# 服务生成网络：架构、技术与展望

黄韬[1,2]，冯立[1]，谢人超[1,2]，唐琴琴[1]，贾庆民[2]，周晓茂[2]，张语嫣[1]，李圆菊[1]，吴双[1,3]，刘韵洁[1,2]
（1.北京邮电大学网络与交换技术全国重点实验室，北京 100876；2.紫金山实验室，江苏 南京 211111
3. 都柏林大学计算机科学学院，都柏林 D04 V1W8）

**摘　要：**针对新兴业务的多维资源需求与差异服务保障，提出服务生成网络（SGN）的创新理念，旨在构建融合感知、推理、决策与优化能力的智能闭环架构，实现网络服务的智能生成、持续优化与全生命周期管理。首先介绍了生成式人工智能技术与算网融合理念；在此基础上提出了一种新型的SGN架构与工作流程；接着阐述了SGN中的关键技术；最后从应用视角出发分析了典型案例，并展望了未来研究方向，为其落地应用与持续演进提供了基础支撑。

**关键词：**生成式人工智能；算网融合；服务生成网络；算力网络；数字孪生
**中图分类号：**TN92
**文献标志码：**A
**DOI：**10.11959/j.issn.1000–436x.2025186

# Service-generated networking: architectures, technologies, and prospects

HUANG Tao[1,2], FENG Li[1], XIE Renchao[1,2], TANG Qinqin[1], JIA Qingmin[2], ZHOU Xiaomao[2], ZHANG Yuyan[1], LI Yuanju[1], WU Shuang[1,3], LIU Yunjie[1,2]
1. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
2. Purple Mountain Laboratories, Nanjing 211111, China
3. School of Computer Science, University College Dublin, Dublin D04 V1W8, Ireland

**Abstract:** To address the multidimensional resource demands and differentiated service guarantees of emerging services, the innovative concept of service-generated networking (SGN) was proposed. SGN aimed to construct an intelligent closed-loop architecture integrating perception, reasoning, decision-making, and optimization capabilities, thereby enabling intelligent generation, continuous optimization, and full-lifecycle management of network services. Firstly, the integration of generative artificial intelligence technologies and computing-network convergence concepts was introduced. Based on this, a novel SGN architecture and workflow were proposed. Then, the key technologies within SGN were elaborated. Finally, typical cases were analyzed from an application perspective, and future research directions were outlined, providing fundamental support for practical deployment and continuous evolution.

**Keywords:** generative artificial intelligence, computing-network convergence, service-generated networking, computing power network, digital twins

## 0　引言

近年来，随着大语言模型（LLM, large language model）[1]、多模态智能体[2]、生成式人工智能（GAI, generative artificial intelligence）[3]等技术的快速发展，网络系统在基础设施与服务能力等方面正面临前所未有的需求与挑战。一方面，此类新兴技术通常具备参数规模大、推理链条复杂的特点，对底层基础设施提出了大规模异构算力、低时

收稿日期：2025–08–19；修回日期：2025–10–14
通信作者：谢人超，Renchao_xie@bupt.edu.cn
基金项目：国家自然科学基金资助项目（No.92367104）
**Foundation Item：**The National Natural Science Foundation of China (No.92367104)

延高吞吐量网络以及自适应快速存储等多维资源需求[4]。另一方面，不同业务类型（如自动驾驶、实时视频生成、个性化大模型服务等）呈现出差异化服务需求，在时延、安全、可靠性保障等方面存在显著区别[5]。这些趋势共同推动了网络系统正从传统的以数据传输为中心加速迈向资源协同与服务感知的新阶段。

然而，传统云计算范式受限于集中式架构引起的拥塞与时延瓶颈，边缘计算范式的有限协同范围则难以兼顾大规模算力调度与多样化服务质量保障[6]。为此，学术界和产业界相继提出了以算力网络[7]、意图网络[8]、服务定制网络[9]为代表的多种新型网络架构，从资源协同、服务定制和意图驱动等不同维度推动了网络体系的发展。其中，算力网络侧重于计算与网络资源的统一管理与跨域协同，能够提升异构算力的可达性与利用效率，但对业务意图的表达与感知支持有限；意图网络则以用户意图为核心驱动力，能够实现意图解析与策略匹配，但缺乏对算网资源的深度协同与跨域编排支持；服务定制网络强调面向服务的差异化保障，根据任务需求下发定制化策略以实现灵活调度，但其生成机制仍依赖预定义规则，难以适应复杂多变的业务环境。

针对上述不足，伴随人工智能（AI，artificial intelligence）技术的发展，业界开始探索智能网络新范式。其中，一类典型方向是面向网络的大模型[10]，通过领域化预训练与参数高效微调，实现自然语言到网络语言的映射，辅助网络规划、配置与运维等复杂任务的自动化处理，但在与实际网络环境的资源约束、跨域协同及闭环演化的结合方面仍显不足。另一类探索是人工智能生成网络（AIGN, AI-generated networking）[11]，强调以生成模型为核心实现网络优化，能够在多目标约束下自动生成多样化的网络方案，但主要关注网络结构的优化，缺乏对服务需求的直接建模。

为此，本文提出一种面向未来的智能网络体系——服务生成网络（SGN, service-generated networking）。SGN 以网络服务的智能生成为核心，旨在深度结合算网融合架构的协同编排机制与 GAI 技术的理解生成能力，实现业务意图解析、服务策略生成、算网协同编排与网络配置执行的全流程智能管控。作为对 SGN 理念的系统化研究，本文首先提出了 SGN 的基础架构与工作流程，随后凝练了意图解析、策略生成与模型优化 3 项关键技术，接着分析了 SGN 技术在典型场景中的应用案例，最后展望了未来研究方向，为 SGN 的持续演进提供了扎实基础。

## 1　概述

近年来，算网融合理念不断演进，为多样化智能任务提供了高效可达的资源保障。与此同时，GAI 技术为网络体系的智能化提供了新的方法论支撑。将 GAI 的推理生成能力融入算网融合体系，以提升服务生成与决策效率，是推动下一代网络架构演进的重要方向。基于此，本节首先介绍算网融合背景下的 3 类新型网络架构，继而梳理 GAI 的发展脉络与网络应用，为 SGN 的提出奠定理论基础。

### 1.1　新型网络架构概述

随着新兴业务在算力、存储与网络资源方面提出多维度需求，传统以传输为中心的网络范式逐渐难以满足复杂动态环境下差异化服务保障的需求。为突破通信与计算分离的局限，产业界提出了算网融合理念，其核心思想在于通过多类资源的统一抽象与协同编排，构建面向业务的全局供给与智能调度能力。在这一理念的驱动下，围绕资源、服务和意图等不同维度，相继涌现出算力网络、意图网络和服务定制网络等新型架构，共同推动网络体系向智能化与服务化演进。

算力网络是算网融合的重要实践，其目标是实现云、边、端间算力的按需可达、协同调度与编排，从而提升异构算力的利用效率与服务响应能力。在学术界，学者提出了基于博弈[12]、负载调节[13]和上下文感知的调度方法[14]，为算力网络中的资源协同提供了技术支撑；在产业界，相关企业已在信息通告[15]、跨域调度[16]和算力路由[17]等方面展开部署，推动其从概念验证逐步走向工程化落地。然而，现有设计仍主要聚焦资源层，关注资源的感知、度量与协同，对业务意图的表达与感知支持有限。

意图网络以用户意图为核心驱动力，其基本思想是通过语义化描述业务需求，并将其解析为可执行的策略与配置。在实践中，意图网络已在意图建模[18]、网络管理[19]和策略生成[20]等方面展现出初步能力，并在部分运维与配置场景中获得应用。但

其发展仍受制于两方面限制：一是缺乏可验证的策略合成与执行保障，难以满足运营级可靠性要求；二是与算网资源的深度协同不足，跨域统一调度与闭环控制能力有限。

服务定制网络强调以用户需求为导向，通过功能模块化与资源柔性编排实现差异化服务保障。其核心在于根据不同业务场景动态定制网络功能链路与资源分配，以满足多目标、多租户环境下的个性化需求。近年来，学术界与产业界在服务匹配[21]、切片定价[22]、跨切片调度[23]及强化学习驱动[24]的优化方法等方面均取得进展。在产业界，其核心成果已在运营商、工业、卫星等典型场景落地，并在国家未来网络试验设施中实现分钟级按需定制网络能力[25]。但其编排逻辑仍依赖预定义规则，缺乏对多模态、多时空业务变化的自适应支撑，且对资源冲突、业务抢占与动态调整的支持不足，难以实现服务生命周期的闭环管理。

总体而言，算力网络以资源协同为核心，强调异构资源的统一管理与跨域调度；意图网络以语义驱动为特征，通过意图解析与策略匹配简化运维并提升用户体验；服务定制网络则以差异化保障为目标，依托策略下发与功能编排实现按需服务的灵活构建。三者分别在资源层、语义层与服务层拓展了网络体系的能力边界，推动网络从资源中心向服务中心转型。然而，它们也存在共性不足：其一，整体设计仍偏向资源或配置导向，尚未真正实现以服务为中心的系统化建模与优化；其二，缺乏从意图解析到资源调度的端到端承接机制，语义与执行之间存在脱节；其三，缺乏可验证的闭环反馈体系，策略生成与执行难以保证持续自适应与鲁棒性。

