

华中科技大学

研究生（文献综述与选题）报告

题目：面向边缘视觉大模型推理的 WebRTC-WiFi 跨层协
同优化机制研究

学 号 M202573151

姓 名 宓呈祥

专 业 信息与通信系统

指 导 教 师 高雅珣

院(系、所) 电子信息与通信学院

华中科技大学研究生院制

填 表 注 意 事 项

- 一、本表适用于攻读硕士学位研究生选题报告、学术报告,攻读博士学位研究生文献综述、选题报告、论文中期进展报告、学术报告等。
- 二、以上各报告内容及要求由相关院(系、所)做具体要求。
- 三、以上各报告均须存入研究生个人学籍档案。
- 四、本表填写要求文句通顺、内容明确、字迹工整。

目 录

1	引言	1
2	国内外研究概况及发展趋势	3
2.1	实时视觉大模型通信需求研究	3
2.2	面向视频类 AI 的视频传输优化研究	4
2.3	WebRTC 优化研究	5
2.4	Wi-Fi 实时业务优化研究	7
2.5	总结与讨论	8
3	研究内容与方法	10
3.1	系统模型	10
3.2	面向实时视觉模型业务的闭环联合仿真平台	10
3.3	Wi-Fi、WebRTC 和视觉模型任务性能的跨层关联分析	13
3.4	面向视觉模型服务质量的 WebRTC-Wi-Fi 联合优化方法	15
3.5	基于联合仿真平台的多场景实验验证	17
4	可行性分析与预期成果	20
4.1	前期工作与研究基础	20
4.2	预期成果	20
5	研究计划	22
6	总结	23
	参考文献	24

一 引言

近年来,生成式人工智能(Generative Artificial Intelligence, GenAI)业务呈现爆发式增长。截止 2026 年 3 月,中国日均 Token 调用量已超过 140 万亿次,较 2025 年底增长约 40%;AI 搜索流量预计将于 2028 年超越传统搜索,对无线网络格局产生深远影响^[1]。与传统多媒体业务相比,GenAI 业务呈现出显著不同的流量特征——上行方向,受多模态大模型发展驱动,持续上传视频、图像等多模态数据,呈现高上行带宽、强突发性和多模态交织的特点;下行方向则主要表现为流式文本或语音回应的短包传输,呈现高频小包特征。上述特征已引起 IEEE 802.11 等无线标准化组织的高度关注^[1-3]。与此同时,智能服务正从传统文本交互向持续感知、在线理解和实时决策演进,基于视觉语言模型(Vision-Language Model, VLM)和视觉语言动作模型(Vision-Language-Action Model, VLA)的实时视觉模型业务逐渐从云端下沉至边缘,支撑具身智能、机器人控制、实时视频对话和远程巡检等应用。作为边缘场景的核心无线接入技术,Wi-Fi 网络的传输能力直接决定了上述业务的部署可行性与服务质量。为此,IEEE 802.11 标准化组织已成立 AI Offload 研究小组,探索将计算密集型 AI 推理任务卸载到边缘 Wi-Fi 接入点及其他支持 Wi-Fi 的边缘计算设备^[4]。由于终端侧算力和能耗受限,此类业务通常需要持续上传摄像头视频至边缘 AP,基于推理结果做出决策^[5,6]。在这一过程中,视频传输质量直接决定模型“看到”了什么、何时看到,从而深刻影响推理准确率和响应及时性;而 WebRTC 作为承载实时音视频交互的关键协议,其与 Wi-Fi 之间的协同效率将直接影响传输效果。因此,研究面向视觉模型推理质量的 WebRTC-Wi-Fi 视频传输优化机制,具有紧迫的理论意义和实际价值。

然而,实时视觉模型推理对视频传输提出了不同于传统视频会议、云游戏和虚拟现实等业务的新要求。传统实时视频应用主要面向人类观看体验,通常关注清晰度、卡顿率、端到端时延和交互流畅性;而实时视觉模型业务的最终目标并不是让人“看得更清楚”,而是让模型在限定时间内获得对任务有用的视觉信息,从而提高任务准确率、响应及时性和推理稳定性。不同视觉任务对视频传输的敏感因素可能存在显著差异:例如,文字识别、细粒度目标识别等任务可能更依赖图像清晰度和空间细节,动态事件理解和机器人感知任务可能更依赖帧率、时序连续性和视频帧新鲜度。因此,传统网络服务质量(Quality of Service, QoS)或面向人类观看体验的体验质量(Quality of Experience, QoE)指标,未必能够准确反映实时视觉模型的服务质量。现有实时视觉大模型研究虽然已经开始关注连续视频输入、在线理解和低时延推理^[7-9],但对无线网络状态、WebRTC 传输行为和视频质量变化如何影响模型推理效果,仍缺少系统分析。

与此同时,现有 WebRTC 优化研究和 Wi-Fi 实时业务优化研究也难以直接满足上述需求。一方面,现在还少有研究面向模型服务质量。另一方面,WebRTC 和 Wi-Fi 协议分属不同网络层,两者存在天然的不透明性。在 Wi-Fi 网络中,信道竞争、无线重传、链路速率变化和队列积压会显著影响时延和丢包表现,但 WebRTC 仅依赖端到端反馈时,难以及时准确地区分真实网络拥塞和无线链路波动,也难以利用 Wi-Fi 内部状态优化拥塞控制和 FEC (Forward Error Correction) 策略。而 WiFi 同样无法感知 WebRTC 的帧类型和帧级截止时间,难以做优先级调度和过期包清理。现有研究大多分别从 WebRTC 侧或 Wi-Fi 侧进行优化,仅有少量工作尝试引入跨层信息。并且现有引入跨层信息的工作通常只引入了单方面的跨层感知,并以一个协议为主体进行优化,尚未形成 WebRTC、Wi-Fi 之间的双向协同机制^[18,20]。也就是说,WebRTC 通常不了解 Wi-Fi 内部的信道竞争和队列状态,Wi-Fi 也不了解 WebRTC 数据包所属视频帧的重要性、截止时间和冗余属性,二者难以围绕视觉模型 QoE 进行联合优化。

针对上述问题,本文拟研究面向实时视觉模型推理质量的 WebRTC-Wi-Fi 跨层协同传输方法。

- 首先,本文拟基于 ns-3 和 ns3-ai 构建闭环联合仿真平台,将真实视频客户端、WebRTC 传输机制、ns-3 Wi-Fi 网络仿真和实际部署的 VLM/VLA 推理服务连接起来,实现视频上传、无线传输、模型推理和结果返回的完整业务闭环。
- 其次,本文拟利用该平台系统分析 Wi-Fi 网络状态、WebRTC 视频传输行为和视觉模型推理质量之间的关联关系,明确不同网络状态和视频传输指标对模型准确率、响应时延和信息新鲜度的影响。
- 在上述分析的基础上,本文拟设计 WebRTC-Wi-Fi 双向跨层协同优化方法:一方面使 WebRTC 能够感知 Wi-Fi 的信道竞争、无线重传、链路速率和队列状态,并据此优化码率选择、发送节奏和 FEC 策略;另一方面使 Wi-Fi 能够感知 WebRTC 提供的视频帧类型、帧级截止时间、数据重要性和冗余属性,并据此进行优先级调度、关键帧保障和过期帧清理。

与传统更关注人类视觉体验的优化机制不同,本文提出的 WebRTC-Wi-Fi 跨层协同优化方法将以视觉模型推理质量为核心指标,并尝试建立两者之间的双向协同机制,旨在提升实时视觉模型推理的准确率和响应速度,从而更好地支撑边缘 AI 业务的发展。

二 国内外研究概况及发展趋势

面向实时视觉模型推理的无线传输优化涉及模型任务、视频传输和无线接入等多个层面的协同。现有研究尚未形成完整的 WebRTC-Wi-Fi-视觉模型闭环优化框架,但已从四个相关方向展开探索:一是实时视觉大模型的流式推理与通信需求研究,主要关注连续视频输入、在线理解和边缘推理;二是面向视频类 AI 的视频传输优化研究,主要关注如何根据视频内容和模型反馈调整帧率、分辨率、传输区域与带宽分配;三是 WebRTC 优化研究,主要围绕拥塞控制、速率调节和 FEC 等传输机制展开;四是 Wi-Fi 实时业务优化研究,主要关注无线状态感知、资源调度和低时延保障。本章按照上述顺序梳理相关工作的主要进展,并在此基础上归纳现有研究在跨层协同和模型任务导向优化方面的不足。

2.1 实时视觉大模型通信需求研究

随着多模态大模型的发展,视觉理解任务正在从离线图像理解和离线视频问答,逐渐转向连续视频流上的在线感知、实时问答和具身交互。在传统离线视频理解任务中,模型通常可以在完整视频已经采集和存储之后再行推理;而在实时视觉大模型业务中,视频数据仍在不断产生,模型需要在有限时延内持续接收最新视觉输入,并结合当前场景、历史上下文和用户指令生成响应。因此,这类业务不再只是模型结构或推理效率问题,而是同时依赖视频采集、编码压缩、无线传输、边缘计算和模型推理之间的协同。本节将从流式模型能力、在线视频理解评测和边缘具身应用三个方面梳理相关研究,并进一步说明实时视觉模型业务对无线视频传输提出的新需求。

