arXiv:2512.13904v1 [cs.MM] 15 Dec 2025

用于视频翻译的生成式人工智能：多语言视频会议的可扩展架构

AmirkiaRafiei Oskooei* ErenCaglar Ibrahim ,Sahin Ayse Kayabay Mehmet S. Aktas

土耳其伊斯坦布尔耶尔德兹技术大学计算机工程系，邮编 34220

摘要

将级联生成式人工智能流水线实时部署于视频翻译等应用，面临着重大的系统层挑战。这些挑战包括顺序模型推理带来的累积延迟，以及使多用户视频会议应用难以扩展的二次（$\mathcal{O}$（N^2））计算复杂度。本文提出并评估了一种实用的系统级框架，旨在缓解这些关键瓶颈。所提出的架构采用轮流发言机制，在多用户场景下将计算复杂度从二次降至线性，并采用分段处理协议来管理推理延迟，从而实现感知上的实时体验。我们实现了一个概念验证流水线，并在多层级硬件配置上进行了严格的性能分析，涵盖商品级（NVIDIA RTX 4060）、云端（NVIDIA T4）和企业级（NVIDIA A100）图形处理器。客观评估表明，该系统在现代硬件上实现了实时吞吐量（$\tau < 1.0$）。进一步开展的主观用户研究验证了这一方法，结果显示，用户非常乐于接受可预测的初始处理延迟，以换取流畅、不间断的播放体验。本研究提出了一种经过验证的端到端系统设计，为在多语言通信平台中部署可扩展的实时生成式人工智能应用提供了切实可行的路线图。

关键词：生　成式人工智能；应用计算机视觉；多媒体；人机交互；深度学习　　g

注： 本手稿是作者已接收的论文版本，该论文发表于 *Applied Sciences* (MDPI)，2025年。最终版本可在出版商网站获取：https://www.mdpi.com/2076-3417/15/23/12691。

1 引言

强大的生成式人工智能（GenAI）与数字通信平台在全球的普及相结合，正在从根本上重塑人类互动。生成式人工智能模型能够创造新颖的高保真内容，包括文本、代码 [1, 2, 3, 4], 音频和视频 [5, 6, 7]，有望让在线环境更具沉浸感和功能性。随着视频会议系统、增强现实/虚拟现实（AR/VR）[8, 9, 10, 11, 12], 以及社交网络 [13, 14, 15] 等平台得到广泛应用，这一技术变革带来了变革性机遇，有望消除全球通信中长期存在的障碍，尤其是语言障碍 [16, 17]。在这一领域，“视频翻译”——也称为视频到视频翻译或面对面翻译——是一种新兴范式，备受关注 [18]。视频翻译旨在通过全面翻译人类表达的各个方面，提供无缝的多语言体验。这个过程包括转换口语内容，

*通讯作者：amirkia.oskooei@std.yildiz.ed　　u.tr

同时保留发言者的语气和风格，并且尤为重要的是，使其唇部动作与译后的语音同步。这种全面的翻译能够促成更自然流畅的对话，为国际商务、全球学术会议和多元文化社交活动带来巨大价值。要实现这一目标，需要构建端到端处理流水线，整合多个生成式人工智能模型，用于自动语音识别（ASR）、机器翻译（MT）、文本转语音（TTS）合成和唇形同步（LipSync）等任务，如图1所示。

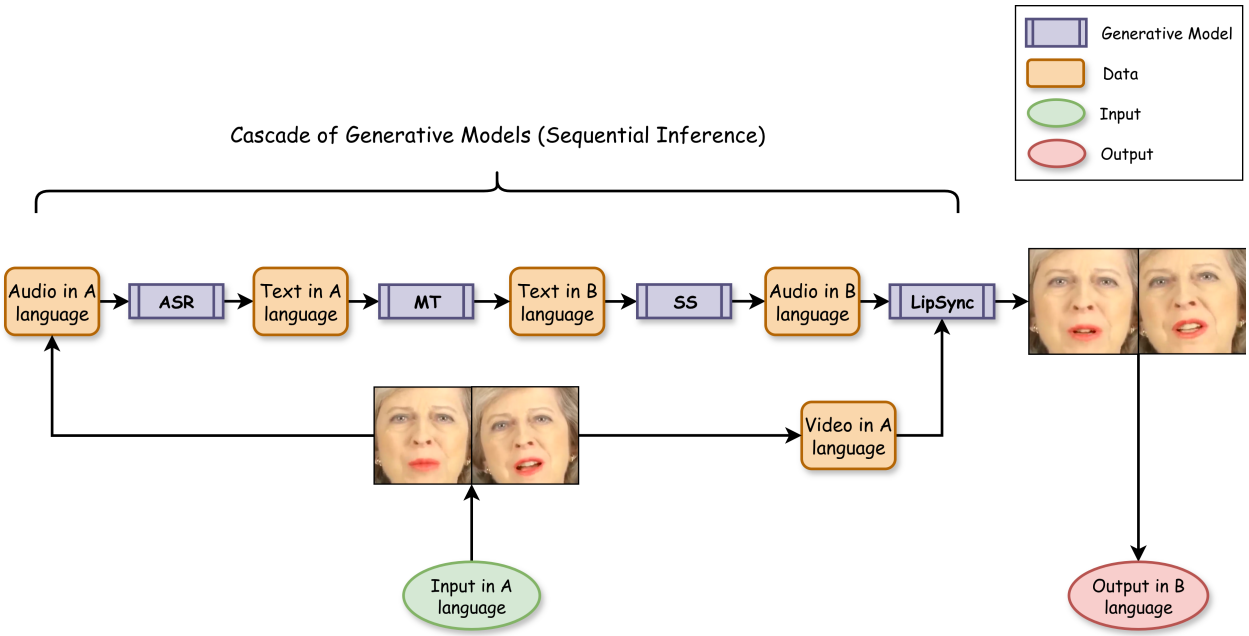


图1： 视频翻译的顺序式四阶段处理流水线。语言A的输入依次经过自动语音识别（ASR）、机器翻译（MT）、语音合成（SS）和唇形同步（LipSync）模型处理，生成语言B的完整翻译结果，并使画面与语音同步。

然而，要在实时、大规模应用中实际部署这些复杂的多阶段处理流水线，仍面临严峻的系统级工程挑战，而现有研究尚未充分解决这些问题。生成式人工智能模型的计算开销很大，需要图形处理器等高性能硬件才能及时运行。在视频会议等实时环境中，这一需求更加突出，由此产生了两个主要瓶颈：

1. 延迟：依次运行多个深度学习模型会造成显著的处理延迟。级联中的每个阶段都会增加总推理时间，因此难以实现流畅、不间断对话所需的低延迟吞吐量。

2. 可扩展性：在多用户视频会议中，若采用简单直接的实现方式，每位参与者都需要同时处理其他所有发言者的视频流。这种方法会导致 O(N2) 的计算复杂度，成本高得难以承受，技术上也难以管理，即使参与者（N）人数很少也是如此。

虽然许多研究聚焦于改进单个模型（模型级优化），但在开发架构与协议层解决方案方面仍存在明显的研究空白，这些解决方案能够使此类系统在实时多用户场景中有效运行（系统级优化）。尽管轮流发言和批处理等架构原语在分布式系统中已广为人知，但如何将它们专门整合并应用于解决实时生成式视频翻译所面临的独特计算与语义挑战，仍是一个尚未解决的

关键研究领域。为弥补这一关键空白，并推动视频翻译及类似生成式人工智能流水线的实际部署，本文提出了一个系统级框架。本研究的主要贡献如下：

- 我们提出了一种全新的系统架构，旨在支持生成式处理流水线在多用户视频会议环境中的可扩展部署。该架构的主要创新包括：
  - 一种用于管理说话者轮次的令牌环机制，可将系统的计算复杂度从 O(N2) 降至线性复杂度 O(N)，从而确保可扩展性。– 一种结合逆吞吐量阈值化的分段批处理协议，为管理延迟并实现近实时性能提供了数学框架。

- 我们通过概念验证实现，对该框架进行了实证验证。本研究对商品级和企业级硬件进行了性能分析，并通过用户研究开展主观评估，以确认系统在实际应用中的可行性及用户感知。

本文是对我们先前发表于第 25 届计算科学及其应用国际会议的初步研究成果的大幅扩展版本 [18]。会议论文介绍了概念验证，而本文则通过构建所提框架完整的理论与实证基础，作出了重大且新颖的贡献。本文的主要进展包括：(1) 为我们的架构与协议提供了形式化的数学方法论和算法伪代码；(2) 在商品级、云端和企业级硬件上进行了全面的多层次客观性能评估，以实证验证我们的实时处理主张；(3) 基于人数增至三倍的参与者群体（N = 30）开展了统计稳健的主观评估，并采用新指标确认用户对系统核心设计的接受度。

本文其余部分安排如下。第2 节综述关键技术和实时多媒体系统的相关文献。第3 节阐述我们提出的系统架构。第4节介绍概念验证的实现和实验设置。第5 节展示并分析技术评估和用户评估的详细结果。第6 节讨论研究发现的意义，第7节总结全文。