### 1.2 生成式人工智能概述

GAI是近年来人工智能领域的重要突破，以学习大规模数据分布为基础，具备自主内容生成能力[26]。其技术脉络经历了从早期规则引擎[27]向深度学习范式[28]的演进，在产业界，自2018年基于Transformer的GPT模型提出以来，大模型的持续迭代推动生成式技术逐步走向工程化应用。与此同时，生成对抗网络[29]、变分自编码器[30]、自回归模型[31]和扩散模型[32]等代表性方法也相继成熟，为GAI奠定了方法论基础。

在网络场景中，GAI的引入已成为学术界和产业界的重要探索方向，其中具有代表性的研究包括网络大模型与人工智能生成网络。网络大模型是将大语言模型的语义建模能力引入网络领域的一种实践，其核心思想是通过领域化预训练与参数高效微调，使模型能够理解用户自然语言意图，并将其映射为网络可识别的策略与配置，实现自然语言到网络配置语言的跨模态转换，从而辅助网络规划、配置与运维的自动化。然而，网络大模型在应用中仍存在不足：从意图到配置的生成过程缺乏形式化约束与验证机制，结果可靠性不足；跨域算网协同与实时性保障能力有限；同时，大模型的规模与推理开销也对边缘场景下的轻量化部署提出了挑战。

与此不同，AIGN更直接面向网络结构与协议层的设计与优化，提出了一种意图驱动的生成范式，旨在在多目标、多约束条件下自动生成可行的网络方案。其原理是在约束条件下利用扩散模型或变分建模方法生成候选解，再通过评估机制筛选最优方案，从而实现对复杂优化空间的高效探索与自适应优化。但在应用层面，AIGN同样面临挑战：缺乏对任务语义与服务需求的建模，生成方案难以与用户意图直接对齐；高维优化空间下的收敛性与稳定性尚待提升；生成结果的可解释性与可验证性不足。

总体来看，网络大模型与AIGN分别代表了GAI在语义驱动与优化驱动的两条技术路径。前者以意图解析与策略生成为核心，提升了网络管理与配置的智能化水平；后者则通过生成建模在资源调度与协议设计中实现自动优化。二者从不同维度展现了生成式方法在网络领域的潜力，共同揭示了网络从被动配置向主动生成演进的趋势。然而，这两类路径在实际应用中仍存在共性局限：一是生成结果缺乏可执行性与可验证性，端到端可靠性不足；二是跨域异构资源的协同与动态适配能力有限，难以支撑复杂业务需求；三是轻量化部署与实时推理面临挑战，闭环演化机制尚未成熟。

综上所述，新型网络架构的探索与GAI的发展为未来网络演进奠定了坚实基础。前者从资源协同、语义驱动与服务定制3个维度拓展了网络体系的能力边界，后者依托语义理解与生成优化展现了赋能网络智能化的潜力。然而，二者均难以满足复杂动态环境下的多样化需求。

## 2　服务生成网络的系统架构

为提升网络对复杂业务的自适应能力，亟须构建一种面向服务生成的新型网络架构，以业务语义为出发点，自动生成可执行的网络配置与运行机制。基于此，本文提出SGN，其核心在于依托算网融合的协同体系，融合GAI的语义理解与优化生成能力，通过深度解析用户意图与业务语义，在多维约束条件下生成算网协同的最优服务方案。通过这一融合，SGN有望突破现有研究的瓶颈，实现从用户意图到服务交付的端到端生成，推动网络体系由资源被动配置向服务主动生成的范式跃迁。本节围绕SGN的设计理念，系统介绍其整体架构与工作流程，并与几类典型范式进行对比。

### 2.1　服务生成网络的设计理念

综上所述，传统网络普遍遵循以资源为中心的设计理念，其目标主要集中于最大化算力、带宽与存储等底层资源的利用效率。该模式虽能够提升系统整体的承载能力，但难以直接映射用户意图与业务目标，导致资源最优分配与服务体验最优保障之间出现偏离。因此，未来网络亟须实现从以资源为中心向以服务为中心的范式转型。本文所讨论的服务，不仅涵盖通信、计算与存储等传统网络功能，还包括模型推理、跨域协作与多模态处理等新型智能服务。其内涵可界定为满足用户意图与实现业务目标的端到端能力组合，外延则体现为由资源抽象、功能组件与优化策略共同支撑的业务交付单元。

在此基础上，本文提出服务生成网络，通过融合资源、意图与服务，并引入GAI技术的语义理解与优化生成能力，形成闭环自治的智能服务体系。本节形象化地将SGN抽象为三元组$\mathrm{SGN}=\{R,P,S\}$，其中$R$表示多域异构资源集合，$P$表示从语义解析到策略优化的生成过程，$S$表示服务需求集合。在该描述下，SGN是一种以用户意图为驱动，通过生成过程$P$将服务需求集合$S$映射到多域异构资源集合$R$，实现从意图解析到服务交付的闭环自治的新型网络体系。

具体来说，SGN聚焦服务生成过程中的智能感知、策略生成与协同调度，具备以下4项核心特性。1)意图认知驱动：基于用户自然语言输入，识别其服务意图与性能诉求，并映射为可执行的资源需求；2)智能服务生成：面向动态网络状态与任务需求，自适应生成个性化网络服务策略，实现按需调度与快速响应；3)全局资源协同：融合多域算网资源状态信息，实现云边端一体化的资源优化与协同管控；4)自主演进优化：通过闭环反馈与持续学习机制，不断提升服务策略的有效性与适应性。

上述能力的融合赋予了SGN更高水平的智能自治与服务适配能力，标志着下一代网络架构正从功能定义向语义驱动与策略生成一体化演进，为支撑复杂多样的业务需求和动态变化的网络环境提供了关键支撑，也为后续的体系架构和工作流程提供了清晰的设计主线。

### 2.2　服务生成网络的基础架构

基于上述设计理念，本文构建了一种分层式模块化SGN体系架构，如图1所示。该架构由基础资源底座、服务生成体系和全局管控中心3个部分构成。底层资源汇聚基础承载与认知驱动资源；服务生成体系由意图驱动、策略生成、孪生映射与服务编排模块组成，实现从用户请求到服务输出的全过程生成与执行；全局管控中心包含闭环优化与安全治理模块，负责系统的动态调控、安全约束与合规审计。三层协同支撑SGN实现从自然语言意图到可验证服务的自治与可信运行。

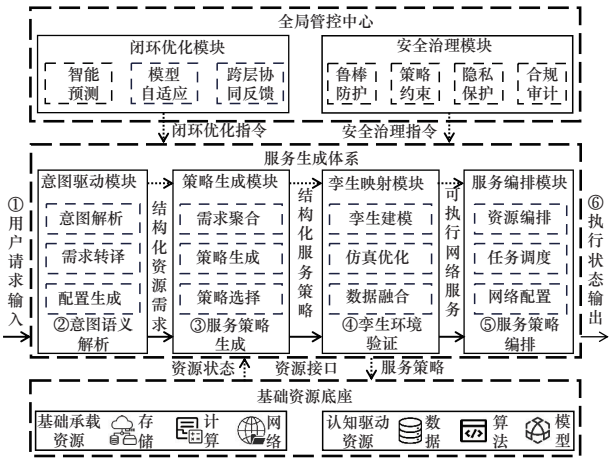


图1　SGN体系架构

1) 基础资源底座。作为SGN运行的底层支撑平台，承担“任务调度”与“服务生成”的双重职能，故可将其划分为基础承载资源与认知驱动资源两个协同子体系。其中，基础承载资源涵盖存储、计算与网络3类传统资源，分别提供多层级的存储服务、多场景的算力保障以及高效灵活的通信支撑。认知驱动资源则主要提供数据、算法

与模型3类智能资源，能够有效保障生成式内容的质量、效率与创造性。在整个SGN架构中，基础承载资源与认知驱动资源通过接口和调用机制与上层模块协同运行，为意图驱动模块提供可调用的语义知识，为策略生成模块提供多域资源的状态输入，为孪生映射模块提供真实数据与环境参数，并为服务编排与闭环优化模块提供执行能力与运行反馈，共同支撑从“物理执行”到“智能决策”的服务生成体系。

2) 意图驱动模块。位于SGN架构的最左侧，作为业务需求与网络资源之间的智能语义映射入口，承担将用户自然语言文本精准转译为结构化资源需求的功能。该模块由意图解析、需求转译与配置生成3个组件构成，融合自然语言处理、知识图谱与大语言模型等关键技术，能够实现文本语义解析、业务需求推理及资源配置生成，保障用户意图的深度理解与精准映射，并为策略生成模块提供结构化资源需求，是整个SGN架构实现资源高效分配与服务智能生成的逻辑起点。

3) 策略生成模块。位于SGN架构的决策中枢，承担多业务需求聚合与服务策略生成功能，是实现服务生成的核心环节。该模块包括需求聚合、策略生成与策略选择3个功能组件，基于时序预测、图神经网络、强化学习及大决策模型（LDM, large decision model）等技术，生成面向全局业务的结构化服务策略。由此可见，策略生成模块是SGN架构中连接抽象意图与具体服务的智能决策核心，输入侧接收来自意图驱动模块的结构化资源需求，输出侧为孪生映射模块提供网络服务策略，为网络服务的稳定性、适应性与智能性提供关键支撑。

