面向视频通信系统实时多语言唇形同步的异步流水线并行

Eren Caglar耶尔德兹技术大学数据科学与大数据系土耳其伊斯坦布尔
eren.caglar@std.yildiz.edu.tr

Amirkia Rafiei Oskooei耶尔德兹技术大学计算机工程系土耳其伊斯坦布尔
amirkia.oskooei@yildiz.edu.tr

Mehmet Kutanoglu*Aktif Investment Bank Inc.*研发中心土耳其伊斯坦布尔
meh
met.kutanoglu@aktifbank.com.tr

Mustafa Keles*Aktif Investment Bank Inc.*研发中心土耳其伊斯坦布尔
mustafa.keles
@aktifbank.com.tr

Mehmet S. Aktas耶尔德兹技术大学计算机工程系土耳其伊斯坦布尔 aktas@yildiz.edu.tr

arXiv:2512.18318v1 [cs.MM] 20 Dec 2025

摘要——**本文介绍了一种并行异步 Transformer 框架，旨在为实时视频会议系统实现高效、准确的多语言唇形同步。所提出的方法采用流水线并行设计，将翻译、语音处理和唇形同步模块集成在一起，并通过基于消息队列的解耦实现模块并发执行；与顺序方法相比，端到端延迟最多可降低 3.1× 。为提升计算效率和吞吐量，我们通过低级图编译、混合精度量化和硬件加速内核融合优化了各模块的推理流程。这些优化在保持模型准确性和视觉质量的同时，显著提升了效率。此外，上下文自适应静音检测组件会在语义连贯的边界处对输入语音流进行分段，从而提升不同语言之间的翻译一致性和时间对齐效果。实验结果表明，所提出的并行架构在处理速度、同步稳定性和资源利用率方面均优于传统顺序流水线。这种模块化、面向消息的设计适用于资源受限的物联网通信场景，包括远程医疗、多语言自助终端和远程协助系统。总体而言，本研究推动了面向下一代人工智能物联网系统的低延迟、资源高效型多模态通信框架的发展。**

索引词——**实时系统、流水线并行、多模态翻译、面向消息的中间件、推理优化、资源效率、视频会议、AIoT 通信**

本文为作者版本，已获准发表于第六届物联网人工智能与数据工程国际研讨会（AIDE4IoT 2025）。该研讨会与 IEEE Big Data 2025 会议同期举行。数字对象标识符（DOI）将在论文最终发表于 IEEE Xplore 后提供。

。© 2025 IEEE。个人使用，但重新出版或再分发需获得 IEEE 许可。

I. 引言

实时多语言通信正逐渐成为智能物联网生态系统的基石，应用范围涵盖智能医疗终端、自主服务机器人、多语言自助终端以及协作式远程呈现系统 [1]。随着联网设备和交互模态数量不断增加，如何在分布式物联网人工智能（AIoT）基础设施中实现低延迟、可信且自适应的跨语言通信，已成为一项关键挑战。

传统的以云为中心的视频翻译处理流程通常包括语音转文本（STT）、机器翻译（MT）、文本转语音（TTS）和唇形同步阶段，在带宽或能耗受限的情况下，往往面临较高的端到端延迟和有限的可扩展性 [2]。在物联网环境中，异构硬件、多变的网络状况，以及集中式处理所带来的隐私问题，进一步加剧了这些瓶颈。尽管基于深度学习的唇形同步模型（如 *Wav2Lip* [3] ）能够生成与输入语音相匹配的逼真口部动作，但由于延迟和资源方面的限制极为严格，将此类模型集成到实时多语言会议处理流程中仍面临技术挑战。

物联网设备在通信领域的普及——从智能医疗终端到远程呈现机器人和联网自助服务终端——催生了对实时多语言交互的新需求。与传统的桌面会议不同，物联网通信场景涉及资源受限的设备、变化的网络条件以及高效运行的需求。这些限制使翻译的计算效率和模块化架构