2 文献综述

在实时无缝翻译口语方面取得突破，长期以来一直是人机交互领域的目标，而现代语音到语音翻译（S2S）系统则代表着实现这一目标的重要一步。传统的 S2S 框架采用级联流水线，首先通过自动语音识别（ASR）模块将语音转写为文本。随后，机器翻译（MT）组件将文本译为目标语言，最后由语音合成（SS）或文本转语音（TTS）系统将其转换为语音。深度学习极大地提升了这一流水线的各个阶段。即使在具有挑战性的声学环境下，神经自动语音识别架构也已实现稳健的性能 [19, 20, 21],；与此同时，神经机器翻译（NMT）已成为标准，其质量远超早期统计方法 [22, 23]。大语言模型（LLMs）的出现进一步推动了最先进技术的发展，使翻译能够更细腻地结合语境，甚至实现端到端多语言能力[2, 24, 25]。与此同时，神经文本到语音系统如今能够生成高度自然且富有表现力的语音，而零样本语音克隆技术则能够保留发言者独特的声音身份 [26, 27]。这些同步取得的进展为高保真音频翻译奠定了坚实基础；目前，人们也正在探索将多模态生成式人工智能用于医学教育等领域中的复杂对话模拟 [28]。

对于基于视频的通信，仅翻译音频是不够的。要实现真正身临其境的体验，必须使发言者的唇部动作与翻译后的音频同步，这项任务由说话人头部生成或唇形同步模型来完成。对于视频翻译流水线中的此类模型，一项关键要求是语言无关性，确保视觉输出完全由音频音素驱动，而非由特定语言的语义决定。该领域的奠基性研究证明了利用神经网络实现语言无关唇形同步的可行性，为后续研究奠定了基础[29]。早期的成功主要由生成对抗网络（GAN）推动，这类网络在生成逼真的面部纹理和动作方面表现出色 [30, 31, 32, 33, 34, 35, 36]。近年来，该领域已转向基于扩散的模型，它们通常能生成质量更高、更稳定的结果 [37, 38, 39, 40, 41], ，以及基于神经辐射场（NeRF）的模型；这类模型擅长创建照片级真实感、视角一致的三维说话人头部 [42, 43, 44, 45, 46, 47]。除了单纯的同步之外，研究还探索了如何增强这些模型的表现力，例如让人们能够控制生成面部的情感表现 [48]。评估这些系统性能时会使用多种指标，范围从视觉保真度到对底层多模态信号的分析，例如唇形同步图像的神经相关物 [49]。这些模型能否泛化到多种不同语言，仍是重要的研究方向，这也印证了其在全球应用中的重要性 [50]。

尽管各个组件模型都已得到充分研究，但将这些部分整合为完整的端到端视频翻译系统，仍是一个探索较少的领域。早期的神经网络前尝试证明了这一概念 [51], ，而后续研究讨论了面对面翻译的潜力，却未提供可用于现实场景的完整系统实现 [29]。一些研究将多语言文本转语音与唇形同步结合起来，但并未关注部署中的实际挑战 [52]。近期一个颇具代表性的方法 TransFace 提出了一种端到端模型，以避免级联的独立模块 [53]。然而，此类端到端系统可能会牺牲独立升级组件和进行精细控制所需的模块化特性。我们此前的工作主要致力于实现低延迟性能，但未涉及多用户应用中固有的更广泛的系统级可扩展性挑战。

对这方面文献的批判性综合揭示了一条清晰的发展轨迹：尽管视频翻译的各项组件技术已经取得了显著成熟，但对整体化、可部署系统的关注仍然不足。研究已经为处理流水线的各个阶段构建了高效模型，从稳健的自动语音识别和大规模多语言机器翻译，到高保真零样本文本转语音 [21, 22, 26]。同样，唇形同步领域也从基础性的基于生成对抗网络的方法，迅速发展到更加逼真的基于扩散模型和 NeRF 的模型 [30, 40, 42]。然而，将这些组件整合为一个功能完整的整体，会带来重大的工程挑战，而这些挑战往往是孤立处理的。例如，一些端到端系统牺牲了轻松升级这些快速演进组件所需的模块化特性 [53], ，而另一些系统级工作则主要着眼于降低单个用户的延迟，却没有解决多用户可扩展性这一关键挑战 [54]。这造成了一个明确的研究空白：需要一种全面的系统级框架，既具备模块化特性，又明确旨在解决实际多用户部署中固有的两个相互关联的问题，即实时延迟和二次可扩展性。

尽管生成模型取得了这些进展，但在系统层面的工程设计方面仍存在显著的研究空白，尤其是如何将这些处理流水线部署到视频会议等实时、多用户环境中。本文通过提出并实证验证一种全新的框架，直接填补了这一空白。该框架专为实现实时视频翻译而设计，旨在使其能够在真实场景中实现可行、可扩展且高效的部署。

3 方法论

本研究的主要贡献在于提出一个全面的系统级框架，旨在弥合生成式人工智能模型的强大能力与其在实时、

多用户通信系统中的实际应用之间的差距。我们的方法论并不侧重于模型级优化，而是着眼于实现此类系统所需的架构与协议层工程，使其具备可行性、可扩展性和高效率。本节将详细阐述这一框架，首先介绍作为我们测试平台的基础生成式流水线，随后深入分析我们的核心架构与协议层创新。

3.1 生成式人工智能流水线

为了通过实证验证我们的系统级框架，我们首先针对视频翻译任务实现了一条模块化的端到端处理流水线。模块化设计是经过深思熟虑的选择；它确保每个功能组件都能在更先进的模型问世后独立升级，从而使架构具备面向未来的适应能力。该处理流水线旨在处理发言者的输入视频，并生成语义等效的输出：其中，语音经过翻译，唇部动作与新的音频同步，同时保留原始音色。如图 1所示，该处理流水线包含四个依次进行的独立阶段：

1. 自动语音识别（ASR）：处理流水线接收输入视频中的原始音频流。自动语音识别模型执行转录，将说出的音素转换为源语言的文本表示。2. 机器翻译（MT）：随后将源语言文本传入多语言机器翻译模型。该组件负责将文本翻译成所需的目标语言，同时保留原意和上下文。3. 语音合成（SS）：文本到语音（TTS）模型具备关键的零样本语音克隆能力，可将翻译后的文本合成为音频波形。语音克隆功能对于保留发言者独特的音色和韵律至关重要，这对维持身份特征并提供自然的用户体验尤为关键。4. 唇形同步（LipSync）：最后，与语言无关的 LipSync 模型接收两个输入：原始的无声视频帧，以及新合成的目标语言音频。随后，该模型通过调整发言者的口部区域，使其与翻译后的音频精确同步，从而生成新视频，完成视频翻译流程。

3.2 系统层挑战

在视频会议应用中直接采用上述处理流水线的简单实现，会因两个根本性且难以容忍的系统层挑战而失败。我们的方法论的核心部分，是首先对这些挑战进行形式化定义，以此为后续解决方案提供依据。

• 挑战 1：累积延迟。依次级联执行四个深度学习模型，会导致显著的累积处理延迟。总推理时间是各阶段延迟之和（$p_{\text{总计}} = p_{\text{自动语音识别}} + p_{\text{机器翻译}} + p_{\text{语音合成}} + p_{\text{唇形同步}}$）。在实时通信中，用户对超过几百毫秒的延迟十分敏感；这种累积延迟会导致服务质量（QoS）降至不可接受的水平，破坏对话流畅性。

• 挑战 2：二次可扩展性。在包含 N 名参与者的多用户环境中，蛮力架构要求每位 N 参与者维护 $N-1$ 条并发处理流，以便翻译来自所有其他发言者的视频。这将需要总计 $N \times (N-1)$ 个处理实例，导致计算复杂度达到 $\mathcal{O}(N^2)$。这种二次增长在计算上难以处理且经济上不可行，因此系统无法扩展到超过少数用户的规模。

3.3 提出的系统架构

我们的解决方案始于一种稳健的系统架构，其基础是解耦原则。我们将面向用户的应用逻辑与繁重的计算工作负载分离。这种设计模式提升了模块化、可维护性和可扩展性。如图2所示，该架构由四个不同的层组成：

- 服务器层：该层代表核心视频会议基础设施（例如信令服务器），负责会话管理并建立客户端之间的通信通道。
- 客户端层：这是面向用户的应用程序（例如任何基于 WebRTC 的视频会议客户端）。其职责包括管理用户交互、处理媒体流编码/解码，以及渲染最终视频。
- 处理层：这一中间编排引擎是我们解决方案的核心。它是一项无状态服务，用于管理动态的GPU资源池。它会拦截媒体流、按需分配流水线实例，并将翻译后的输出路由至正确的接收方。
- 用户层：这一抽象层代表与客户端层交互并指定所需翻译目标语言的人类参与者。