4) 孪生映射模块。该模块通过构建真实网络的虚拟镜像，提供策略验证与仿真优化能力，由孪生建模、仿真优化与数据融合3个组件构成。依托数字孪生、故障模拟与异常检测等关键技术，在孪生环境中完成对策略生成模块输出的生成式网络服务的有效验证与优化，为服务编排模块提供可执行策略，有效降低实际网络运行风险，是SGN架构中实现理论策略向实际部署转化的重要保障环节。需要注意的是，因孪生验证开销较大，需根据业务风险等级选择性触发，保障关键场景获得风险缓释与服务保障，从而在开销与部署风险间实现最优折中。具体来说，系统基于实时的意图分析与性能监测，当策略漂移度超过10%、性能指标退化超过5%或代价函数高于阈值时，自动划定业务风险等级并触发孪生验证流程。其中，高风险业务启用全量仿真验证，中低风险业务采用抽样验证以降低开销。同时，引入A/B安全发布机制，在A版本中执行验证部署，B版本平行监测运行效果并可一键回滚，确保策略切换的安全与稳定。验证过程中产生的仿真数据将经由噪声过滤与置信度加权处理后反馈至模型更新环节，用于增量微调与闭环优化，实现策略、验证、学习的持续演进。

5) 服务编排模块。作为SGN架构中面向实际网络的关键执行模块，该模块负责将孪生映射模块输出的可执行策略转化为具体的资源配置与服务部署方案，由资源编排、任务调度与网络配置3个功能组件组成，依托资源虚拟化、网络切片、流量工程、动态路由与自动化配置等关键技术，实现对可执行服务的高效转译与部署落地，支持资源的精细化调度与业务的高性能执行。该模块在保障服务质量、提升执行效率方面发挥着核心作用，是实现SGN服务落地与运行保障的关键桥梁。

6) 闭环优化模块。位于SGN架构顶层，作为全局控制中枢，负责网络状态的动态感知与服务策略的智能优化。模块由智能预测、模型自适应与跨层协同反馈3个核心组件构成，结合流量预测、模型微调、迁移学习与跨层控制等技术，实现网络状态的实时监测、业务趋势的前瞻预测以及服务生成策略的动态调整与持续优化。整体而言，闭环优化模块对外提供对网络运行态势与业务演化的精准感知能力，对内支撑认知驱动资源的灵活更新与服务策略的智能增强，是SGN架构实现自适应调控与持续演进的核心引擎，显著提升了系统应对动态网络环境与复杂服务需求的智能化水平与鲁棒性。

7) 安全治理模块。作为系统的安全中控与合规守护单元，负责对SGN全过程进行可信约束与可解释监管。该中枢由鲁棒防护、策略约束、隐私保护与合规审计4个核心组件构成，融合语义一致性检测、约束求解、零信任访问控制与可解释审计日志等关键技术，实现服务生成全流程的安全可信运行与风险自适应防控。整体而言，安全与治理模块对外提供意图与策略的合规审查与可信验证能力，对内支撑跨域资源的安全隔离与模型行为的可追溯审计，可有效防御对抗样本与幻觉意图引发的

策略偏移，防止跨域资源抢占与策略越权行为，并通过数据主权保护与故障域隔离机制保障多域协同环境下的隐私安全与系统韧性，从而显著提升SGN在运营级场景下的安全性、可控性与可信度。

### 2.3　服务生成网络的工作流程

图1中，虚线箭头表示数据传输与反馈路径，实线箭头表示从用户请求输入到执行状态输出的全流程智能执行链路，具体包括以下步骤。

1) 用户请求输入。系统首先接收来自终端用户的自然语言请求或服务调用指令，并结合上下文信息、业务元数据等，完成初步业务意图的识别与输入规范化处理。

2) 意图语义解析。通过大模型驱动的语义理解与知识推理机制，将用户自然语言输入转化为结构化的资源需求表达，实现从抽象业务意图到可执行资源配置参数的连续映射。

3) 服务策略生成。融合全网多用户的结构化资源需求，结合当前网络状态，通过智能推理与调度算法生成具备全局最优性的服务策略，覆盖资源配置、任务部署、路径规划等关键要素。

4) 孪生环境验证。根据用户意图与风险级别选择触发，若执行，则在数字孪生环境中对生成策略进行预部署验证，模拟资源分配、任务调度与业务响应等执行过程，评估可行性、稳定性、鲁棒性等性能，并对服务策略进行优化调整。

5) 服务策略编排。将经过仿真优化的服务策略映射为底层网络配置指令，通过编排引擎完成资源调度与服务部署的自动化执行，确保服务按需上线并具备可管可控能力。

6) 执行状态输出。实时监测服务运行状态与网络性能指标，作为系统内部反馈用于后续模型优化与策略调整，并以结构化形式将执行结果反馈给用户，实现系统与用户的交互闭环。

除上述核心执行流程外，闭环优化模块基于实时监测数据与执行反馈，实现策略的自适应调整与模型的持续优化；安全与治理模块对各关键环节进行安全约束与合规审计，保障策略生成与执行的可信性；同时，系统配置了异常分支与回滚机制，当策略不可行、仿真未通过或部署未达预期时，可自动触发回退与修复流程。三者协同构成了SGN的全局控制闭环，为网络服务的智能生成、稳定运行与安全可控提供系统级保障。接下来将重点介绍支撑上述流程的3项关键技术，包括基于LLM的意图解析方法、基于LDM的策略生成机制，以及模型微调与自适应迁移技术。

### 2.4　服务生成网络与邻近范式对比

基于上述对SGN的描述，可见其在体系结构与运行逻辑上实现了范式重构，通过“感知–推理–决策–优化”闭环管控体系，打通了从业务意图到网络执行的端到端生成链路。为进一步明确SGN与算力网络、意图网络、服务定制网络、AIGN等邻近范式间的边界与互补关系，表1给出了这些范式在研究目标、输入输出、跨域协同、可验证性、闭环机制及工程接口等6个维度的系统对比。

从表1可以看出，算力网络强调算网资源的统

**表1　SGN与邻近范式对比**

| 网络范式 | 研究目标 | 输入 / 输出 | 跨域协同 | 可验证性 | 闭环机制 | 工程接口 |
|---|---|---|---|---|---|---|
| 算力网络 | 异构资源高效管控 | 输入：资源状态、任务请求<br>输出：资源分配策略 | 支持跨域管控 | 缺乏语义级与策略一致性校验 | 弱闭环，主要依赖监测反馈 | 南向接口基于算网应用程序接口，缺乏语义层标准 |
| 意图网络 | 意图到资源的映射 | 输入：用户意图描述<br>输出：任务所需资源 | 聚焦单域或局部 | 缺乏形式化验证与执行一致性 | 局部闭环，反馈限于策略匹配 | 具备北向意图接口，无标准化中间层 |
| 服务定制网络 | 差异化服务编排 | 输入：任务需求模板<br>输出：定制化服务策略 | 支持跨切片编排 | 依赖人工配置，验证机制弱 | 开环执行，缺少自适应优化 | 提供网络即服务接口，偏向静态配置 |
| AIGN | 自动生成网络结构 | 输入：网络状态、性能约束<br>输出：网络拓扑或配置 | 跨域任务感知弱 | 生成结果可行性与可解释性不足 | 离线训练，无实时反馈闭环 | 工程接口不完善，模型难嵌入运行面 |
| SGN | 智能生成网络服务 | 输入：业务意图与资源状态<br>输出：可执行的服务策略 | 云边端全域协同 | 引入孪生映射，实现策略级验证 | 全流程闭环 | 提出中间表示，支持多范式互操作 |

一调度，意图网络着重于语义到资源的转译，服务定制网络依赖规则与模板生成，AIGN关注网络结构的生成式优化。相比之下，SGN并非上述范式的增量增强，而是面向语义生成与服务自演化的体系重构，在范式层实现了跨资源、跨语义、跨模型的统一抽象与映射，其从意图到策略的规范化中间表示（I2PIR，intent to policy intermediate representation）可作为多范式协同的标准化接口，实现SGN与AIGN、SCN等体系的互操作与协同优化。

## 3　服务生成网络的关键技术

面对动态网络环境与复杂业务需求，SGN在语义解析、策略生成与模型优化等方面面临严峻挑战，为实现各层级功能模块的高效协同与动态优化，亟须引入具备强表达性、强泛化性与强自适应性的智能技术体系。基于此，本节围绕SGN架构中从用户意图解析到服务生成与闭环优化的完整流程，系统梳理了支撑SGN的3项具有代表性的关键技术。

### 3.1　基于LLM的意图解析技术

意图解析是指将用户以自然语言表达的抽象服务意图自动识别、理解并转化为结构化、可执行的完备服务需求的过程，是SGN架构中实现智能服务生成的关键起点。LLM是一类基于大规模语料库预训练的深度神经网络模型，具备强大的上下文语义建模与泛化能力，已广泛应用于语义解析与意图识别任务中。然而，与通用语义解析任务相比，SGN环境下的意图解析具有更高的复杂性与特殊需求。一方面，用户请求不仅包含显式的服务目标，还隐含资源、性能与安全等多维约束，呈现出语义跨层、意图多域的复合特征，使传统LLM难以建立业务语义与网络资源的有效映射关系；另一方面，SGN中的意图解析需具备动态环境感知能力，能够随网络状态变化自适应调整语义权重，实现从语义要素到策略参数的动态映射。

为此，针对SGN的多域、动态特征，本文引入两处优化：一是引入知识图谱增强的语义嵌入机制，融合网络拓扑、资源属性与业务本体，实现跨域语义的一致性建模；二是设计自适应Prompt模板体系，根据网络状态与任务类型动态调整语义引导结构，确保模型在多场景、多约束条件下保持高效而稳定的语义解析性能。