从模型能力演进来看,近期研究已经开始从离线视频理解转向连续视频输入和在线响应能力。VideoLLM-online 提出 Learning-In-Video-Stream 框架,使视频大语言模型能够在持续接收视频流的过程中进行时间对齐理解和实时对话^[7]。LiveVLM 和 StreamingVLM 等工作进一步面向实时或近无限视频流提出流式 KV 缓存和紧凑记忆机制,以支持模型在长时间视频输入下维持较低响应延迟^[22,23]。VideoScan 和 Mobile-VideoGPT 则分别从帧级语义压缩、关键帧选择和视觉 token 裁剪等角度降低视频理解过程中的计算和显存开销^[24,25]。这些研究共同说明,实时视觉大模型已经不再只关注“能否理解完整视频”,而是开始关注“能否在视频持续到达的过程中及时理解和响应”。相应地,模型对输入视频帧的新鲜度、连续性和有效信息保留提出了更高要求。

从评测基准和数据集的发展来看,实时视觉大模型的研究目标也正在由离线长视频理解转向在线、流式和时间敏感的视频理解。Video-MME、LongVideoBench 和 MLVU 等基准主要推动多模态大模型处理更长、更复杂的视频输入,要求模型能够在长时间、多事

件和多语义线索的视频中进行检索与推理^[26–28]。在此基础上，OVBench、OVO-Bench 和 StreamingBench 进一步强调在线视频理解中的当前场景感知、历史信息记忆、未来事件预测和时间点触发响应能力^[8,29,30]。这些基准的变化表明，实时视觉模型的能力评估已经不再局限于“看完整段视频后回答问题”，而是逐渐转向“边看边理解、边交互边响应”。因此，视频帧是否能够及时、连续、稳定地传输到模型侧，将直接影响模型在线感知和实时响应能力的发挥。

从应用场景来看，实时视觉大模型正在向机器人、移动终端、可穿戴设备和具身智能系统扩展，这进一步强化了其通信需求。RT-2 和 OpenVLA 将视觉语言模型扩展到机器人动作决策和通用操作策略，使模型能够根据视觉观测和语言指令生成控制行为^[31,32]；EgoLife 面向 AI 眼镜和第一人称日常生活助手构建长时间、多模态视频数据集，反映了视觉模型持续感知环境并提供实时辅助的应用趋势^[33]。由于 VLM/VLA 通常具有较高计算和显存开销，许多应用需要借助边缘或云端算力。Ahmad 等人在 O-RAN/MEC 环境中部署 VLM，并设计基于 WebRTC 的多模态数据流传输管线，将机器人采集的视频传输到边缘节点进行实时感知推理^[5]。这些应用说明，实时视觉模型的输入往往需要经过无线网络传输到边缘或云端，传输时延、连续性和视频质量会直接影响模型获得的视觉信息及其输出质量。

总体而言，实时视觉大模型正在从离线视频理解走向连续视频流上的在线感知、实时问答和具身交互，并越来越多地依赖边缘或云端推理。与面向人类观看的视频业务相比，这类业务更关注任务准确率、响应及时性和输入信息新鲜度，不同任务对清晰度、帧率和时序连续性的敏感程度也可能不同。因此，传统网络 QoS 或人类视频 QoE 指标未必能够准确反映模型任务质量。现有研究的重点仍主要集中在模型结构、视觉 token 压缩、流式记忆和推理效率，对无线网络状态及传输自适应行为如何影响模型任务质量缺少系统分析。

2.2 面向视频类 AI 的视频传输优化研究

除实时视觉大模型本身的流式推理研究外，近年来也有一类工作开始从“机器理解视频”而非“人类观看视频”的角度重新设计视频传输机制。早期面向视频分析的研究已经指出，传统视频流传输通常以人类观看体验为目标，难以直接适配目标检测、识别和场景理解等 DNN 推理任务。Pakha 等人较早提出应面向分布式视觉分析重新设计视频流传输协议，使其优化目标从人眼感知质量转向分析准确率与带宽开销之间的权衡^[34]。AWSStream 进一步面向广域流式分析场景，通过在帧率、分辨率和数据压缩等多个维度上进行自适应降级，在带宽受限时维持应用级分析准确率和数据新鲜度^[35]。Chameleon 则针对大规模视频分析系统，利用时间和跨摄像头相关性动态选择视频分辨率、帧率和模

型配置,以在资源消耗和分析准确率之间取得平衡^[36]。这些工作表明,视频类 AI 业务中的传输优化不能简单等同于传统视频 QoE 优化,而应结合模型任务对输入质量的敏感性进行设计。

在边缘视频分析场景中,后续研究进一步将视频传输与服务端模型推理过程结合起来。DDS 提出服务端驱动的视频流传输机制,先传输低质量视频流,由服务端 DNN 根据检测结果决定哪些区域需要以更高质量重新传输,从而在降低带宽消耗的同时保持或提升推理准确率^[37]。CASVA 则针对实时视频分析中的配置选择问题,根据视频内容、网络带宽和推理准确率之间的关系,自适应选择分辨率、帧率和模型配置,以提升实时分析任务在带宽变化下的稳定性^[38]。Runespoor 面向机器人边缘-云视觉分析提出内容自适应数据率控制机制,在受限带宽下改善机器人视觉任务的尾部准确率^[6]。这些研究共同说明,面向视频类 AI 的传输优化已经逐渐从单纯压缩视频码流,发展为结合视频内容、模型反馈和任务准确率的联合优化问题。

近期,多模态大模型视频助手的出现进一步推动了面向 MLLM 的实时视频通信优化研究。Wu 等人指出,在 AI Video Chat 场景中,实时通信的一端不再是人类用户,而是部署在云端或边缘侧的多模态大模型,因此传统 RTC 中以人眼感知质量、播放流畅性和卡顿率为核心的 QoE 指标不再完全适用,系统更应关注 MLLM 响应准确率和端到端交互时延^[39]。在此基础上,Artic 提出面向 MLLM Video Assistant 的 AI-oriented RTC 框架,通过响应能力感知的自适应码率控制限制不必要的视频码率,并利用上下文感知流式传输将有限码率分配给与当前问题最相关的视频区域,同时构建 DeVIBench 评估 RTC 引起的视频退化对 MLLM 准确率的影响^[40]。这类工作与传统 WebRTC 优化的差异在于,其优化目标已经从“让人看得清楚和流畅”转向“让模型及时获得足够有用的视觉信息”。

总体来看,面向视频类 AI 的传输优化已经从视频参数自适应发展到结合内容、模型反馈和任务准确率的联合控制。然而,现有工作大多侧重应用层视频配置或端到端传输策略,对 Wi-Fi 信道竞争、链路层重传和队列积压等底层状态考虑不足,因而尚未充分揭示无线传输行为对视觉模型任务质量的影响。

2.3 WebRTC 优化研究

WebRTC 是当前低时延实时音视频通信中的重要技术框架,被广泛应用于低延迟流媒体传输、视频会议多码率优化、边缘计算通信和实时监控等场景^[41–45]。与传统视频点播业务不同,WebRTC 业务通常需要在严格时延约束下持续传输音视频流,因此其优化目标不仅包括提高链路利用率和视频清晰度,还包括控制排队时延、抖动、丢包恢复开销和帧级传输稳定性。围绕这些问题,现有 WebRTC 优化研究主要可以分为三类:一是以

GCC 为代表的端到端拥塞控制研究,二是面向无线动态网络的状态感知与速率控制增强研究,三是面向视频 QoE 的码率、分辨率、帧率和 FEC 等多维联合优化研究。这些研究为实时视频传输提供了重要基础,但其优化目标主要仍然面向视频传输质量和人类观看体验,尚未充分考虑实时视觉模型推理业务的特殊需求。

WebRTC 优化的基础是端到端拥塞控制机制,其代表性方案为谷歌拥塞控制 (Google Congestion Control, GCC)。Carlucci 等人系统研究了 GCC 机制,指出 GCC 通过端到端单向时延变化、丢包反馈和带宽估计判断网络拥塞状态,并结合接收端延迟控制器和发送端丢包控制器对视频发送速率进行调节^[10]。该类方法的优点在于不依赖网络内部设备支持,具有较好的端到端部署性,因而成为 WebRTC 实时媒体传输中的典型拥塞控制方案。然而, GCC 的核心输入仍然是端到端传输反馈。当网络状态快速变化,尤其是在 Wi-Fi 和蜂窝等无线网络中出现信道竞争、链路速率变化和突发排队时,仅依赖端到端时延和丢包信号容易产生状态判断滞后或误判,从而影响视频码率调整的及时性和稳定性。

针对无线网络环境中端到端反馈不足的问题,已有研究开始尝试增强 WebRTC 或实时视频传输系统对动态网络状态的适应能力。Carlucci 等人研究了 GCC 在 Wi-Fi 网络中的表现,指出 Wi-Fi 信道中断结束后的分组突发到达会影响 GCC 延迟估计,并提出基于分组聚合的改进方法,以降低无线突发行为对拥塞判断的干扰^[11]。Dai 等人进一步引入数据驱动方法,通过学习端到端时延、发送速率和网络状态之间的关系,提高实时视频通信在无线链路中的适应能力^[12,13]。Yu 等人针对移动网络中 RTT 和丢包反馈难以准确反映视频播放体验的问题,提出 Mustang 以改善高抖动场景下的实时视频控制效果^[46]。Karimi 等人提出 Vidaptive,指出 GCC 等实时视频速率控制机制在网络快速变化时可能存在响应滞后,并通过解耦传输控制与编码器目标码率调节,提高动态网络中的链路利用率和时延表现^[14]。此外, PBE-CC、ABC 和 Mortise 等工作虽然并非全部专门针对 WebRTC 设计,但从无线跨层拥塞控制和网络感知参数优化角度说明,引入物理层测量、显式网络反馈或网络状态感知参数调优,有助于减少仅依赖端到端探测带来的控制滞后^[47-49]。这些研究共同表明,无线状态感知能够改善动态网络中的实时视频传输性能,但多数方法仍主要从传输层或单向网络状态感知角度出发,尚未形成 WebRTC 与 Wi-Fi 之间的双向跨层协同。

