$、$o$分别代表主语、谓语和宾语等语言逻辑。评分函数可等价表示为：

$$\phi(s,p,o)=\mathrm{Re}\left\{\left\langle \boldsymbol{w}_p,\boldsymbol{e}_s,\overline{\boldsymbol{e}_o}\right\rangle\right\} \tag{9}$$

其中，$\boldsymbol{e}_s$、$\boldsymbol{e}_o$和$\boldsymbol{w}_p$是k维复空间嵌入向量，$\boldsymbol{e}_s,\boldsymbol{w}_p,\boldsymbol{e}_o\in\mathbf{C}^k$，$\mathbf{C}^k$为$k$维向量空间，其中对复向量$\boldsymbol{e}_o$求共轭，计算埃尔米特共轭积，从而取实部计算评分，复数域的埃尔米特共轭积避免了实数域向量数量积可交换的弊端。

（3）图神经网络模型

图神经网络是构建知识图谱能力的重要工具。关系图卷积网络（relational graph convolutional network，R-GCN）作为一种特殊的图神经网络避免了图结构中所有邻居节点聚合过程具有等价性的弊端，为知识图谱的各类关系赋予不同权重，所有邻居节点的信息通过权重矩阵变换后再聚合，避免了不同关系的语义重叠。

R-GCN在面对数量庞大的关系时容易导致模型参数量激增，因此在建立网络用户知识图谱时引入分类器实现用户关系的预分类，将成千上万的关系$\boldsymbol{r}$映射到C个关系簇，假设神经网络分类器的函数为f_θ，针对特征嵌入关系向量$\boldsymbol{e}_r$，经过分类后可以得到分类分数为c_r，如下：

$$c_r=\arg\max_{c\in\{1,2,\cdots,C\}}\left[f_\theta(\boldsymbol{e}_r)\right]_c \tag{10}$$

分类簇集合为$c_r\in\{1,2,\cdots,C\}$，$|C|$为该集合的阶数，远小于关系向量空间的维度。由此可大幅降低R-GCN需要处理的关系数量。另外由于关系分类器的输出不连续，选择过程不可微，无法反向传播梯度，本文利用softmax函数实现可微化完善，此时权重矩阵引申为阶为$|C|$的权重矩阵簇集合$\{\boldsymbol{W}_1^{(l)},\boldsymbol{W}_2^{(l)},\cdots,\boldsymbol{W}_C^{(l)}\}$，对于任何一个给定的关系$\boldsymbol{r}$，可以确定概率如下：

$$p(c|\boldsymbol{r})=\mathrm{softmax}\left(f_\theta(\boldsymbol{e}_r)\right)_c=\frac{\exp\left(f_\theta(\boldsymbol{e}_r)_c\right)}{\sum_{k=1}^{C}\exp\left(f_\theta(\boldsymbol{e}_r)_k\right)} \tag{11}$$

由上述概率计算对权重矩阵加权求和可得关系$\boldsymbol{r}$的有效权重矩阵如下：

$$\boldsymbol{W}_r^{\mathrm{e},(l)}=\sum_{c=1}^{C}p(c|\boldsymbol{r})\cdot\boldsymbol{W}_c{}^{(l)} \tag{12}$$

最后可得该神经网络的消息传递函数：

$$\boldsymbol{h}_i^{(l+1)}=\sigma\left(\sum_{r\in\mathbf{R}}\sum_{j\in N_i^r}\frac{1}{c_{i,r}}\boldsymbol{W}_r^{\mathrm{e}(l)}\boldsymbol{h}_j^{(l)}+\boldsymbol{W}_0^{\mathrm{e},(l)}\boldsymbol{h}_i^{(l)}\right) \tag{13}$$

其中，$\boldsymbol{h}_i^{(l+1)}$是节点i在第$l+1$层的嵌入向量，$\sigma(\cdot)$为非线性激活函数，本文选择ReLU函数以增加模型的表达能力，$\boldsymbol{W}_r^{\mathrm{e},(l)}$为有效关系权重矩阵，由基于softmax关系分类器所构建，负责语义的线性变换，$\boldsymbol{W}_0^{\mathrm{e},(l)}$为自环权重矩阵，由模型训练时通过学习得到，$c_{i,\boldsymbol{r}}$为将聚合信息归一化的常数因子，表示节点$i$的分类分数，$N_i^r$表示在关系$\boldsymbol{r}$下，节点$i$的所有邻居节点。

