arXiv:2602.12758v1 [eess.IV] 13 Feb 2026

VINEETVC：在严苛带宽限制下，利用音频驱动的说话人头部重建实现自适应视频会议

Vineet Kumar Rakesh
工程科学，霍米·巴巴国家研究所培训学校综合楼，Anushaktinagar，孟买，马哈拉施特拉邦 400094，印度；计算机与信息学组，可变能量回旋加速器中心，1/AF，Bidhannagar，加尔各答，西孟加拉邦 700064，印度 vineet@vecc.gov.in

Soumya Mazumdar计算机科学与商业系统系，Gargi Memorial Institute of Technology，Baruipur，加尔各答，西孟加拉邦 700144，印度 reachme@soumyamazumdar.com

Tapas Samanta
计算机与信息学组，可变能量回旋加速器中心，1/AF，Bidhannagar，加尔各答，西孟加拉邦 700064，印度；工程科学，霍米·巴巴国家研究所培训学校综合楼，Anushaktinagar，孟买，马哈拉施特拉邦 400094，印度 tsamanta@vecc.gov.in

Hemendra Kumar Pandey
能源回旋加速器中心 1/AF，Bidhannagar，加尔各答，西孟加拉邦 700064，印度 工程科学，霍米·巴巴国家研究所培训学校综合楼，Anushaktinagar，孟买，马哈拉施特拉邦 400094，印度 hkpandey@vecc.gov.in

Amitabha Das核研究与应用学院 贾达普大学 盐湖城，加尔各答，西孟加拉邦 700106，印度 amitabhad.snsa@jadavpuruniversity.in

Sarbajit Pal工程科学，霍米·巴巴国家研究所培训学校综合楼，Anushaktinagar，孟买，马哈拉施特拉邦 400094，印度 sarbajit@vecc.gov.in

2026年2月16日

摘要

消费级网络和受限网络中的带宽严重耗尽，可能会损害实时视频会议的稳定性：编码器码率管理达到饱和，丢包加剧，帧率下降，端到端延迟显著增加。本文介绍了一种自适应会议系统，该系统将 WebRTC 媒体传输与辅助的音频驱动说话人头部重建流程以及由遥测数据驱动的模式调节相结合。系统由 WebSocket 信令服务、用于多方传输的可选选择性转发单元、能够实时提取 WebRTC 统计信息并导出 CSV 遥测数据的浏览器客户端，以及一个 AI REST 服务组成；该服务处理参考人脸图像和录制音频，生成合成 MP4；浏览器可将输出摄像头轨道替换为合成流，其中位带宽为 32.80 千比特每秒。该方案采用带宽模式切换策略和客户端模式状态记录器。

关键词：视频会议、WebRTC 遥测、带宽模式、说话人头部合成、音频驱动的重建、选择性转发单元

1 引言

实时视频通话如今已成为商务、教育和医疗保健领域的重要工具。在 COVID-19 疫情期间，远程办公、远程会议和在线学习变得普遍得多，而许多此类习惯在疫情结束后仍延续至今 [1]。远程医疗也有所发展，当难以与患者当面会诊时，视频问诊成了联系患者的常用方式 [2]。因此，越来越多的消费者希望视频通话在各种网络环境下都能流畅运行，例如移动网络和共享家庭网络连接。
另一方面，当带宽极低或不稳定时，视频会议仍会无法正常工作。对热门平台进行的现实环境测试表明，要获得通常所说的“舒适”画质，一般需要数兆比特的上行/下行速率。延迟和丢包可能很快导致视频变得模糊、卡顿，并造成音画不同步 [3]。网络问题对压缩视频的影响尤其严重，因为即使只丢失少量数据包，也可能导致当前帧乃至后续帧（由于帧间预测）都无法解码，从而产生明显的画面卡顿 [4]。这些挑战在农村或偏远地区以及带宽有限、移动电话覆盖可能不稳定的发展中国家更为常见。

WebRTC 是大多数基于浏览器的视频会议解决方案的基础。它使媒体能够在对等端之间传输、穿越网络地址转换，并以安全的方式实时传送 [6]。WebRTC 通常采用优质的音频编解码器（如 Opus）和视频编解码器（如 VP8/H.264/VP9/AV1，具体取决于配置方式）。Opus 专为交互式通信而设计，即使在极低的比特率下，也能让语音保持清晰可懂 [18]。WebRTC 还提供一个名为 getStats 的标准统计接口，使应用能够跟踪比特率、丢包、延迟及其他重要的通话指标 [16]。选择性转发单元（SFU）是提高多方通话分发能力的标准方式，无需对所有内容重新编码 [17]。尽管有这些改进，传统系统仍需传输压缩视频帧。当可用带宽低于数百千比特每秒时，视频质量和连续性往往会迅速恶化。

一种值得期待的替代方案，是将“视频临场感”视为可以逐步构建，而非直接传输的内容。近期的音频驱动说话人头部生成技术，能够利用参考人脸和语音信号 [19, 20, 21]，生成逼真的嘴部运动（有时也包括头部运动）。这提供了一种节省带宽策略：系统无需持续传输像素视频，而是可以广播低码率音频以及对面部动作的简要描述，并在接收端重建说话人头部视频流。由于有效载荷远少于完整视频，这种模式即使在带宽严重受限的情况下也可能仍然可用。

在本研究中，我们介绍了 Versatile Intelligent Network Enhanced Efficient Telepresence，这是一种将 WebRTC 传输与辅助音频驱动重建路径相结合的自适应视频会议平台。其核心理念很简单：网络状况良好时，通话以传统 WebRTC 视频会话的方式运行；当带宽骤降时，系统切换至 AI重建模式，在保持类似视频的体验的同时，将发送速率维持在极低水平。模式选择由通过 WebRTC 统计 API [16] 获取的实时遥测数据决定。在我们的原型系统中，客户端可以将传出的摄像头轨道替换为根据用户的参考人脸和语音生成的合成说话人头部视频流。这种方法针对的是极低带宽环境：在此情况下，进一步进行视频压缩所带来的收益逐渐降低，而用户体验则主要受到画面冻结和不稳定性的影响 [4]。具体而言，我们当前的部署在统一服务 API 后端集成了多个最先进的音频驱动的说话人头部生成器。除非另有说明，本文中的所有定量结果均使用 Wav2Lip[19] 作为默认后端；这是一个性能强大且广泛采用的唇形同步模型。为证明我们的系统并不局限于单一

架构，我们还为 MakeItTalk[20], SadTalker[21], 和 VadicTHG 模型实现了后端，这些后端可以在运行时切换，而无需更改视频会议管线的其余部分。这项工作连接了两个研究方向：(i) 主流视频会议系统在受限网络条件下的表现；以及 (ii) 说话人头部生成如何在无需持续传输完整视频帧的情况下提供类似视频的体验。

本工作的主要贡献如下：

- **自适应会议设计：**一个基于 WebRTC 的平台，利用标准化浏览器统计数据进行遥测驱动的模式控制 [6, 16]。
- **超低比特率“视频临场感”模式：**一种音频驱动的说话人头部重建方案，其设计借鉴了唇形同步和说话人头部生成领域的最新进展 [19, 20, 21]。
- **工程实现：**一个可运行的原型系统，支持信令、可选的选择性转发单元（RTP 拓扑），以及客户端侧对合成视频的轨道替换 [17]。
- **真实约束下的证据：**长期带宽日志记录，用于量化正常视频模式与人工智能重建模式之间的带宽差距（结果部分报告了该数据，并在表1中进行了总结）。

2 相关工作

带宽受限的视频会议面临的一项主要困难，是各个平台缺乏统一性：它们的媒体栈、编解码器配置、拥塞管理机制、服务器端规则，以及向用户和研究人员提供的透明度各不相同。如果研究没有采用共同的刺激材料和标准化的测量指标来确保可复现性，这种差异就会使跨平台研究结果难以比较 [5]。一项新近开展的带宽感知型基准测试研究表明，许多系统在网络条件中等至良好时能够保持稳定，但随着带宽下降，质量会变得极不稳定，各平台的适应策略也会出现分歧。在带宽极为有限的环境中，平台可能无法启动会话、会停用视频，或无法切换回纯音频功能，这揭示了实际的下阈值，而这些阈值往往被面向用户的规格说明所掩盖 [15]。重要的是，该研究指出，即使可用带宽较高，系统行为仍可能受供应商设定的策略上限限制，而非网络约束；因此，解读结果时必须区分由策略驱动的行为与真正由网络驱动的自适应 [14]。研究还指出，帧率随时间变化、分辨率随时间变化等时间分辨指标，比单独考察吞吐量更能洞察用户感知到的质量下降。此外，在极端场景下，可能还需要在连续指标之外纳入类别型结果 [13]。现代基于浏览器的视频会议解决方案通常依赖 WebRTC，它提供标准化的统计接口，可通过浏览器API公开传输层和RTP层级指标。该基准测试研究指出，WebRTC统计模型（通过 getStats）为在基于浏览器的应用中收集类似指标提供了可行框架，包括吞吐量趋势、帧动态、抖动和丢包指标。研究还指出，可观测性因客户端类型而异：浏览器实现通常提供更丰富、标准化程度更高的指标，而原生客户端可能只显示部分诊断信息，或隐藏内部自适应状态 [12]。这些限制促使人们设计这样的系统架构：(i) 能够使用客户端可靠可获取的遥测数据；(ii) 即使标准码率自适应已不足以应对，也能确保对话连续性。说话人头部生成（THG）已发展成为一种可行方案，无需持续传输摄像头画面即可呈现“视觉存在感”。说话人头部生成综述根据输入模态和生成范式梳理了既有研究，涵盖利用图像、音频、文本或驱动视频来驱动人脸动画的方法，以及基于关键点运动迁移、显式三维建模、神经渲染、基于扩散模型的生成和动画管线的方法 [11]。这些方法的共同目标，是在保留个体特征的同时，生成与语音同步且时间连贯的运动。综述指出，一个明显趋势是从纯二维运动迁移转向三维和基于神经渲染的表征，以更好地处理姿态变化并提升逼真度；进一步的发展方向则是混合式设计，将神经渲染与更强的生成先验相结合，以改善时间平滑度和可控性 [11]。该研究还强调了面向实时渲染的说话人头部生成，其架构旨在服务于交互式应用场景，并力求在确保身份特征和同步性的同时降低延迟 [7]。说话人头部合成的评估通常集中于少数关键标准：身份保持、视觉质量、视听同步和运动自然度。综述强调，研究既采用客观测量，也采用主观评估；新提出的感知度量则试图比单纯的像素级相似度更准确地反映人类感知[11]。由于即使面部在物理表现上令人信服，只要唇部动作与语音不一致，也可能显得不自然，因此以同步性为重点的评估至关重要 [11]。在部署方面，综述讨论了若干反复出现的局限：依赖大型预训练生成器（可能存在数据集引入的偏差）、难以应对极端头部姿态和遮挡、由于训练数据受语言和数据集偏差影响而难以适应多语言环境，以及长序列中的时间漂移或闪烁 [11]。综述还强调了采用负责任部署技术的必要性，包括