除拥塞控制外,WebRTC 优化研究还逐渐从单一发送速率控制扩展到面向用户体验质量 (Quality of Experience, QoE) 的多维联合优化。传统 GCC 主要依据端到端反馈调节目标码率,但实际视频质量还与分辨率、帧率、量化参数、FEC 冗余和视频内容复杂度等因素密切相关。Li 等人提出 Mamba,将 WebRTC 中的自适应码率问题扩展为多维决策问题,通过多智能体强化学习协同调整量化参数、分辨率和帧率,以提升视频会议场景下的用户体验质量^[50]。Smirnov 等人面向 5G 网络下的 WebRTC 视频流,利用深度强化学

习进行实时速率控制,以改善动态网络条件下的 QoE^[51]。在丢包恢复方面, Lee 等人提出 R-FEC, 利用强化学习联合调整视频码率和 FEC 冗余, 在丢包恢复能力、带宽开销和视频质量之间进行权衡^[52]; Cheng 等人提出 ABRF, 将自适应码率控制与 FEC 控制联合建模, 根据网络状态和丢包模式动态调整码率与纠错冗余^[53]。这些工作说明, WebRTC 优化已经不再局限于拥塞控制本身, 而是逐渐发展为码率、分辨率、帧率、冗余和视频 QoE 之间的多维联合控制问题。

综上所述, 现有 WebRTC 研究提供了码率调整、发送节奏控制、FEC 配置和编码参数调节等控制手段, 也表明更准确的网络状态有助于改善动态网络中的实时视频传输。然而, 现有方法主要依赖端到端反馈或单向网络状态感知, 难以直接获得 Wi-Fi 内部的信道竞争、无线重传和队列积压等状态; 其优化目标也主要是清晰度、流畅度、卡顿率和端到端时延, 尚未充分考虑模型任务准确率、输入信息新鲜度和响应及时性。

2.4 Wi-Fi 实时业务优化研究

Wi-Fi 是移动终端、机器人、增强现实、边缘智能和室内实时视频传输中常见的无线接入方式。与有线网络相比, Wi-Fi 链路更容易受到信道竞争、无线干扰、终端数量变化、物理层速率自适应、链路层重传和队列积压等因素影响。总体来看, 相关研究主要可以分为三类: 一是面向实时业务的 Wi-Fi QoS 保障机制, 二是 Wi-Fi 配置、视频业务特征与应用 QoE 之间关系的测量分析, 三是通过 Wi-Fi 与应用层的跨层感知进行优化设计, 分为将 Wi-Fi 状态引入应用层和将应用层信息引入 Wi-Fi 两大类。

早期 Wi-Fi 实时业务优化主要依托 IEEE802.11e 中的增强型分布式信道接入机制 (Enhanced Distributed Channel Access, EDCA)。EDCA 通过为不同业务类型配置不同的 AIFS、竞争窗口和 TXOP 参数, 使语音、视频等实时业务能够获得高于普通数据流的信道接入优先级。然而, 传统 EDCA 参数多为静态配置, 难以适应实时视频流的码率波动、帧大小变化和无线竞争环境变化。随着 Wi-Fi6/7 的发展, 研究者进一步从 OFDMA 资源分配、TXOP 限制、预先信道接入以及多链路传输等角度支持实时业务。Bankov 等人面向 Wi-Fi7 实时应用研究了限制非实时业务信道占用时间和预先信道接入等机制, 以满足实时业务对时延和丢包率的严格要求^[54]。Adame 等人则从时间敏感网络 (Time-Sensitive Networking, TSN) 角度分析 IEEE802.11be 支持低时延高可靠业务的机制, 为 Wi-Fi7/8 面向实时业务的标准演进提供了参考^[55]。这类研究说明, Wi-Fi 侧已经具备通过接入优先级、空口资源分配和低时延机制保障实时业务的技术基础, 但其优化粒度通常仍停留在业务类别或链路层参数层面, 难以直接反映具体视频帧或模型推理任务的实时价值。

在基础 QoS 机制之外, 相关研究进一步关注 Wi-Fi 底层配置、视频业务特征和应用 QoE 之间的关系。Kulkarni 等人研究了信道宽度、接入类别、优先级队列和硬件配置等

Wi-Fi 参数对体积视频流应用 QoE 的影响,指出底层 Wi-Fi 配置会显著影响视频业务的稳定性和最终体验^[15]。Casasnovas 等人构建了基于 Unity Render Streaming、WebRTC 和 Wi-Fi6 的边缘/云端虚拟现实实验系统,分析 WebRTC 视频帧发送批次、Wi-Fi 分组聚合和空口占用之间的关系,发现 WebRTC 视频帧的周期性批量发送会影响 Wi-Fi 空口效率和完整视频帧到达时延^[16]。这类工作表明,实时视频业务的性能不仅由平均吞吐量决定,还与视频帧发送节奏、MAC 层聚合方式、信道占用和完整帧到达时延密切相关。因此, Wi-Fi 实时业务优化需要从单纯链路层参数配置,逐步转向结合应用层帧结构和业务体验质量的跨层分析。

近年研究进一步开始为应用层引入 Wi-Fi 物理层和 MAC 层状态,以辅助实时业务传输控制。Du 等人提出 Cupid,重新审视 Wi-Fi 网络中的拥塞控制问题,利用接收端 Wi-Fi 物理层测量信息感知空口占用、并发传输和物理层速率等状态,从而更准确地判断 Wi-Fi 瓶颈并进行速率调节^[17]。Yi 等人提出 Athena,通过从物理层到应用层的高分辨率跨层测量,分析无线网络事件对视频会议 QoE 的影响,并指出当前视频会议拥塞控制算法看到的往往只是网络状态的表象,缺乏对无线底层事件的可解释感知^[18]。这些研究说明,仅在应用层被动观察端到端反馈已经难以满足低时延实时业务的稳定性需求,为应用层主动引入 Wi-Fi 底层状态有助于解释和改善无线环境下的视频传输性能。

除向上层提供 Wi-Fi 状态外,部分研究通过给 Wi-Fi 引入业务信息,以实现更加精准的优先级调度和拖尾时延控制。Shen 等人提出 LAW,引入 WebRTC 的尾帧信息,并通过重塑队列结构和细粒度传输机会控制降低 WebRTC 应用中的尾帧时延和卡顿率^[20]。Guo 等人提出 BLADE,面向云游戏和 XR 等下一代实时通信业务,通过动态调整竞争窗口来降低 Wi-Fi 分组尾时延和视频卡顿率^[19]。Gong 等人提出 Flick,在 Wi-Fi 侧引入视频帧边界等信息,以进行帧感知分组调度,从而降低完整视频帧的到达时延^[21]。这些工作共同表明,为 Wi-Fi 层引入应用层的帧级信息,可以实现更细粒度的调度和优先级控制,从而改善实时视频业务的传输稳定性和体验质量。

总体来看,Wi-Fi 实时业务优化已经从静态 QoS 配置发展到无线状态感知和帧级跨层调度。现有工作证明了信道竞争、链路层重传、发送节奏和队列管理会显著影响实时业务,也说明 Wi-Fi 侧可以利用部分上层信息进行优先级控制和队列管理。不过,多数方法仍面向视频会议、云游戏和 VR/XR 等人类感知业务;已有跨层方法通常只引入局部、单向的信息,尚未同时结合 WebRTC 传输状态与视觉模型任务需求形成双向闭环。

2.5 总结与讨论

综上所述,四类相关研究分别为本文问题提供了不同基础。实时视觉大模型研究推动了连续视频理解、在线响应和边缘推理的发展,明确了新型智能业务对视频时效性和

任务质量的需求；面向视频类 AI 的传输研究证明了视频内容、传输参数和模型任务质量之间存在耦合关系；WebRTC 研究提供了拥塞控制、速率调节和冗余配置等端到端自适应机制；Wi-Fi 实时业务研究则揭示了无线配置、发送节奏和 MAC 层竞争对应用体验的影响，并提供了网络侧调度与低时延保障手段。

然而，从本文研究问题出发，现有工作之间仍缺少进一步的协同。首先，已有研究一方面尝试将 Wi-Fi 物理层或 MAC 层状态提供给 WebRTC 及其他实时传输机制，以改善拥塞判断和速率控制；另一方面也尝试将视频帧边界、尾帧状态等应用层信息提供给 Wi-Fi，以指导队列管理和资源调度。但这两类工作通常分别优化信息传递的单一方向，尚未形成 WebRTC 与 Wi-Fi 相互感知、联合决策的双向协同机制。其次，面向视频 AI 的研究已经提出了多种由视频内容、模型反馈和任务准确率驱动的端到端传输优化方法，但其控制对象大多集中在帧率、分辨率、区域选择和带宽分配等应用层参数，较少进一步结合 WebRTC 的拥塞控制、发送节奏与 FEC 机制以及 Wi-Fi 的信道竞争、链路层重传和队列调度。因而，现有研究尚未将模型任务需求、WebRTC 传输控制和 Wi-Fi 底层状态贯通起来，形成以视觉模型任务质量为目的的跨层闭环优化框架。

针对上述不足，本文拟开展三个方面的研究：首先，基于 ns-3 和 ns3-ai 构建闭环联合仿真平台，连接真实视频客户端、WebRTC 传输、Wi-Fi 网络仿真和实际 VLM/VLA 推理服务；其次，系统分析 Wi-Fi 状态、WebRTC 传输行为和视觉模型任务性能之间的关联，明确传统 QoS 与视频 QoE 指标在实时视觉模型业务中的适用边界；最后，设计面向视觉任务需求的 WebRTC-Wi-Fi 跨层协同传输方法，综合利用无线链路状态、端到端传输状态和模型任务反馈，对视频传输和无线资源进行联合调整。