（4）知识图谱构建

对于用户画像的知识图谱构建，NIRCN引入由R-GCN模型与ComplEx模型相结合的方式实现。NIRCN知识图谱训练流程如图15所示，整体流程可以分为3个阶段。

阶段1 阶段1主要包含动态子图的构建，由庞大的知识图谱训练集中构建事实三元组，而后通过反向邻居采样完成子图构建，逐层向邻居节点进行固定次数的随机采样，直到完成预设图神经网络的层数，形成计算子图。

阶段2 阶段2主要包含权重计算，分析计算子图中出现的所有关系类型$\boldsymbol{r}$，调用关系分类器f_θ对关系特征$\boldsymbol{e}_r$计算概率分布$p(c|\boldsymbol{r})$，生成数量为C的对角化权重矩阵簇$\{\boldsymbol{W}_c\}$，由概率分布加权求和，计算有效权重矩阵$\boldsymbol{W}_r^{\mathrm{e},(l)}$。

针对式（13）中的权重矩阵，为减小模型的参数量及计算复杂度，在表达能力损失可接受的情况下将权重矩阵对角化，生成维度为B（$B<C$）

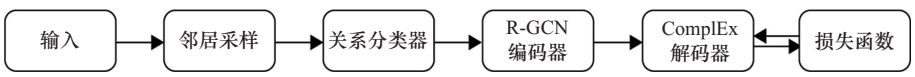


图15 NIRCN知识图谱训练流程

的对角矩阵 $\{\boldsymbol{W}_c=\mathrm{diag}\left(Q_{1c},\cdots,Q_{Bc}\right)$，加快模型收敛速度与推理速度，可以得到：

$$\boldsymbol{W}_{\boldsymbol{r}}^{\mathrm{e}(l)}=\sum_{c=1}^{C}p\left(c|\boldsymbol{r}\right)\cdot\boldsymbol{W}_c^{(l)} \tag{14}$$

其中，有效权重矩阵进一步展开可以得到：

$$\boldsymbol{W}_{\boldsymbol{r}}^{\mathrm{e}}=\mathrm{diag}\left(\sum_{c=1}^{C}p\left(c|\boldsymbol{r}\right)Q_{1c},\sum_{c=1}^{C}p\left(c|\boldsymbol{r}\right)Q_{2c},\cdots,\sum_{c=1}^{C}p\left(c|\boldsymbol{r}\right)Q_{Bc}\right) \tag{15}$$

阶段 3　阶段 3 为图神经网络编码，由 R-GCN 编码完成计算节点的嵌入表示，单层单节点聚合邻居节点信息完成信息更新，经过 L 层的信息传递与聚合后，形成顶层目标节点的实域嵌入向量。对任意一个 i 节点，在第 $l+1$ 层的嵌入向量 $\boldsymbol{h}_i^{(l+1)}$ 为：

$$\boldsymbol{h}_i^{(l+1)}=\sigma\left(\sum_{r\in\mathbf{R}}\sum_{j\in\left(N_i^r\cap N_l\right)}\frac{1}{\left|N_i^r\cap N_l\right|}\boldsymbol{W}_{\boldsymbol{r}}^{\mathrm{e}(l)}h_j^{(l)}+\boldsymbol{W}_0^{\mathrm{e}(l)}h_i^{(l)}\right) \tag{16}$$

其中，N_l 表示在邻居采样过程中被选入第 l 层计算子图的节点集合，两者的交集 $N_i^r\cap N_l$ 则表示节点 i 在关系 $\boldsymbol{r}$ 下被采样的邻居节点。

阶段 4　阶段 4 为评分计算与参数更新，将实数嵌入向量作为 ComplEx 解码器的输入的实部，虚部作为独立参数由参数库中获取，形成复数嵌入向量。通过模型后完成置信度评分，并且由损失函数计算比较模型评分与真实标签，计算总误差。最后完成端到端的反向传播优化，其中损失函数如下：

$$L=-\sum_{(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t},y)\in D}\left(y\log\left(\sigma\left(\phi\left(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}\right)\right)\right)+\left(1-y\right)\log\left(1-\sigma\left(\phi\left(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}\right)\right)\right)\right) \tag{17}$$