针对合成媒体采取保护措施并建立明确的治理机制，尤其是在教育和视频通信等交互式环境中 [9]。综合来看，视频会议基准测试文献表明，在严重限制下，典型的会议技术栈可能退化为不稳定状态，或硬性切换至备用模式；同时，依托遥测数据开展可复现评估，对于理解这些情形至关重要 [10]。与此同时，说话人头部生成文献表明，音频驱动的人脸重建日益能够生成同步且保持身份特征的人脸视频，适用于交互式应用，但在鲁棒性和长时稳定性方面仍面临实际挑战 [8]。本文基于这些见解，将“视觉存在感”视为一种自适应模态：当网络条件导致像素流传输不可靠时，系统可以将传输内容从视频帧切换为语音（以及轻量级控制提示），并据此重建说话人头部，同时利用客户端可见的遥测数据来指导系统稳定运行。

3 系统概述

图 1 展示了一种混合式视频会议（VC）架构，它支持两种互补的传输路径：（i）适用于正常网络条件的传统 WebRTC 像素流传输路径；以及（ii）当带宽受到严重限制时，以可重建表示取代像素视频传输的辅助音频驱动重建路径。其主要目的是通过改变传输内容的含义来保持对话连续性。在带宽充足时，系统的运行方式与普通视频会议应用相同：采集摄像头和麦克风信号，对其进行编码并实时传输。当带宽受到严重限制时，系统会继续传输实时音频以保持通话，同时根据音频和人脸参考图生成合成的说话人头部视频流，即使无法再持续发送像素视频，也能维持类似视频的临场感。媒体路径依赖 WebRTC 实现实时传输。信令服务会交换建立 WebRTC 会话所需的会话描述和连接候选项。信令通道仅承载建立会话和连接所需的控制信息；会话建立后，音频和视频媒体则通过实时 RTP 流传输。对于多方会话，可以插入一个 SFU，将流转发给多名参与者。在这种情况下，每个端点只向 SFU 上传一个流，再由 SFU 有选择地将相应的流转发给其他参与者。这样，无需发送方针对每个接收方分别上传流的副本，即可支持可扩展的群组通话。该系统的一个核心要素是在端点持续提取遥测数据。设 Δt 为遥测数据的采样间隔。端点会监测已发送和已接收总字节数的累计计数器。设 $B_{\mathrm{tx}}(t)$ 为截至时间 t 的累计已发送字节数， $B_{\mathrm{rx}}(t)$ 为截至时间 t的累计已接收字节数。通过计算这些计数器在最近一个时间间隔内的差值，并将字节换算为比特，即可得出以千比特每秒为单位的即时上行链路和下行链路吞吐量估计值：

$$\hat{R}_{\mathrm{tx}}(t) = \frac{8\,(B_{\mathrm{tx}}(t) - B_{\mathrm{tx}}(t-\Delta t))}{1000\,\Delta t}, \tag{1}$$

$$\hat{R}_{\mathrm{rx}}(t) = \frac{8\,(B_{\mathrm{rx}}(t) - B_{\mathrm{rx}}(t-\Delta t))}{1000\,\Delta t}. \tag{2}$$

在式 (1)–(2) 中， $B_{\mathrm{tx}}(t) - B_{\mathrm{tx}}(t-\Delta t)$ 表示长度为 Δt 的上一个时间间隔内发送的字节数；乘以 8 可将字节转换为比特；除以 Δt 可将数值转换为比特/秒；最后除以 1000 可转换为千比特每秒。这两个估计值足以表征在时间 t 呼叫双方各自使用了多少网络容量，并提供实时信号，用于判断哪种运行模式可行。仅凭吞吐量无法充分描述有损链路下的呼叫稳定性，因此端点也会根据实时传输协议计数器估计丢包相关指标。设 $P_{\mathrm{lost}}(t)$ 和$P_{\mathrm{recv}}(t)$ 分别为丢失数据包数和接收数据包数的累计计数。区间丢包率估计如下

$$\hat{p}_{\mathrm{loss}}(t) = \frac{\Delta P_{\mathrm{lost}}(t)}{\Delta P_{\mathrm{lost}}(t) + \Delta P_{\mathrm{recv}}(t)}, \quad \text{where} \quad \Delta P(t) = P(t) - P(t-\Delta t). \tag{3}$$

这个估计简单且易于计算：它衡量上一个时间间隔内丢失的数据包占比。考虑丢包情况的一种实用容量代理指标是有效吞吐量，其定义为根据式 4，预计能够成功到达的数据占发送比特率的比例。丢包率上升时，有效吞吐量 $G(t)$ 会降低，即使发送方尝试以相同速率发送数据也是如此。由于网络测量值会波动，控制器使用平滑估计值来避免不稳定的切换。根据式 5，采用标准的指数移动平均进行计算

$$\hat{G}(t) = \hat{R}_{\mathrm{tx}}(t)\,(1 - \hat{p}_{\mathrm{loss}}(t))\,. \tag{4}$$

$$\tilde{G}(t) = \alpha\,\tilde{G}(t-\Delta t) + (1-\alpha)\,\hat{G}(t), \qquad 0 < \alpha < 1. \tag{5}$$

其中， α 控制估计值对变化作出反应的速度。如果 α 接近1，估计值变化缓慢，且能抵御短时峰值。如果 α 较小，估计值则能迅速响应突发下降。系统采用三种运行模式，如公式21所示。每种模式由传输的数据决定，因此也对应不同的

总比特率分解方式。设总传输比特率为 $R_{\text{tot}}(t)$，并将其分解为公式6中提到的四个组成部分。

$$R_{\text{tot}}(t) = R_{\text{aud}}(t) + R_{\text{vid}}(t) + R_{\text{ctrl}}(t) + R_{\text{ref}}(t). \tag{6}$$

在式（6）中， $R_{\text{aud}}(t)$ 是音频比特率， $R_{\text{vid}}(t)$ 是传统压缩像素视频的比特率， $R_{\text{ctrl}}(t)$ 是用于重建的紧凑控制信息（例如面部运动描述符）的比特率， $R_{\text{ref}}(t)$ 是由偶尔进行的参考更新产生的平均比特率。这种分解十分重要，因为所提出的方法主要通过在 AI 模式下将 $R_{\text{vid}}(t)$ 降至接近零，同时保留 $R_{\text{aud}}(t)$ 和低开销控制信息项，从而提升带宽限制极严时的性能。在普通模式下，系统传输传统 WebRTC 音频和视频，不依赖重建信号：

$$R_{\text{tot}}^{(\text{Normal})}(t) \approx R_{\text{aud}}(t) + R_{\text{vid}}(t), \qquad R_{\text{ctrl}}(t) \approx 0,\ R_{\text{ref}}(t) \approx 0. \tag{7}$$

此模式可提供最高的视觉质量，因为 $R_{\text{vid}}(t)$ 分配给了编码像素视频。然而，随着带宽降低，维持 $R_{\text{vid}}(t)$ 变得困难，并会导致严重的压缩伪影、丢帧和画面冻结。在低比特率视频模式下，系统仍采用像素流传输，但通过降低空间分辨率和/或帧率来减少 $R_{\text{vid}}(t)$。一种实用的工程模型是，视频比特率近似随每秒编码的像素数量变化。设空间缩放因子为 $\eta \in (0, 1]$，宽度和高度均按 η 缩放；时间缩放因子为 $\gamma \in (0, 1]$，帧率按 γ 缩放。由于每帧的像素数量随面积变化，面积按 η^2 缩放。因此，降低后的视频比特率可用公式 8 近似表示

$$R_{\text{vid}}^{(\text{LB})}(t) \approx \eta^2 \gamma\, R_{\text{vid}}^{(\text{Normal})}(t), \tag{8}$$

因此

$$R_{\text{tot}}^{(\text{LB})}(t) \approx R_{\text{aud}}(t) + R_{\text{vid}}^{(\text{LB})}(t). \tag{9}$$

公式（8）表达了一个直观的想法：将宽度和高度都减半，像素数大约会减少到四分之一；降低帧率则会进一步减少每秒的像素数。此模式旨在延缓像素流传输变得不可用的时刻，但仍依赖于持续发送视频帧。在人工智能重建模式下，系统停止传输像素视频，转而传输可重建表示。传输流主要由语音音频、轻量级控制信号，以及公式10中偶尔进行的参考更新构成

$$R_{\text{tot}}^{(\text{AI})}(t) \approx R_{\text{aud}}(t) + R_{\text{ctrl}}(t) + R_{\text{ref}}(t), \qquad R_{\text{vid}}(t) \approx 0. \tag{10}$$