上述研究将把模型任务目标、WebRTC 传输控制与 Wi-Fi 底层状态连接起来，使无线传输优化由传统网络质量或人类观看体验进一步面向模型任务效果，并为多模态智能体、具身智能和边缘视觉推理场景下的无线网络优化提供实验与方法基础。

三 研究内容与方法

本文围绕 Wi-Fi 网络中实时视觉模型推理业务的传输优化问题展开研究。具体而言, 本文研究内容包括四个层次: 首先构建能够运行真实视觉模型业务的系统级联合仿真平台, 为后续实验提供可控、可重复的环境; 其次利用该平台分析 Wi-Fi 网络状态、WebRTC 传输行为和视觉模型任务性能之间的跨层关联; 然后在关联分析的基础上设计面向视觉任务需求的 WebRTC-Wi-Fi 联合优化方法; 最后通过在不同场景下的实验验证所提出方法的有效性。接下来将详细介绍每个层次的研究内容和方法。

3.1 系统模型

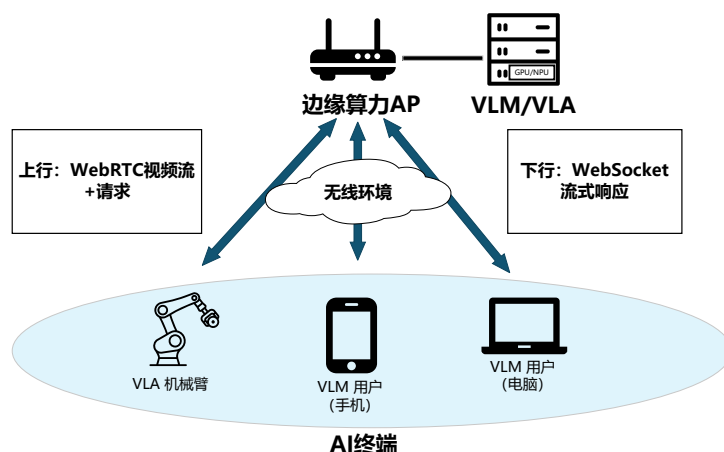


图 3-1 面向实时视觉模型推理的 WebRTC-Wi-Fi 联合优化系统模型

3.2 面向实时视觉模型业务的闭环联合仿真平台

3.2.1 研究内容

实时视觉模型业务具有视频持续输入、模型在线推理和结果实时返回等特点, 如果仅使用预先生成的流量模型或离线网络轨迹, 难以准确反映网络变化对模型任务结果的影响, 并且考虑到可复现性和可分析性, 本研究最终决定使用半实物仿真来支撑后续的分析研究。现有仿真平台难以同时满足真实业务闭环、Wi-Fi 细粒度仿真和可复现实验需求, 因此本研究决定首先建立一个能够同时包含真实业务逻辑、WebRTC 传输过程和 Wi-Fi 网络仿真的闭环仿真实验平台。

该平台以 ns-3 作为 Wi-Fi 网络仿真基础,用于模拟客户端与服务端之间的无线通信过程。ns-3 能够较细粒度地描述 Wi-Fi 网络中的信道竞争、调制编码方式、无线重传、队列变化和多终端接入等现象,适合用于研究无线网络状态对实时业务的影响。同时,平台使用 ns3-ai 模块提供的共享内存通信机制,将 ns-3 仿真进程与外部真实应用进程连接起来,使仿真网络不再只是处理预设模型产生的流量,而是可以承载真实应用产生的视频流和业务请求。本研究力求将该框架构建成一个通用的联合仿真平台,因此并不假定 GenAI 客户端和 GenAI 服务端的具体实现,而是提供标准化的接口。这意味着 GenAI 客户端可以是任何能够发送请求的仿真器(例如 Maniskill 等机器人仿真器)或实物,GenAI 服务端可以是任何能够接收请求并返回结果的模型——这意味着本研究既可以部署 LLM/VLM 等真实模型,也可以使用现成的流量模型,甚至可以支持 LLM ServingSim 等专门的模型推理仿真器。通过这种设计,平台能够适应不同的研究需求和实验场景,具有较强的灵活性和扩展性。

3.2.2 技术路线

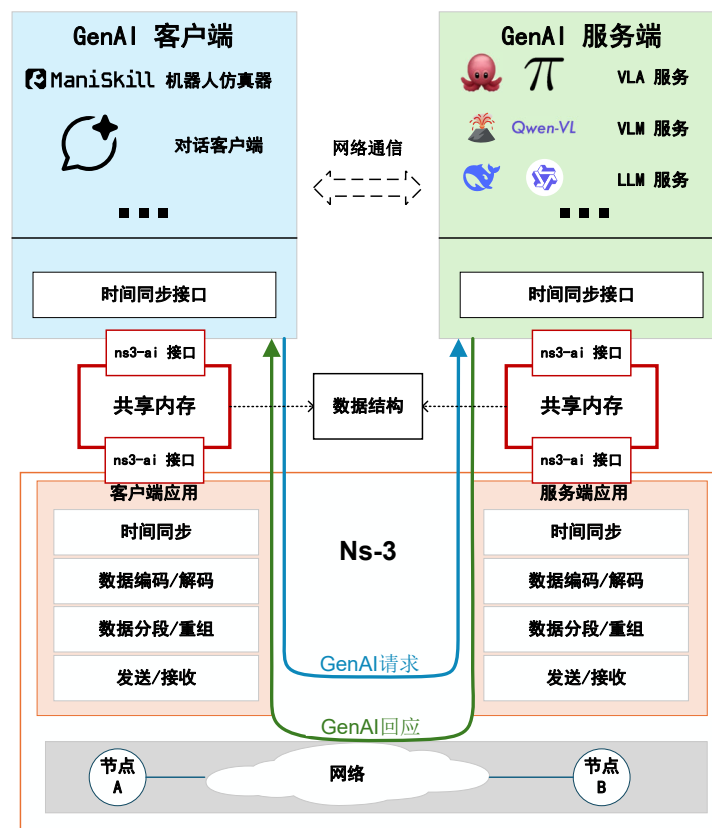


图 3-2 面向 GenAI 业务的联合仿真平台

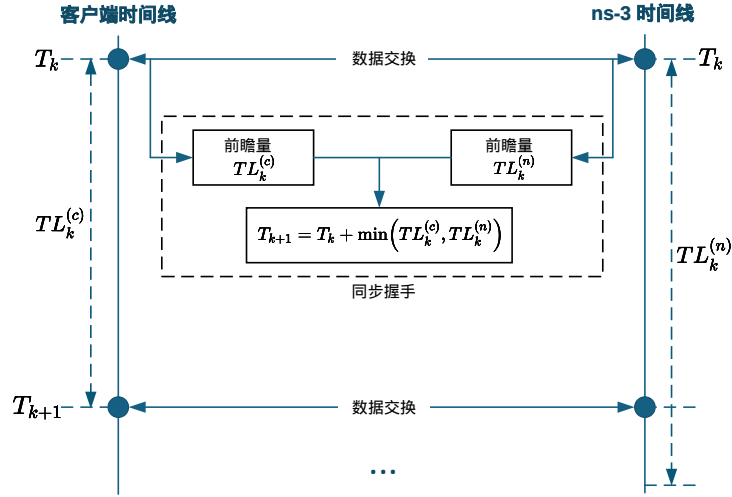


图 3-3 联合仿真时间同步示意图

如图 3-2 所示，在平台运行过程中，客户端连接视频采集与请求程序，将实时视频流以及任务请求送入 ns-3 仿真的网络中。ns-3 将视频流通过 WebRTC 协议发送到 Wi-Fi 网络中，并模拟视频分组在无线链路中的传输过程，包括排队、竞争、重传、丢包和时延变化。服务端连接已经部署好的视觉语言模型或视觉语言动作模型服务，在接收到视频流和请求后完成推理，并将模型生成的文本响应、决策结果或控制指令通过同一仿真网络返回客户端。由此，平台形成从客户端输入、无线网络传输、模型推理到结果返回的完整闭环。

该平台的关键难点在于联合仿真的时间同步和时序因果一致性。ns-3 采用离散事件仿真机制，内部按照仿真时间推进，而外部视频客户端和模型服务器运行在真实时间中。如果两者之间缺少严格的同步机制，就会出现事件的时序因果错误。以 GenAI 客户端与 ns-3 网络之间的二分交互为例：假设初始时间点为 t_0 ，客户端计划在未来的时间点 t_1 向 ns-3 发送请求。为避免因果性违规，客户端必须在 t_1 准确暂停，并确保 ns-3 的仿真时间尚未推进到 t_1 之后，然后在客户端和 ns-3 的时间都到达 t_1 时，客户端才能将请求发送到 ns-3 中。

作为联合仿真的一个经典问题，本研究参考 High Level Architecture(HLA) 同步机制的实现，设计了一种基于前瞻量的保守时间同步机制。此机制要求 GenAI 客户端/服务端和 ns-3 仿真器在每次同步事件中交换当前时间和前瞻承诺 (Lookahead) 参数，以独立计算下一个安全同步点，从而确保双方都能安全地推进时间而不违反时序因果性。所谓前瞻量可以理解为一个严格的时间承诺，即在当前时间点 T_k 加上前瞻值 TL_k 之前，相关仿真器绝对不会生成任何面向另一个仿真器的新事件或数据。

如图 3-3 所示，在每次同步事件中，外部仿真器（如 GenAI 客户端）与 ns-3 网络环境之间会进行一次双向数据交换。客户端将其当前仿真时间记为 T_k ，并将其前瞻承诺记为