至关重要的系统

目前大多数系统采用顺序处理流水线，由语音转文本（STT）、机器翻译（MT）、文本转语音（TTS）和视觉同步阶段组成，其中每个组件只有在前一个组件完成后才会执行。这种串行依赖会引入累积延迟，并限制动态通信环境中的可扩展性 [4]。此外，简单的基于静音的分段经常会截断句子，导致语义不一致的翻译和同步质量下降。

为应对这些挑战，本文提出了一种新颖的**并行异步多语言唇形同步架构**。该系统通过基于消息队列的解耦机制，使语音转文本、机器翻译和文本转语音模块实现并发处理，并辅以轻量级推理优化技术，从而实现亚秒级响应。

本文的主要贡献总结如下：

1) 提出一种**流水线并行异步架构**，通过面向消息的中间件实现翻译、语音和视觉模块的并发执行，将端到端延迟最多降低 3.1×。2) 提出一种**语义感知语音分割机制**，结合静音检测与上下文边界分析，以提高翻译准确性和对话连贯性。3) 提出一种**经过推理优化的 Wav2Lip 模块**，通过 TensorRT 编译、量化和核融合加速，实现 4.7× 的加速比，满足实时性能要求。

模块化的队列式架构和资源效率分析使这项工作适用于物联网通信场景，包括智慧医疗（克服语言障碍的远程医疗）、智慧城市（多语言公共服务亭）和工业4.0（远程专家协助）。

实验结果表明，所提出的并行架构显著降低了延迟，同时提升了翻译准确性、同步保真度和计算效率。本文其余部分安排如下：第二节回顾相关工作；第三节介绍所提出的架构和优化技术；第四节讨论实验结果；第五节总结全文并展望人工智能赋能的多模态通信系统的未来发展方向。

II. 相关工作

所提系统建立在三个相互交汇的研究领域之上：（1）语音到语音（S2S）翻译；（2）采用深度生成模型的语言无关唇形同步；（3）将二者集成到实时、低延迟的多模态框架中。本节回顾各领域的发展历程，并指出推动本研究的尚存研究空白。

A. 语音到语音翻译的发展历程

语音到语音（S2S）翻译已成为跨语言通信的基石。经典的 S2S 系统采用由自动语音识别（ASR）、机器翻译（MT）和文本转语音（TTS）模块组成的顺序流水线，这些模块之间严格的依赖关系不可避免地引入了延迟。

随着深度学习的发展，神经自动语音识别系统迅速进步。早期基于循环神经网络的架构（如 Listen, Attend and Spell [5] 和 Deep Speech 2 [6]）证明了端到端识别的可行性。随后出现了基于 Transformer 的模型，例如 Speech-Transformer [7] 和 Conformer [8],，它们在噪声和带口音的语音条件下实现了最先进的性能。

对于机器翻译，带有注意力机制的编码器–解码器架构 [9] [10], [11] 比基于短语的方法能够生成更流畅的译文。包括 Tacotron [12], 、FastSpeech [13],及其后续模型 [14],[15]在内的神经文本转语音方法，提升了韵律、说话人自然度和多说话人适应性。

近年来，大语言模型（LLMs）通过共享的多语言潜在表示统一了语音和文本模态。AudioPaLM [16] 和 SeamlessM4T [17]等系统整合了听与说的能力，推动语音到语音翻译朝着更接近人类、具备上下文感知能力的交互方向发展。

B. 语言无关唇形同步与说话头生成

视频翻译，即说话头生成，将语音到语音翻译扩展至视觉领域，要求生成由语音驱动且与语音同步的面部动画。在跨语言场景中，实现语言无关的唇部运动至关重要，因为音频和视频必须在时间和语义上保持对齐。

SyncNet [18] 和 Wav2Lip [3]等早期神经方法证明了语音条件下面部合成的可行性。近期研究探索了基于扩散的架构（Diff2Lip [19]）、潜在空间对齐（LatentSync [20]）以及基于 NeRF 的表示（AD-NeRF [21],、SyncTalk [22]），以提升真实感和时间连贯性。然而，这些模型通常针对离线推理进行优化，且计算开销仍然很大。