重要的是， R_{ctrl}（t）和 R_{ref}（t）的设计规模可以远小于像素视频。这使人工智能模式在即使经过激进压缩的视频也不稳定的情况下仍然可行。紧凑的控制流可以建模如下。假设控制信号以每秒 f_{ctrl} 个控制帧的速率发送。每个控制帧包含 N 个运动描述符（例如关键点），每个描述符经过量化/打包后用 d比特编码。那么，控制比特率约为

$$R_{\text{ctrl}} \approx \frac{f_{\text{ctrl}}\, N\, d}{1000}. \tag{11}$$

(11) 中的所有项都有直接的物理意义： f_{ctrl} 表示发送控制信息的频率， N 表示每个控制帧发送的描述符数量， d 表示每个描述符使用的比特数。除以 1000 可将比特/秒转换为千比特每秒。该公式也清楚地体现了主要的设计权衡：增加 f_{ctrl} 可以提高时间平滑度，增加 N 可以改善几何细节，增加 d 可以降低量化误差；这三者都会提高比特率。参考更新偶尔进行，其开销会随时间摊销。设每次参考更新需要 S_{ref} 比特，更新间隔为 T_{ref} 秒。平均参考比特率为

$$R_{\text{ref}} \approx \frac{S_{\text{ref}}}{1000\, T_{\text{ref}}}. \tag{12}$$

公式 (12) 是一种平均速率计算：参考值越大或更新越频繁， R_{ref} 就越大；而不频繁更新则会使其保持较小。模式选择使用平滑有效吞吐量 $G(t)$ 作为可用容量和稳定性的代理指标。使用两个阈值 τ_{low} 和 τ_{high} 来实现滞回，其中 $\tau_{\text{low}} < \tau_{\text{high}}$。设模式为 $m(t) \in \{$正常，LB，人工智能$\}$。一种简单的滞回规则如下：

$$m(t) = \begin{cases} \text{AI (e.g., Wav2Lip) (BR3)}, & \tilde{G}(t) < \tau_{\text{low}}, \\ \text{Normal WebRTC (e.g., VP8) (BR1)}, & \tilde{G}(t) \geq \tau_{\text{high}}, \\ \text{Low Bitrate (e.g., AV1) (BR2)}, & \tau_{\text{low}} \leq \tilde{G}(t) < \tau_{\text{high}}. \end{cases} \tag{13}$$

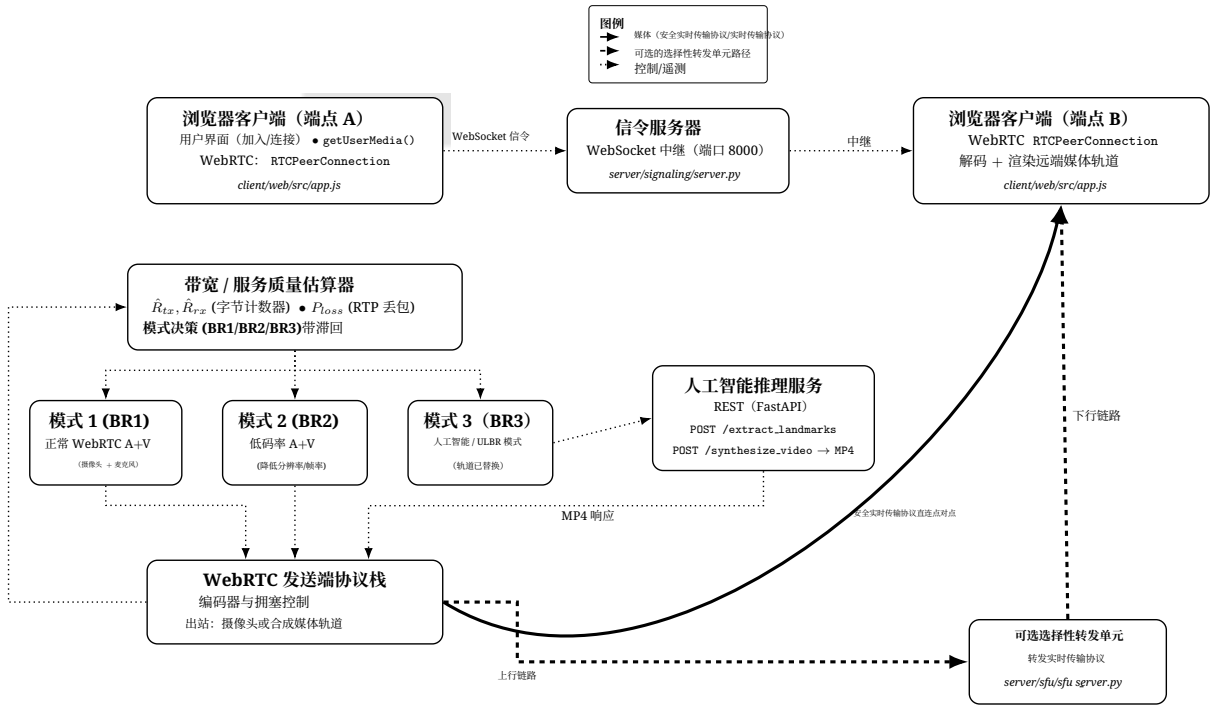

图 1：所提出的带宽自适应视频会议框架的系统概述。完成基于浏览器的信令和 WebRTC 会话建立后，发送端会定期评估带宽/服务质量，并采用三模式策略：BR1（标准音视频）、BR2（码率受限音视频）和 BR3（纯音频上行链路，结合服务器端说话人头部合成和轨道替换）。内容通过安全实时传输协议点对点传输，或通过可选的选择性转发单元传输。

滞回的作用是在网络围绕单个阈值波动时防止快速振荡。简单来说，只有当网络状况明显改善并超过 τ_{high} 时，系统才会升回正常模式；只有当网络状况明显恶化并低于 τ_{low} 时，系统才会降至人工智能模式。一个必要的可行性条件是，所选˜模式必须在估计容量范围内。使用 $G(t)$作为容量代理指标，可将可行性表示为

$$R_{\text{tot}}^{(m)}(t) \leq \tilde{G}(t), \tag{14}$$

其中， $R^{(m)}$ 总 (t) 根据选定模式，由式（7）、（9）–（8）或（10）计算得出。如果普通模式下违反式（14），控制器应降低视频负载（LB），或使用重建（AI）替代视频。如果 AI 模式下仍违反式（14），剩余的可行回退方案就是纯音频，因为在式（6）中，音频项通常具有最小且最稳定的比特率。人工智能重建路径改变了传输视频模态的含义，同时保留了会话的交互性。音频仍以实时方式传输，并驱动生成的嘴部运动；视觉流则变为根据参考内容和语音合成的渲染输出。这种设计在极低带宽环境下尤为实用，因为它避免了持续发送像素视频；像素视频是式（7）中的主要项，也是受限链路下最难保留的项。数学上的码率分解（6）–（10）明确体现了核心优势：AI 模式移除了 $R_{\text{vid}}(t)$，并将其替换为可控项 $R_{\text{ctrl}}(t)$ 和 $R_{\text{ref}}(t)$，其平均码率可通过式（11）和（12）按设计设定。基于遥测的容量估计（1）–（5）和切换规则（13）提供了一种简单稳定的机制，可在网络条件变化时切换模式。

4 自适应带宽控制器

自适应带宽控制器采用闭环运行，反复根据实时 WebRTC 遥测数据估算稳定的可用网络容量，并利用滞回机制选择运行模式（正常、低码率或人工智能重建），以避免链路波动时通话模式快速振荡。控制器在离散时刻 $t_k = k\Delta t$运行，其中 Δt 表示遥测采样间隔。瞬时上行链路和下行链路

吞吐量分别记为 $\hat{R}_{\mathrm{tx}}(t_k)$ 和 $\hat{R}_{\mathrm{rx}}(t_k)$，根据端点观测到的累计字节计数器计算得出。在公式 15 中，$B_{\mathrm{tx}}(t_k)$ 和 $B_{\mathrm{rx}}(t_k)$ 分别表示截至时间 t_k 测得的累计发送字节数和累计接收字节数，而差值 $B_{\mathrm{tx}}(t_k) - B_{\mathrm{tx}}(t_{k-1})$ 和 $B_{\mathrm{rx}}(t_k) - B_{\mathrm{rx}}(t_{k-1})$ 分别表示最近一个时间间隔内传输的字节数。公式 15 中的系数 8 用于将字节转换为比特，除以 $1000\Delta t$ 后即可得到以千比特每秒为单位的吞吐量：

$$\hat{R}_{\mathrm{tx}}(t_k) = \frac{8\big(B_{\mathrm{tx}}(t_k) - B_{\mathrm{tx}}(t_{k-1})\big)}{1000\,\Delta t}, \qquad \hat{R}_{\mathrm{rx}}(t_k) = \frac{8\big(B_{\mathrm{rx}}(t_k) - B_{\mathrm{rx}}(t_{k-1})\big)}{1000\,\Delta t}. \tag{15}$$

由于公式 15 中的原始估计值可能因短期排队、调度和突发式分组化而波动，˜控制器采用指数平滑来获得更稳定的数值。在公式 16 中，$\tilde{R}_{\mathrm{tx}}(t_k)$ 和 $\tilde{R}_{\mathrm{rx}}(t_k)$ 是经过平滑处理的吞吐量，而 $\alpha \in (0,\ 1)$ 控制响应性与稳定性之间的权衡，其中较大的 α会保留更多历史数据，降低对突发波动的敏感度：

$$\tilde{R}_{\mathrm{tx}}(t_k) = \alpha\,\tilde{R}_{\mathrm{tx}}(t_{k-1}) + (1-\alpha)\hat{R}_{\mathrm{tx}}(t_k), \qquad \tilde{R}_{\mathrm{rx}}(t_k) = \alpha\,\tilde{R}_{\mathrm{rx}}(t_{k-1}) + (1-\alpha)\hat{R}_{\mathrm{rx}}(t_k). \tag{16}$$