$TL_k^{(c)}$ 发送给 ns-3。前瞻值 $TL_k^{(c)}$ 作为一项严格保证, 确保客户端在 $T_k + TL_k^{(c)}$ 之前绝对不会生成任何面向 ns-3 网络的新事件或数据。与此同时, ns-3 仿真器也将其当前时间 T_k 及其前瞻值 $TL_k^{(n)}$ 发送给客户端。基于这些交换的参数, 系统使用以下公式独立计算所有交互域之间的下一个安全同步点 T_{k+1} :

$$T_{k+1} = T_k + \min \left(TL_k^{(c)}, TL_k^{(n)} \right), \quad (3.1)$$

从而确保两个仿真器可以安全地将各自本地时钟推进至 T_{k+1} , 而不会违反时序因果性。

该研究产出一个面向通用智能体 AI 工作流的跨平台联合仿真框架, 支持真实视觉模型业务的闭环仿真。该平台不仅能够用于分析 Wi-Fi 网络状态、WebRTC 传输行为和视觉模型任务性能之间的关系, 还能够为后续设计多种 GenAI 业务场景提供支持。因此后续本文可以利用该平台模拟出具身智能、AI 视频问答等多种 GenAI 业务场景, 并在此基础上分析不同网络状态下视觉模型任务的性能表现, 为后续的优化方法设计提供数据支持和理论依据。

3.3 Wi-Fi、WebRTC 和视觉模型任务性能的跨层关联分析

在完成联合仿真平台的基础上, 本文进一步研究 Wi-Fi 网络状态、WebRTC 视频传输行为和视觉模型任务性能之间的关联关系。该部分的核心目标不是立即提出优化算法, 而是先回答一个基础问题: 动态 Wi-Fi 网络中的链路变化究竟会通过什么机制影响 WebRTC 视频传输, 并进一步影响实时视觉模型的 QoE 指标。在本文中, 模型 QoE 指标用于描述实时视觉模型业务的最终服务质量, 为了定量描述, 本文将模型 QoE 定义为由准确性、TTFT (首 token 响应时间)、ITL (Token 间隔) 和输入帧新鲜度 (以视频帧的传输时延衡量) 共同构成的综合指标。

3.3.1 研究内容

本文拟首先分析 Wi-Fi 状态对 WebRTC 传输行为的影响。在 Wi-Fi 网络中, 信道竞争、链路速率变化、无线重传和队列积压会改变端到端时延、丢包率、抖动和可用带宽, 从而影响 WebRTC 的带宽估计、码率调整、FEC 冗余配置和视频帧传输时延。本文将通过实验分析不同 Wi-Fi 底层状态变化是否会导致 WebRTC 拥塞控制产生滞后、误判或过度保守等现象。

其次, 本文拟分析 WebRTC 视频传输行为对视觉模型任务性能的影响。对于实时视觉模型业务, 视频质量下降并不一定直接等价于模型性能下降, 不同任务对视频分辨率、帧率、关键帧丢失、帧完成时延和输入帧新鲜度的敏感程度可能不同。例如, 文字识别和细粒度目标识别任务可能更依赖空间细节, 异常事件识别任务可能更依赖时间连续性,

实时视频问答和机器人感知任务则可能更加关注输入帧新鲜度和模型响应及时性。因此, 本文将比较不同视频传输质量和不同任务类型下模型准确率、任务成功率、TTFT 和 ITL 分布的变化规律。

最后, 本文拟分析传统网络 QoS、视频 QoE 指标与模型 QoE 之间的对应关系。本文将重点判断吞吐量、RTT、丢包率、抖动、视频码率、帧率和帧完成时延等传统指标, 是否能够稳定解释模型任务准确率和响应时延变化。如果发现传统指标与模型 QoE 之间存在不一致现象, 则进一步说明面向视觉模型业务进行专门跨层优化的必要性。

3.3.2 技术路线

本文将基于联合仿真平台开展多因素控制实验。实验中将 Wi-Fi 网络条件、WebRTC 传输配置和视觉任务类型作为主要实验变量, 并同步采集 Wi-Fi 层、WebRTC 层和模型任务层的观测指标。具体实验变量与观测指标如表 3.1 所示。

表 3.1 跨层关联分析的实验变量与观测指标

层次	实验变量或观测指标	分析目的
Wi-Fi 层	信道质量、背景流量强度、竞争终端数量、信道忙碌比例、MCS、无线重传率、队列长度、竞争时延等	分析无线链路状态和接入竞争对端到端传输行为的影响
WebRTC 层	发送码率、目标码率、带宽估计值、RTT、丢包率、抖动、FEC 冗余率、视频帧大小、关键帧比例、帧完成传输时延、发送队列状态等	分析 WebRTC 自适应传输行为及其对视频帧质量、时延和新鲜度的影响
模型任务层	任务准确率、任务成功率、TTFT、ITL 均值、ITL 高分位数、模型响应完成时延、输入帧新鲜度、过期帧比例等	分析网络和视频传输变化对最终模型服务质量的影响

在数据分析方法上, 本文拟采用统计分析和数据驱动建模相结合的方式。首先, 通过相关性分析、分组对比和敏感性分析, 识别哪些 Wi-Fi 状态和 WebRTC 传输指标与模型准确率、TTFT 和 ITL 变化关系最为密切。其次, 通过多变量回归、决策树或轻量级机器学习模型, 建立从跨层状态到模型 QoE 的预测关系, 用于判断不同网络状态和传输行为对模型任务效果的影响程度。最后, 针对不同视觉任务分别分析关键影响因素, 形成任务类型与传输敏感因素之间的对应关系。

该部分研究预期形成三类结果。第一,获得 Wi-Fi 底层状态对 WebRTC 拥塞控制、码率调整和视频帧传输时延的影响规律;第二,获得视频传输质量、帧新鲜度和模型任务准确率、TTFT、ITL 之间的关联关系;第三,筛选出对模型 QoE 影响显著的跨层状态和控制变量,为后续强化学习优化方法中的状态空间、动作空间和奖励函数设计提供依据。通过该部分研究,本文将从实验数据层面说明为什么需要进行 WebRTC-Wi-Fi 双向跨层协同优化,以及联合优化应重点关注哪些网络状态、视频传输参数和模型服务质量指标。

3.4 面向视觉模型服务质量的 WebRTC-Wi-Fi 联合优化方法

在完成跨层关联分析之后,本文将进一步研究面向视觉模型服务质量的 WebRTC-Wi-Fi 双向跨层协同优化方法。与传统实时视频传输优化主要关注网络吞吐量、端到端时延、丢包率或人类观看体验不同,本文拟将视觉模型任务的 QoE 作为最终优化目标。在此基础上,本文拟从研究内容和技术路线两个方面开展工作。

3.4.1 研究内容

本文拟建立 WebRTC 与 Wi-Fi 之间的双向跨层信息交互机制。

(1)Wi-Fi 状态辅助的 WebRTC 拥塞控制与 FEC 策略: 使 WebRTC 能够感知 Wi-Fi 层的信道竞争、无线重传、链路速率变化和队列积压等底层状态,从而辅助其进行码率调整、发送节奏控制和 FEC 冗余配置。相比仅依赖端到端 RTT、丢包率和带宽估计的传统 WebRTC 控制方式,引入 Wi-Fi 显式状态有助于减少无线环境下拥塞判断的滞后和误判。

(2)WebRTC 帧级信息驱动的 Wi-Fi 队列管理: 使 Wi-Fi 层能够感知 WebRTC 提供的视频帧级信息,包括帧类型、帧级截止时间、数据重要性、FEC 冗余属性和过期状态等。基于这些信息,Wi-Fi 层可以在队列管理和分组调度过程中对不同数据进行差异化处理。例如,对关键帧和高重要性数据包给予更高传输优先级,对已经超过任务截止时间的过期帧进行及时清理,对 FEC 冗余数据根据当前链路状态和主数据包传输情况动态调整优先级。通过这种方式,Wi-Fi 层不再无差别传输所有视频数据,而是能够围绕视觉模型任务需求进行无线资源分配。

(3)模型 QoE 驱动的跨层联合决策: 在双向跨层信息交互的基础上,本文拟将 Wi-Fi 状态感知、WebRTC 帧级信息下传和模型服务质量反馈结合起来,形成面向实时视觉模型推理业务的跨层协同优化机制,使系统能够在视频质量、传输时延、帧新鲜度和无线资源利用之间进行联合权衡,从而提升模型任务准确率并降低模型任务响应时延。

3.4.2 技术路线

本文拟将 WebRTC-Wi-Fi 联合优化问题建模为一个面向模型 QoE 的序贯决策问题。系统以固定控制周期运行,在每个控制周期内采集 Wi-Fi 层、WebRTC 层和模型任务层的跨层状态,并根据当前状态选择下一阶段的传输控制和无线调度动作。该问题可表示为马尔可夫决策过程,其中状态空间、动作空间和奖励函数分别设计如下。

首先,状态空间 s_t 由 Wi-Fi 层状态、WebRTC 层状态和模型任务层状态共同组成:

$$s_t = [s_t^{wif}, s_t^{rtc}, s_t^{model}] \quad (3.2)$$

其中, Wi-Fi 层状态 s_t^{wif} 用于描述无线链路和接入竞争情况,主要包括信道忙碌比例、物理层速率或 MCS、无线重传率、发送队列长度、队列等待时延、竞争时延和背景流量强度等。WebRTC 层状态 s_t^{rtc} 用于描述端到端视频传输行为,主要包括当前发送码率、目标码率、带宽估计值、RTT、丢包率、抖动、FEC 冗余率、视频帧大小、关键帧比例、视频帧完成传输时延和发送队列状态等。模型任务层状态 s_t^{model} 用于描述视觉模型业务的服务质量,主要包括近期模型任务准确率、首 token 响应时间 (Time to First Token, TTFT)、输出 token 间隔 (Inter-Token Latency, ITL) 的均值和高分位数等。