具体而言，该技术由意图驱动模块中的3个核心组件协同完成，如图2所示。

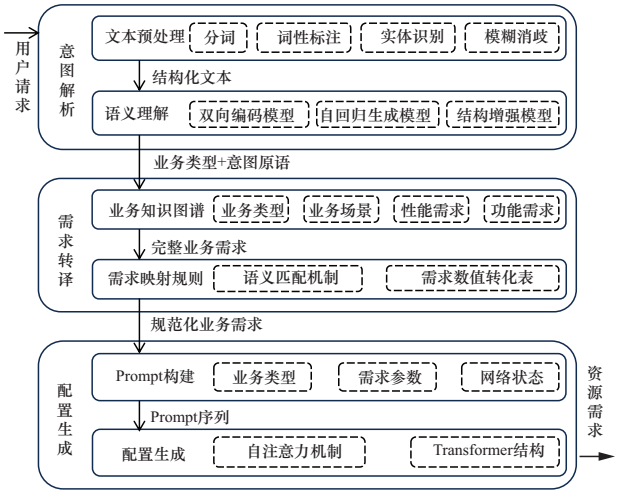


图2　基于LLM的意图解析技术

1) 意图解析。该组件主要负责对用户输入的自然语言文本进行预处理与语义理解，以抽取结构化的业务信息。首先，系统对原始文本执行分词、词性标注、命名实体识别及模糊消歧等预处理操作，提取初步的结构化文本特征；随后，基于多种预训练语言模型开展深层语义理解，从中识别业务类型与关键意图原语。例如，对于输入“我需要一个不卡顿、支持千人在线的视频会议”，系统能够识别“高清视频会议”为业务类型，“不卡顿、千人在线”为关键意图原语。该过程为后续的需求推理与服务生成提供了语义层面的基础支撑。

2) 需求转译。该组件的核心在于结合预构建的业务知识图谱与需求映射规则，生成全面且规范化的业务需求表达。系统首先基于已识别的业务类型与场景信息，推理出对应的性能需求与功能需求集合，实现对原始用户意图中隐含或遗漏信息的语义扩展；随后，调用需求映射规则，结合语义匹配机制与数值转化表，将非结构化、模糊化或非定量描述转译为形式规范、可度量的业务需求。例如，在“高清视频会议”场景下，系统可自动补全其在带宽、负载均衡等方面的需求维度，并将“不卡顿”转译为“网络时延<100 ms”，将“千人在线”转译为“总带宽≥1 Gbit/s”。

3) 配置生成。该组件以LLM技术为核心，负责将多维业务需求自动转化为结构化的资源配置方案。系统首先基于业务类型、需求参数与网络状

态，构建自适应的 Prompt 序列，明确生成任务的格式、目标维度及输出约束；随后，利用基于自注意力机制的 Transformer 结构对输入信息建模，生成可直接执行的资源配置参数。例如，在高清视频会议场景中，系统可结合具体的时延、带宽需求与当前网络状态，生成如“边缘 GPU 节点 2 个、云端 GPU 节点 4 个、边缘缓存 500 GB、网络带宽 7 Gbit/s”的配置方案，为后续的全局资源调度与服务编排提供结构化资源视图支撑。

整体而言，该技术主要基于 LLM，通过精细化的 Prompt 工程设计，实现用户自然语言输入向后续服务生成可直接调用的结构化需求的高效转化。相较于传统依赖规则匹配或关键词提取的解析方法，LLM 在语义表达与泛化能力方面表现更优，显著提升了意图解析在多样化、模糊性与隐式表达场景下的识别准确性，为 SGN 在复杂需求条件下的准确解析与智能服务生成提供了关键支撑。

### 3.2　基于 LDM 的策略生成技术

策略生成模块主要负责聚合多用户的结构化资源需求，智能生成任务调度、路由规划与资源分配等网络服务策略，是实现服务生成的核心环节。传统策略生成方法（如启发式算法与博弈算法）难以实时响应网络状态变化与用户意图更迭，近年来兴起的强化学习虽具备一定的动态适应与自主学习能力，但仍存在训练周期长、泛化能力弱、对非平稳环境敏感等问题。此外，SGN 场景下的策略生成还需同时满足服务质量、资源约束、安全隔离与多域协同等多维目标，任务间的资源竞争与依赖关系复杂，策略空间呈指数级增长，亟须在全局最优与局部可行之间实现动态平衡。

为应对上述挑战，SGN 架构引入了大决策模型作为网络服务的初始策略生成器。LDM 是一类面向策略生成任务的模型体系，其核心技术包括多模态输入融合、自适应策略生成与结构化输出表达，旨在支撑复杂、高维、非结构化的智能决策场景。在 SGN 场景中，该技术的创新主要体现在 3 个方面：一是引入资源竞争图谱建模机制，将多用户任务的冲突关系以图结构形式表达，并通过图神经网络（GNN，graph neural network）提取任务耦合特征与资源冲突路径，实现策略生成前的结构化状态建模；二是设计基于时序预测的环境感知模块，通过预测算网资源的动态演化趋势，使 LDM 具备前瞻性决策能力；三是提出经验注入机制，将 LDM 生成的可执行策略作为强化学习阶段的初始经验，为智能体提供高质量先验，缓解强化学习在冷启动、样本效率低与训练不稳定等方面的典型挑战。

基于 LDM 的策略生成技术如图 3 所示，该技术以 LDM 为核心，在策略生成模块内部由下述 3 个组件协同完成。

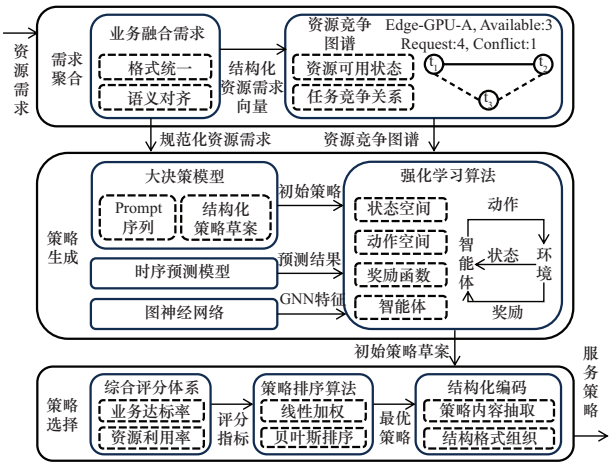


图3　基于 LDM 的策略生成技术

1) 需求聚合。该组件主要负责对意图解析模块输出的多源结构化资源需求进行统一处理与冲突建模，为后续策略生成提供语义完备的输入。首先，系统对多个用户的结构化需求进行格式规范与语义对齐，确保输入一致性；随后，构建任务间资源竞争图谱，将每个任务建模为图中的节点，任务间由于共享资源产生的冲突关系则以带权边表示。例如，若任务 $t_1$ 与 $t_2$ 均请求同一边缘 GPU 节点，且该节点资源不足，则在两任务之间建立一条表示 GPU 冲突的实边，并附加冲突强度、剩余资源等信息（如 Edge-GPU-A，Available:3，Request:4，Conflict:1），从而量化任务间的资源竞争关系与冲突程度。

2) 策略生成。该组件以大决策模型和强化学习算法为核心，负责生成并优化网络服务策略。首先，系统对融合后的业务需求与资源竞争图谱进行统一编码，结合资源状态、约束条件与历史策略经验等多模态信息构造 Prompt 序列，驱动 LDM 生成涵盖任务部署、资源分配与路径选择的结构化策略草案，并进行可行性验证。随后，引入时序预测模型（如 Informer、LSTM）预测算网资源负载的变化趋势，结合 GNN 从资源竞争图中提取任务耦合

特征与冲突路径。最终，系统将策略草案作为强化学习智能体的初始策略输入，结合预测结果与图结构特征嵌入状态空间，通过“状态–动作–奖励”机制开展策略训练与演化，生成多种可适配不同性能需求与资源约束的候选策略集合。

3) 策略选择。该组件负责从候选策略集合中选取用于孪生验证与服务部署的最优服务策略。系统首先构建综合评分体系，依据业务达标率、资源利用率等多个关键指标对策略进行定量评估；随后结合线性加权、贝叶斯排序等多策略排序算法选取较优策略；最终将选中策略编码为结构化格式，输出供孪生仿真模块调用执行。

在上述服务生成过程中，状态空间规模可近似表示为$|S| = O(NR)$，其中$N$为任务数，$R$为资源维度；动作空间规模为$|A| = O(TK)$，其中$T$表示时间片数，$K$表示可选策略数。由此，训练样本复杂度约为$O(T|S||A|)$。在采用自适应学习率与经验回放机制的情况下，模型通常可在$10^5$步迭代内达到稳定收敛。协同计算开销方面，图结构建模的复杂度约为$O(N^2)$，时序预测模块的计算开销约为$O(T\log T)$，整体在吞吐率与时延之间可实现可控权衡。上述结果表明，LDM与强化学习协同生成机制在中等规模网络中具备良好的可扩展性与训练稳定性，为后续的在线决策与闭环优化提供依据。

总体来看，相较于专注语言理解的LLM，LDM更关注策略生成的结构合理性与可部署性，具备复杂状态感知、多目标策略优化与全局搜索能力。基于LDM生成的服务策略不仅能够直接应用于孪生映射模块进行验证，确保在真实网络中的可行性与稳定性，还可为服务编排模块提供结构完备、决策可靠的策略支持，实现从感知到部署的高效联动。

### 3.3　模型微调与自适应迁移技术

在SGN架构中，大模型的部署与调用贯穿于意图解析、服务生成、仿真优化与服务编排等多个核心环节。面对网络环境的动态变化、算网资源的高度异构以及用户需求的多样化，大模型需具备良好的泛化能力，以实现对不断演进的网络状态与服务请求的快速适应。然而，传统的模型训练方法通常依赖大量的标注数据、充足的计算资源和较长的训练周期，在网络频繁变动、业务快速迭代的实际场景中，易出现训练效率低、泛化能力弱等问题。同时，在SGN环境下，模型还需在动态网络态势下保持策略稳定与快速收敛，兼顾跨域资源的差异性与反馈噪声的干扰。