控制器还会考虑丢包，因为丢包会直接影响可解码性、冻结事件以及音频和视频的质量。最近一个时间间隔内的丢包情况通过累积 RTP 计数器估算。在公式 17 中，$P_{\mathrm{lost}}(t_k)$ 表示丢失数据包的累积数量，$P_{\mathrm{recv}}(t_k)$ 表示接收数据包的累积数量。公式 17 中的运算符 Δ 表示时间间隔差值，因此 $\Delta P(t_k) = P(t_k) - P(t_{k-1})$，该比率衡量了上一个遥测窗口内丢失的数据包所占比例：

$$\hat{p}_{\mathrm{loss}}(t_k) = \frac{\Delta P_{\mathrm{lost}}(t_k)}{\Delta P_{\mathrm{lost}}(t_k) + \Delta P_{\mathrm{recv}}(t_k)}, \qquad \Delta P(t_k) = P(t_k) - P(t_{k-1}). \tag{17}$$

一个有用的单一数值代理指标，用于衡量预期能够成功到达的已传输比特率比例，即有效˜有效吞吐量。如公式18所定义，$G(t_k)$按区间丢包率对平滑后的上行吞吐量进行折减，因此即使发送方以相同速率传输，更高的丢包也会降低估计的可用预算：

$$\tilde{G}(t_k) = \tilde{R}_{\mathrm{tx}}(t_k)\big(1 - \hat{p}_{\mathrm{loss}}(t_k)\big). \tag{18}$$

交互式会议质量还会受到时延和时延变化的影响。因此，保守的容量代理指标

˜可以使用测得的往返时延和抖动对 $G(t_k)$进行惩罚。在公式19中，$RTT(t_k)$表示ˆ测得的往返时延，$J(t_k)$表示测得的抖动，而 κ和 λ是非负权重，用于决定这些项对有效容量的削弱强度：

$$\tilde{C}(t_k) = \frac{\tilde{G}(t_k)}{1 + \kappa\,\widehat{RTT}(t_k) + \lambda\,\hat{J}(t_k)}. \tag{19}$$

˜公式19中的量 $C(t_k)$充当控制器对网络在时间 t_k所能支持能力的稳定性感知估计。

˜一旦控制器获得 $C(t_k)$，便使用滞回阈值选择运行模式。这些阈值记为τ_{low} 和τ_{high}，且满足 $\tau_{\mathrm{low}} < \tau_{\mathrm{high}}$；滞回通过要求阈值条件在切换模式前持续达到最短时长来实现。这种持续性通过计数器实现。在公式20中，$n_{\downarrow}(t_k)$统计 $C(t_k)$保持低于 τ_{low}的˜连续采样，$n_{\uparrow}(t_k)$统计 $C(t_k)$保持不低于τ_{high}的连续采样。指示函数 $\mathbb{I}[\cdot]$ 在其条件为真时等于1，否则等于0，从而确保相应条件不成立时每个计数器都会重置：

$$n_{\downarrow}(t_k) = \mathbb{I}[\tilde{C}(t_k) < \tau_{\mathrm{low}}]\,(n_{\downarrow}(t_{k-1}) + 1), \qquad n_{\uparrow}(t_k) = \mathbb{I}[\tilde{C}(t_k) \geq \tau_{\mathrm{high}}]\,(n_{\uparrow}(t_{k-1}) + 1). \tag{20}$$

最小稳定持续时间 T_{stable} 对应于最少数量的连续采样 $N_{\mathrm{stable}} = \lceil T_{\mathrm{stable}}/\Delta t \rceil$。公式21中的模式判定规则使用这些计数器来防止模式快速切换。模式变量 $m(t_k)$ 的取值属于集合 {正常、LB、人工智能}，其中 LB 表示低比特率像素流传输。如公式21所述，只有当网络低于下阈值并持续足够长时间后，控制器才会切换到人工智能重建；只有当网络高于上阈值并持续足够长时间后，控制器才会返回正常模式；在其他情况下则保持在中间低比特率模式：

$$m(t_k) = \begin{cases} \mathrm{AI}, & n_{\downarrow}(t_k) \geq N_{\mathrm{stable}}, \\ \mathrm{Normal}, & n_{\uparrow}(t_k) \geq N_{\mathrm{stable}}, \\ \mathrm{LB}, & \text{otherwise}. \end{cases} \tag{21}$$

选择模式后，控制器会确保预期传输需求不超过估计容量。公式22将总出站比特率需求分解为多个组成部分。在该公式中，$R_{\mathrm{aud}}(t_k)$ 表示音频比特率，$R_{\mathrm{vid}}(t_k)$ 表示编码像素视频比特率，$R_{\mathrm{ctrl}}(t_k)$ 表示用于重建的控制流比特率，$R_{\mathrm{ref}}(t_k)$ 表示偶尔进行参考更新所产生的摊销比特率：

$$R_{\mathrm{tot}}(t_k) = R_{\mathrm{aud}}(t_k) + R_{\mathrm{vid}}(t_k) + R_{\mathrm{ctrl}}(t_k) + R_{\mathrm{ref}}(t_k). \tag{22}$$

可行性条件在公式 23 中表示，其中 $R^{(m)}{}_{\text{总计}}(t_k)$ 表示选定模式 $m(t_k)$ 所隐含的需求。公式 23 中的不等式表明，所选模式的需求不应超过链路能够稳定提供的持续比特率：

$$R_{\mathrm{tot}}^{(m)}(t_k) \le \tilde{C}(t_k). \tag{23}$$

表达式 $R^{(m)}{}_{\text{总计}}(t_k)$ 是根据公式 22 得出的，方法是确定各模式中哪些项处于激活状态。在普通模式下，主要项为 $R_{\mathrm{aud}}(t_k)$ 和 $R_{\mathrm{vid}}(t_k)$，因此需求约为 $R_{\mathrm{aud}}(t_k) + R_{\mathrm{vid}}(t_k)$。在人工智能模式下，像素视频项被抑制，因此需求约为 $R_{\mathrm{aud}}(t_k) + R_{\mathrm{ctrl}}(t_k) + R_{\mathrm{ref}}(t_k)$。在低比特率模式下，控制器通过降低分辨率和/或帧率来减少像素视频需求。公式 24 给出了视频比特率随空间缩放和时间缩放变化的一种简单工程近似。在该公式中，$\eta \in (0,1]$ 是同时应用于宽度和高度的空间缩放因子，$\gamma \in (0,1]$ 是应用于帧率的时间缩放因子。因子 η^2 反映了每帧像素数随图像面积变化，因子 γ 反映了每秒帧数的减少：

$$R_{\mathrm{vid}}^{(\mathrm{LB})}(t_k) \approx \eta^2 \gamma\, R_{\mathrm{vid}}^{(\mathrm{Normal})}(t_k). \tag{24}$$

因此，低码率需求约为 $R_{\mathrm{aud}}(t_k) + R^{(\mathrm{LB})}{}_{\text{视频}}(t_k)$，控制器可以选择 $(\eta,\ \gamma)$，使等式 23 得以满足。

在 AI重建模式下，取代像素视频的比特率贡献是可以明确控制的。控制流比特率可以用更新速率和有效载荷大小建模。如等式 25 所定义，f_{ctrl} 表示每秒控制更新的次数，N 表示每次更新发送的描述符数量，d表示量化/打包后用于编码每个描述符的比特数。除以 1000 可将比特/秒转换为千比特每秒：

$$R_{\mathrm{ctrl}} \approx \frac{f_{\mathrm{ctrl}}\, N\, d}{1000}. \tag{25}$$

类似地，参考更新贡献是低频身份刷新所产生的摊销成本。如等式 26 所述，S_{ref} 是一次参考更新的大小（以比特计），T_{ref} 是两次更新之间的时间（以秒计）：

$$R_{\mathrm{ref}} \approx \frac{S_{\mathrm{ref}}}{1000\, T_{\mathrm{ref}}}. \tag{26}$$

˜等式 25 和 26 明确列出了控制器的可调参数。当 $C(t_k)$ 降低时，控制器可以降低 f_{ctrl}、降低N、通过采用更粗的量化来降低 d，并提高 T_{ref}，从而在保持˜$R_{\mathrm{aud}}(t_k)$实时运行的同时满足等式 23。当 $C(t_k)$ 增加时，控制器可以反向调整这些参数，以提升视觉稳定性和身份保持。

总体而言，该控制器由可测量的量和简单、可审计的规则驱动。公式 15 中的吞吐量估计、公式 16 中的平滑处理、公式 17 和 18 中的丢包与有效吞吐量计算，以及公式 19 中考虑稳定性的容量代理指标，共同实时估算链路所能承载的流量。公式 20 和 21 中的滞回与持久性逻辑可防止不稳定切换。公式 23 中的可行性约束将所选运行模式与测得的容量关联起来。其关键优势在于，当严重的带宽限制使 $R_{\mathrm{vid}}(t_k)$ 不可行时，切换到人工智能重建便可从公式 22 中移除占主导地位的像素视频需求项，并将其替换为由公式 25 和 26 控制的紧凑、可设计调控的项 R_{ctrl} 和 R_{ref}，从而在极度受限的网络条件下保持对话连续性。所有研究均采用 Wav2Lip 作为说话人头部后端。我们选择 Wav2Lip，是因为它应用广泛、开源，并且为唇形同步质量提供了可靠的基础。不过，该系统完全兼容其他 THG 后端（MakeItTalk、SadTalker 和 VadicTHG），我们已通过同一个推理 API 将这些后端集成其中。需要注意的是，后端切换不会影响控制器、SFU 和 WebRTC 流程。