其次,动作空间 a_t 由 WebRTC 侧控制动作和 Wi-Fi 侧调度动作组成:

$$a_t = [a_t^{rtc}, a_t^{wif}] \quad (3.3)$$

其中, WebRTC 侧动作 a_t^{rtc} 主要包括目标码率调整、发送节奏调整、FEC 冗余率调整、视频帧率调整和分辨率等级调整等。Wi-Fi 侧动作 a_t^{wif} 主要包括关键帧优先级调整、普通帧优先级调整、FEC 冗余包优先级调整、过期帧清理阈值调整以及不同队列的调度权重调整等。考虑到连续动作空间会增加训练难度,本文初步拟采用离散化动作集合,例如对码率、FEC 冗余率、帧率和队列优先级分别设置“提高、保持、降低”三类动作,并根据后续实验结果进一步扩展为更细粒度的连续控制。

最后,奖励函数以模型 QoE 为核心进行设计。本文重点关注模型任务准确率和模型任务响应时延,其中响应时延包括 TTFT 和 ITL 的分布特征。对于第 t 个控制周期,奖励函数可设计为:

$$r_t = \alpha \tilde{A}t - \beta_1 \tilde{L}t^{TTFT} - \beta_2 \tilde{L}t^{ITL} - \beta_3 \tilde{F}t^{stale} - \beta_4 \tilde{C}_t^{overhead} \quad (3.4)$$

其中, $\tilde{A}t$ 表示归一化后的模型任务准确率或任务成功率, $\tilde{L}t^{TTFT}$ 表示归一化后的 TTFT 时延惩罚, $\tilde{L}t^{ITL}$ 表示归一化后的 ITL 时延惩罚, $\tilde{F}t^{stale}$ 表示输入视频帧过期或新鲜度下降带来的惩罚, $\tilde{C}_t^{overhead}$ 表示过高码率、过高 FEC 冗余或过度占用无线资源带来的系统开销。 $\alpha, \beta_1, \beta_2, \beta_3, \beta_4$ 为权重参数,可根据不同任务类型进行设置。

由于 ITL 的尾部时延会直接影响模型输出的连续性和交互稳定性, 本文不仅考虑 ITL 的平均值, 也将考虑其高分位数, 例如 90 分位数或 95 分位数。因此, $\tilde{L}_t^{ITL}$ 可以设计为 ITL 的均值和高分位数的加权组合, 以同时反映整体时延水平和长尾时延风险。

基于上述建模, 本文拟采用强化学习方法学习跨层控制策略。强化学习智能体以跨层状态 s_t 作为输入, 以联合控制动作 a_t 作为输出, 并通过奖励函数 r_t 引导策略优化。训练阶段将在联合仿真平台中构建多种 Wi-Fi 网络条件和视觉任务场景, 使智能体学习不同网络状态和任务需求下的传输控制策略; 验证阶段将与原生 WebRTC/GCC 和 BBR (瓶颈带宽与往返传播时间) 基线方法进行对比, 并通过消融实验分析 Wi-Fi 状态反馈、WebRTC 帧级信息下传和模型 QoE 反馈对系统性能提升的贡献。

3.5 基于联合仿真平台的多场景实验验证

3.5.1 研究内容

本文将基于前述闭环联合仿真平台, 对所提出的 WebRTC-Wi-Fi 双向跨层协同优化方法进行实验验证。实验将围绕多类实时视觉模型业务场景展开, 以验证所提出方法相比其他方案的有效性和泛化能力。具体而言, 本文将设计以下几个典型的实时视觉模型业务场景:

- **具身智能场景:** 在该场景中, 机器人作为客户端, 通过 Wi-Fi 连接到 VLA 服务端。机器人将上传自身摄像头拍摄的第一视角视频, 服务端模型根据当前画面生成环境理解、动作建议或控制指令, 使得机器人可以完成各种任务。
- **实时视频对话场景:** 在该场景中, 用户使用的对话终端作为客户端, VLM 作为服务端。用户通过摄像头与模型进行连续交互, 模型需要根据当前视频内容及时生成文本回答。
- **异常事件识别场景:** 在该场景中, 监控摄像头作为客户端, VLM 作为服务端。系统需要在连续视频流中及时发现关键事件并返回结果, 例如识别异常行为、检测安全隐患或监测设备状态。

3.5.2 技术路线

本文拟使用对比实验和消融实验开展实验验证。本文将基于联合仿真平台构建可控的实时视觉模型业务实验环境, 构建上述三个不同的场景, 并统计模型 QoE 指标。

在对比实验方面, 本文拟将所提出的 WebRTC-Wi-Fi 双向跨层协同优化方法与典型的 GCC 和 BBR 类拥塞控制基线方案进行比较, 如表 3.2 所示。通过该组实验, 可以评估本文方法相较于当前经典方案的整体优势。

表 3.2 对比实验设计

实验方案	方法说明	对比目的
GCC	采用 WebRTC 默认拥塞控制机制, 仅依据端到端 RTT、丢包率和到达时间反馈进行码率调整	评估仅依赖端到端反馈的传统实时视频传输方法在动态 Wi-Fi 环境下的性能
BBR	采用基于瓶颈带宽和往返时延估计的拥塞控制方法, 通过估计网络可用带宽和最小 RTT 来调节发送速率	作为模型式拥塞控制基线, 评估不依赖传统丢包驱动机制的速率控制方法在实时视觉模型业务中的适用性
本文方法	同时引入 Wi-Fi 状态反馈、WebRTC 帧级信息下传和模型 QoE 驱动联合决策机制	验证双向跨层协同优化相较于传统方法的综合优势

在消融实验方面, 本文将分别去除或简化所提出方法中的关键模块, 以分析不同机制对最终性能提升的贡献。具体消融实验设计如表 3.3 所示。

通过这种方式, 可以判断性能提升究竟主要来自 Wi-Fi 显式状态感知, 还是来自 WebRTC 向 Wi-Fi 提供的帧级时限和重要性信息, 或者来自模型 QoE 对联合决策的引导, 从而解释各个跨层模块对最终性能提升的具体作用。

表 3.3 消融实验设计

消融方案	去除或简化的模块	分析目的
去除 Wi-Fi 状态反馈	WebRTC 不再获取信道忙碌比例、无线重传率、链路速率和队列状态等 Wi-Fi 底层信息	分析 Wi-Fi 显式状态感知对拥塞判断、码率控制和 FEC 调整的贡献
去除 WebRTC 帧级信息下传	Wi-Fi 层不再获取视频帧类型、帧级截止时间、数据重要性和冗余属性等信息	分析帧级语义信息对 Wi-Fi 队列管理、关键帧保障和过期帧清理的贡献
去除模型 QoE 反馈	优化目标退化为传统网络 QoS 或视频传输质量指标, 不再直接考虑模型准确率、TTFT 和 ITL	分析模型 QoE 驱动优化相较于传统网络指标优化的必要性
完整方法	保留 Wi-Fi 状态反馈、WebRTC 帧级信息下传、模型 QoE 反馈、过期帧清理和 FEC 自适应机制	作为消融实验中的完整方案, 用于衡量各模块叠加后的综合效果

四 可行性分析与预期成果

4.1 前期工作与研究基础

本研究已经开展了一系列前期工作,包括实验平台搭建以及研究设备的采购等。

在 Wi-Fi 研究和 ns-3 仿真平台的使用上,本文作者已经有了一定的研究基础,在 ICNS3 2025 会议上发表了题为“ns3-WiMacStat: Enabling In-Depth WiFi MAC Performance Analysis via RadioTap-Based Statistics”的论文,介绍了一个基于 ns-3 的 Wi-Fi 性能分析工具,能够提供丰富的 MAC 层统计信息,为本研究中 Wi-Fi 状态感知和优化提供了技术积累^[56]。

另外,本文作者已经设计并完成了本文研究内容中说明的联合仿真平台的初步版本,能够支持真实视觉模型业务的闭环仿真,并实现了基于前瞻量的保守时间同步机制,确保了仿真过程中事件的时序因果一致性。这为后续的跨层关联分析和联合优化方法设计提供了坚实的实验基础。

最后,本课题组已经专门采购了边缘算力平台和 Wi-Fi 开发板,为本课题的实验提供了必要的硬件支持。这些设备将用于部署和测试实时视觉模型推理业务,并验证所提出的 WebRTC-Wi-Fi 联合优化方法在实际场景中的效果。

综上所述,本文在研究基础与实验准备方面已具备扎实条件,且在 ns-3 仿真平台与 Wi-Fi 研究上取得了阶段性成果,足以支撑本课题的开展。

4.2 预期成果

本研究的预期成果主要包括以下几个方面:

- 本文预期形成一套面向 GenAI 业务无线网络分析的通用闭环联合仿真平台。该平台能够在 ns-3 仿真 Wi-Fi 网络中接入真实视频客户端和实际部署的 VLM/VLA 服务,支持视频流上传、模型推理和结果返回的完整业务闭环。与传统使用预设流量模型的网络仿真相比,该平台能够直接观察无线网络状态变化对真实智能业务性能的影响,可为后续研究 GenAI 业务、实时视觉模型推理、具身智能和边缘多模态服务的无线网络性能提供可靠实验基础。
- 本文预期形成面向实时视觉模型推理业务的跨层传输优化研究成果。相关成果有望推动无线网络优化目标从传统网络 QoS 和视频 QoE 进一步扩展到模型服务质量层面,为实时视觉模型、具身智能和多模态智能体等新型 GenAI 业务提供更加有效的网络支撑。同时,相关研究将有助于揭示 Wi-Fi 网络状态、实时视频传输行为