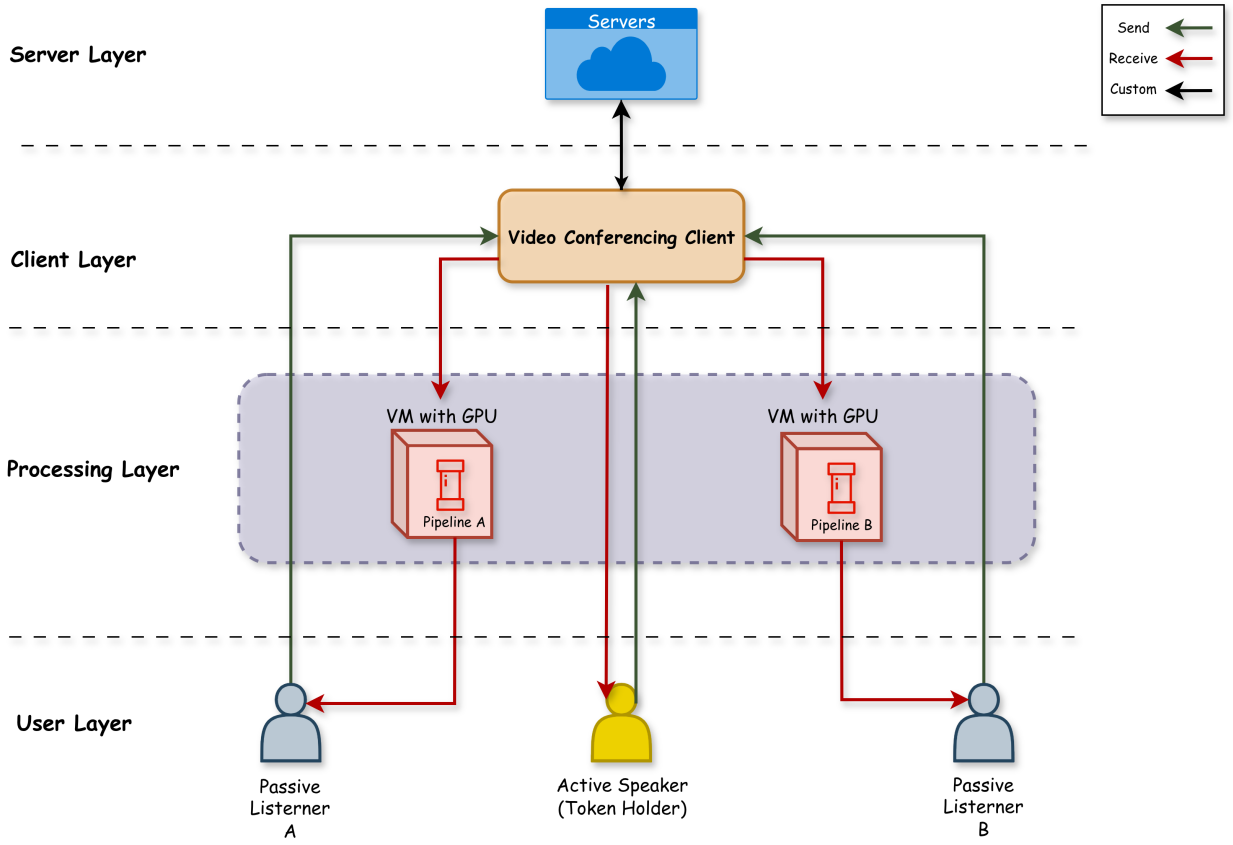


图2： 所提出的四层系统架构。该设计将面向用户的客户端与后端处理层解耦，后者充当 GPU资源池的编排引擎。

3.4 “令牌环”机制

为解决第 3.2 节中定义的二次可扩展性问题，我们提出一种新颖的轮流发言协议，称为“Token Ring”机制。该协议并非仅仅是一种启发式方法，而是一种结构化的资源管理方法，从根本上改变了系统的计算复杂度。

3.4.1 形式化成本建模与复杂度分析

为对该机制提供严谨论证，我们首先定义相关术语，然后对系统的计算成本进行形式化建模。

定义

- 设 $N \in \mathbb{Z}^+$ 为会议中的参与者总数，其中 $N \geq 2$。
- 设 C 为单个完整视频翻译处理流水线实例的计算成本（例如 GPU资源分配）常数。
- 令 P 表示系统总成本，定义为所有并发运行的流水线实例的总成本。

场景 1：朴素（蛮力）系统 在未管理的系统中，每位 N 参与者都必须具备处理来自其他所有 $N-1$ 参与者的输入流的能力。因此，系统总成本 P_{naive}等于参与者数量与每位参与者必须处理的流数量的乘积:

$$P_{\text{naive}} = C \cdot N \cdot (N-1) = C(N^2 - N) \tag{1}$$

由于 $N \to \infty$，成本主要由二次项决定，计算复杂度为$\mathcal{O}(N^2)$。这种二次增长使系统在计算上难以处理，对于任何非平凡数量的参与者而言，在经济上也不可行。

场景 2：广义令牌环系统 我们提出的机制指定一位参与者为当前发言者，其余 $N-1$ 参与者为被动听众。关键在于，系统成本不再取决于参与者总数，而是取决于所请求的目标语言的多样性。

广义模型

- 设 L 为被动听众所选目标语言的集合，其中各语言互不相同。
- 设 $k = |L|$ 为该集合的基数，表示不同目标语言的数量。
- k 的值受限于 $1 \leq k \leq N-1$。

所需处理流水线实例的总数现在等于 k，因为请求相同目标语言的所有听众都可以由一个共享流水线实例提供服务。因此，总系统成本 P_{token} 为:

$$P_{\text{token}} = C \cdot k, \quad \text{where } k \in [1, N-1] \tag{2}$$

边界条件复杂度分析

- 最坏情况复杂度：当每个被动听众都选择不同的目标语言时，就会出现最坏情况。在这种情形下，$k = N - 1$。成本变为 $P\text{worsttoken} = C \cdot (N - 1)$，从而确定了一个明确的上界，其线性计算复杂度为 $\mathcal{O}(N)$。
- 最好情况复杂度：当所有被动听众选择相同的目标语言时，即为最好情况。在此情形下，$k = 1$。开销变为 $P\text{besttoken} = C$，这是一个与用户数量无关的常数开销。因此，其复杂度下界为 $\mathcal{O}(1)$。根据渐近符号，最好情况复杂度也可表示为 $\Omega(1)$，这意味着它具有常数时间复杂度。

这项形式化分析表明，令牌环机制将一个难以处理的 $\mathcal{O}(N^2)$ 问题转变为一个易于管理的线性问题；在许多实际场景中，该问题趋近于常数时间 $\mathcal{O}(1)$ 解决方案。

为形式化描述令牌环机制的运行逻辑，我们在算法1中给出了其核心编排算法。每当当前发言者发生变化时，该算法便会在处理层中执行，管理图形处理器流水线实例的分配、复用和释放，以将计算资源占用降至最低。

算法1令牌环流编排协议

```
输入: P ={p1. .... pN} 参与者.                    s 发言者,      G GPU池
状态: Pi: L ->G（语言到流水线的映射）
输出: 更新后的路由和分配

函数: UpdateOrchestration（P. s.          G):
  路由(s.video, s.id, bypass=true) # 发言者绕过处理

  L  <- {p.lang  | p in P- {s}} # 所需语言
  k  <-  |L| # 计数: k in  [0  , N-1]

  对于每个 ell in L: # 分配新管道
    如果ell 不在dom(Pi):
      如果 G 非空:
        g  <- 分配(G)
        初始化(g,  ell,  s.语言)
        Pi(ell)    <- g
      否则:
        记录错误(ell); 继续

  对每个 p 在 P - {s}: # 路由到监听器
    如果 p.lang 属于 dom(Pi):
      g <- Pi(p.lang)
      路由(s.video,        g.中); 路由(g.out,       p.id)

  Ls  <- dom(Pi) - L # 释放过时的流水线
  对于每个 ell 在 Ls 中:
    g <- Pi(ell)
    停用(g):                释放(g.        G)
    Pi  <- Pi - {(ell,         g)}

  断言 P  = C*  |dom(Pi)| = C * k # O(N) 开销
```

算法1 实现了令牌环机制，并在每次说话者切换时执行。三个不变量确保正确性：(1) 发言者 s 接收旁路流（不进行自我翻译）；(2) L 确定

最小化处理流水线分配（无冗余），(3) 过时释放（第 20–24 行）维持 $|\mathrm{dom}(\Pi)| = k \leq N-1$，保留 $\mathcal{O}(N)$ 第 3.4 节中的复杂度3.4。关键洞见：从以参与者为中心的分配（$N \times (N-1)$ 个实例）转变为以语言为中心（仅k 个实例）。在语言同质性较高的情况下（例如双语会议），这趋近于 $\mathcal{O}(1)$，因为 $k \ll N$。

3.4.2 论证与设计原理

强制仅允许一名当前发言者发言，乍看之下可能会限制对话自然且流畅的动态过程，因为其中可能出现打断和重叠语音。然而，这一设计选择是经过慎重考虑且不可或缺的权衡，旨在确保翻译语境中的可理解性。在专业、教育或正式多语言交流领域，有效的信息交换本就依赖清晰、按顺序进行的轮流发言。当多方同时讲话并被翻译成多种语言时，音频流将混杂成一片嘈杂声，使所有听众都无法理解对话。因此，令牌环机制并非施加了不自然的限制；相反，它将现有的社会协议形式化，以实现连贯交流，并将其作为实现计算可处理性和确保高体验质量（QoE）的架构基石。

3.5 分段批处理

为了解决延迟问题，我们开发了一种协议，用于管理处理流水线固有的高处理延迟，从而提供感知上的实时用户体验。我们的分析有意将计算延迟与网络延迟分离，这是对系统性能进行建模时的标准做法。

3.5.1 系统性能表征

我们的协议基于对处理流水线性能的严格表征。

定义

- 设 $t \in \mathbb{R}^+$ 为输入视频片段的时长，单位为秒。
- 设 $p: \mathbb{R}^+ \to \mathbb{R}^+$ 为描述处理流水线处理时长为 t 的数据段所需总处理时间的函数。