5 实验设置

本节介绍如何以一种可重复、对所有基线方法公平且与在严苛带宽限制下保持交互式通信这一目标紧密相关的方法评估该系统。评估围绕四个要点展开：(i) 所有方法采用相同的视频会议任务和相同的测试内容；(ii) 控制

网络，使带宽、丢包、时延和抖动能够以已知方式变化；(iii) 同时测量网络层面的服务质量（QoS）和面向用户的体验质量（QoE）；(iv) 将离线模型质量与端到端视频会议行为区分开来。在整个测试过程中，所有算法都以实时方式运行，通话时长和音频内容相同，所有测量均按固定采样间隔 Δt 进行，以便开展时间对齐的比较。

实验系统由两个端点（呼叫方和被叫方）组成，并在需要时配备一台作为选择性转发单元（SFU）运行的转发服务器。每个端点都是正常的 WebRTC 参与者，负责采集麦克风音频以及（启用视频时）摄像头视频。信令通道仅用于创建会话；所有内容均通过实时安全通道传输。这些研究既包括点对点通话，也包括经 SFU 路由的通话，因为多方系统通常会为双人连接部署 SFU，而且 SFU 转发可能会改变丢包模式和时延特征。对于每种配置，均遵循相同的测试流程：建立通话，留出一段较短的预热时间以稳定编码器和拥塞控制，施加网络条件，然后让系统在固定的评估窗口内运行，并在此期间记录所有遥测数据和质量测量结果。

评估输入采用两种形式。第一种是实时交互式通话，参与者进行即兴交谈。该选项能够捕捉停顿、轮流发言和音量变化等真实的对话动态，还有助于测试系统在实际使用条件下能否正常运行。第二种是受控播放配置，使用预录音轨（以及可选的参考人脸视频）作为输入。第二种方式十分重要，因为它使测试具备可重复性：每种方法在相同的时间内获得完全相同的语音材料，从而避免将不同运行之间说话风格的变化误认为系统效果。当采用受控播放时，所有人工智能重建也使用相同的参考人脸图像，以确保身份标准保持一致。

这些实验旨在考察与本研究相关的完整带宽范围，并特别关注严苛带宽条件。通过流量整形调控网络，以便直接配置可用带宽、延迟、抖动和丢包。带宽控制通过限制发送端上行链路的最大速率来实现，因为上行链路通常是视频会议中的主要瓶颈。令 $B_{\max}$ 表示某次运行设定的带宽限制。在每次运行中，限速器会强制要求传输数据速率在短时间窗口内不超过 $B_{\max}$，从而促使编码器和拥塞控制进行调整。除了带宽限制外，我们还会施加独立的网络损伤，以营造真实且严苛的情境。可以设置固定的单向时延来模拟较长的传输路径，也可以添加随机抖动来模拟排队变异性。丢包可以采用均匀随机丢包或突发丢包，其中突发丢包更能反映无线干扰和队列拥塞。这些可控的网络损伤使研究能够涵盖“低带宽但网络稳定”和“低带宽且网络不稳定”两种情况，因为实际网络通常兼有这两种特征。每种网络损伤配置都会重复运行多次，以便观察变异性；同时，对所有基准应用相同的配置，确保比较公平。

所有基线方法都在完全相同的网络条件和输入内容下进行评估。第一个基线是普通的 WebRTC 视频会议，语音和像素视频均通过传统编码与传输方式交付。该基线展示了基于浏览器的视频会议在依靠编解码器码率控制和拥塞管理应对带宽下降时的正常运行方式。第二个基线是纯音频模式，此时视频已停用，仅传输麦克风音频。这个基线很重要，因为纯音频通常是实际产品采用的备用方案；它还提供了带宽使用量的下限，以及在没有视觉信号时维持对话连续性的参考。第三个基线是低码率像素视频，通过主动降低分辨率和／或帧率，尝试在带宽额度范围内继续传输视频帧。这个基线将“像素的激进压缩”与“通过重建进行语义替换”区分开来，后者正是本研究的核心贡献。我们既评估所提出系统的完整自适应形式（系统根据遥测数据选择模式），也评估固定模式设置（系统在整个运行期间被强制保持在一种模式）。固定模式运行用于确定有多少性能提升来自人工智能重建本身，而有多少来自切换逻辑。

实验设置的一个重要方面是日志记录与测量管线。端点以 Δt 的时间间隔捕获时间分辨遥测数据，生成吞吐量、丢包、抖动、往返时延和视频渲染信息的同步时间序列。令 $t_k = k\Delta t$ 为采样时间。上行链路和下行链路的吞吐量按照公式 15，使用累计字节计数器计算。丢包按照公式 17 计算。每个 t_k 都会将这些指标记录在一行中，从而能够绘制吞吐量随时间变化、丢包随时间变化等图表。除了传输计数器外，端点还会跟踪媒体层信息，例如传输帧率、有效分辨率以及视频冻结发生的时间。当通话仍在进行时，如果显示帧时间戳超过预设的冻结阈值 T_{freeze} 仍未前进，则判定发生了一次冻结事件。因此，一次运行中的总冻结时长是所有冻结时长之和，而冻结率则是评估窗口内处于冻结状态的时间所占比例。冻结率是一个直观且易于理解的指标

用于衡量观看者遭遇“卡顿视频”的频率。对于音频，通过监测播放缓冲区中的间隙，或留意实时传输协议音频时间戳的不连续情况来检测中断；整个音频中断时长也会以类似方式记录。

对于所提出的人工智能重建模式，我们加入了额外的检测机制，以量化端到端生成延迟及其对对话延迟的影响。设 t_k^{cap} 为采集参考帧和音频片段的时间，t_k^{req} 为发送推理请求的时间，t_k^{resp} 为收到合成片段的时间，t_k^{play} 为其首次作为外发视频轨一部分进行渲染的时间。合成流水线延迟可分为：（i）采集/封装延迟；（ii）网络请求/响应延迟；（iii）推理计算时间；以及（iv）解码/渲染延迟。端到端合成延迟的定义见公式 27。

$$T_{\mathrm{synth}}(k) = t_k^{\mathrm{play}} - t_k^{\mathrm{cap}}. \tag{27}$$

记录构成 $T_{\mathrm{synth}}(k)$ 的各个组件，可以清楚地看出延迟主要由计算还是传输造成，并有助于调整系统架构以适用于交互式使用。此外，我们还评估模式切换延迟。如果控制器选择在时间 t_k 切换模式，那么模式切换完成时间就是输出轨道完全处于新模式的最早时刻。切换时延即这两个时刻之差。这项测量至关重要，因为即使平均体验质量良好，频繁切换仍会损害用户体验。评估指标旨在同时体现效率和用户满意度。网络效率通过评估期间的平均吞吐量、百分位吞吐量以及数据使用总量来衡量。设 $R_{\mathrm{tx}}(t_k)$ 为公式 15 中的瞬时上行吞吐量。对于包含 K 个样本的一次运行，其平均上行吞吐量按公式 28 计算。

$$\overline{R}_{\mathrm{tx}} = \frac{1}{K}\sum_{k=1}^{K} \hat{R}_{\mathrm{tx}}(t_k), \tag{28}$$

总传输数据量通过将整个运行过程中的累计字节计数器差值换算为兆字节来确定。稳定性通过丢包率、抖动和冻结率衡量。响应性通过端到端延迟以及从带宽损失中恢复所需的时间衡量。恢复时间定义为从$\tilde{}(t_k)$大幅下降 C到选定模式根据公式 23 再次具备可行性的时间，这意味着通话恢复到稳定稳态。对于视频质量，我们考察两类指标。第一类是在有像素视频时，对解码像素帧计算的典型感知指标和结构指标。第二类是重建专用标准，例如身份一致性，以及音频与生成的嘴部运动之间的唇形同步一致性。唇形同步可通过评估音频特征与口部区域运动之间的对齐情况来衡量；身份特征则可通过比较参考身份与生成帧之间的人脸嵌入来评估。这些指标基于远端接收到的输出计算，因为用户实际看到的正是该输出。为确保比较公平，每次运行均采用相同的控制方法。所有方法使用相同的带宽配置、相同的时长和预热时间，并以相同的时间间隔进行日志记录。丢包等随机网络劣化使用固定随机种子生成，以便在每种方法中重复相同的丢包模式。每种场景均重复多次，所报告的结果既反映集中趋势，也体现变异性。评估主观质量时，人工评估均在一致的观看和聆听条件下进行，并采用随机呈现顺序以避免偏差。最后，实验区分了两种部署条件：一种是在本地计算机上进行推理（代表设备端或边缘推理），另一种是在远程服务器上进行推理（代表云端推理）。这种划分有助于研究合成时延和网络开销如何随人工智能模块所在位置而变化，同时保持人工智能模式固有的节省带宽行为不变。总体而言，实验设置旨在明确回答三个问题。第一，在带宽限制严苛的情况下，系统能否比像素视频基线更好地保持对话连续性，这可通过冻结率、音频稳定性和模式可行性体现。第二，人工智能重建模式能否提供实用的类视频体验，这可通过基于接收输出计算的唇形同步和身份一致性指标来证明。第三，自适应控制器能否可靠地切换模式且仅造成轻微干扰，这可通过切换时延以及带宽变化时不发生振荡来体现。本文其余部分将基于这些受控条件及相关记录数据呈现实验结果。

6 结果

本节通过以下内容展示会议系统的观测行为：(i) 时间分辨吞吐量轨迹；(ii) 包含明显模式切换事件的长期带宽日志；以及 (iii) 将时间序列汇总为易于解读的数据的汇总表。本次调查旨在以透明且可重复的方式说明：系统以传统 WebRTC 视频模式运行时的表现，以及切换到人工智能重建模式后这种表现如何变化。所有图表和表格均采用从记录的计数器中提取的相同基础数值。