其中，L 为二元交叉熵损失函数（ϕ 是评分函数，在复空间中为 $\phi\left(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}\right)=\mathrm{Re}\left(\left\langle\boldsymbol{w}_{\boldsymbol{r}},\boldsymbol{e}_{\boldsymbol{h}},\overline{\boldsymbol{e}_{\boldsymbol{t}}}\right\rangle\right)$，$y$ 为与数据集中三元组 $\left(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}\right)$ 相关联的真实标签。模型通过让损失函数值尽可能小，从而达到在批次训练数据 D 的范围内尽可能逼近真实情况，对 ComplEx 模型完成反向传播和梯度更新，保持模型的训练方向正确。

4.3.2　数据过滤

数据过滤是网络安全的第一道保险，由网络获取的数据包含多种有害信息，不仅占用存储空间，同时也挑战网络法律红线，严重影响客户的服务质量。另外，用户群体的知识图谱与用户的隐私数据强关联，需要严格保护，因此统一数据面需要强有力的数据过滤能力，保障数据可靠性，真正做到敏感数据不出场，重要数据不出网。

5　智能内生实时通信网络应用实践

基于上述研究，中国移动与中兴通讯针对 NIRCN 边缘智能体原型进行了联合实践与验证，搭建融合 AI 网络平台实现网络的边缘 AI 能力上线，引入插件化的开源大模型作为能力基础，验证 NIRCN 的 AI 内生方案，检验该方案在提供个人智能助理、安全助理等创新业务时是否具有可行性。

部署 AI 模型、智能体、RAG 及 ASR 等功能组件插件化集成至边缘媒体面，由流水线插件化实现实时传输媒体流（real time transport protocol，RTP）的封装、解封、编解码等媒体处理过程，达到基础媒体处理功能与 AI 功能组合编排，为未来实现智能客服等业务提供了验证基础。

EAMP 原型平台如图 16 所示。EAMP 的插件化平台，将音/视频编解码、背景替换、虚拟头像、图片叠加、字幕合成等能力抽象为 Native 基础功能插件；将 AI 模型、ASR、RAG、智能体等 AI 能力封装为服务化插件，以容器方式配合边车管理，实现 AI 服务能力的插件化集成，并通过 Adaptor 代理插件集成到流水线。插件引擎调度执行流水线，接收 RTP 媒体流，经各插件实现 AI 加工，为业务赋能。

基于上述验证方案，本文对安全助理和智能客服两个典型场景进行验证。网络模型训练资源参数见表 1，网络模型推理资源参数见表 2。

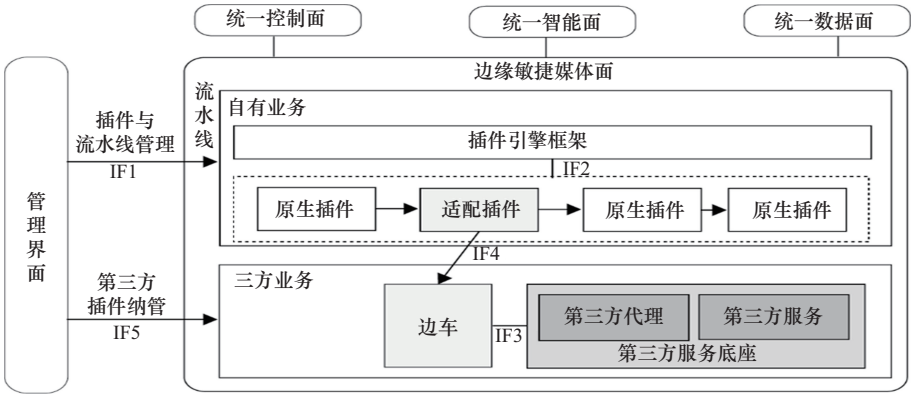


图16　EAMP原型平台

表1　网络模型训练资源参数

开源7B蒸馏模型GPU参数	模型采样率	训练时长/h
共256 GB显存（单块48 GB GDDR6 ECC显存，带宽864 Gbit/s）每块48 GB GDDR6	10 000+	2.5