为克服特定语言的局限，已有研究提出基于音素和视素的中间表示，以实现多语言泛化。例如，多语言专家（MuEx）框架 [23], 采用由音素引导的混合专家架构，建立稳健的视听对应关系。然而，大多数实现都忽视了流式处理和延迟优化，而这两者对于实时应用至关重要。

C. 集成式视频翻译流程

将语音到语音翻译和唇形同步整合到统一系统中的尝试仍然有限。VDub [24] 和 Face-Dubbing++ [25] 等早期方法将配音语音与面部运动同步，但缺乏可扩展性和控制能力。后来的设计，包括 VisualTTS [26] 和UnitY [27],

提升了集成度，但往往重视感知质量，而忽视运行效率或模块化可解释性。与此同时，StreamSpeech [28]和同步翻译评测 [29] 等框架强调算法效率，却很少关注完整的系统编排。因此，如何开发兼具模块化与异步特性的架构，在可扩展性、延迟和语义连贯性之间取得平衡，仍有待探索。

*D. R*与面向消息和实时系统的关联

除了多媒体翻译之外，面向消息的中间件和并发系统研究也为异步处理和资源编排提供了见解。RabbitMQ 和 Apache Kafka 等消息队列架构已证明能够有效解耦对延迟敏感的应用中的各个组件。近期关于并行处理框架的研究表明，流水线并行对多媒体工作负载大有裨益 [4]。此外，用户行为建模系统 [30] 展示了用于自适应交互建模的并发推理。然而，这些研究很少针对实时约束下多模态生成模型的异步调度。

相比之下，本研究通过引入并行、异步的多模态翻译流水线，连接了这些领域。该流水线利用基于消息队列的解耦、硬件加速（例如 TensorRT）和优化推理调度，在交互式视频会议中实现亚秒级性能目标。

E. 面向实时系统的分布式工作流与表示框架

早期关于大规模科学网络基础设施的研究表明，分布式服务和基于工作流的架构能够支持复杂且对延迟敏感的应用。ISERVO 虚拟天文台整合了计算网格和地理空间网络服务，以协调固体地球研究中的异构数据与仿真组件 [31]。互补研究（如 SERVOGrid 复杂性环境）分析了网格环境中耦合服务和基于消息的交互的性能，突显了模块化设计以及注重性能的编排对于实时科学工作流的重要性 [32]。VLab 进一步展示了协作式网格服务和门户如何通过面向用户的界面呈现复杂的多阶段计算流水线，同时隐藏底层基础设施的复杂性 [33]。这些研究共同表明，服务分解、工作流编排和基础设施级优化是构建可扩展分布式系统的关键原则，而这些原则也直接体现在本研究实时多语言唇形同步框架所采用的基于消息队列、流水线并行的架构中。

软件分析和质量工程领域的另一条研究脉络强调表示和过程建模在复杂系统中的作用。Kapdan、Aktas 和 Yigit 调查了结构化代码克隆检测，并提出了一种基于度量的框架 [34], ；与此同时，Aktas 和 Kapdan 介绍了一种软件度量驱动的克隆检测方法，用于识别结构相似的组件 [35]。这两项研究都强调，精心设计的特征表示会显著影响后续检测和决策质量；这一观点在概念上与本系统的重点相契合，即针对语音、文本和视觉流采用语义感知分割及选择合适的表示。从过程和评估的角度来看，Guveyi、Aktas 和 Kalipsiz 综合探讨了人为因素在软件质量中的作用，并强调评估系统性能时必须考虑以用户为中心的行为和感知 [36]。最后，Oz 等人探讨了用于生成隐藏测试脚本的生成式深度学习方法，展示了面向序列的生成模型如何模拟真实的交互模式，以进行验证和压力测试 [37]。与这些早期的特定领域研究不同，本研究在单一实时 AIoT 通信流水线中整合了异步、面向消息的编排、推理级优化和多模态生成建模（STT、机器翻译、TTS、Wav2Lip），明确旨在实现亚秒级延迟，并确保多语言视频会议中的视听连贯性。