图2所示的发送和接收吞吐量，是根据累计字节计数器采用公式15中所述的常规差分方法估算得出的。同样，图2中的长期上行链路轨迹采用按给定采样间隔汇总的记录上行吞吐量 $R_{\mathrm{tx}}(t)$，这是衡量会话随时间消耗多少上行链路容量的主要指标。图2展示了单次会话中吞吐量变化的细粒度视图。横轴表示时间（秒），纵轴表示吞吐量（千比特每秒）。两条曲线分别对应发送端吞吐量和接收端吞吐量，在图中标记为“发送 kbps”和“接收 kbps”。发送曲线近似表示瞬时应用发送流量（包括实时传输协议开销），因而反映了上行链路所承受的需求。接收曲线近似表示对端的瞬时接收流量，由于丢包、重传行为和测量不对称性，其数值通常略低于发送曲线。会话稳定时，两条曲线应平滑增长并彼此接近，表明网络正在以指定的媒体码率传输数据，且未出现严重丢包或拥塞。网络受限时，发送吞吐量往往会呈现阶梯式下降或振荡等适应性变化模式，反映出编码器码率控制与拥塞管理的综合作用。与此同时，如果丢包增加，接收吞吐量可能会更快下降，因为传输到接收端的数据包会减少。因此，两条曲线之间的差距本身就是一种间接的不稳定信号：持续存在的较大差距通常意味着丢包更多，或过程中码率下降更剧烈；而两条曲线几乎重合则表明数据传输良好。在本文中，图2主要用于验证日志记录流水线能否正确捕获时间分辨的发送/接收带宽，并展示系统在分钟以下的时间尺度上对条件变化作出反应的速度，这对任何实时控制器都至关重要。图2的解读也与控制器量密切相关。如公式15所述，时间 t_k 的吞吐量通过对跨越 Δt 的累计字节数作差得出；因此，时间间隔内的短时突发会表现为局部振荡，而整体趋势则体现有效发送速率。当控制器采用经过平滑处理的容量代理指标，例如公式19中的 $\tilde{}$（C）时，其目的在于避免对这些细微波动作出反应，而是针对持续性变化进行响应。因此，吞吐量随时间变化图为控制器采用平滑处理和持续性判定提供了直观依据：它表明即使在稳定运行期间，测得的速率也并非完全恒定，并展示了变化发生的时间尺度。图2还有助于检验模式转换是否造成明显中断。如果切换实现正确，图中应显示发送吞吐量从较高的运行模式（像素视频）平稳过渡到较低的运行模式（重建开销），既不会出现长时间吞吐量为零的情况（这意味着卡顿），也不会在不同水平之间反复切换（这意味着切换不稳定）。因此，该图可作为初步诊断工具，并为后续长期研究提供帮助。图2展示了覆盖一段长时间会话的上行链路带宽轨迹：实验前半段系统以正常像素视频模式运行，随后在指定时刻切换至人工智能重建模式。图中包含一条表示上行吞吐量的主曲线，以及一条标记人工智能模式启用时刻的垂直虚线。该图有两方面用途。首先，它直观对比了不同运行模式：常规视频模式通常需要更多上行链路容量，因为发送的有效载荷除音频外，还包含压缩像素帧。其次，它表明人工智能模式转换后出现的下降并非短暂现象，而是在较长时间内保持稳定；这一点至关重要，因为短期测试可能掩盖不稳定、漂移或速率持续上升等问题。图2中的虚线是研究设计的重要组成部分，为前后对比提供了参照。虚线之前的区域对应正常模式下上行链路数值的分布，之后的区域则对应人工智能模式下的分布。由于日志每秒记录一次，每个时段都包含大量数据，因此即使不训练模型，这种比较也具有统计意义。最主要的定性发现是，上行链路曲线在模式切换时刻骤然下降，随后一直保持在狭窄的低位区间。这一表现正是码率分解所预测的结果。在正常模式下，总发送速率主要由公式22所示分解中的像素视频项决定。在人工智能模式下，像素视频项被抑制，并由公式10所述的较小控制项和参考项取代；因此，只要音频和控制信号保持稳定，预测的上行链路需求就会降低并维持在低位。长期图表表明，这种表现确实能在数小时内实现，这有力支持了该技术适用于受限场所的持续视频会议，而不仅仅是短暂演示的论点。表1将长期轨迹归纳为汇总数据，便于直接比较不同模式。上行链路均值和下行链路均值量化了测量时段内两个方向所用带宽的平均值。总均值是上行链路均值与下行链路均值之和；在关注整体网络成本时，这是一个有用的单一指标。总中位数则稳健地评估“典型”运行点，排除罕见峰值的影响。总计 $p95$给出第95百分位数，反映接近最坏情况的需求；这对于容量规划，以及判断该方法是否会产生可能压垮弱连接的突发流量至关重要。“每小时数据量”和“5小时数据量”两列将平均吞吐量换算为总传输数据量，这对于使用按量计费或受限连接的客户而言，是一项切合实际的指标。该表支持两项独立发现。首先，正常模式需要大量带宽，因为它会持续发送像素视频。这与公式7相符，其中主要成本来自视频项。其次，人工智能模式显著降低了上行链路和下行链路的消耗，因为像素视频项被删除，并由音频和紧凑的控制表示取代，如

公式10。这一下降不仅体现在平均值上，也体现在中位数和 $p95$上，这意味着降幅是持续的，并非依赖于少见的低速率时段。综合解读各列至关重要。某种技术的平均值较低，但 $p95$较高，表面上看平均效率不错，却仍可能产生导致抖动或缓冲的峰值；相比之下，同时保持中位数和高百分位数较低的方法，更有可能在弱连接上保持稳定。就此而言，表1证明，AI模式带来的不仅是较低的平均速率，还有在较长时间内更可预测的速率范围。每一行对应一种带宽限制和一种技术。WebRTC视频基准反映的是在相同限制下常见的像素流传输技术。纯音频基准反映了许多系统采用的最低带宽回退方案，并为没有视觉呈现时的对话连续性提供了参考点。标为“我们的方法（AI）”的建议技术体现了语义替代方法，即通过音频和紧凑控制信号重建视觉呈现，而不是传输像素。表1中的各列旨在同时体现用户感知和系统稳定性。体验质量列描述感知质量（例如通过平均意见分或类似定义的分数衡量）。唇形同步列表示嘴部动作与音频的匹配程度。身份识别列评估重建人脸与参考身份的相似程度。卡顿列评估视频不稳定性；数值越低越好，因为这表示画面冻结次数更少或持续时间更短。延迟列衡量端到端延迟，数值越低越有利于自然轮流发言。表1的结构有意与模式设计相匹配。在严格带宽上限下，传统WebRTC视频通常会变得不稳定，因为它仍需要最低速率来传输视频帧，而且其视觉流容易受到丢包和预测链中断的影响，从而增加画面冻结。纯音频或许能够保持对话连续性，却无法提供视觉呈现，因此唇形同步和身份识别不适用。在严格带宽状态下，建议的AI方法很可能表现出色，因为其传输负载主要是音频和紧凑控制信息，因此在限制严苛时更实用，同时仍能提供类视频输出。如果表中提供具体指标，应显示“我们的方法（AI）”在最低带宽上限下比WebRTC视频具有更高的体验质量和更少的画面冻结，同时在唇形同步和身份识别方面取得显著分数，而这些是纯音频无法实现的。对于论文的论述，表格最重要的解读方式是逐一比较各带宽上限下的表现；在64–256 kbps范围内，最佳方法应是所提出的方法，因为这正是语义替代旨在胜过高强度像素压缩的工作区间。每个后续行都会移除或削弱一个组件，例如断开控制信号、降低参考更新频率、删除时间平滑，或采用更小的模型。平均意见分/体验质量列衡量整体感知质量，唇形同步列反映视听同步，而闪烁列记录时间不稳定性，例如面部动画抖动、帧间不一致或身份漂移。该表旨在避免结论含糊不清。如果完整系统表现最佳，但较小的替代方案效果也几乎相当，这就表明部署时可以简化哪些元素。反之，如果移除时间平滑导致闪烁显著增加，就说明平滑并非可有可无的修饰，而是实现稳定观看所必需的。同样，如果删除控制信号会降低唇形同步或身份稳定性，就能支持即使在低带宽模式下也发送紧凑描述符的决策。对于审稿人而言，消融实验尤为重要，因为它们能够证明性能来自所提出的系统架构，而不是某个模型配置意外产生的伪影。只有快速切换且不干扰对话，模式切换才有益。切换延迟的测量方式是：从控制器识别到满足切换条件的时刻起，到接收端看到新模式生成的第一帧稳定帧为止。这一定义体现了用户感知的中断概念，因为只有当接收端的新视觉输出趋于稳定时，切换才算完成。在控制器中，检测时间对应于首个样本 t_k：此时公式20中的持续性计数器满足稳定性标准，且公式21中的模式选择发生状态变化。完成时间则对应于输出轨道和渲染管线开始生成新模式下的帧，且不存在超出短暂稳定性容差窗口的间隔的首个时刻。测试通过受控流量整形对带宽进行阶跃调整，从而反复触发阈值交叉。整个过程中音频保持连续，因此观测到的任何中断主要是视觉上的，可归因于切换机制，而非完整的会话重启。报告切换延迟时，必须同时纳入集中趋势和离散程度，因为即使平均值较低，长尾仍可能偶尔造成严重干扰。因此，延迟表应包含平均值和中位数，以及变异性和极值的衡量指标。分析还应区分不同的转换类型，因为切换到AI模式可能需要初始化重建和控制流处理，而切回像素视频则可能需要编码器启动。如果系统避免了会话重新协商，那么延迟应主要反映本地管线重新配置和缓冲。成功的切换机制应具有有界延迟，且不应造成音频中断、通话掉线或反复振荡。结合图2来看，切换分析不仅说明切换后带宽消耗有所下降，也表明这一降幅能够以最小的瞬态干扰实现，这对实时会议至关重要。