表2　网络模型推理资源参数

模型	推理参数
LLM	由单张GPU完成（24 GB LPDDR5显存，带宽307.2 Gbit/s）
ASR	由单张GPU完成（24 GB LPDDR5显存，带宽307.2 Gbit/s）

（1）安全助理

安全助理业务中被叫终端B用户开通业务，主叫终端A发起音频呼叫，EAMP识别用户并开通业务，调度创建流水线。终端A上行音频复制流和终端B上行音频复制流作为流水线输入，经解封装、解码、ASR、大模型插件处理，识别呼叫通话是否存在涉诈风险。当识别到涉诈风险时，流水线触发插播流程，向终端B拉起视频呼叫，根据安全风险类型，进行针对性音频插播和字幕安全提示，安全助理处理流程如图17所示。

安全助理推理边缘时延与集中推理时延见表3，在多路并发请求下，安全助理推理的时延测量结果中，对话上下文字数为50。安全助理推理时延对比如图18所示，测试结果包含边缘推理和集中推理的对比，集中推理主要增加网络时延，通常为10~50 ms。

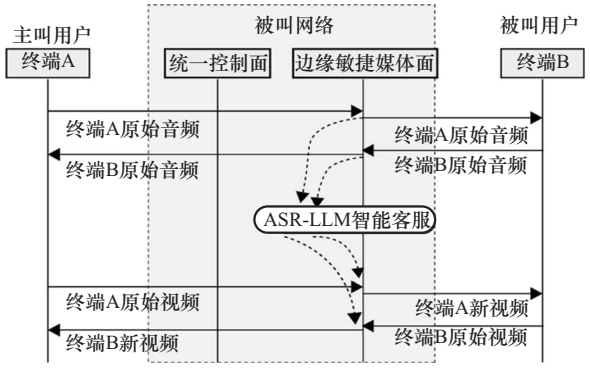


图17　安全助理处理流程

表3　安全助理推理边缘时延与集中推理时延

并发请求数	边缘部署推理时延/s	集中部署推理时延/s
1	0.21	0.25
3	0.36	0.41
5	0.49	0.54
7	0.57	0.60
9	0.77	0.81
11	0.85	0.89
13	1.02	1.05
15	1.15	1.20
17	1.28	1.33
19	1.43	1.49
21	1.51	1.56
23	1.63	1.67
25	1.72	1.76
27	1.85	1.89
29	1.97	2.03
31	2.17	2.21
33	2.29	2.34
35	2.43	2.48
37	2.69	2.75
39	2.86	2.91
41	3.13	3.19