与模型任务性能之间的关系,为后续开展面向智能业务的跨层协同优化研究提供参考。

- 本课题预期完成两篇学术论文。其中一篇围绕面向实时视觉模型业务的联合仿真平台展开,重点介绍平台设计、时间同步机制、闭环业务接入和无线网络实验能力;另一篇围绕 WebRTC-Wi-Fi 双向跨层协同优化方法展开,重点介绍跨层信息交互机制、模型 QoE 驱动的传输控制方法以及多场景实验验证结果。
- 本文预期申请发明专利一项,内容可围绕面向实时视觉模型推理的 WebRTC-Wi-Fi 跨层协同传输方法。

五 研究计划

本课题研究计划如下：

第一阶段(2026 年 6 月-2026 年 8 月)

完善并测试本项目所用的联合仿真平台，撰写其设计与实现相关的学术论文，并投稿至相关会议或期刊，同时准备开源事宜。

第二阶段(2026 年 9 月-2027 年 2 月)

基于联合仿真平台，完成 Wi-Fi 网络状态、WebRTC 传输行为和视觉模型任务性能之间的跨层关联分析，本阶段预期明确 Wi-Fi 状态如何影响 WebRTC 拥塞控制和视频传输行为，以及 WebRTC 传输质量变化如何进一步影响实时视觉模型服务效果。通过该阶段研究，确定后续联合优化方法需要重点关注的跨层状态、控制对象和评价指标。

第三阶段(2027 年 3 月-2027 年 8 月)

本阶段将在跨层关联分析结果基础上，调研并探索其实现路径，设计面向视觉模型服务质量的 WebRTC-Wi-Fi 联合优化方法。本阶段预期完成优化方法设计、主要功能模块实现和初步调试。

第四阶段(2027 年 9 月-2028 年 2 月)

本阶段将对所提出的联合优化方法进行全面实验验证，评估其在不同 Wi-Fi 网络条件、不同视觉任务类型和不同模型服务质量指标下的性能表现，并与现有方法进行对比分析。本阶段预期形成完整的实验结果和分析，并撰写相关学术论文和相关专利。

第五阶段(2028 年 3 月-2028 年 6 月)

完成毕业论文撰写并准备毕业答辩。

表 5.1 研究计划

时间	任务
2026 年 6 月-2026 年 8 月	完善联合仿真平台，撰写论文并投稿
2026 年 9 月-2027 年 2 月	跨层关联分析，确定优化重点
2027 年 3 月-2027 年 8 月	设计并实现联合优化方法
2027 年 9 月-2028 年 2 月	实验验证与对比分析，撰写论文和专利
2028 年 3 月-2028 年 6 月	完成毕业论文与答辩准备

六 总结

本文面向实时视觉模型推理业务中的无线传输问题,围绕 WebRTC 视频传输、Wi-Fi 链路状态与视觉模型 QoE 之间的跨层关系展开研究。本文首先构建基于 ns-3 和 ns3-ai 的闭环联合仿真平台,为真实 VLM/VLA 业务的无线网络分析提供可控实验环境。在此基础上,本文拟通过实验分析 Wi-Fi 状态、WebRTC 传输行为与模型服务质量之间的关联,并提出一种面向视觉模型 QoE 的 WebRTC-Wi-Fi 双向跨层协同优化方法,使 Wi-Fi 状态能够辅助 WebRTC 拥塞控制和 FEC 策略,同时使 WebRTC 帧级信息能够指导 Wi-Fi 队列管理、优先级保障和过期帧清理。预期本文能够为 GenAI 业务的无线网络分析提供可靠平台,为实时视觉模型推理提供有效的跨层传输优化方案,并形成相关论文和专利成果。

参考文献

- [1] Huawei Technologies Co., Ltd. 802.11bx for ai: 11-26-1012-00-0wng[R/OL]. HUAWEI, 2026. <https://mentor.ieee.org/802.11/dcn/26/11-26-1012-00-0wng-802-11bx-for-ai.pptx>.
- [2] Huawei Technologies Co., Ltd. Striding towards the intelligent world white paper: Wireless network[R/OL]. Huawei, 2025. <https://www-file.huawei.com/admin/asset/v1/pro/view/9a6d45ad388446e29a1ad513b1b30788.pdf>.
- [3] Huawei Technologies Co., Ltd. Post-802.11bn: 11-26-0180-00-0wng[R/OL]. HUAWEI, 2026. <https://mentor.ieee.org/802.11/dcn/26/11-26-0180-00-0wng-post-802-11bn.pptx>.
- [4] IEEE 802.11 AI Offload Study Group. Ai offload study group may 2026 meeting agenda: 11-26-0779-04-0aio[R/OL]. IEEE 802.11, 2026. <https://mentor.ieee.org/802.11/dcn/26/11-26-0779-04-0aio-aio-sg-may-2026-meeting-agenda.pptx>.
- [5] Ahmad S, Hafeez M, Zaidi S A R. Vision-Language Models on the Edge for Real-Time Robotic Perception[M/OL]. arXiv, 2026[2026-06-11]. <http://arxiv.org/abs/2601.14921>. DOI: 10.48550/arXiv.2601.14921.
- [6] Wang Y, Wang W, Liu D, et al. Enabling edge-cloud video analytics for robotics applications[C/OL]//IEEE INFOCOM 2021 - IEEE Conference on Computer Communications. 2021: 1-10. DOI: 10.1109/INFOCOM42981.2021.9488801.
- [7] Chen J, Lv Z, Wu S, et al. VideoLLM-online: Online Video Large Language Model for Streaming Video[C/OL]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 18407-18418[2026-06-11]. <https://ieeexplore.ieee.org/document/10657274/>. DOI: 10.1109/CVPR52733.2024.01742.
- [8] Niu J, Li Y, Miao Z, et al. OVO-Bench: How Far is Your Video-LLMs from Real-World Online Video Understanding?[C/OL]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2025: 18902-18913[2026-06-11]. <https://ieeexplore.ieee.org/document/11094230/>. DOI: 10.1109/CVPR52734.2025.01761.
- [9] Zou Y, Chen Y, Chen W, et al. CodecSight: Leveraging Video Codec Signals for Efficient Streaming VLM Inference[M/OL]. arXiv, 2026[2026-06-11]. <http://arxiv.org/abs/2604.06036>. DOI: 10.48550/arXiv.2604.06036.
- [10] Carlucci G, De Cicco L, Holmer S, et al. Congestion Control for Web Real-Time Communication[J/OL]. IEEE/ACM Transactions on Networking, 2017, 25(5): 2629-2642[2026-06-11]. <http://ieeexplore.ieee.org/document/7938641/>. DOI: 10.1109/TNET.2017.2703615.
- [11] Carlucci G, De Cicco L, Holmer S, et al. Making Google Congestion Control robust over Wi-Fi networks using packet grouping[C/OL]//Proceedings of the 2016 Applied Networking Research Workshop. Berlin Germany: ACM, 2016: 74-80[2026-06-11]. <https://dl.acm.org/doi/10.1145/2959424.2959431>.
- [12] Dai T, Zhang X, Guo Z. Learning-based Congestion Control for Internet Video Communication over Wireless Networks[C/OL]//2018 IEEE International Symposium on Circuits and Systems (ISCAS). Florence: IEEE, 2018: 1-5[2026-06-11]. <https://ieeexplore.ieee.org/document/8351530/>. DOI: 10.1109/ISCAS.2018.8351530.
- [13] Dai T, Zhang X, Zhang Y, et al. Statistical Learning Based Congestion Control for Real-Time Video Communication[J/OL]. IEEE Transactions on Multimedia, 2020, 22(10): 2672-2683[2026-06-11]. <https://ieeexplore.ieee.org/document/8931660/>. DOI: 10.1109/TMM.2019.2959448.