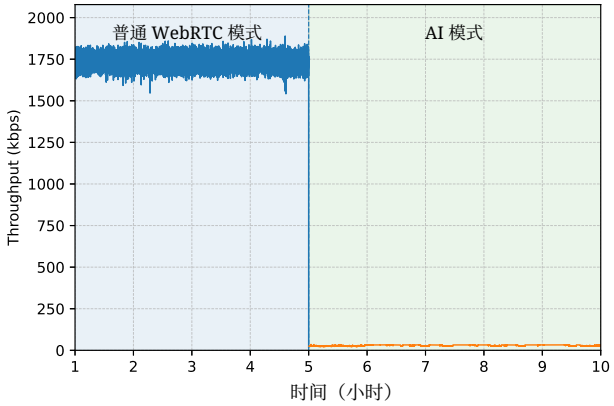

图2：标准 WebRTC 与 AI 合成的吞吐量（千比特每秒）对比分析。请注意，与标准 WebRTC 基准相比，AI 合成阶段实现了显著的带宽节省。

表1：带宽和数据用量

Mode	上行链路 Mean （千比特每秒）	下行链路 Mean （千比特每秒）	总计 Mean （千比特每秒）	总计 中位数 （千比特每秒）	总计 p95 （千比特每秒）	每小时数据量 （吉比字节）	5小时数据 （吉比字节）
AV1	1486.33	1418.04	2904.38	2972.10	3503.70	1.31	12.76
人工智能模式	28.23	15.42	43.65	32.80	64.70	0.02	0.19

7 讨论

研究结果表明，以音频驱动的说话人头部重建方法取代像素视频传输，或许能在极低带宽下维持类视频的对话体验。这一趋势背后的关键工程原因，是传输负载中的主导部分发生了变化。在传统视频会议中，编码视频项往往是总带宽的最大组成部分，也最容易受到网络拥塞的影响，因为帧间预测会使视频流容易遭受丢包和时延波动。带宽降低时，编码器必须降低画质和帧率，而传输层也可能面临更严重的丢包，从而造成画面冻结和播放不稳定。在所提出的技术中，AI重建模式会停止持续传输像素视频，转而传输实时语音音频、紧凑控制信息和不频繁的参考更新。这使带宽需求从高变异性、对丢包敏感的像素流，转变为规模更小且更易控制的负载，从而解释了表1所总结的大幅带宽节省。吞吐量轨迹进一步表明，AI模式启动后，系统可能持续处于低速率状态；对于长期受限而非仅短暂突发受限的网络而言，这对实际部署至关重要。尽管有这些优势，这项技术也存在必须明确指出的严格局限。首先，重建视频流并非对真实摄像头画面的忠实像素级呈现，而是一种合成近似，其准确性取决于说话人头部模型的鲁棒性、参考身份的质量以及输入音频的特征。如果参考图像光线不足、部分被遮挡，或不能代表用户的日常外貌，身份一致性就可能受到影响。同样，如果说话人频繁更换或存在多个声音，除非系统具备说话人分离和身份切换功能，否则重建人脸可能无法对应当前说话人。这可能导致听觉与视觉身份识别不匹配，在正式会议或临床交流等特定场合可能并不合适。其次，对于模型训练分布未充分涵盖的头部姿态、情绪或遮挡情况，重建质量可能会下降。极端偏航角、快速头部运动、手部遮挡面部、麦克风遮住部分嘴唇，以及低光照噪声，在实际通话中都很常见，并可能造成伪影、时间闪烁或偶发的帧不稳定。第三，长时稳定性面临挑战：生成结果可能在外观上缓慢漂移，也可能根据模型处理时间条件控制的方式出现周期性伪影。即便平均唇形同步表现稳定，当语音包含快速音素转换、背景噪声或笑声、咳嗽、说话人重叠等非语音噪声时，仍可能偶尔出现不同步。第四，AI模式依赖计算能力。尽管带宽需求有所降低，端点或推理服务仍必须在实时限制内完成模型推理。如果计算能力不足或发生拥塞，系统可能会以更高延迟或更低帧率换取带宽节省。这意味着对于某些设备，

表2：音频/视频抖动、往返时延和冻结次数。数值越低越好（↓）。

方法	音频抖动（毫秒）↓	视频抖动（毫秒）↓	往返时延（毫秒）↓	卡顿↓
模式 1	2.414	4.186	2.729	14
模式 2	2.138	5.090	18.910	17
模式 3	0.256	3.000	2.333	0

尤其对于低功耗终端，本地推理可能并不可行，因此服务器端或边缘推理变得至关重要，但这会带来额外的网络延迟和潜在的隐私问题。故障模式可分为传输层故障、控制器层错误和合成层问题。传输层故障包括极高的丢包率、明显的抖动或短暂断连，这些都会影响实时传输；在这种情况下，即使是低码率控制信号也可能无法稳定送达，而最佳可行的备用方案可能只有纯音频。控制器层错误包括网络状况长时间徘徊在阈值附近时模式间反复振荡、平滑处理过慢导致切换延迟，或阈值设置过于激进导致过早切换。滞回和持久性逻辑可以降低这些风险，但系统仍容易受到参数选择和实际网络异质性的影响。合成层问题包括参考输入有误、音频采集困难、模型运行时故障，或出现非人类声音等意外内容，这些都可能导致不理想的视频片段。因此，系统应纳入防护措施，例如输入验证、对地标/控制信息提取的置信度估计、回退到静态化身或最近的良好帧，以及在合成无法达到最低质量要求或延迟约束时迅速切回纯音频。从实时传输的角度看，一项关键约束是切换不应要求重新协商底层 WebRTC 会话，因为重新协商会引入很大且不可预测的延迟；因此，在保持会话不变的同时替换外发视频轨，是一项重要的部署特性，但也需要谨慎处理时序和缓冲，避免接收端将这一变化误判为流故障。将合成媒体用于通信时，隐私、伦理和安全都是重要议题。即使 AI模式可以减少广播的原始视频量，系统仍会使用生物特征信息，也可能涉及敏感内容。参考照片和音频包含可识别个人身份的线索，而重建视频则是对真实人物的合成呈现。这会带来遭到滥用的风险，例如冒充、虚假录音录像或捏造误导性信息。因此，负责任的部署要求明确取得用户许可，并保持公开透明。用户应知晓其外发视频是合成的，接收端也应能够识别该视频流为合成内容，例如通过视觉标识或元数据。此外，还应实施访问控制和数据最小化。参考图像和音频片段的保存时间不应超过合成所需时长；如果采用服务器端推理，默认应避免存储，或依据严格的数据保留规定妥善保护数据。从传输角度看，应通过常见的安全实时媒体通道确保基本的机密性和完整性。从分析角度看，所有遥测数据收集都必须符合隐私标准；带宽和质量日志可能反映用户行为和连接状况的趋势，因此应将日志作为敏感运营数据处理。如果系统传输精简的面部运动描述符，应说明这些描述符是否可能被用于重建身份或预期用途之外的表情；此类表征仍可能属于生物特征信息，因此其处理应遵循与视频相同的预防措施。部署方面的考量包括兼容性、可扩展性、延迟和运营韧性。兼容性不可或缺，因为视频会议生态系统各不相同；并非所有客户端都支持相同的编解码器，也并非所有环境都支持相同的网络穿越行为。一种切实可行的方式是确保 AI模式无需对接收端进行常规 WebRTC 客户端无法支持的修改即可运行，例如将合成视频作为标准视频轨提供。多方会话中，可扩展性尤为重要。在像素视频模式下，选择性转发单元（SFU）可以通过转发流来扩展，但在 AI 模式下，系统必须决定在何处进行合成。如果每个发送端都在本地合成后再传输，服务器成本较低，但对客户端计算能力的要求会提高。如果在边缘或云服务上进行合成，客户端需求会降低，但服务器成本和网络延迟会上升。应根据部署限制确定这些权衡，例如边缘端图形处理器的可用性、预估参与者数量，以及有关生物特征处理获准地点的法律考量。延迟对对话质量至关重要；因此，合成流水线的设计必须确保 AI模式带来的额外延迟处于可接受的交互范围内。这需要谨慎选择音频分块大小、缓冲技术和帧生成频率，并降低参考信息更新成本。运营鲁棒性包括监控和回退系统。系统应持续核验当前模式在观测到的容量条件下是否仍可行，并保留安全的回退阶梯，例如正常 →低码率 →人工智能重建 →纯音频。此外，系统还应能够应对推理服务的短暂故障，例如保存近期参考数据、在可行时预热模型，并采用超时机制，避免错误中断请求。

8 结论

本研究展示了一种自适应视频会议解决方案：它通过实时遥测数据引导，从基于像素的视频传输切换至音频驱动的说话人头部重建模式，从而在带宽严重受限时仍能正常运行。长期带宽日志显示，启用 AI模式后，持续网络需求大幅降低，如表1所示。这表明，即使仍提供视觉流，系统也能在远低于传统视频会议典型需求的条件下运行。结果框架还进一步明确了在特定带宽限制下应如何评估性能，因为带宽使用量的中位数为 32.80kbps，同时兼顾稳定性。从工程角度看，核心贡献在于视频的语义替代：系统不再试图将像素压缩到实用极限之外，而是传输语音音频和精简的视觉控制信息，并重建出可信的说话人头部输出，从而降低带宽需求并提升受限网络中的稳定性。研究还强调，这种能力必须配合严格的模式切换逻辑、安全回退机制，以及负责任的隐私和披露程序，因为合成媒体会改变通信的安全与信任模型。未来工作可通过提升对姿态和遮挡的鲁棒性、扩展对多说话人和多方身份管理的支持、降低低功耗设备上的合成延迟，以及集成更完善的保障措施和透明度机制，来增强系统，使其适用于敏感应用领域。