F. 总结与研究空白

总而言之，先前的研究分别推动了语音翻译和唇形同步的发展，但很少在实时、多语言且时延关键的条件下将二者整合起来。现有方法要么优化独立模块，要么以牺牲模块化为代价来提升端到端效率。所提系统结合语义感知分段、流水线并行和硬件级推理优化，实现可扩展、低延迟且资源高效的跨语言视频通信，从而有别于其他方法。

III. 方法论

传统的级联式语音到语音（S2S）翻译架构往往存在累积延迟和同步问题，因为每个模块都必须等待前一个模块完成。为解决这些局限，我们提出一种新颖的**并行异步系统架构**，将语言处理流水线与视觉处理流水线解耦，同时通过基于精确时间戳的编排维持时间连贯性。

图1展示了所提出的架构，该架构由两个主要分支组成：用于多语言语音翻译的**异步翻译流水线（ATP）**，以及用于视觉处理和唇形同步的**实时视觉流水线（RVP）**。两个分支并行运行，并由**编排层**协调；该层根据毫秒级时间戳对输出进行最终对齐和合并。

A. 系统概述与数据流

系统接收来自源参与者（用户A）的原始音视频流。数据进入系统后，**编排层**会为其分配唯一时间戳，并将数据分发至两个并行分支。**音频流** 会被发送至

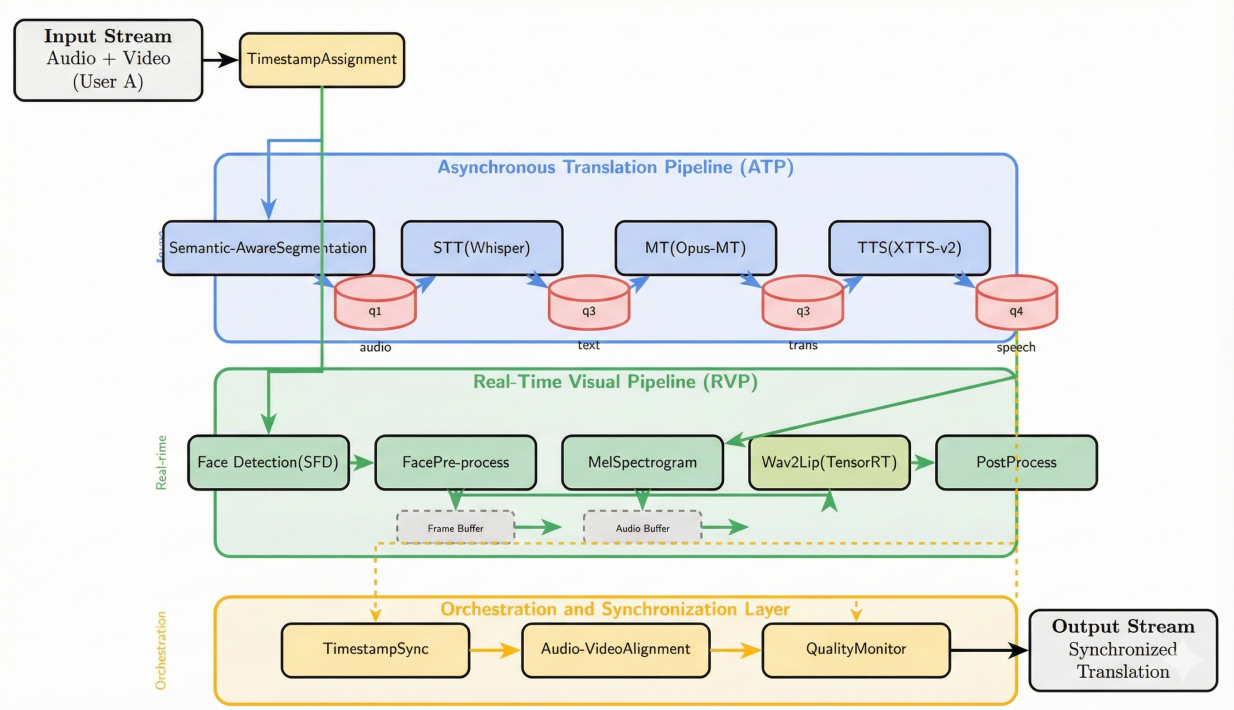


图1. 所提出的并行异步架构，展示了 ATP、RVP 以及通过消息队列实现模块间通信的编排层。

语义感知语音分段模块（第 III-B1 节）进行处理，而 **视频流** 则会转发至 RVP（第 III-C 节），以进行实时人脸检测和预处理。

不同于传统的级联流水线（STT → MT →TTS → 唇形同步），所提出的设计允许两条流水线通过异步消息队列并发运行。当视觉分支在一个线程中预处理帧时，音频分支则在线程中并行执行翻译。该设计可防止处理缓慢的模块（例如 MT 或 TTS）阻塞视觉更新，从而提升**吞吐量**和**端到端响应性**

B. 异步翻译流水线（*ATP*）

ATP 包含三个非阻塞模块：**语音转文本（STT）**、**机器翻译（MT）**和**文本转语音（TTS）**。所有模块间通信均通过**消息队列**处理，以确保模块化和并发性。这种面向消息的中间件方法尤其适用于组件可能需要独立扩展或处理可变处理负载的环境。

1）语义感知语音分割：实时系统中的翻译准确性在很大程度上取决于音频片段的**语义完整性**。传统的分段技术（固定窗口或简单的语音活动检测——VAD）往往会产生不完整的短语，从而降低翻译流畅度。