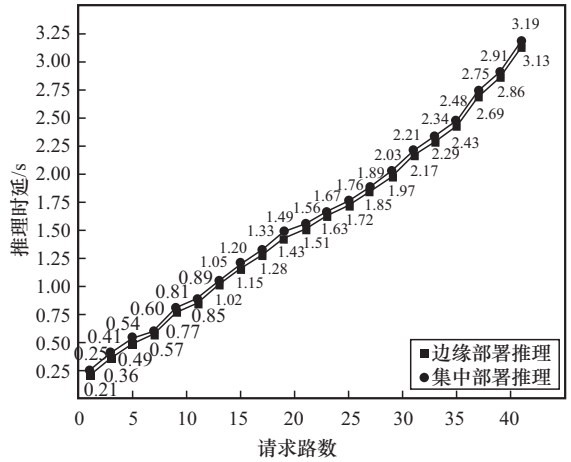


图18　安全助理推理时延对比

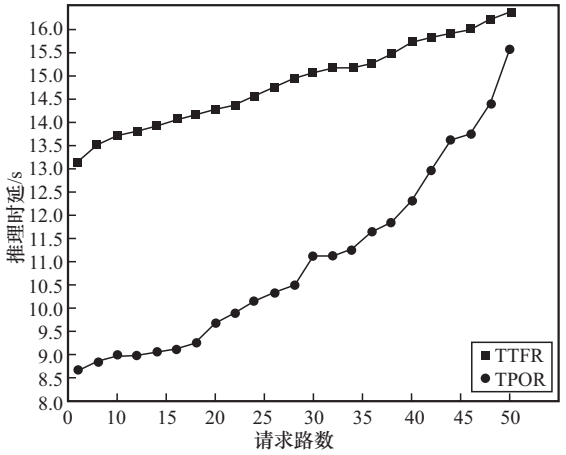


图19　安全助理端到端评估关键指标测试结果

反诈AI模型推理评估指标与含义见表4，安全助理业务中影响用户体验的指标可概括为反诈模型首次回复时间（time to first reply，TTFR）及反诈模型回复间隔时间（time per output reply，TPOR），这些指标影响系统检测用户对话是否涉嫌诈骗的时效性。

表4　反诈AI模型推理评估指标与含义

评估关键指标	含义
TTFR	首次判断对话内容是否涉嫌反诈的时间
TPOR	首次外每次判断对话内容是否涉嫌反诈的时间

设静音包数量大于N_1时，触发一次ASR识别。累计收到语句为N_2时，将对话内容送给反诈模型；后续每次收到对话数量达到N_3则向反诈模型反馈一次，其中上下文语段数量为N_4。

参数配置为N_1=8、N_2=3、N_3=1、N_4=20场景的安全助理端到端评估关键指标测试结果如图19所示。由图19可知，资源一定时，反诈模型的推理随并发请求数的增加呈现相关增长，总体相对平稳。另一方面，并发数的增加亦引起服务时延增加，边缘部署的推理系统在实验环境下具有一定优势，实际部署环境由于全国地理跨度较大且各地区网络架构存在天然区别，网络时延的增长更加明显。

（2）智能客服

用户B签约个人智能客服助聊业务。智能客服处理流程如图20所示，用户A语音起呼用户B，通话过程中触发智能客服助聊业务，拉起主被叫双向插播，同时向用户A、B插播视频，视频为智能客服提醒界面。系统实时分析用户A、B的语音，将语音转为文本，并识别文本内是否存在唤醒词、停止词，如果存在，则将用户A、B的语音文本汇总为段落提交大模型；大模型完成问题应答，输出文本，文本转为字幕叠加到插播视频上。