- [14] Karimi P, Fouladi S, Sivaraman V, et al. Vidaptive: Efficient and Responsive Rate Control for Real-Time Video on Variable Networks[M/OL]. arXiv, 2024[2026-06-11]. <http://arxiv.org/abs/2309.16869>. DOI: 10.48550/arXiv.2309.16869.
- [15] Kulkarni U, Diab K, Aggarwal S, et al. Understanding the Impact of Wi-Fi Configuration on Volumetric Video Streaming Applications[C/OL]//Proceedings of the 2023 Workshop on Emerging Multimedia Systems. New York NY USA: ACM, 2023: 41-47[2026-06-11]. <https://dl.acm.org/doi/10.1145/3609395.3610599>.
- [16] Casasnovas M, Michaelides C, Carrascosa-Zamacois M, et al. Experimental evaluation of interactive Edge/Cloud Virtual Reality gaming over Wi-Fi using unity render streaming[J/OL]. Computer Communications, 2024, 226-227: 107919[2026-06-11]. <https://linkinghub.elsevier.com/retrieve/pii/S0140366424002585>. DOI: 10.1016/j.comcom.2024.08.001.
- [17] Du X, Li J, Shao Y, et al. Revisiting congestion control for wifi networks[C/OL]//APNet '24: Proceedings of the 8th Asia-Pacific Workshop on Networking. New York, NY, USA: Association for Computing Machinery, 2024: 88-94. <https://doi.org/10.1145/3663408.3663421>.
- [18] Yi F, Wan H, Jamieson K, et al. Athena: Seeing and mitigating wireless impact on video conferencing and beyond[C/OL]//HotNets '24: Proceedings of the 23rd ACM Workshop on Hot Topics in Networks. New York, NY, USA: Association for Computing Machinery, 2024: 103-110. <https://doi.org/10.1145/3696348.3696889>.
- [19] Guo F, Zhou Y, Jiang L, et al. BLADE: Adaptive Wi-Fi Contention Control for Next-Generation Real-Time Communication[M/OL]. arXiv, 2026[2026-06-11]. <https://arxiv.org/abs/2603.16119>. DOI: 10.48550/ARXIV.2603.16119.
- [20] Shen Y, Meng Z. Law: Towards consistent low latency in 802.11 home networks[C/OL]//23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI 26). Renton, WA: USENIX Association, 2026: 1741-1757. <https://www.usenix.org/conference/nsdi26/presentation/shen-yibin>.
- [21] Gong Q, Xu J, Shi J, et al. Flick: Frame-Perceptive Packet Scheduling for Low-Latency Video Services in Wi-Fi Networks [C/OL]//2025 IEEE 33rd International Conference on Network Protocols (ICNP). Los Alamitos, CA, USA: IEEE Computer Society, 2025: 1-13. <https://doi.ieeecomputersociety.org/10.1109/ICNP65844.2025.11192371>.
- [22] Ning Z, Liu G, Jin Q, et al. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval[A]. 2025. arXiv: 2505.15269.
- [23] Xu R, Xiao G, Chen Y, et al. Streamingvlm: Real-time understanding for infinite video streams[A/OL]. 2026. arXiv: 2510.09608. <https://arxiv.org/abs/2510.09608>.
- [24] Li R, Tan Y, Shi Y, et al. Videoscan: Enabling efficient streaming video understanding via frame-level semantic carriers[A/OL]. 2025. arXiv: 2503.09387. <https://arxiv.org/abs/2503.09387>.
- [25] Shaker A, Maaz M, Gou C, et al. Mobile-videogpt: Fast and accurate model for mobile video understanding[A/OL]. 2026. arXiv: 2503.21782. <https://arxiv.org/abs/2503.21782>.
- [26] Fu C, Dai Y, Luo Y, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2025: 24108-24118.
- [27] Wu H, Li D, Chen B, et al. Longvideobench: A benchmark for long-context interleaved video-language understanding[J]. Advances in Neural Information Processing Systems, 2024, 37: 28828-28857.
- [28] Zhou J, Shu Y, Zhao B, et al. Mlvu: Benchmarking multi-task long video understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025: 13691-13701.

- [29] Huang Z, Li X, Li J, et al. Online video understanding: Ovbench and videochat-online[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 3328-3338.
- [30] Lin J, Fang Z, Chen C, et al. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding[C]//ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2026: 12147-12151.
- [31] Zitkovich B, Yu T, Xu S, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control[C]//Tan J, Toussaint M, Darvish K. Proceedings of Machine Learning Research: volume 229 Proceedings of The 7th Conference on Robot Learning. PMLR, 2023: 2165-2183.
- [32] Kim M J, Pertsch K, Karamcheti S, et al. Openvla: An open-source vision-language-action model [C]//Agrawal P, Kroemer O, Burgard W. Proceedings of Machine Learning Research: volume 270 Proceedings of The 8th Conference on Robot Learning. PMLR, 2025: 2679-2713.
- [33] Yang J, Liu S, Guo H, et al. Egolife: Towards egocentric life assistant[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 28885-28900.
- [34] Pakha C, Chowdhery A, Jiang J. Reinventing video streaming for distributed vision analytics[C/OL]//10th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 18). Boston, MA: USENIX Association, 2018. <https://www.usenix.org/conference/hotcloud18/presentation/pakha>.
- [35] Zhang B, Jin X, Ratnasamy S, et al. Awstream: adaptive wide-area streaming analytics[C/OL]//SIGCOMM '18: Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication. New York, NY, USA: Association for Computing Machinery, 2018: 236 - 252. <https://doi.org/10.1145/3230543.3230554>.
- [36] Jiang J, Ananthanarayanan G, Bodik P, et al. Chameleon: scalable adaptation of video analytics[C/OL]//SIGCOMM '18: Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication. New York, NY, USA: Association for Computing Machinery, 2018: 253-266. <https://doi.org/10.1145/3230543.3230574>.
- [37] Du K, Pervaiz A, Yuan X, et al. Server-driven video streaming for deep learning inference[C/OL]//SIGCOMM '20: Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication. New York, NY, USA: Association for Computing Machinery, 2020: 557 - 570. <https://doi.org/10.1145/3387514.3405887>.
- [38] Zhang M, Wang F, Liu J. Casva: Configuration-adaptive streaming for live video analytics[C/OL]//IEEE INFOCOM 2022 - IEEE Conference on Computer Communications. 2022: 2168-2177. DOI: 10.1109/INFOCOM48880.2022.9796875.
- [39] Wu J, Ren Z, Liu L, et al. Chat with ai: The surprising turn of real-time video communication from human to ai[C/OL]//HotNets '25: Proceedings of the 24th ACM Workshop on Hot Topics in Networks. New York, NY, USA: Association for Computing Machinery, 2025: 402-410. <https://doi.org/10.1145/3772356.3772390>.
- [40] Wu J, Ren Z, Zhong J, et al. Artic: Ai-oriented real-time communication for mllm video assistant [A/OL]. 2026. arXiv: 2602.12641. <https://arxiv.org/abs/2602.12641>.
- [41] 陈震, 卢金勤, 朱奇, 等. 基于 WebRTC 的低延迟流媒体传输方法研究[J]. 机电产品开发与创新, 2024, 37(06): 29-31.
- [42] 高朝龙, 张景波, 蔡星娟. WebRTC 视频会议的高维多目标多码率优化研究[J/OL]. 小型微型计算机系统, 2023, 44(08): 1735-1742. DOI: 10.20009/j.cnki.21-1106/TP.2021-0899.
- [43] 牛冠冲, 刘飞翔, 杨雯, 等. 基于边缘计算和 WebRTC 的元宇宙教育通信技术方案的实现[J]. 计算机科学, 2024, 51(10): 94-104.

- [44] 徐嵩, 王尔申, 郭倩文, 等. 基于 5G 及 WebRTC 的车辆行驶状态实时监控系统设计[J]. 电子器件, 2025, 48(04): 863-868.
- [45] 钟育伙. WebRTC 的实时会议系统设计与实现[J/OL]. 福建电脑, 2025, 41(04): 80-84. DOI: 10.16707/j.cnki.fjpc.2025.04.015.
- [46] Yu E, Zhou J, Li Z, et al. Mustang: Improving qoe for real-time video in cellular networks by masking jitter[J/OL]. ACM Trans. Multimedia Comput. Commun. Appl., 2024, 20(9). <https://doi.org/10.1145/3672399>.
- [47] Xie Y, Yi F, Jamieson K. PBE-CC: Congestion Control via Endpoint-Centric, Physical-Layer Bandwidth Measurements[C/OL]//Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication. Virtual Event USA: ACM, 2020: 451-464[2026-06-11]. <https://dl.acm.org/doi/10.1145/3387514.3405880>.
- [48] Goyal P, Agarwal A, Netravali R, et al. ABC: A Simple Explicit Congestion Controller for Wireless Networks[M/OL]. arXiv, 2021[2026-06-11]. <http://arxiv.org/abs/1905.03429>. DOI: 10.48550/arXiv.1905.03429.
- [49] Shen Y, Chen R, Wang B, et al. Mortise: Auto-tuning congestion control to optimize QoE via Network-Aware parameter optimization[C/OL]//23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI 26). Renton, WA: USENIX Association, 2026: 2303-2321. <https://www.usenix.org/conference/nsdi26/presentation/shen-yixin>.
- [50] Li Y, Zhang Z, Chen H, et al. Mamba: Bringing multi-dimensional abr to webrtc[C/OL]//MM '23: Proceedings of the 31st ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2023: 9262-9270. <https://doi.org/10.1145/3581783.3611915>.
- [51] Smirnov N, Tomforde S. Real-time rate control of webrtc video streams in 5g networks: Improving quality of experience with deep reinforcement learning[J/OL]. J. Syst. Archit., 2024, 148(C). <https://doi.org/10.1016/j.sysarc.2024.103066>.
- [52] Lee I, Kim S, Sathyanarayana S, et al. R-fec: RL-based fec adjustment for better qoe in webrtc[C/OL]//MM '22: Proceedings of the 30th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2022: 2948-2956. <https://doi.org/10.1145/3503161.3548370>.
- [53] Cheng S, Hu H, Zhang X. Abrf: Adaptive bitrate-fec joint control for real-time video streaming[J/OL]. IEEE Trans. Cir. and Sys. for Video Technol., 2023, 33(9): 5212-5226. <https://doi.org/10.1109/TCSVT.2023.3248164>.
- [54] Bankov D, Chemrov K, Khorov E. Tuning channel access to enable real-time applications in wi-fi 7 [C/OL]//2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). 2020: 20-25. DOI: 10.1109/ICUMT51630.2020.9222409.
- [55] Adame T, Carrascosa-Zamacois M, Bellalta B. Time-sensitive networking in ieee 802.11be: On the way to low-latency wifi 7[J/OL]. Sensors, 2021, 21(15). <https://www.mdpi.com/1424-8220/21/15/4954>. DOI: 10.3390/s21154954.
- [56] Mi C, Gu J, Zhang C, et al. ns3-wimacstat: Enabling in-depth wifi mac performance analysis via radiotap-based statistics[C/OL]//ICNS3 '25: Proceedings of the 2025 International Conference on Ns-3. New York, NY, USA: Association for Computing Machinery, 2025: 29-37. <https://doi.org/10.1145/3747204.3747215>.

研究生签字 _____

指导教师签字 _____

院 (系、所) 领导签字 _____

年 月 日