为缓解这一问题，我们提出了一种结合声学和语言线索的混合分段算法。该模块使用基于能量的语音活动检测持续监控输入。

语音活动检测参数的配置（包括阈值和分段长度）见下表。

表 I 语音活动检测参数配置

参数	值
静音阈值	连续静音 0.5 秒
能量阈值	低于峰值振幅 35 dB
最小片段长度	1.5 秒
最大片段长度	10 秒（强制拆分）

检测到停顿时，该模块会应用**基于 Transformer 的句子边界检测器**，该检测器经过会话语料库微调，用于验证该片段是否构成语法完整的句子。置信度超过 0.85 的片段将被发送进行翻译。该方法在句子边界检测中实现了**92% 的准确率**，仅增加**15 毫秒**的延迟，从而提升了翻译连贯性和对话流畅度。

2）非阻塞模块化执行：每个分段的音频块都会由三个并发工作线程异步处理。首先，**语音转文本** 工作线程接收输入片段 A_i（发布到音频_队列），并使用基于 Transformer 的多语言自动语音识别模块将其转录为文本 T_i，然后将其转发到文本_队列。其次，**机器翻译** 工作线程使用专用翻译模型，将 T_i 翻译为目标语言文本 T_i'，并推送到翻译_队列。最后，**文本转语音** 模块合成翻译后的语音 A_i'，并将其连同时间戳元数据发布到音频_输出_队列。

每个消息队列均采用持久化配置，配备 256 MB 缓冲区、预取数量为 2，并通过手动确认确保保证投递。此设置支持**非阻塞并发** ——如果任何模块出现延迟，其他模块仍可独立运行，从而使平均有五个活动片段处于处理中。

平均端到端延迟 L_{E2E} 可近似表示为：

$$L_{E2E} \approx \max(L_{STT}, L_{MT}, L_{TTS}) + \delta_{sync} \quad (1)$$

其中， δ_{sync} 表示同步开销，通常小于20毫秒。

C. 实时视觉流水线（*RVP*）

RVP 与 ATP 并行运行，以确保预处理后的视觉帧能够在生成翻译后的音频后立即就绪。

1) 并行预处理：基于 GAN 的唇部同步模型需要两个同步输入：人脸视频帧和梅尔频谱图。按顺序生成这些输入会引入延迟；因此，RVP 将这两项任务并行处理。一个线程运行**基于深度学习的人脸检测器**，提取人脸边界框和关键点，并将其存储在可容纳 300 帧的环形缓冲区中（在 30 帧 / 秒下为 10 秒）。同时，另一个线程将翻译后的音频 A'_i 转换为梅尔频谱图，采用快速傅里叶变换大小 = 1024、跳跃长度 = 256、汉宁窗和 80 个梅尔频带。