通过实证分析，我们将 $p(t)$ 的行为建模为由硬件相关常数定义的分段函数，即系统阈值 (T_{thresh})。该阈值标志着两种运行状态之间的转换点：

$$p(t) = \begin{cases} p(t) > t & : t < T_{thresh} \quad \text{(System Lag Regime)} \\ p(t) \leq t & : t \geq T_{thresh} \quad \text{(Real-Time Regime)} \end{cases} \tag{3}$$

在系统滞后状态下，系统开销占主导地位，处理流水线无法跟上实时要求。在实时状态下，初始开销得以摊销，处理时间 $p(t)$呈次线性增长；我们通过实证将这一行为建模为 $p(t) \approx \log(t)$。

3.5.2 实时可行性条件

为使该模型具备可操作性，我们定义倒数吞吐率， $\tau(t)$，这是一种无量纲指标，用于将处理时间相对于实时持续时间进行归一化：

$$\tau(t) = \frac{p(t)}{t} \tag{4}$$

$\tau(t)$的值直接表明系统可行性： $\tau(t) > 1$表示系统正在落后，而$\tau(t) \leq 1$ 表示系统达到或领先于实时。只有系统持续处于 $\tau(t) < 1$ 状态，才可能实现连续、不间断的播放。这引出了我们的核心运行要求：

$$\textbf{Condition for Smooth Playback:} \exists T \text{ such that } \tau(t) < 1.0 \text{ for all } t \geq T \tag{5}$$

我们将最优片段时长 (T_{opt})定义为满足此条件的 t 的最小值，从而确保稳定性并尽量降低初始延迟。

$$T_{opt} = \min(\{t \mid \tau(t) < 1.0\}) \tag{6}$$

3.5.3 重叠缓冲

一旦 T_{opt} 针对给定的硬件配置确定后，该协议便通过将输入流划分为时长固定为5的数据块 T_{opt} ，并采用重叠缓冲策略，将公式 ( )付诸实施。这样，听众只需经历一次初始缓冲，之后便能持续接收流，从而将固有延迟较高的系统转变为能够提供感知上近乎实时体验的系统。

分段处理协议的运行逻辑详见算法2。该算法在处理层中的专用图形处理器实例上运行。它持续从活动发言者的流中读取时长最优的数据段（T_{opt}），并使用队列异步处理这些数据段，以缓冲输出，确保在初始启动延迟之后，听众能够流畅、不间断地播放。

算法 2采用重叠缓冲协议的分段处理

```
输入：S 输入流，Phi 处理流水线，T 分段时长
输出：S' 输出流
前提条件：存在 T、使得 tau(T) =D(T)/T < 1.0（实时可行）

函数ProcessStream(S       , Phi, T):
  Q  <- 空 队列 # 先进先出作业队列
  缓冲 <- 假 # 缓冲状态
  k  <- 0 # 分段计数器

  当 S.HasData():

    vk <- 读取(S,     T) # 获取长度为 T 的数据段
    k <- k + 1
    thetak  <- 异步作业(Phi, vk) # 非阻塞调用
    加入队列(Q,     thetak)

    如果未缓冲: # 初始缓冲
      Wait(Front(Q)) # 阻塞于第一个作业
      已缓冲 <- true

    当 Q 不为空 且 Done(Front(Q)): # 流式处理已完成的作业
      theta  <- 从 Q 出队
      Write(S' ' Result(theta))

  当 Q 非空时: # 清空剩余作业
    theta  <- Dequeue(Q)
    Wait(theta)
    Write(S' ' Result(theta))

  关闭(S')
  断言 tau(T)  < 1.0 表示启动后不会发生缓冲
```

算法2 通过采用带有显式初始缓冲的生产者—消费者模式，实现了分段批处理协议。算法的正确性在很大程度上取决于这一前提： T_{opt} 满足实时可行性条件（$\tau(T_{\text{opt}}) < 1.0$），如第5节中的实证结果所确立。第12—15行的一次性阻塞等待构成了我们在方法论中确定的“可预测的初始延迟”；在此之后，重叠机制确保持续输出：当数据段 v_i流式传输给听众时，数据段 v_{i+1} 正从输入端并发读取，而数据段 v_{i+2} 则进行流水线处理。之所以能够实现这种三阶段重叠，正是因为 $p(T_{\text{opt}}) < T_{\text{opt}}$，即处理速度快于实时播放速度。队列排空阶段（第23—28行）确保流终止过程平稳完成且不会丢失数据。重要的是，无论流持续多长时间，总体感知延迟始终被限制在 $t_{\text{startup}} = p(T_{\text{opt}})$，因为稳态条件 $\tau(T_{\text{opt}}) < 1.0$ 消除了后续所有缓冲事件，从而带来经主观评估（5.2）验证的流畅体验。

论证与设计 rationale

虽然任何实时系统的理想状态都是零延迟，但级联生成模型对计算资源的巨大需求，使得这一点在当前技术条件下无法实现。因此，系统设计者面临着一项关键的工程权衡：(a) 尝试实现连续、低延迟的流，但处理流水线难以跟上，极易频繁出现卡顿、缓冲和不同步；或 (b) 在每轮发言开始时引入一次可预测的延迟，以换取此后流畅、不间断的播放。从用户体验（用户体验）和心理语言学视角来看，后者显然优越得多。用户很快就能适应可预测的初始延迟，而间歇且不可预测的打断则会严重干扰用户，并破坏对话流畅性的感知。我们的协议有意将这项无法避免的计算成本前置，从而确保发言者整个发言过程中都能获得高质量、可靠且感知流畅的体验。

4 概念验证与实验设置

本节详细介绍了用于验证第 3 节所述理论方法的实证框架。其主要目标并非构建可投入生产、达到商业级标准的应用程序，而是开发一个简化但功能完善的 概念验证（*PoC*）。该概念验证具有两个关键目的：首先，作为测试平台，在受控条件下严格刻画我们所提出的架构和协议的性能；其次，用于证明我们方案的可行性。通过保持简洁并利用广泛可用的技术，该概念验证旨在提供一个基础性的通用框架，供开发者和企业在此基础上进行构建。

4.1 软件

该概念验证以完整系统的形式实现，涵盖后端生成式处理流水线和前端用户界面，用于模拟真实世界的视频会议环境。

4.1.1 模型与库

第 3.1 节所述的四阶段视频翻译流水线，采用根据性能和易用性选定的先进开源模型实现。模块化设计使每个组件都能独立进行基准测试和更新。为确保完全可复现性——这是高质量科学研究的基石——本文在表 1 中详细列出了所用的确切模型和库及其版本号。

表 1： 可复现性模型与库规范。

阶段	模型/库	版本
ASR	OpenAI Whisper	3.1.1
MT	Meta SeamlessM4T	1.0.0
语音合成（文本转语音）	Coqui XTTS	2.0.2
唇形同步	Wav2Lip（生成对抗网络）	1.0.0
框架	**库**	**版本**
深度学习	PyTorch	2.1.0
GPU 加速	CUDA 工具包	12.1
编程语言	Python	3.10

4.1.2 图形用户界面

为便于开展主观评估并模拟真实的用户体验，我们使用 React.js、CSS 和 HTML 开发了一个简单的网页用户界面（UI）。用户可通过该界面发起模拟视频通话，从下拉菜单中选择翻译的目标语言，并查看唇部动作同步的最终译制视频。客户端之间的实时信令和通信由 Socket.io 管理，而点对点视频和音频流传输则通过 Simple-Peer（WebRTC）库实现。

至关重要的是，该原型采用了平台无关的设计。这确保核心处理逻辑只需极少修改，即可集成到任何现有的视频会议平台中，从而进一步彰显了我们所提出架构的普适性。用户界面的简洁设计是有意为之，重点在于测试后端系统的可行性和感知质量。

4.2 硬件

本研究的一项关键目标是评估系统在不同计算能力条件下的表现，范围涵盖消费级硬件到企业级基础设施。此方法既展现了框架的易用性，也验证了其在生产环境模拟条件下的性能。为此，我们建立了三个不同的硬件测试平台：

• 消费级（商用笔记本电脑）：使用 NVIDIA RTX 4060 笔记本电脑 GPU，代表典型的高性能消费级或开发者设备。在此级别进行的实验均在本地机器上运行。

• 云端／数据中心级（Google Colab）：使用 NVIDIA T4 GPU（一款普遍可用且经济高效的数据中心显卡）来模拟常见的云计算环境。T4 代表云端性能的一种保守基线。

• 企业级（Google Colab Pro）：使用 NVIDIA A100 GPU（一款专为严苛人工智能工作负载设计的高性能显卡）来模拟企业级生产环境。

这种多层级硬件策略使我们能够直接检验系统阈值（T_{thresh}）对硬件的依赖性，并验证第3.5节所定义协议的实时可行性。

4.3 数据集

为进行客观评估，我们构建了一个标准化测试数据集，以确保测量结果一致且可比较。该数据集包含来自不同语言背景的发言者的 8 秒视频片段，素材取自公有领域访谈。为系统分析输入时长与处理时间（p（t））之间的关系，我们将这些 8 秒样本精细切分为五种不同长度的数据段：1、2、3、5 和 8 秒。（在所有客观推理时间分析中，我们采用同一基准测试场景——将德语视频翻译成英语——以获得一致的性能概况，并排除语言对差异这一混杂因素。）