为此，SGN引入了模型微调与自适应迁移两类关键优化技术，以提升LLM与LDM模型的快速更新、高效泛化与持续适应能力。考虑到SGN的动态多样性，本文进一步提出模型自适应优化机制：一是动态优化触发机制，通过实时监测输入分布漂移、资源负载变化与策略性能退化信号等多类指标，确保模型更新的及时性与必要性；二是策略判定与分级优化机制，依据任务扰动程度与性能退化幅度，动态选择微调或迁移策略，实现模型在多场景条件下的高效适应与持续进化。

模型微调与自适应迁移技术如图4所示，该技术部署于闭环优化模块中的模型自适应组件内，主要由优化触发、策略判定与优化执行3个子组件构成，具体机制如下所述。

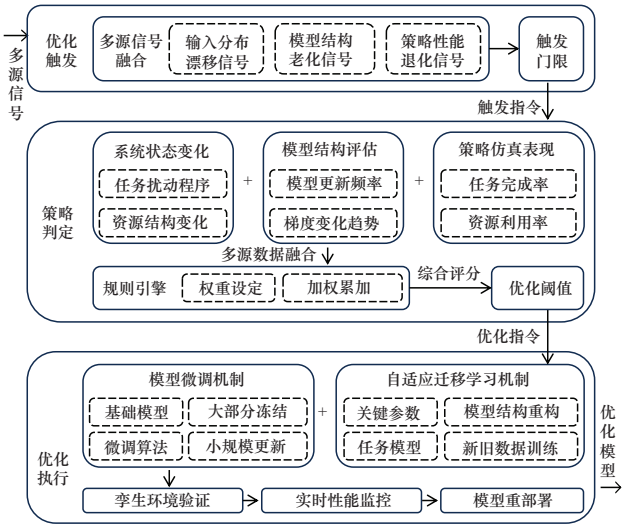


图4　模型微调与自适应迁移技术

1) 优化触发。该组件负责实时监测模型运行状态，并在检测到潜在性能退化风险时触发优化流程，以保障模型的持续有效性与服务的稳定性。系统主要感知3类信号：其一，来自意图解析模块的输入分布漂移信号，用于识别由任务类型变更、用户表达调整或资源结构重构所引起的输入语义偏移；其二，源自策略生成模块的模型结构老化信号，反映模型因参数饱和、泛化能力下降或多任务干扰增强所导致的性能退化趋势；其三，由孪生映射模块提供的策略性能退化信号，用于感知服务执

行失败、预测偏差扩大等异常行为。系统基于预设门限对上述信号进行持续感知与动态判定，以确定是否进入模型优化阶段。

2) 策略判定。当触发模型优化流程后，系统将进入策略判定阶段，以确定是采用微调机制还是采用迁移学习机制。首先，系统汇总当前业务输入与资源状态信息，评估任务扰动程度与资源结构变化，判断是否存在显著的输入偏移或拓扑重构。随后，分析模型参数更新频率、梯度变化趋势及相关性指标，判断是否存在容量瓶颈或表达能力下降。同时，对比现有策略在真实部署环境与孪生仿真环境中的执行效果，重点考察任务完成率、资源利用率等关键性能指标，量化性能退化程度。最终，系统通过规则引擎，依据各类指标对服务质量的影响程度进行加权评分。当综合评分低于阈值时，表示模型仅发生轻度退化，系统将执行快速微调策略以实现低开销、高效率的适应；当评分高于阈值时，表示模型能力已无法支撑当前业务复杂度，系统将启动迁移学习机制以进行结构重构与能力重建。

3) 优化执行。根据策略判定结果，系统将启动对应的优化执行流程。若采用快速微调机制，系统将加载模型的当前基础版本，冻结大部分参数，仅对策略输出层或关键子模块进行轻量更新，训练数据来自近期收集的样本与反馈信息，训练方法采用LoRA、QLoRA等参数高效微调技术，以降低资源消耗并降低优化时延。若采用迁移学习机制，系统将综合应用领域适应与元学习等技术，重构模型结构与关键模块，完成能力迁移与表达重建。训练数据涵盖原始知识与新场景数据，以提升模型在复杂环境下的泛化能力。优化完成后，系统将在孪生环境中验证模型性能，确保其在稳定性与可控性方面满足部署要求，最终实现策略的逐步上线部署，保障服务连续性与系统安全性。

可见，通过模型高效微调与自适应迁移机制的动态协同，SGN能够在不破坏模型基础能力的前提下，实现对环境变化的快速响应与策略能力的持续进化。该技术不仅是服务生成智能化的关键支撑，更是SGN具备环境感知、风险响应与持续优化能力的核心引擎，对实现高效、稳定、自适应的下一代智能网络架构具有重要战略意义。

综上所述，本节提出的3项关键技术共同构成了SGN智能技术体系的核心能力结构：基于LLM的意图解析技术赋予系统强表达性，能够精准理解多层次语义与差异化业务意图；基于LDM的策略生成技术体现出强泛化性，可在多域任务与复杂约束下生成可迁移的决策策略；模型微调与自适应迁移技术则强化了系统的自适应性，保障SGN中各类模型在动态环境中持续优化与稳健进化。通过三者的协同作用，SGN可以在不破坏模型基础能力的前提下，实现对环境变化的快速响应与策略能力的持续进化，形成从理解到决策再到进化的闭环智能体系，为构建高效、稳定、自演化的下一代智能网络架构奠定基础。

## 4　服务生成网络的典型应用

随着生成式服务在各大领域的快速普及与应用，车联网、工业互联网、卫星互联网等新兴场景对网络系统的智能部署与算网协同能力提出了更高要求。从算力、算法及智能体发展的维度来看，可将新型智能服务的需求归纳为3类典型应用：多域异构资源管理、AI模型弹性部署与智能体协同优化。其一，算力、网络、存储一体化的发展趋势推动了跨域资源的动态编排与统一调度；其二，模型中心化的智能服务生态，使模型训练、推理与演进的全生命周期弹性部署成为关键；其三，具身智能与群体智能的规模化落地，要求网络能够支撑多主体的实时协作与韧性控制。上述3类典型应用总结如表2所示。

可以看到，现有研究多采用特定场景下的优化策略，如基于强化学习的单域资源分配、利用混合架构提升训练效率或通过多智能体学习实现局部协同控制，虽在特定性能指标上取得显著提升，但总体上仍存在3个方面的局限：1)架构割裂，多数方法聚焦于任务层或资源层的局部优化，缺乏面向服务目标的语义抽象与统一建模；2)协同不足，难以实现跨域、跨模型、跨智能体的全局策略生成与动态适配；3)演化受限，缺乏持续反馈与自我优化机制，无法支撑生成式服务的长期自演进。

相较之下，SGN并非在现有方法之上进行功能增强，而是通过语义驱动的自适应策略生成与闭环优化机制，根本性地重构问题建模与决策生成的范式，展现出跨域、跨场景的泛化能力。本节围绕3类典型应用场景，结合具体示例系统阐述SGN的核心能力与实践价值。

表2 典型应用总结

| 应用类别 | 研究问题 | 研究方法 | 不足 | SGN优势 |
|---|---|---|---|---|
| 异构资源灵活编排 | 车联网资源分配[33] | 资源联合优化 | 场景局限，难泛化跨域 | 通过意图解析实现跨域资源语义统一，通过策略生成实现多目标动态调度，通过孪生验证与闭环优化实现持续演化，实现按需、可解释、自进化的全局资源编排 |
| | 工程设备间资源竞争[34] | 分层强化学习 | 缺乏语义建模与抽象 | |
| | 边缘网络任务卸载[35] | 知识定义边缘计算 | 静态知识库，自适应弱 | |
| | 异构任务的资源匹配[36] | 大模型增强框架 | 无闭环自优化机制 | |
| | 车载网络的任务分片[37] | 强化学习优化 | 局部最优，无全局优化 | |
| 训推模型弹性部署 | Transformer训练能耗[38] | 混合训练架构 | 缺乏系统级部署策略 | 通过LLM理解任务语义，LDM生成弹性部署方案，并结合模型微调与迁移优化实现跨阶段、跨域自适应训练与推理，突破现有方法的静态部署与被动调度瓶颈 |
| | 协同推理资源受限[39] | 联合缓存与推理 | 模型迁移效率低 | |
| | 模型推理高开销[40] | 查表推理 | 静态查表，难适配变化 | |
| | 大模型训练通信瓶颈[41] | 自动张量并行 | 缺少资源自适应机制 | |
| | 协同推理通信开销大[42] | 管线并行 | 阶段划分固定，泛化差 | |
| 多智能体高效协同 | 智能体通信负载高[43] | 异步共识算法 | 缺乏复杂任务适应性 | 通过认知语义建模实现多模态共享理解，通过LDM生成协作策略，通过闭环演化实现任务动态重构，支撑多智能体系统的自治协同与韧性演化 |
| | 多信号控制协作优化[44] | 多智能体强化学习 | 训练成本高，泛化差 | |
| | 车联网能效协同优化[45] | 多智能体强化学习 | 缺乏异构智能体建模 | |
| | 任务过程自动化[46] | 协同架构 | 无全局协同优化 | |
| | 多模态信息检索协作[47] | 智能体协同搜索 | 任务域单一 | |