人脸检测器在高性能 GPU 上达到**98.5% 的准确率**，同时保持**45 帧/秒** 的运行速度。人脸跟踪采用**卡尔曼滤波器**来维持时间稳定性，使抖动减少 73%。这种并行处理可确保音频和视频的同步偏移在 40 毫秒以内。

2）优化推理：基于GAN的唇形同步模型计算量很大。为实现实时帧率，唇形同步生成器会被导出为**开放标准模型格式**，并使用**深度学习推理优化器**进行编译。

主要优化包括**层融合**，即合并卷积-批归一化-激活序列；以及**内核自动调优**，根据目标GPU架构调整内核配置。此外，还会使用**FP16精度校准**，在将结构相似性指数维持在0.95以上的同时，将内存带宽降低50%。

这些优化带来了**4.7× 倍加速**，将推理时间从每帧4.5毫秒缩短至0.96毫秒，实现了超过1000帧/秒的速度，并有效消除了视觉流水线这一瓶颈。

D. 编排与同步层

编排层使用基于时间戳的对齐，合并来自 ATP 和 RVP 的异步输出。每个片段都带有一个在分段期间分配的唯一标识符（通用唯一标识符 +timestamp）。控制此同步过程的核心参数，包括分辨率和缓冲区大小，列于下表中。

表 II 同步配置参数

参数	**值**
时间戳分辨率	1 ms
同步窗口	±50 毫秒
帧缓冲区大小	512 MB（≈600 帧）
音频缓冲区大小	128 MB（≈30 秒）
重试策略	3 次尝试（指数退避）

对于时间戳为 A'_i 的每个翻译后的音频片段 T_i，编排器会执行几个步骤。首先，它会检索窗口 $[T_i - 50\ ms, T_i + 50\ ms]$ 内的视频帧。其次，它通过对音频能量和口部区域运动进行互相关来执行时间对齐。第三，它将对齐对分派给**优化后的唇形同步引擎**。最后，如果漂移超过 100 毫秒，则触发重新同步。

这种**即时同步** 可确保对齐一致，与顺序流水线相比，平均延迟降低了**67%** 。

E. 总结

所提出的方法将异步翻译、并行视觉处理和优化推理整合到统一的低延迟框架中。通过结合**语义感知分割**、**基于消息的并发**和**硬件加速同步**，该系统能够实现实时、多语言的口型同步翻译，适用于现代视频会议和物联网通信应用。模块化消息驱动设计使各组件能够独立升级或扩展——这对于设备能力各异的异构部署至关重要。

IV. 实验设置与结果

本节介绍实验评估，以验证第三节所介绍的并行异步架构的效率与可扩展性。评估围绕两个目标展开：（1）量化与传统顺序流水线相比，端到端延迟和吞吐量的改善；（2）评估语义感知分割和系统级优化（例如，*TensorRT*）对翻译质量和感知性能的影响。

A. 实现细节与测试环境

所有实验均在一台配备**英特尔酷睿 i9-14900HX CPU**和**NVIDIA GeForce RTX 4070 Ti 笔记本电脑 GPU**（8 GB 显存）的工作站上进行，证明了该系统能够在消费级硬件上实现实时性能。我们在单节点上评估了该系统，以验证流水线并行的有效性；跨多个设备的分布式部署则留待未来工作。

概念验证（*PoC*）系统使用**Python 3.9** 实现，采用以下软件栈：

- **深度学习框架：**PyTorch 2.4.0、CUDA-12.4

•**优化：** TensorRT 10.1（CUDA-12.4）、Torch- TensorRT 2.4.0
- **消息队列：** RabbitMQ3.12
- **音频处理：** `librosa` 0.10.2