4.4 评估

我们的评估方案分为两个互补部分：对计算性能进行客观分析，以及对用户体验进行主观评估。本节汇总了评估方法论的所有细节，旨在清晰、全面地概述我们的实验流程。

4.4.1 客观评估

本次评估旨在通过实证测量我们在方法论（第 3.5 节）中定义的性能指标。我们测量了两个主要指标：(1) 推理时间 ($p(t)$)，即整个处理流水线处理时长为 t 的视频数据段所需的实际耗时（秒）；(2)倒数吞吐量($\tau(t)$)，按公式（4）（$\tau(t) = p(t)/t$）计算。我们记录了整个处理流水线以及各个独立模块的推理时间，以识别瓶颈。为确保统计稳健性并考虑细微变化，每项实验均在每个硬件测试平台上、针对每种分段时长重复进行三次。最终报告的数值为这三次运行结果的均值和标准差。

4.4.2 主观评估

本次评估旨在从终端用户视角量化系统输出的感知质量。我们招募了30名参与者，以确保研究结果具有统计稳健性。研究共有30名志愿参与者（18名男性、12名女性），年龄在22至48岁之间（平均年龄：31.5岁）。参与者来自不同的地理和语言背景，包括北美洲、欧洲和中东，涵盖英语、德语和土耳其语的母语者。所有参与者均表示，他们非常熟悉标准的视频会议工具，但在生成式人工智能方面的专业水平各不相同，因此他们代表的是普通用户群体，而非人工智能专家小组。提供这些人口统计信息是为了提升研究结果的透明度和泛化能力。

参与者通过概念验证的用户界面进行交互，观看了在第5节所述特定硬件条件下生成的翻译视频片段。随后，他们使用 5 分制李克特量表（1 = 差， 2 = 较差， 3 =一般， 4 = 良好，5 = 优秀）对体验进行评分。汇总结果采用平均意见分数（MOS）评估方案进行了分析。这五项评估标准旨在全面评估用户体验，详见表2。

表 2：主观用户体验评估标准。

指标	缩写	描述
唇形同步准确度	LSA	发言者的唇部动作与译文的匹配程度如何? 生成的音频?
动作自然度	MN	生成的唇部和嘴部动作看起来有多自然、逼真, 不考虑同步情况?
视觉质量	VIQ	视频的整体视觉质量如何，包括任何 口部区域是否存在伪影或模糊?
语音质量	VOQ	合成语音听起来自然吗？它在多大程度上保留了 原说话者的音色?
启动延迟可接受性	SDA	考虑到流畅、不间断的视频所带来的好处，您认为 播放开始前的初始延迟是否可以接受?

这一全面的实验设置涵盖功能完备的概念验证、多层级硬件以及严谨的双管齐下的评估方案，为下一节所呈现的结果与分析奠定了实证基础。

5 结果与分析

本节介绍第4节中详细说明的综合评估方案所得的实证结果。结果分为两大部分。首先，我们呈现客观性能分析，以实证验证系统架构和处理协议的理论模型。其次，我们呈现主观用户体验分析，对概念验证的感知质量进行量化，并从终端用户视角确认其实际可行性。

5.1 客观性能分析

本次客观评估旨在严格分析视频翻译流水线在三个不同硬件层级上的计算性能。主要目标是：(1) 验证该处理流水线的分段性能模型及其推理时间的次线性扩展行为， $p(t)$；以及 (2) 通过实证确定满足“实时可行性条件”($\tau(t) < 1.0$)的条件。

表 3 展示了五种视频片段时长下，各硬件测试平台的总处理流水线推理时间（$p(t)$）以及计算得出的倒数吞吐量（$\tau(t)$）的完整汇总结果。

表3： 不同硬件层级的客观性能指标。推理时间 ($p(t)$)以秒为单位。倒数吞吐量 ($\tau(t)$)为无量纲比率。数值以三次运行的均值 ± 标准差表示。

硬件	视频时长 (t)	推理时间 ($p(t)$)	倒数吞吐量 ($\tau(t)$)
NVIDIA T4 （云端基线）	1 s	8.99 ± 0.11	8.99
	2 s	10.27 ± 0.15	5.14
	3 s	10.92 ± 0.09	3.64
	5 s	12.01 ± 0.13	2.40
	8 s	12.70 ± 0.10	1.59
NVIDIA RTX 4060 （商品级）	1 s	4.52 ± 0.08	4.52
	2 s	4.81 ± 0.09	2.41
	3 s	5.10 ± 0.08	1.70
	5 s	5.68 ± 0.10	1.14
	8 s	6.55 ± 0.07	**0.82**
NVIDIA A100 （企业级）	1 s	1.87 ± 0.04	1.87
	2 s	2.08 ± 0.05	1.04
	3 s	2.29 ± 0.03	**0.76**
	5 s	2.71 ± 0.05	**0.54**
	8 s	3.34 ± 0.04	**0.42**

原始数据清楚地显示了两种趋势。首先，对于给定的视频时长，推理时间与图形处理器的计算能力成反比：A100 的速度明显快于 RTX 4060，而 RTX 4060 又快于 T4。其次，对于所有硬件而言，推理时间的增长幅度小于视频时长的增长幅度。

为了结合我们的方法论更直观地呈现这些结果，图3 绘制了倒数吞吐量（$\tau(t)$）随视频时长变化的曲线。该图是验证我们分段处理协议的主要工具。

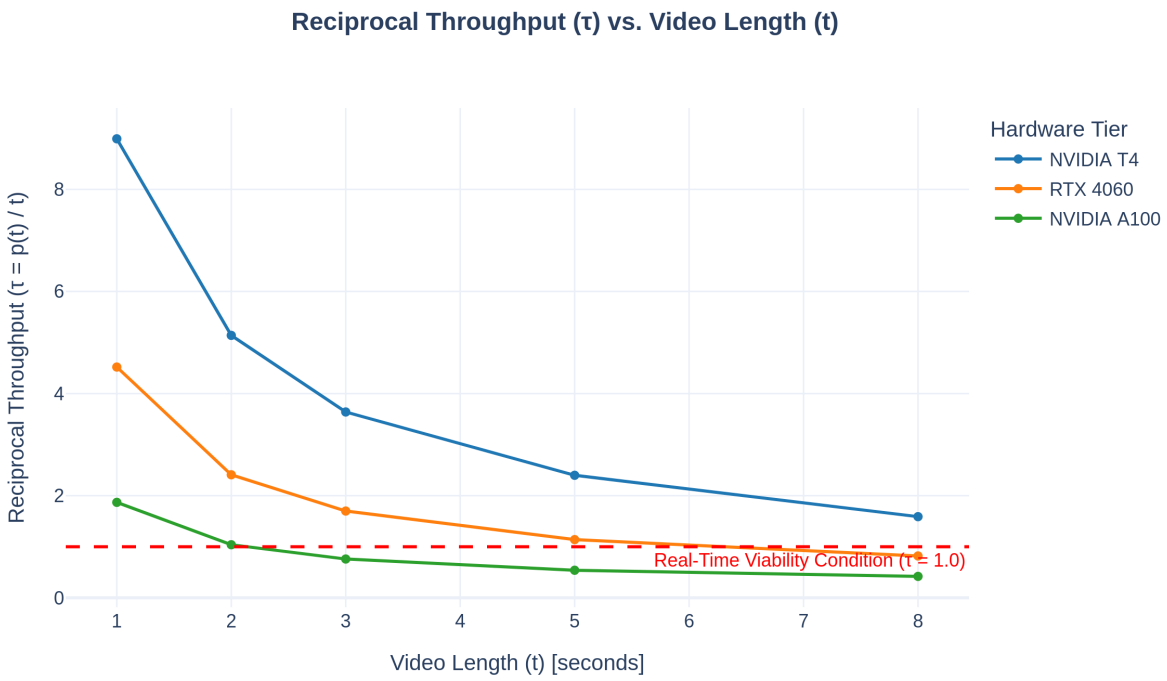


图 3： 三个硬件层级下，倒数吞吐量（$\tau(t)$）与视频时长（t）的关系。位于 $\tau = 1.0$ 处的水平虚线表示实时可行性条件，即处理时间等于播放时间。系统仅在 $\tau < 1.0$时保持稳定。

如我们的模型（方程（5））所预测，图3 提供了对我们协议的直接实证验证。企业级 A100 GPU 在 $\tau = 1.0$ 阈值处于 $t = 2$ 和 $t = 3$ 秒之间，因此该硬件的最优片段时长（T_{opt}）约为 3 秒。消费级 RTX 4060 在 $t = 5$和 $t = 8$秒之间达到这一条件，因此 T_{opt} 约为 8 秒。相比之下，基准 T4 GPU 在测试的 8 秒范围内未能达到 $\tau < 1.0$，始终处于“系统延迟区间”。这直接证实，该协议是否可行取决于硬件，并且在现代消费级和企业级 GPU 上轻松即可实现流畅的实时播放。