参考文献

[1]K. Mouratidis 和 A. Papagiannakis，《COVID-19、互联网与出行：远程办公、远程医疗、电子学习和网上购物的兴起》，《可持续城市与社会》，第 74 卷，文章编号 103182，2021年。doi:10.1016/j.scs.2021.103182。[2] S. Omboni 等，《COVID-19 期间远程医疗的全球影响：当前证据与未来建议》，《互联健康与远程医疗》，综述文章，2022年1月。doi:10.20517/ch.2021.03。[3]A. Jones、P. Sevcik 和 R. Wetzel，《支持居家办公和电子学习的高效视频会议互联网连接要求》，NetForecast 报告 NFR5137，2021年2月15日。在线获取：https://www.netforecast.com/wp-content/uploads/NFR5137-Videoconferencing_Internet_Requirements.pdf。[4] M. Rudow、F. Y. Yan、A. Kumar、G. Ananthanarayanan、M. Ellis 和 K. V. Rashmi，《Tambur：通过流式编码实现高效的视频会议丢包恢复》，载于《第*20*届 *USENIX* 网络系统设计与实现研讨会（*NSDI*）论文集》，2023年。在线获取：https://www.usenix.org/system/files/nsdi23-rudow.pdf。[5] H. Zhang、P. Wu、G. Li、Y. An 和 H. Zhang，《一种流式可变神经语音编解码器》，《人工智能工程应用》，第 162 卷，文章编号 112418，2025年。doi:10.1016/j.engappai.2025.112418。[6] W3C，《WebRTC：浏览器中的实时通信》，W3C 技术报告。在线获取：https: //www.w3.org/TR/webrtc/。[7] Y.-J. Choi 和 B.-G. Kim，《用于视频超分辨率的分层循环变换器网络》，《人工智能工程应用》，第 166 卷，文章编号 113714，2026年1月。doi:10.1016/j.engappai.2025.113714。[8] Y. Zhang、T. Zhang、S. Wang 和 P. Yu，《一种基于深度学习辅助视频显著性和恰可察觉失真的高效感知视频压缩方案》，《人工智能工程应用》，第 141 卷，文章编号 109806，2024年12月。doi: 10.1016/j.engappai.2024.109806。[9] S. Mazumdar，《6G技术对社会5.0的教育影响》，载于爱思唯尔电子书，2025年1月，第 255–265 页。doi:10.1016/b978-0-443-27434-3.00017-9。[10]M. Z. Naser，《机器学习认知无知、隐藏悖论及其他矛盾指南》，《*Wiley*跨学科评论：数据挖掘与知识发现》，第 15 卷，第 3 期，2025年。doi: 10.1002/widm.70038。[11] V. K. Rakesh、S. Mazumdar、S. Pal、A. Das 和 T. Samanta，《说话人头部生成的进展：技术、指标与挑战综述》，《视觉计算机》，第 42 卷，第 1 期，2025年12月。doi:10.1007/s00371-025-04232-w。[12]M. Riveiro 和 S. Thill，《“这不是我预期的输出！”：终端用户期望在创建人工智能系统解释中的作用》，《人工智能》，第 298 卷，文章编号 103507，2021年4月。doi:10.1016/j.artint.2021.103507。

[13] E. L. Turner、J. E. Dobson 和 S. J. Pocock，《流行病学出版物中连续风险因素的分类：当前实践调查》，*Epidemiologic Perspectives& Innovations*，第 7 卷，第 1 期，文章编号 9，2010年10月。doi: 10.1186/1742-5573-7-9。[14] P. Megyesi、A. Botta、G. Aceto、A. Pescap´e 和 S. Moln´ar，《软件定义网络中可用带宽测量的挑战与解决方案》，*Computer Communications*，第 99 卷，第 48–61 页，2016年12月。doi: 10.1016/j.comcom.2016.12.004。[15] C. W. F. Parsonson、J. L. Benjamin 和 G. Zervas，《用于数据中心网络基准测试的流量生成》，*Optical Switching and Networking*，第 46 卷，文章编号 100695，2022年6月。doi: 10.1016/j.osn.2022.100695。[16] W3C，《WebRTC 统计 API 的标识符》，W3C 候选推荐标准草案，2025年9月25日。在线获取：https://www.w3.org/TR/webrtc-stats/。[17]M. Westerlund 和 S. Wenger，《RTP 拓扑》，RFC 7667，2015年11月。在线获取：https://datatracker.ietf.org/doc/html/rfc7667。[18] J.-M. Valin、K. Vos 和 T. Terriberry，《Opus 音频编解码器定义》，RFC 6716，2012年9月。在线获取：https://datatracker.ietf.org/doc/html/rfc6716。[19] K. R. Prajwal、R. Mukhopadhyay、V. P. Namboodiri 和 C. V. Jawahar，《只需一位唇形同步专家，即可实现野外语音唇形生成》，载于《*ACM* 多媒体（*MM*）论文集》，2020年。doi:10.1145/3394171.3413532。[20] Y. Zhou、X. Han、E. Shechtman、J. Echevarria、E. Kalogerakis 和 D. Li，《MakeItTalk：具备说话者感知能力的说话人头部动画》，载于《*ACM* 图形学汇刊（*TOG*）》，2020年。doi:10.1145/3414685.3417774。[21] X. Zhang 等，《SadTalker：为风格化音频驱动单图说话人脸动画学习逼真的三维运动系数》，载于《*IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）论文集》，2023年。doi:10.1109/CVPR52729.2023.00836。

作者贡献

VineetKumarRakesh: 概念化、方法学、软件（后端开发和服务部署）、可视化、写作（审阅与编辑）、监督。**Soumya Mazumdar:** 软件（前端开发及部分 JavaScript 贡献）、验证、调查、写作（初稿）。**Tapas Samanta:** 验证。**Hemendra KumarPandey:** 资源、调查、验证。**AmitabhaDas:** 验证。**Sarbajit Pal:**资源、验证。所有作者均审阅并批准了稿件的最终版本，并同意对工作的各个方面负责。

致谢

本研究得到了可变能量回旋加速器中心（VECC）和霍米·巴巴国家研究所（HBNI）的支持；两者均隶属于印度政府原子能部（DAE），并提供了必要的设施和技术支持。作者感谢同行评审人提出的深刻意见和建设性建议。我们还要感谢 VECC图书馆的工作人员在本研究过程中给予的宝贵协助。

利益冲突声明

作者声明，他们不存在可能看似影响本研究所述工作的已知经济利益冲突或个人关系。

数据可用性声明

经合理请求，可向通讯作者获取支撑本次调查结论的数据。本出版物中的所有研究结果均在调查期间即时产生，因此无需额外许可即可访问或使用这些数据。

作者简介



Vineet Kumar Rakesh 是印度原子能部可变能量回旋加速器中心（VECC）的技术官员（科学类），在软件工程、数据库系统和人工智能领域拥有超过22年的经验。他的研究重点是说话人头部生成、唇读以及用于实时视频会议的超低码率视频压缩。他正在孟买的霍米·巴巴国家研究所攻读博士学位。Rakesh先生曾参与VECC的办公自动化、OCR系统和数字化转型项目。他是印度工程师学会的准会员，也是原子能部集团成就奖获得者。



Soumya Mazumdar 正在攻读双学位：在 Gargi Memorial Institute of Technology 攻读计算机科学与商业系统工学学士学位，并在印度理工学院马德拉斯分校攻读数据科学理学学士学位。他参与了跨学科研究，在爱思唯尔、施普林格、IEEE、威立和 CRC出版社出版的期刊及编著文集中发表了 25 篇以上作品。他的研究兴趣包括人工智能、机器学习、6G 通信、医疗保健技术和工业自动化。



Tapas Samanta博士 是可变能量回旋加速器中心（VECC）、原子能部（印度）计算机与信息学组的高级科学家兼负责人。他拥有二十多年的工作经验，研究领域涵盖人工智能、工业自动化、嵌入式系统、高性能计算和加速器控制系统。他还负责 VECC 的技术转移计划和公众科学传播工作。



Hemendra Kumar Pandey 是印度加尔各答原子能部可变能量回旋加速器中心（VECC）放射性离子束设施组的科学官。他在印度理工学院克勒格布尔分校获得博士学位，在阿拉哈巴德大学获得技术硕士学位。他于1999年加入巴巴原子研究中心，并自2000年起任职于 VECC，参与了粒子加速器射频和微波系统相关工作，包括放射性离子束设施的开发工作。他还担任霍米·巴巴国家研究所副教授。他的研究兴趣包括粒子加速器射频系统、束流诊断、高功率射频放大器开发、混合信号射频集成电路设计，以及基于加速器的技术中的抗辐射器件。



Amitabha Das博士 是加尔各答贾达普大学核研究与应用学院的主任兼负责人。他的研究兴趣包括核仪器、嵌入式系统、反应堆控制系统以及基于FPGA的实时数据采集。他还参与了唇读和手语识别等人工智能驱动的应用，并指导了核反应堆控制方法方面的前沿研究。



Sarbajit Pal博士 是印度政府原子能部可变能量回旋加速器中心（VECC）控制与仪表组的退休高级科学家及前负责人。他拥有电子工程博士学位，并为粒子加速器的控制与仪表系统作出了重要贡献，其中包括K500超导回旋加速器。他的专业领域包括嵌入式系统、实验物理以及基于EPICS的控制架构。