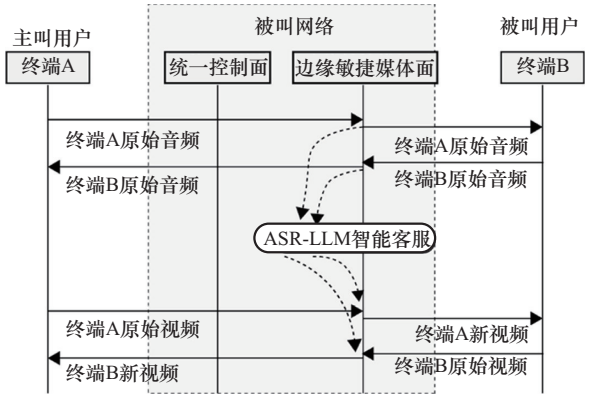


图20　智能客服处理流程

在智能客服业务流程中，唤醒词识别、ASR识别、LLM推理、字幕叠加处理是智能客服流程中时延的主要构成，提问响应首个token输出

时延。在多路并发请求下，智能客服推理时延对比如图21所示，智能客服推理时延对比数据见表5，其中对话上下文字数为150，控制输出文字数200。

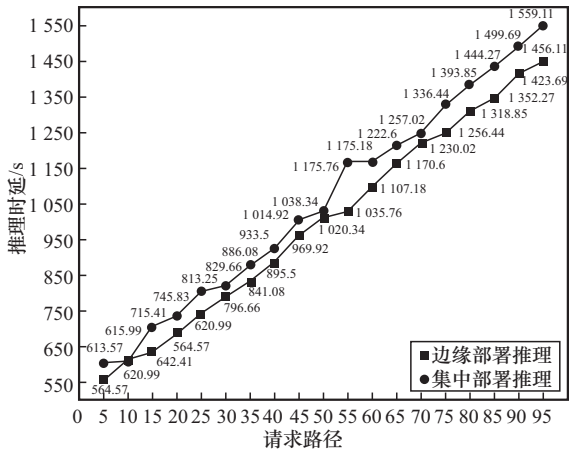


图21　智能客服推理时延对比

表5　智能客服推理时延对比数据

并发请求数	边缘部署推理时延/ms	集中部署推理时延/ms
5	564.57	613.57
10	620.99	615.99
15	642.41	715.41
20	694.83	745.83
25	752.25	813.25
30	796.66	829.66
35	841.08	886.08
40	895.50	933.50
45	969.92	1 014.92
50	1 020.34	1 038.34
55	1 035.76	1 175.76
60	1 107.18	1 175.18
65	1 170.60	1 222.60
70	1 230.02	1 257.02
75	1 256.44	1 336.44
80	1 318.85	1 393.85
85	1352.27	1 444.27
90	1 423.69	1 499.69
95	1 456.11	1 559.11

媒体引擎革新是加速实时通信与AI能力融合创新的重点，此次搭载开源大模型的实时通信网络，被赋予了语义分析、意图识别等先进AI能力。该实践验证了以边缘智能体为核心的网络智能内生，验证了NIRCN网络架构对未来用户提供应用服务的可行性，实验表明边缘网络直接向用户提供AI服务相比于传统模式存在整体优势。

6　结束语

实时通信网络边缘智能是未来6G沉浸式通信的重要发展方向和趋势，具备虚实融合、智能化、敏捷化、分布化、开放化等多个特性。本文从实时通信边缘智能角度，对6G边缘智能的架构、基本策略、智能体内生、边云推理协同等进行了详细的阐述，并且针对其中典型场景的个人用户安全助理智能体进行了探索和应用实践，未来可以进一步结合远程医疗、银行客服等行业应用进行场景拓展实时通信网络未来商用，但是相关研究也面临着诸多的挑战，还需要继续不断地探索和实践，包括未来巨量智能体如何管理、资源如何支撑、安全可靠如何保障，以及如何确保智能体通信是可信的，结果是可预期的、可解释的等。

参考文献：

[1] LIN X Q. 3GPP evolution from 5G to 6G: a 10-year retrospective[J]. Telecom, 2025, 6(2): 32.

[2] XU F M, ZHANG L Y, ZHOU Z. Interworking of wimax and 3GPP networks based on IMS [IP multimedia systems (IMS) infrastructure and services[J]. IEEE Communications Magazine, 2007, 45(3): 144-150.

[3] 金宁, 王庆扬. 基于聚类算法的6G典型应用场景研究[J]. 电信科学, 2022, 38(1): 121-131.
JIN N, WANG Q Y. Research on clustering algorithm based 6G typical usage scenarios[J]. Telecommunications Science, 2022, 38(1): 121-131.

[4] BESSIS T. Improving performance and reliability of an IMS

network by co-locating IMS servers[J]. Bell Labs Technical Journal, 2006, 10(4): 167-178.

[5] 翟振辉, 郝倩, 刘蕾, 等. 下一代新通话网络媒体面演进及关键技术研究[J]. 电信科学, 2025, 41(4): 191-198.
ZHAI Z H, HAO Q, LIU L, et al. Research on the evolution and key technology of the next generation new calling network media node[J]. Telecommunications Science, 2025, 41(4): 191-198.

[6] 杨震, 赵建军, 黄勇军, 等. 基于网络演进的人工智能技术方向研究[J]. 电信科学, 2022, 38(12): 27-34.
YANG Z, ZHAO J J, HUANG Y J, et al. Study on the direction of artificial intelligence technology based on network evolution[J]. Telecommunications Science, 2022, 38(12): 27-34.

[7] 唐博恒, 柴鑫刚. 基于云边协同的计算机视觉推理机制[J]. 电信科学, 2021, 37(5): 72-81.
TANG B H, CHAI X G. Cloud-edge collaboration based computer vision inference mechanism[J]. Telecommunications Science, 2021, 37(5): 72-81.

[8] 陈新宇, 牛娇红, 陆光辉. 基于 6G AIaaS 的分布式网络框架及技术方案[J]. 移动通信, 2023, 47(6): 110-114.
CHEN X Y, NIU J H, LU G H. Distributed network framework and technical solution based on 6G AIaaS[J]. Mobile Communications, 2023, 47(6): 110-114.