此外，表 3 中的数据直接证实了次线性扩展行为，这是我们的协议赖以成立的基础。对于较短的 1 秒数据段，处理时间异常长，凸显了模型加载和初始化产生的固定开销所造成的显著影响。然而，随着视频时长增加，这些初始成本显然得到了摊销。例如，在 A100 上，视频时长从 1 秒加倍至 2 秒，处理总时间仅增加 11%（从 1.87 秒增至 2.08 秒）；而最后 3 秒的视频（从 5 秒到 8 秒）仅使总时间增加 23%。性能曲线趋于平缓，正是我们的分段协议得以奏效的关键：通过处理更长的数据块，系统能够有效地“赶上进度”，并提前于实时进度运行。

5.2 主观用户体验分析

在对客观性能进行表征后，我们邀请 30 名参与者开展主观评估，以评估系统输出的感知质量。为检验分段处理协议的核心前提，我们向参与者展示了在特定延迟条件下生成的视频片段，这些条件分别对应不同的硬件层级。

对于从未达到实时条件的基准硬件 NVIDIA T4，我们向用户展示了时长为 8 秒、对应处理延迟为 12.7 秒的视频片段，以了解用户对系统滞后的容忍度。对于 RTX 4060 和 A100，我们则向参与者展示了时长满足实时可行性条件（$\tau < 1.0$）的视频片段：RTX 4060 对应 8 秒片段（延迟 6.6 秒），A100 对应 3 秒片段（延迟 2.3 秒）。这样便能在保证流畅播放的条件下，直接评估“启动延迟可接受性”（SDA）。

表4 总结了五项标准的平均意见分数（MOS）。虽然 LSA、动作自然度、视觉质量和语音质量的模型输出在各平台间技术上完全相同，但我们仍报告了各条件下收集的分数，以反映不同用户体验可能带来的感知差异或偏差。

表 4： 主观评估的平均意见分数（MOS）（N = 30）。分数采用 5 点李克特量表。结果以均值 ± 标准差表示，括号内为 95% 置信区间。

硬件	LSA	MN	VIQ	VOQ	SDA
NVIDIA T4 （*8* 秒视频，延迟 *12.7* 秒）	3.91 ± 0.88 [3.58, 4.24]	3.76 ± 0.93 [3.41，4.11]	3.25 ± 1.12 [2.83, 3.67]	4.55 ± 0.62 [4.32, 4.78]	4.15 ± 0.79 [3.85, 4.45]
RTX 4060 （*8* 秒视频，延迟 *6.6* 秒）	3.97 ± 0.85 [3.64, 4.30]	3.81 ± 0.90 [3.46，4.16]	3.29 ± 1.12 [2.87，3.71]	4.58 ± 0.60 [4.35, 4.81]	4.60 ± 0.61 [4.38, 4.82]
NVIDIA A100 （*3* 秒视频，延迟 *2.3* 秒）	4.02 ± 0.83 [3.70，4.34]	3.86 ± 0.89 [3.52, 4.20]	3.33 ± 1.10 [2.91，3.75]	4.62 ± 0.58 [4.40，4.84]	4.85 ± 0.42 [4.70，5.00]

为清晰呈现用户对最关键感知指标的评分，图 4展示了启动延迟可接受性（SDA）均值的柱状图。

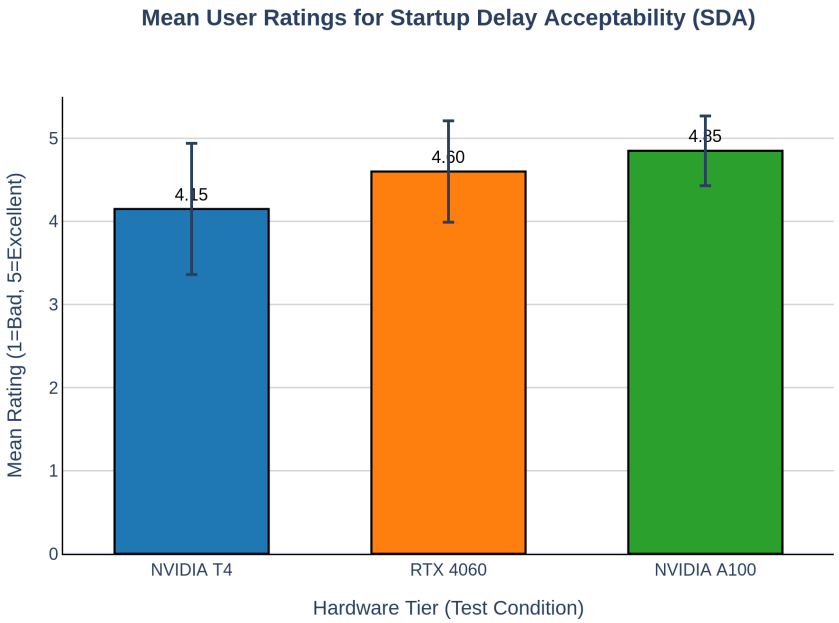


图 4： 启动延迟可接受性（SDA）的用户平均评分（N = 30）柱状图。每个柱子的高度表示相应硬件条件下的均值，误差线表示用户评分的标准差。

分析揭示了两个关键发现。首先，无论底层硬件如何，生成模型本身的感知质量都保持一致。语音质量（VOQ）的评分很高，均值始终高于 4.5；唇形同步准确度（LSA）和动作自然度（MN）的结果也很理想，这证实了所选模型的有效性。平均意见分数显示，视觉质量（VIQ）在所有测试中始终是评分最低的指标，这表明 LipSync 模型产生的视觉伪影是感知质量的主要瓶颈。

其次，也是最关键的一点，图4 中的柱状图通过展示我们的分段处理协议对硬件的依赖性，为其提供了直接的用户验证。硬件性能与用户对启动延迟的满意度之间存在明显且强烈的正相关关系。基线 T4 的延迟较长，为 12.7 秒，其平均启动延迟评分为 4.15，虽属良好但相对较低。RTX 4060 将延迟减半至 6.6 秒，其评分显著提升至 4.60。最后，A100 将启动延迟降至几乎难以察觉的 2.3 秒，获得了近乎满分的 4.85 平均分。

这是一个关键发现。它为我们协议中的核心权衡提供了直接的实证验证：用户认为，预先存在且可预测的启动延迟，是换取随后流畅、不中断的播放体验所需付出的可接受代价。此外，这也证明，随着硬件能力提升，这种感知上的权衡会逐渐减弱，直至不再构成问题，从而确认我们的协议不仅在技术上切实可行，还能在现代硬件上带来高度令人满意且以用户为中心的体验质量。

6 讨论

上一节呈现的结果为我们提出的系统级框架提供了有力的实证验证。本节将超越原始数据，对这些发现进行解读，讨论其对实时生成式人工智能系统领域的更广泛影响，并客观评估当前工作的局限性，以指导未来研究。

6.1 系统架构与协议的实证验证

我们的主要贡献，是针对多用户生成式人工智能应用中的延迟和可扩展性挑战，提出了一套架构与协议层面的解决方案。客观性能分析为这一框架提供了三项关键验证依据。

首先，推理时间呈次线性增长（$p(t)$），如表 3中的数据所示。这并非单纯的性能伪影，而是我们的分段处理协议旨在利用的一项基本特性。短时长片段一开始便产生较高开销，这证实了朴素的连续流式传输方式效率极低且容易失败。通过证明这些开销会随着数据段变长而得到摊薄，我们从实证层面验证了协议的核心假设：采用经过优化的定长分块进行处理，显然更为高效。

其次，多层级硬件评估（图3）证实，我们的“实时可行性条件”（$\tau(t) < 1.0$）是一个切实可行且能够实现的目标。结果清楚地揭示了不同硬件层级之间的差异，并为行动提供了明确依据：系统在 NVIDIA T4 这类基础云端 GPU 上具备可行性，而在现代商用硬件（RTX 4060）和企业级基础设施（NVIDIA A100）上则已可用于生产。这一发现意义重大，因为它为部署提供了清晰的路线图：开发者可以先在易于获取的硬件上进行原型开发，并确信系统扩展到生产环境后能够实现真正的实时性能，从而降低采用此类技术的风险。

第三，我们的主观评估中对启动延迟可接受性（SDA）的高度正面评价，为我们的工程权衡提供了至关重要、以用户为中心的验证。数据证实，用户认为，系统只要具有可预测的一次性初始延迟，随后能够实现流畅且不中断的播放，就能提供高质量的体验。这一心理视觉研究发现挑战了“所有延迟都必须

不惜一切代价加以最小化”的传统观念。相反，它表明，对于计算密集型任务而言，流的可预测性和可靠性，比绝对的初始延迟更能影响用户的体验质量。这一洞见的适用范围远不止视频翻译，还涵盖其他实时生成任务。

此外，我们的多层次主观评估可以视为对协议主要优势的一项隐式消融研究。NVIDIA T4 测试用例未能满足实时可行性条件（$\tau > 1.0$），代表一个“对照”系统：用户会经历完整的原始处理延迟，且无法保证流畅播放。RTX 4060 和 A100 测试用例满足该条件（$\tau < 1.0$），代表按设计运行的“处理”系统。启动延迟可接受性（SDA）得分从对照条件到处理条件显著且大幅提升，这为该协议在改善用户体验质量方面的有效性提供了直接的定量证据。

6.2 用户感知与处理流水线质量分析

如今，主观评估纳入了更多参与者，因此能够更有把握地分析终端用户体验。最突出的是语音质量（VOQ）评分极高，均值始终高于 4.5，这表明现代零样本语音克隆技术已足够成熟，可用于实际应用。这一成果至关重要，因为保留说话者的声音身份对于维持对话临场感和真实性极为重要。