### 4.1 异构资源灵活编排

在算力网络等融合基础设施中，异构资源编排旨在实现对计算、存储与网络资源的统一管理与按需调度，以提升系统资源利用率与服务质量。图5展示了一体化调度平台示例，聚合公有云、私有云、边缘节点和专用计算集群等异构算力，以及多域网络与分布式存储资源。传统编排依赖调度中心按预设权重分配资源，缺乏意图感知与策略自适应，难以应对数据与网络状态的动态变化。

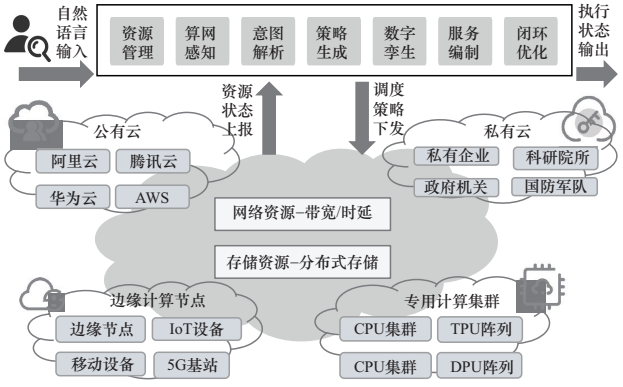


图5 基于SGN的算力网络一体化调度平台示例

针对这一问题，基于SGN的调度平台构建了七大功能模块：资源管理模块用于资源发现与注册；算网感知模块实时采集运行状态；意图解析模块理解用户需求；策略生成模块生成资源编排方案；数字孪生模块用于策略验证；服务编排模块完成任务部署与运行监控；闭环优化模块实现策略自适应演进与模型持续优化。各模块协同运行，实现了从用户自然语言输入到执行状态输出的全流程自动化。

以某企业在多地部署数据密集型AI训练任务为例，系统目标是提升训练效率并最小化跨域传输时延。用户仅需以自然语言输入需求，如“在北京、上海、广州部署高性能训练任务，要求最短时长与最低时延”，SGN系统即可自动解析任务类型、区域与性能约束。策略生成模块基于LDM生成候选方案（如侧重算力或时延优化），并通过孪生仿真评估资源利用率、通信负载与训练时间，识别瓶颈并优化参数。最终，系统结合用户偏好与反馈自动选择最优策略，下发调度指令，实现AI训练的跨域高效部署与动态优化。

综上，SGN架构下的异构资源编排不仅简化了用户操作流程，使用户不需要感知底层资源异构性，即可实现资源获取；同时有效提升了资源利用率与部署效率，降低了人工配置成本与运维复杂度。

### 4.2 训推模型弹性部署

随着LLM技术的快速演进，AI模型在训练与推理阶段的任务规模持续扩大，对算力资源的调度能力、部署效率及服务稳定性提出了差异化与动态

化的要求。训练阶段通常面向海量数据处理，需占用大量计算、存储与通信资源，而推理阶段则更加关注低时延、高并发与能效比。现有的大部分 AI 模型训练与推理往往采用中心化架构，且模型更新周期固定，易导致通信负载过重、边缘推理时延高，且难以随场景变化快速调整模型结构或更新频率。

针对该问题，以智慧交通场景中的端边云协同部署为例，如图 6 所示，端侧部署摄像头等传感设备进行原始视频采集；边缘侧由道路侧单元（RSU, road side unit）构成，集成计算与存储资源以支撑本地推理与模型更新；云侧提供高性能算力进行全量训练与策略生成。系统集成多类 AI 模型：在训练侧，云端执行全量训练构建通用模型，边缘进行区域增量训练以适应局部变化；在推理侧，RSU 部署轻量卷积网络识别目标，多模态模型检测异常，剪枝或量化模型支撑低功耗实时推理。部署管控器基于 SGN 架构解析意图、生成策略并动态优化。

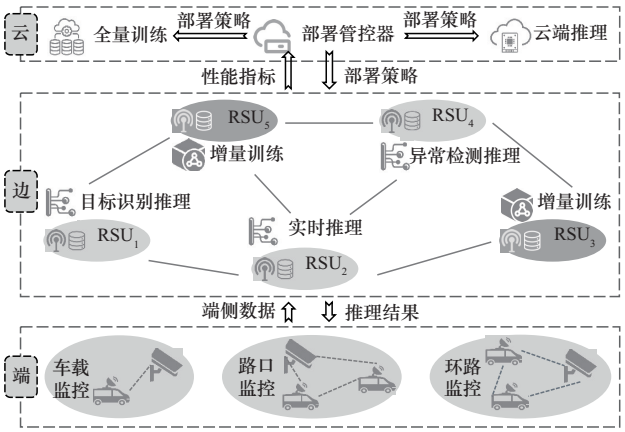


图 6　基于 SGN 的交通训推模型部署示例

在 SGN 理念下，以某城市在多个路口部署异常检测模型为例，用户目标为“确保模型准确率> 95%，当精度下降时自动更新并优化资源使用”。用户仅需通过自然语言输入需求，系统即通过 LLM 解析出结构化的训推部署意图，并结合边缘设备运行状态、网络负载与当前模型表现，生成多组候选部署策略。例如，策略 A 倾向于采用边缘增量训练以快速提升局部模型适应性，策略 B 侧重云端重训练以保障整体识别精度。候选策略将输入数字孪生模块，在仿真环境中对推理时延、训练成本、模型精度等关键指标进行评估，辅助识别资源瓶颈与潜在风险点。最终，系统结合用户偏好与仿真反馈自动选择最优部署方案，完成策略下发与模型调度，并在模型运行过程中持续迭代优化部署策略，实现 AI 模型的高效弹性部署与服务质量保障。

总体而言，SGN 通过智能化策略生成与动态自适应机制，显著缓解资源限制、策略失效与环境动态变化所带来的挑战，为大规模 AI 应用提供高性能、低时延、自演化的部署保障机制。

### 4.3　多智能体高效协同

智能体指具备感知、推理与自主行动能力的智能实体，能够在物理或虚拟环境中完成指定任务。相较于单一智能体，多智能体系统具备更强的系统覆盖能力与任务协同能力，可通过分布式协作与知识共享应对复杂的全局性任务需求。

如图 7 所示，在集成多类智能体的智能制造车间中，系统包括自动导引运输车（AGV, automated guided vehicle）、机械臂、感知单元、仓储控制与管理单元等多类型工业智能体。AGV 智能体负责运输与路径规划，机械臂智能体执行装配搬运，感知智能体采集环境状态，仓储智能体负责库存与调度，管理智能体统筹任务分发与决策。传统方法在插单、故障或物料短缺时需人工干预，响应慢且易冲突。在基于 SGN 的系统中，作为系统控制中枢的管理智能体能够依托 SGN 技术解析生产意图，自动生成协作方案，经数字孪生仿真与优化后输出最优策略，实现多智能体在资源分配、路径规划与执行反馈中的闭环协同。

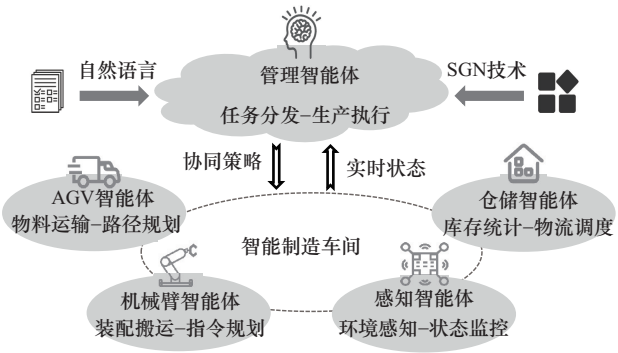


图 7　基于 SGN 的多智能体协同生产示例

以智能制造场景中的紧急插单为例，当生产管理人员提出如“保障产品 A 的紧急订单按时交付”的自然语言服务请求时，SGN 系统可自动解析出最小化订单时延的优化目标与资源约束条件。

结合车间状态数据，策略生成模块基于LDM构造多个候选协同策略。例如，策略A优先优化AGV路径以满足时限要求，策略B调整工序顺序以缓解加工瓶颈。候选策略将被输入数字孪生模块，在虚拟环境中完成多智能体仿真验证，并结合用户偏好自动遴选最优策略完成部署，涵盖任务重分配、路径重构与协作计划下发等执行环节。在运行过程中，系统持续感知各智能体状态变化，若检测到协同效率下降或资源瓶颈复现，将自动触发策略更新与微调，实现任务调度的稳定性保障与系统柔性增强。

可见，SGN通过智能策略生成与仿真验证机制，实现了多智能体系统间的感知融合、协同优化与动态重构，为构建分布式、自治型群体智能网络提供了可行路径，推动智能体系统从“单点智能”向“系统智能”的演进。

综上，SGN不仅在性能上具备可扩展性、在场景上具备泛化性，更是在体系结构层实现了网络智能的根本性重构，其核心特征包括：1)统一性，以语义建模统一异构要素，实现任务到执行的语义贯通；2)自演化性，基于生成式决策与反馈闭环，实现任务变化、策略重构、服务再生的动态演化；3)跨域协同性，构建跨域、跨场景的统一决策基座，支撑异构资源管理与智能体协同优化。值得一提的是，尽管当前示例尚未实测，但从理论视角，这种先生成后验证机制与Networks of networks[48]同源，已被证明可显著提升系统的收敛性与鲁棒性，为SGN的有效性提供了坚实依据。