[9] 王晴天, 刘洋, 刘海涛, 等. 面向 6G 的网络智能化研究[J]. 电信科学, 2022, 38(9): 151-160.
WANG Q T, LIU Y, LIU H T, et al. Research on network intelligence for 6G[J]. Telecommunications Science, 2022, 38(9): 151-160.

[10] 李琴, 李唯源, 孙晓文, 等. 6G 网络智能内生的思考[J]. 电信科学, 2021, 37(9): 20-29.
LI Q, LI W Y, SUN X W, et al. Thinking of native artificial intelligence in 6G networks[J]. Telecommunications Science, 2021, 37(9): 20-29.

[11] STRINATI E C, DI LORENZO P, SCIANCALEPORE V, et al. Goal-oriented and semantic communication in 6G AI-native networks: the 6G-GOALS approach[C]//Proceedings of the 2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit). Piscataway: IEEE Press, 2024: 1-6.

[12] 侯祥鹏, 兰兰, 陶长乐, 等. 边缘智能与协同计算: 前沿与进展[J]. 控制与决策, 2024, 39(7): 2385-2404.
HOU X P, LAN L, TAO C L, et al. Edge intelligence and collaborative computing: frontiers and advances[J]. Control and Decision, 2024, 39(7): 2385-2404.

[13] TIAN Y Q, ZHANG Z Y, YANG Y Z, et al. An edge-cloud collaboration framework for generative AI service provision with synergetic big cloud model and small edge models[J]. IEEE Network, 2024, 38(5): 37-46.

[14] YAO J C, ZHANG S Y, YAO Y, et al. Edge-cloud polarization and collaboration: a comprehensive survey for AI[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(7): 6866-6886.

[15] YANG Z, YANG Y, ZHAO C,et al.PerLLM: personalized inference scheduling with edge-cloud collaboration for diverse LLM services[J].arXiv preprint, 2024, arXiv:2405.14636.

[16] LI M, LI Y, TIAN Y, et al. AppealNet: an efficient and highly-accurate edge/cloud collaborative architecture for DNN inference[C]//Proceedings of the 2021 58th ACM/IEEE Design Automation Conference (DAC). Piscataway: IEEE Press, 2021: 409-414.

[17] KARACHALIOS O A, ZAFEIROPOULOS A, KONTOVASILIS K, et al. Distributed machine learning and native AI enablers for end-to-end resources management in 6G[J]. Electronics, 2023, 12(18): 3761.

[18] TEERAPITTAYANON S, MCDANEL B, KUNG H T. Distributed deep neural networks over the cloud, the edge and end devices[C]//Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). Piscataway: IEEE Press, 2017: 328-339.

[19] YUAN Y L, JIAO L, ZHU K L, et al. AI in 5G: the case of online distributed transfer learning over edge networks[C]//Proceedings of the IEEE INFOCOM 2022-IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2022: 810-819.

[20] 张天成, 田雪, 孙相会, 等. 知识图谱嵌入技术研究综述[J]. 软件学报, 2023, 34(1): 277-311.
ZHANG T C, TIAN X, SUN X H, et al. Overview on knowledge graph embedding technology research[J]. Journal of Software, 2023, 34(1): 277-311.

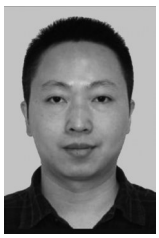
[作者简介]

王辰（1993—），男，中国移动通信有限公司研究院研究员，主要研究方向为 5G-Advanced/6G 网络架构与功能设计、实时通信网络 AI 系统功能与架构等。

白雪茜（1985– ），女，现就职于中国移动通信有限公司研究院，主要研究方向为5G-Advanced/6G、实时通信网络新技术等。

宋月（1984– ），男，中国移动通信有限公司研究院主任研究员，3GPP CT4主席，主要研究方向为5G/6G核心网及IMS网络技术。

魏彬（1983– ），男，中国移动通信有限公司研究院网络与IT技术研究所副所长，主要研究方向为5G/6G核心网标准化及商用技术、5G行业网、流量经营等。



张强（1976– ），男，中兴通讯股份有限公司高级工程师、实时通信核心网架构师、IPR总监，主要研究方向为5G/6G、实时通信新技术及核心网产品规划等。