相反，得分最低的指标——视觉质量（VIQ）——表明这是未来亟待改进的一个方面。VIQ 得分分布较广，这说明虽然部分用户并不在意视觉伪影，但相当一部分用户认为这些伪影会分散注意力。这表明，所选的唇形同步模型（Wav2Lip-GAN）虽然能够正常运行，却是感知质量的主要瓶颈。这一发现凸显了我们所提出处理流水线的模块化特性：系统架构本身稳健；随着更先进的扩散模型或基于 NeRF 的唇形同步模型问世，可将其替换进来，直接解决这一弱点，从而提升整体用户体验。

6.3 局限性与未来工作

尽管本研究成功展示了一个可行的系统级框架，但为了明确未来研究方向，我们仍有必要承认其局限性。

首先，我们的性能分析有意将计算延迟与网络延迟分开，以便清晰地评估系统的处理能力。在实际部署中，需要将这一架构与稳健的网络协议相结合，以应对抖动、丢包和可变带宽，确保流畅的端到端体验。我们的工作为构建此类系统奠定了计算基础。

其次，本文从架构层面介绍了“Token Ring”机制。本研究并未规定令牌传递逻辑的具体实现方式（例如手动“举手”功能、自动语音活动检测或主持人控制）。开发并评估这些不同的令牌管理策略，是系统设计与人机交互（HCI）交叉领域中值得深入探索的未来研究方向。我们工作的关键贡献在于，从形式上证明了任何 此类轮流发言机制一经实施，便能从根本上解决二次可扩展性问题。

最后，尽管我们的评估全面验证了框架的内部性能主张和用户可接受性，但本研究并未与其他端到端系统架构进行直接的定量比较。我们的评估旨在通过倒数吞吐量和启动延迟可接受性等系统级指标，证明特定架构贡献的有效性。随着该领域日趋成熟，制定标准化的系统级基准测试协议将成为至关重要的下一步，使研究人员能够直接比较不同架构方法的性能、可扩展性和效率。

7 结论

将生成式人工智能集成到实时通信系统中，有望从根本上消除语言障碍，但延迟和可扩展性方面严峻的系统层挑战阻碍了这一进程。本文通过提出一种用于部署实时多语言视频翻译的新型综合框架，应对了这些挑战。我们的主要贡献有两方面：一种可扩展的令牌环 系统架构，可将多用户会议的计算复杂度从难以处理的 $\mathcal{O}(N^2)$ 降至可控的 $\mathcal{O}(N)$；以及一种分段批处理 协议，结合逆吞吐量阈值化，旨在管理生成式处理流水线较高的固有延迟。

通过开发概念验证系统，并开展严谨的多层次实证评估，我们验证了该框架的有效性。客观分析的实验结果表明，该系统在现代商品级和企业级硬件上均满足关键的实时可行性条件（$\tau < 1.0$），证实了其大规模部署的可行性。此外，统计结果稳健的主观评估表明，用户高度认可该系统以可预测的初始延迟换取流畅、不间断播放的权衡，因此整体用户体验良好。

尽管网络集成和视觉质量方面的局限仍有待未来从模型层面和实现层面进一步研究，但本研究提供了一个经过验证的基础系统架构。通过解决可扩展性和延迟管理方面的关键瓶颈，该框架为开发者和研究人员提供了切实可行的路线图，助力构建新一代包容、高效且真正面向全球的通信平台。

未来工作将聚焦于两个关键方向。首先，将更先进的非自回归生成模型集成到处理流水线的瓶颈环节——唇形同步阶段。新兴的基于扩散的模型和 NeRF 驱动的说话人头部模型虽然计算开销较大，却能提供更高的视觉保真度；将其集成到我们的框架中，有望针对用户研究中发现的视觉质量（VIQ）局限进行改进。我们的架构为量化这些新一代模型在系统层面的延迟权衡提供了理想的测试平台。其次，我们将探索动态自适应分段协议。当前协议采用固定的最优片段时长（T_{opt}），而更复杂的方法则可根据网络条件和源内容的语言复杂度，实时动态调整块大小，从而进一步优化延迟与计算吞吐量之间的平衡，实现真正无缝的用户体验。

伦理声明

已获得所有参与研究对象的知情同意。生成式人工智能工具仅用于提升语言表达的清晰度。作者对内容承担全部责任。

数据可用性

作者同意应读者要求分享数据集。

利益冲突

作者声明不存在利益冲突。

参考文献

参考文献

[1] Doropoulos, S.; Karapalidou, E.; Charitidis, P.; Karakeva, S.; Vologiannidis, S. 超越人工媒体编码：评估大语言模型和智能体在新闻内容分析中的应用。《应用科学》2025年, 15, 8059.[2] Oskooei, A.R.; Babacan, M.S.; Yaˇgcı, E.; Alptekin, Ç.; Buˇgday, A. 超越合成基准：评估近期大语言模型的代码生成能力。载于第14届国际计算机科学与工程研讨会（WCSE 2024）论文集，泰国普吉岛，2024年6月19日至21日；第290–296页。[3] Acosta-Bermejo, R.; Terrazas-Chavez, J.A.; Aguirre-Anaya, E. 利用未审查大语言模型生成恶意软件源代码，并规避对审查模型的检测。《应用科学》2025年, 15, 9252.[4] Liu, M.; Wang, J.; Lin, T.; Ma, Q.; Fang, Z.; Wu, Y. 使用大语言模型生成安全关键软件代码的实证研究。《应用科学》2024年, 14, 1046.[5] Kim, S.; Lee, K. DanceCaps：使用大语言模型为舞蹈视频生成伪字幕。《应用科学》2024年, 14, 10116.[6] Yan, J.; Chen, Z.; Cai, J.; Xian, W.; Wei, X.; Qin, Y.; Li, Y. 视频驱动人工智能在抗菌肽生成预测建模中的应用：进展与挑战文献综述。《应用科学》2025年, 15, 7363.[7] Oskooei, A.R.; Caglar, E.; Yakut, S.; Tuten, Y.T.; Aktas, M.S. 基于情绪加权的面部压力与疲劳识别：一种深度学习方法。载于国际计算科学及其应用会议论文集，土耳其伊斯坦布尔，2025年6月30日至7月3日；施普林格：柏林/海德堡，德国，2025年；第193–211页。[8] Partarakis, N.; Evdaimon, T.; Katsantonis, M.; Zabulis, X. 通过基于虚拟现实的情境数字孪生培训急救人员。《计算机》2025年, 14, 274.[9] Todino, M.D.; Pitri, E.; Fella, A.; Michaelidou, A.; Campitiello, L.; Placanica, F.; Di Tore, S.; Sibilio, M. 融合传统与创新：人工智能和虚拟现实赋能博物馆变革性教育实践。《计算机》2025年, 14, 257.[10] Johnson, D.; Mamani, B.; Salas, C. Collabvr：通过虚拟现实测试增进大学生之间的社交互动。《计算机》2024年, 13, 40.[11] Jover, V.; Sempere, S. 在元宇宙中创建数字孪生以庆祝纪念活动。《计算机》2025年, 14, 273.[12] Mohamed, A.; Faisal, R.; Al-Gindy, A.; Shaalan, K. 人工智能与沉浸式技术：面向特殊需求学习者的增强现实/虚拟现实虚拟助手。《计算机》2025年, 14, 306.[13] Jun, S. 在数字疗法技术领域运用贝叶斯因子分析和社交网络可视化进行专利关键词分析。《计算机》2025年, 14, 78.

[14] Huitema, I.R.; Oskooei, A.R.; Akta̧s, M.S.; Riveni, M. 去中心化社交网络中的回音室研究：Mastodon案例分析。载于《复杂网络及其应用国际会议论文集》，土耳其伊斯坦布尔，2024年12月10日至12日；施普林格：德国柏林/海德堡，2024年；第316–328页。[15] Wang, H.; Wang, C.; Qi, H. 在线健康社区中医生的网络结构特征及其对推荐率的影响机制研究：基于“好大夫在线”平台的多维视角。应用科学。**2025**，*15*，4583。[16] Soares, N.F.; Marin, A.; Rodrigues, D.L.; Duarte, F.J. 工业4.0背景下的数字孪生与元宇宙。载于《国际计算科学及其应用会议论文集》，土耳其伊斯坦布尔，2025年6月30日至7月3日；施普林格：德国柏林/海德堡，2025年；第254–271页。[17] Leung, C.; Xu, Z. 大语言模型与情绪演化预测。载于《国际计算科学及其应用会议论文集》，越南河内，2024年7月1日至4日；施普林格：德国柏林/海德堡，2024年；第3–19页。[18] Rafiei Oskooei, A.; Caglar, E.; ̧Sahin, I.; Kayabay, A.; Aktas, M.S. 低语、翻译、发声、同步：利用生成式人工智能实现多语言视频会议的视频翻译。载于《国际计算科学及其应用会议论文集》，土耳其伊斯坦布尔，2025年6月30日至7月3日；施普林格：德国柏林/海德堡，2025年；第217–234页。[19] Malik, M.; Malik, M.K.; Mehmood, K.; Makhdoom, I. 自动语音识别：综述。多媒体工具与应用。**2021年**，*80*，9411–9457。[20] Prabhavalkar, R.; Hori, T.; Sainath, T.N.; Schlüter, R.; Watanabe, S. 端到端语音识别：综述。*IEEE/ACM*音频、语音与语言处理汇刊。**2023年**，*32*，325–351。[21] Kheddar, H.; Hemis, M.; Himeur, Y. 采用先进深度学习方法的自动语音识别：综述。信息融合**2024年**，*109*，102422。[22] Team, N.; 等. 将神经机器翻译扩展至200种语言。《自然》**2024年**，*630*，841。[23] Wang, H.; Wu, H.; He, Z.; Huang, L.; Church, K.W. 机器翻译的进展。《工程》**2022年**，*18*，143–153。[24] Zhu, W.; Liu, H.; Dong, Q.; Xu, J.; Huang, S.; Kong, L.; Chen, J.; Li, L. 大语言模型多语言机器翻译：实证结果与分析。*arXiv***2023年**，arXiv:2304.04675。[25] Oskooei, A.R.; Yukcu, S.; Bozoglan, M.C.; Aktas, M.S. 通过层次化摘要实现大语言模型的代码仓库级代码理解：提升代码搜索与缺陷定位。载于国际计算科学及其应用会议；施普林格：德国柏林/海德堡，2025年；第88–105页。[26] Casanova, E.; Davis, K.; Gölge, E.; Göknar, G.; Gulea, I.; Hart, L.; Aljafari, A.; Meyer, J.; Morais, R.; Olayemi, S.; 等. XTTS：大规模多语言零样本文本转语音模型。*arXiv* **2024年**，arXiv:2406.04904。[27] Li, R.; Pu, D.; Huang, M.; Huang, B. Unet-tts：改进单样本语音克隆中的未见说话人与风格迁移。载于《ICASSP 2022—2022年IEEE声学、语音与信号处理国际会议论文集》，新加坡，2022年5月23日至27日；第8327–8331页。