## 5 未来研究方向与技术展望

SGN的提出标志着网络体系正加速从资源驱动向意图驱动、从预设编排向生成式优化、从集中调度向分布式协同演进，但SGN在复杂异构环境下的大规模部署与持续演化仍面临挑战。未来研究将聚焦三大方向：多智能体协同旨在提升策略生成与执行一致性，推动体系由单体智能向群体智能演化；大模型优化将通过模型压缩、持续学习与可解释性增强，实现低算力、高时效推理；绿色低碳节能则融合能耗感知与能源调度，协同优化性能、资源与碳排放。总体而言，这3个方面将从协同智能、生成智能与绿色智能3个维度，为SGN带来架构、算法与运行层的系统提升。

### 5.1 多智能体协同

在当前SGN架构中，分布于各功能模块之间的智能模型尚缺乏高效协同机制，信息传递存在不对称与时延，易导致对用户需求与网络状态的认知偏差，进而影响服务生成的效率与质量[49]。

因此，有必要构建面向SGN的智能体协同机制，使智能体具备信息共享、联合决策与行动协调能力。具体来说，协同机制可提升服务质量，智能体通过信息共享与联合推理更准确地识别用户意图、制定匹配策略，降低失效率与性能波动；同时强化系统的自动化部署与自适应优化，智能体可协同感知网络状态、预测业务变化并动态配置资源，提升利用率并控制成本；此外，还能促进SGN实现分布式、自组织运行，增强在云、边、端融合及大规模物联网环境下的可扩展性与灵活性。

智能体协同的当前研究主要聚焦于异构智能体间的通信与协作协议[50]，在SGN中引入协同仍面临挑战：一是信息共享与隐私保护矛盾突出，需强化访问控制与加密机制；二是缺乏统一评估体系，难以实现策略全局优化与行为量化对比。未来应重点研究高可扩展协同架构、安全高效共享机制、完善评估体系及多场景泛化能力。

### 5.2 大模型优化

SGN中的大模型在意图解析、策略推理与任务调度中发挥核心作用，但仍面临计算开销大、响应时延高、更新困难和可解释性不足等问题，限制了系统的响应速度、可扩展性和部署灵活性。

为此，大模型优化技术通过结构重构、训练机制改进与推理加速[26]，能有效提升SGN在多场景、多约束下的性能与实用性。一方面，可更高效地响应用户请求。例如，通过引入模型压缩与加速机制，意图解析与策略生成过程得以显著提速，有效支撑多样化业务需求的即时响应与动态适配。另一方面，可显著降低对计算与存储资源的依赖，从而提升整体资源利用率并降低运维成本。此外，增强的可解释性也有助于定位异常与提升系统稳定性。

然而，在边缘计算资源受限、业务状态频繁变化的场景下，模型压缩需在精度与效率之间权衡，增量更新面临稳定性风险，可解释性也尚难满足面向运维与策略评估的透明性需求。因此，未来优化可聚焦高效压缩、稳定更新与轻量解释3个方面，构建适配SGN需求的智能模型支撑体系，推动其

在复杂业务环境中的高效演进与可靠部署。

### 5.3　绿色低碳节能

在SGN支持复杂网络服务智能生成的过程中，系统整体能耗仍面临严峻挑战，主要体现在节点能源利用效率偏低、网络切片中资源空置严重以及可再生能源接入比例不足等方面。

因此，亟须引入绿色低碳设计理念，构建面向能源感知与智能调控的运行机制。将SGN与绿色低碳网络[51]结合，可通过对可再生能源与负载需求的智能匹配，有效降低系统运行能耗与碳排放，同时促进能耗感知、资源建模与任务调度等关键技术的协同演进。为进一步量化能效表现并实现优化可控，SGN在策略生成与服务编排过程中采用能耗与服务等级目标的联合优化目标函数 $\min(\alpha\text{Delay}+\beta\text{Power}+\gamma\text{Carbon})$，其中 $\alpha$、$\beta$ 和 $\gamma$ 是用于平衡时延、功耗与碳排放目标的权重。系统通过多层能耗计量接口实时感知策略变更与边云迁移的能耗变化，建立能效基线以评估不同调度方案的节能收益。例如，在异构算力调度与模型推理迁移等典型场景中，基于能耗感知的资源复用与低负载迁移策略可显著提升整体能效，为SGN的绿色演化与可持续运行提供量化支撑。

面向未来，绿色低碳理念在SGN中的落地仍面临多项技术挑战。首先，需建立高精度能耗感知体系。其次，能源调度机制应兼顾业务负载动态性、设备能耗特性与可再生能源波动。最后，在保障性能的前提下，需平衡服务质量与绿色能源利用率。未来研究可聚焦精细化能耗监测、智能能源调度算法与绿色能源融合接入，推动SGN迈向高效、低碳、智能的下一代网络体系。

## 6　结束语

本文首先设计了SGN的系统架构，随后凝练了3项关键技术，接着围绕3类典型场景阐述了SGN的核心能力与应用价值，最后展望了SGN的未来发展路径与挑战。总体而言，SGN的提出不仅是对未来网络的体系创新，更标志着网络服务范式的根本转变。通过将生成式智能深度嵌入网络底座，SGN推动网络由信息传输基础设施向具备理解、生成与优化能力的智能系统演进，为智能、高效、可持续的网络服务提供新范式支撑。未来，SGN有望成为支撑分布式AI基础设施建设、促进多智能体自治协同、引领网络系统智能化演进的核心平台。后续研究将聚焦4个方向：跨模态、多语言下的鲁棒意图解析；复杂业务目标的策略生成可解释与安全保障；高动态环境下的异构资源高效调度；面向隐私、安全与能效的多目标联合优化，为SGN的持续演进与工程落地提供理论与技术支撑。

## 参考文献：


[1] ARISOY E, CHEN S F, RAMABHADRAN B, et al. Converting neural network language models into back-off language models for efficient decoding in automatic speech recognition[J]. ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(1): 184-192.

[2] XIA B Y, XIE P J, WANG J K. Smart mobility with agent-based foundation models: towards interactive and collaborative intelligent vehicles[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(7): 5130-5133.

[3] BANH L, STROBEL G. Generative artificial intelligence[J]. Electronic Markets, 2023, 33(1): 63.

[4] WEI W T, GU H X, WANG K, et al. Multi-dimensional resource allocation in distributed data centers using deep reinforcement learning[J]. IEEE Transactions on Network and Service Management, 2023, 20(2): 1817-1829.

[5] HAN W D, WANG X B. Diverse and differentiated QoS provisioning for 6G communications via demand-aware prioritization and DEI-based resource allocation[J]. IEEE Transactions on Wireless Communications, 2024, 23(12): 18346-18362.

[6] REN J K, YU G D, HE Y H, et al. Collaborative cloud and edge computing for latency minimization[J]. IEEE Transactions on Vehicular Technology, 2019, 68(5): 5031-5044.

[7] 贾庆民, 胡玉姣, 张华宇, 等. 确定性算力网络研究[J]. 通信学报, 2022, 43(10): 55-64.
JIA Q M, HU Y J, ZHANG H Y, et al. Research on deterministic computing power network[J]. Journal on Communications, 2022, 43(10): 55-64.

[8] MEKRACHE A, KSENTINI A, VERIKOUKIS C. Intent-based management of next-generation networks: an LLM-centric approach[J]. IEEE Network: the Magazine of Global Internetworking, 2024, 38(5): 29-36.

[9] 黄韬, 张晨, 肖玉明, 等. 服务定制网络体系架构的设计与思考[J]. 通信学报, 2024, 45(2): 1-17.
HUANG T, ZHANG C, XIAO Y M, et al. Design and research of service customized networking architecture[J]. Journal on Communications, 2024, 45(2): 1-17.

[10] HUANG Y D, DU H Y, ZHANG X Y, et al. Large language models for networking: applications, enabling techniques, and challenges[J]. IEEE Network, 2025, 39(1): 235-242.

[11] HUANG Y D, XU M R, ZHANG X Y, et al. AI-generated network design: a diffusion model-based learning approach[J]. IEEE Network, 2024, 38(3): 202-209.

[12] ZHANG Y D, WANG P, WANG Q, et al. Evolutionary game-based adaptive DT association and transfer for wireless computing power networks[J]. IEEE Transactions on Green Communications and Networking, 2025, 9(2): 670-683.

[13] GAO D X, XIA N, LIU X Q, et al. Joint load adjustment and sleep management for virtualized gNBs in computing power networks[J]. IEEE Transactions on Wireless Communications, 2024, 24(3): 2067-2082.

[14] 马云霄, 吴忠辉, 徐祖云, 等. 算力网络中面向计算重用的任务调度优化[J]. 通信学报, 2023, 44(11): 129-142.

MA Y X, WU Z H, XU Z Y, et al. Optimization of task scheduling for computing reuse in computing power network[J]. Journal on Communications, 2023, 44(11): 129-142.

[15] 卫敏, 赵倩颖, 唐静, 等. 算力网络信息通告技术研究综述[J]. 通信学报, 2025, 46(9): 17-31.
WEI M, ZHAO Q Y, TANG J, et al. Summary of research on information notification technology of computing power network[J]. Journal on Communications, 2025, 46(9): 17-31.

[16] 王路, 董昕, 高允翔. 中国联通算网一体调度技术创新与实践[J]. 通信世界, 2025(11): 24-26.
WANG L, DONG X, GAO Y X. Innovation and practice of integrated scheduling technology for computing and network in China Unicom[J]. Communications World, 2025(11): 24-26.