[28] Chu, S.N.; Goodell, A.J. 合成患者：利用多模态生成式人工智能模拟医学教育中的困难对话。*arXiv* **2024年**，arXiv:2405.19941。[29] KR, P.; Mukhopadhyay, R.; Philip, J.; Jha, A.; Namboodiri, V.; Jawahar, C. 迈向自动化面对面翻译。载于第27届ACM国际多媒体会议论文集，法国尼斯，2019年10月21日至25日；第1428–1436页。[30] Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V.P.; Jawahar, C. 野外语音到唇形生成只需一个唇形同步专家。载于第28届ACM国际多媒体会议论文集，美国华盛顿州西雅图，2020年10月12日至16日；第484–492页。[31] Bao, H.; Zhang, X.; Wang, Q.; Liang, K.; Wang, Z.; Ji, S.; Chen, W. MILG：利用音频调制图像修复生成逼真的唇形同步视频。视觉信息**2024年**，*8*，71–81。[32] Kim, M.; Hong, J.; Ro, Y. M. 利用视觉上下文注意力生成对抗网络实现唇形到语音合成。神经信息处理系统进展。**2021年**，*34*，2758–2770。[33] Vougioukas, K.; Petridis, S.; Pantic, M. 利用时序生成对抗网络实现端到端语音驱动逼真面部动画。载于CVPR研讨会论文集，美国加利福尼亚州长滩，2019年6月16日至20日；第887卷，第37–40页。[34] Das, D.; Biswas, S.; Sinha, S.; Bhowmick, B. 利用级联生成对抗网络学习运动与纹理，实现语音驱动面部动画。载于《计算机视觉–ECCV 2020：第16届欧洲计算机视觉会议论文集》第XXX部分16；施普林格：德国柏林/海德堡，2020年；第408–424页。会议于2020年8月23日至28日在英国格拉斯哥举行。[35] Yin, F.; Zhang, Y.; Cun, X.; Cao, M.; Fan, Y.; Wang, X.; Bai, Q.; Wu, B.; Wang, J.; Yang, Y. Styleheat：通过预训练StyleGAN实现单样本高分辨率可编辑说话人脸生成。载于《欧洲计算机视觉会议论文集》，以色列特拉维夫，2022年10月23日至27日；施普林格：德国柏林/海德堡，2022年；第85–101页。[36] Hong, F.T.; Zhang, L.; Shen, L.; Xu, D. 用于说话头视频生成的深度感知生成对抗网络。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，美国路易斯安那州新奥尔良，2022年6月18日至24日；第3397–3406页。[37] Lin, G.; Jiang, J.; Yang, J.; Zheng, Z.; Liang, C. OmniHuman-1：重新思考单阶段条件人类动画模型的规模化。*arXiv***2025**，arXiv:2502.01061。[38] Li, C.; Zhang, C.; Xu, W.; Xie, J.; Feng, W.; Peng, B.; Xing, W. LatentSync：用于唇形同步的音频条件潜扩散模型。*arXiv***2024**，arXiv:2412.09262。[39] Xu, S.; Chen, G.; Guo, Y.X.; Yang, J.; Li, C.; Zang, Z.; Zhang, Y.; Tong, X.; Guo, B. Vasa-1：实时生成栩栩如生的音频驱动说话人脸。*arXiv***2024**，arXiv:2404.10667。[40] Shen, S.; Zhao, W.; Meng, Z.; Li, W.; Zhu, Z.; Zhou, J.; Lu, J. DiffTalk：为通用音频驱动肖像动画打造扩散模型。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，加拿大不列颠哥伦比亚省温哥华，2023年6月18日至22日；第1982–1991页。[41] Stypułkowski, M.; Vougioukas, K.; He, S.; Zi˛eba, M.; Petridis, S.; Pantic, M. 扩散头部：扩散模型在说话人脸生成方面优于生成对抗网络。载于《IEEE/CVF冬季计算机视觉应用会议论文集》，美国夏威夷州怀科洛阿，2024年1月3日至8日；第5091–5100页。

[42] Ye, Z.; He, J.; Jiang, Z.; Huang, R.; Huang, J.; Liu, J.; Ren, Y.; Yin, X.; Ma, Z.; Zhao, Z. Gene- face++：通用且稳定的实时音频驱动三维说话人脸生成。arXiv 2023, arXiv:2305.00787.[43] Sun, X.; Zhang, L.; Zhu, H.; Zhang, P.; Zhang, B.; Ji, X.; Zhou, K.; Gao, D.; Bo, L.; Cao, X. Vividtalk：基于三维混合先验的单次音频驱动说话头部生成。arXiv 2023, arXiv:2312.01841.[44] Ye, Z.; Zhong, T.; Ren, Y.; Jiang, Z.; Huang, J.; Huang, R.; Liu, J.; He, J.; Zhang, C.; Wang, Z.; et al. MimicTalk：几分钟内模仿个性化且富有表现力的三维说话人脸。arXiv 2024, arXiv:2410.06734.[45] Ye, Z.; Zhong, T.; Ren, Y.; Yang, J.; Li, W.; Huang, J.; Jiang, Z.; He, J.; Huang, R.; Liu, J.; et al. Real3d-portrait：单次逼真三维说话肖像合成。arXiv 2024, arXiv:2401.08503.[46] Shin, A.H.; Lee, J.H.; Hwang, J.; Kim, Y.; Park, G.M. Wav2NeRF：通过基于小波的神经辐射场实现音频驱动的逼真说话头部生成。《图像与视觉计算》2024, 148, 105104.[47] Song, W.; Liu, Q.; Liu, Y.; Zhang, P.; Cao, J. 用于音频驱动说话头部生成的多层级特征动态融合神经辐射场。《应用科学》2025, 15, 479.[48] Zhao, Z.; Zhang, Y.; Wu, T.; Guo, H.; Li, Y. 根据任意情感肖像生成情感可控说话人脸。《应用科学》2022, 12, 12852.[49] Naebi, A.; Feng, Z. 唇形同步图像模型的性能、信号的新组合、补充键合图分类器，以及作为脑电图和脑机接口提取与根分类器的深度公式检测。《应用科学》2023, 13, 11787.[50] Rafiei Oskooei, A.; Yahsi, E.; Sungur, M.; S. Aktas, M. 一个模型能适用于所有情况吗？探索 Wav2Lip 在文化差异显著的语言中的唇形同步泛化能力。载于《国际计算科学及其应用会议论文集》，越南河内，2024年7月1日至4日；施普林格：德国柏林/海德堡，2024年；第149–164页。[51] Ritter, M.; Meier, U.; Yang, J.; Waibel, A. 面部翻译：一种多模态翻译代理。载于《AVSP’99-国际视听语音处理会议论文集》，美国加利福尼亚州圣克鲁斯，1999年8月7–10日。[52] Song, H.K.; Woo, S.H.; Lee, J.; Yang, S.; Cho, H.; Lee, Y.; Choi, D.; Kim, K.w. 使用多语言文本转语音生成说话人脸。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，美国路易斯安那州新奥尔良，2022年6月18–24日；第21425–21430页。[53] Cheng, X.; Huang, R.; Li, L.; Jin, T.; Wang, Z.; Yin, A.; Li, M.; Duan, X.; Zhao, Z.; et al. Trans- Face：用于说话头部翻译的基于单元的视听语音合成器。arXiv 2023, arXiv:2312.15197.[54] Rafiei Oskooei, A.; Akta.s, M.S.; Kele. s, M. Seeing the Sound：用于实时面对面翻译的多语言唇形同步。《计算机》2024, 14, 7。