[17] 鲍雨, 肖明坤. 算力网络技术在天翼视联的应用研究[J]. 江苏通信, 2025,41(3): 19-22, 45.
BAO Y, XIAO M K. Research on the application of computing power network technology in TianyiVision[J]. Jiangsu Communication, 2025, 41(3): 19-22, 45.

[18] 罗森林, 邵思源, 赵智洋, 等. 网络流量对抗样本行为意图建模防御方法研究[J]. 北京理工大学学报, 2025, doi: 10.15918/j.tbit1001-0645.2025.059.
LUO S L, SHAO S Y, ZHAO Z Y, et al. Behavior intent modeling of network traffic adversarial examples for defense[J]. Transactions of Beijing Institute of Technology, 2025, doi: 10.15918/j.tbit1001-0645.2025.059.

[19] NIE J, LI H C, LI Y, et al. Towards intent-based network management: intent-optimized cross-shard transactions and malicious node detection in blockchain system[J]. IEEE Internet of Things Journal, 2025, PP(99): 1.

[20] 郭令奇, 张蕾, 杨红伟, 等. 知识定义的意图网络策略生成技术[J]. 北京邮电大学学报, 2024,47(3): 36-41.
GUO L Q, ZHANG L, YANG H W, et al. Intent-based network policy generation for knowledge definition[J]. Journal of Beijing University of Posts and Telecommunications, 2024, 47(3): 36-41.

[21] HUANG T, TAN S, TANG Q Q, et al. Coordinating services and networks with NaaS tickets towards service customization in distributed clouds[J]. IEEE Transactions on Mobile Computing, 2024, 23(12): 10984-10999.

[22] FAN W H, LI X W, TANG B H, et al. MEC network slicing: Stackelberg-game-based slice pricing and resource allocation with QoS guarantee[J]. IEEE Transactions on Network and Service Management, 2024, 21(4): 4494-4509.

[23] RAVINDRAN S, CHAUDHURI S, BAPAT J, et al. Novel adaptive multi-user multi-services scheduling to enhance throughput in 5G-advanced and beyond[J]. IEEE Transactions on Network and Service Management, 2024, 21(2): 2323-2338.

[24] 陈赓, 齐书虎, 沈斐, 等. 面向B5G多业务场景基于D3QN的双时间尺度网络切片算法[J]. 通信学报, 2022, 43(11): 213-224.
CHEN G, QI S H, SHEN F, et al. Dual time scale network slicing algorithm based on D3QN for B5G multi-service scenarios[J]. Journal on Communications, 2022, 43(11): 213-224.

[25] 刘韵洁, 黄韬, 陶高峰, 等. 面向未来的服务定制网络(SCN)架构[J]. 科技纵览, 2025(2): 50-52.
LIU Y J, HUANG T, TAO G F, et al. Future-oriented service customized network (SCN) architecture[J]. Technology Review, 2025(2): 50-52.

[26] HE Y, FANG J C, YU F R, et al. Large language models (LLMs) inference offloading and resource allocation in cloud-edge computing: an active inference approach[J]. IEEE Transactions on Mobile Computing, 2024, 23(12): 11253-11264.

[27] CAVNAR W B, TRENKLE J M. N-gram-based text categorization[C]// Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. S.l.:s.n., 1994: 1-14.

[28] LIU Y D, GOEBL J, YIN Y D. Templated synthesis of nanostructured materials[J]. Chemical Society Reviews, 2013, 42(7): 2610-2653.

[29] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014(2): 2672-2680.

[30] KIPF T N, WELLING M. Variational graph auto-encoders[J]. arXiv Preprint, arXiv: 1611.07308, 2016.

[31] YU J, XU Y, KOH J Y, et al. Scaling autoregressive models for content-rich text-to-image generation[J]. arXiv Preprint, arXiv: 2206.10789, 2022.

[32] CROITORU F A, HONDRU V, IONESCU R T, et al. Diffusion models in vision: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10850-10869.

[33] LIU Y F, BI Y G, LIU Y H, et al. Service satisfaction-aware adaptive service migration and resource allocation in vehicular edge computing[J]. IEEE Transactions on Mobile Computing, 2025, PP(99): 1-17.

[34] NING P F, WANG H W, TANG T, et al. Diffusion-based deep reinforcement learning for resource management in connected construction equipment networks: a hierarchical framework[J]. IEEE Transactions on Wireless Communications, 2025, 24(4): 2847-2861.

[35] ZHANG C C, HE Q, LI F L, et al. Intelligent task offloading and resource allocation in knowledge defined edge computing networks[J]. IEEE Transactions on Mobile Computing, 2024, 24(5): 4312-4325.

[36] SUI L Y, ZHANG K, WU F, et al. Large models for resource allocation in edge computing power networks[J]. IEEE Network, 2025, 39(4): 82-89.

[37] CAO D, HUANG S R, GU N, et al. Co-optimization of partial offloading and resource allocation for multi-user tasks in vehicular edge networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2025, 36(12): 2537-2548.

[38] DAHAL S, DHINGRA P, THAPA K K, et al. HpT: hybrid acceleration of spatio-temporal attention model training on heterogeneous many-core architectures[J]. IEEE Transactions on Parallel and Distributed Systems, 2025, 36(3): 407-421.

[39] XU X Y, FENG G, LIU Y J, et al. Joint inference offloading and model caching for small and large language model collaboration[J]. IEEE Transactions on Mobile Computing, 2025, PP(99): 1-16.

[40] WANG Q P, JIANG S Q, YANG Y F, et al. Efficient and adaptive diffusion model inference through lookup table on mobile devices[J]. IEEE Transactions on Mobile Computing, 2025, 24(9): 8729-8746.

[41] LI S W, LAI Z Q, LI D S, et al. Oases: efficient large-scale model training on commodity servers via overlapped and automated tensor model parallelism[J]. IEEE Transactions on Parallel and Distributed Systems, 2025, 36(9): 1828-1840.

[42] YE S Y, OUYANG B, ZENG L K, et al. Jupiter: fast and resource-efficient collaborative inference of generative LLMs on edge devices[C]// Proceedings of the IEEE INFOCOM 2025 - IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2025: 1-10.

[43] CHEN R F, LI J, CHEN Y T, et al. An asynchronous consensus method with low communication traffic and high efficiency for distributed multi-agent scheduling[J]. IEEE Transactions on Mobile Computing, 2025, PP(99): 1-14.

[44] LI L L, LI Y F, ZHANGS H, et al. Efficient cooperative mechanism for distributed multi-agent traffic signal control[J]. IEEE Transactions on Mobile Computing, 2025, PP(99): 1-17.

[45] LIN Y, XIAO L Q, TAO Y Y, et al. Multi-agent computing-energy-efficiency optimization in vehicular edge computing: non-cooperative versus cooperative solutions[J]. IEEE Transactions on Wireless Communications, 2025, 24(7): 5461-5476.

[46] GUAN Y C, WANG D, CHU Z X, et al. Intelligent agents with LLM-based process automation[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2024: 5018-5027.

[47] SHI Y X, XU M, ZHANG H M, et al. A learnable agent collaboration network framework for personalized multimodal AI search engine[C]// Proceedings of the 2nd International Workshop on Deep Multimodal Generation and Retrieval. New York: ACM Press, 2024: 12-20.
[48] DAVIS J Q, HANIN B, CHEN L, et al. Networks of networks: complexity class principles applied to compound AI systems design[J]. arXiv Preprint, arXiv: 2407.16831, 2024.
[49] YANG Y P, SHEN B, GE X H, et al. Dynamic event-triggered cluster consensus of multi-agent systems via PSO-GA co-design[J]. IEEE Transactions on Automation Science and Engineering, 2025, 22: 11505-11518.
[50] WANG Y, GUO S, PAN Y, et al. Internet of agents: fundamentals, applications, and challenges[J]. arXiv Preprint, arXiv: 2505.07176, 2025.
[51] ZHANG Y M, SUN P K, JI X Q, et al. Low-carbon economic dispatch of integrated energy systems considering full-process carbon emission tracking and low carbon demand response[J]. IEEE Transactions on Network Science and Engineering, 2024, 11(6): 5417-5431.

[作者简介]

**黄韬**（1980– ），男，重庆人，博士，北京邮电大学教授，主要研究方向为网络系统架构、算网融合、确定性网络等。

**冯立**（1999– ），女，福建莆田人，北京邮电大学博士生，主要研究方向为算力网络、网络人工智能、边缘计算等。

**谢人超**（1984– ），男，福建南平人，博士，北京邮电大学教授、博士生导师，主要研究方向为信息中心网络、工业互联网、算力网络、边缘计算、无服务器计算等。

**唐琴琴**（1994– ），女，广西桂林人，博士，北京邮电大学副研究员、硕士生导师，主要研究方向为算力网络、网络人工智能、网络孪生等。

**贾庆民**（1990– ），男，山东泰安人，博士，紫金山实验室研究员，主要研究方向为算力网络、确定性网络、边缘智能、工业互联网等。

**周晓茂**（1993– ），男，安徽阜阳人，博士，紫金山实验室研究员，主要研究方向为边缘智能、算网自智、生成式人工智能等。

**张语嫣**（2002– ），女，山西长治人，北京邮电大学硕士生，主要研究方向为算力网络、边缘智能等。

**李圆菊**（2003– ），女，云南昆明人，北京邮电大学硕士生，主要研究方向为算力网络、智能体协同等。

**吴双**（2000– ），女，江苏盐城人，北京邮电大学博士生，都柏林大学访问学者，主要研究方向为算力网络、数字孪生。

**刘韵洁**（1943– ），男，山东烟台人，中国工程院院士，主要研究方向为未来网络体系架构、网络融合与演进等。