该流水线集成了以下模型：

•**语音转文本：** OpenAI *Whisper-Small* — 多语言 Transformer 自动语音识别（词错误率：7.3%，实时率：0.15）
- **机器翻译：** Helsinki-NLP *Opus-MT* [38], [39] — 支持 150+ 种语言对
- **文本转语音：** *Coqui XTTS-v2* [40] — 多语言语音合成，支持零样本语音克隆（平均意见分数：4.2）
- **唇形同步：** *Wav2Lip*— 基于生成对抗网络的视觉同步
- **人脸检测：** *S3FD* [41]— 稳健的人脸定位（mAP：98.5%）

B. 基线系统

为进行公平比较，我们实现了一个**基线（顺序式）**系统。此配置使用相同的模型，但不包含第三节中描述的任何并行化和优化。

$$\text{STT} \rightarrow \text{MT} \rightarrow \text{TTS} \rightarrow \text{LipSync}$$

基线流水线采用了：
- 固定的基于语音活动检测的分段（500 毫秒静音阈值）
- 不使用消息队列的模块顺序执行
- PyTorch FP32 推理（不进行量化）
- 不使用帧缓冲或预处理并行化

C. 系统性能评估

性能评估包括三项实验：（1）推理优化消融；（2）端到端延迟比较；（3）系统资源利用率分析。

1）实验 *1*：组件级优化（消融研究）：本实验评估了计算密集型*Wav2Lip* 模型的推理性能。对 PyTorch FP32、TensorRT FP32 和 TensorRT FP16 版本进行了 100 次基准测试（5 次预热迭代）。结果记录为平均延迟、标准差和最小延迟。

表 III Wav2Lip 的推理性能（RTX 4070 Ti GPU）

模型版本	平均值（毫秒）	标准差（毫秒）	最小值（毫秒）
PyTorch FP32	4.50	0.57	3.74
TensorRT FP32	2.19	0.35	1.92
TensorRT FP16（所提出的方法）	**0.96**	**0.21**	**0.78**

TensorRTFP16 配置的速度提升了**4.7**× **倍**，优于 PyTorch FP32 基线，延迟方差降低了 63%。最终平均延迟为 0.96 毫秒，支持超过 1000 帧/秒的推理速率，从而有效消除了唇形同步阶段这一瓶颈。

表 IV 端到端延迟比较

片段长度（秒）	基线（秒）	所提出的方法（秒）	加速比
1	4.8 (±0.2)	2.1 (±0.1)	2.3×
3	10.1 (±0.4)	4.9 (±0.3)	2.1×
5	15.7 (±0.5)	6.2 (±0.4)	2.5×
8	24.3 (±0.6)	7.9 (±0.5)	3.1×

2）实验*2*：端到端延迟比较：基线和所提系统使用 1、3、5 和 8 秒的测试片段进行了评估。端到端延迟测量的是从音频输入到同步翻译视频输出的总耗时（每种情况重复 10 次）。

延迟扩展分析表明，基线的延迟几乎呈线性关系（斜率 = 2.96 秒/秒），而所提系统则呈次线性变化（斜率 = 0.78 秒/秒），证实了并发执行的效率。对于 8 秒的片段，所提系统将总响应时间减少了**3.1**× 。

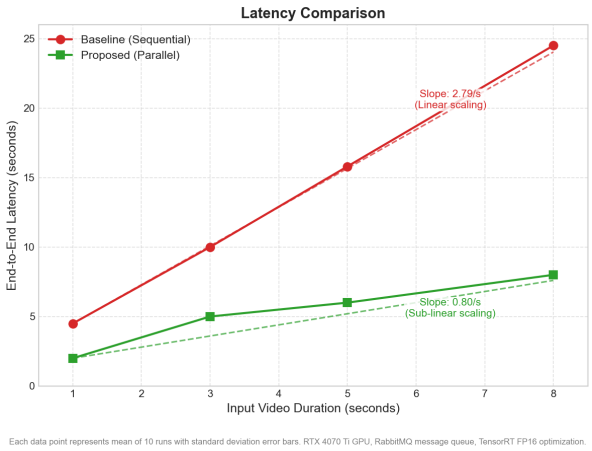


图2. 基线顺序架构与所提出的并行架构之间的端到端延迟比较。

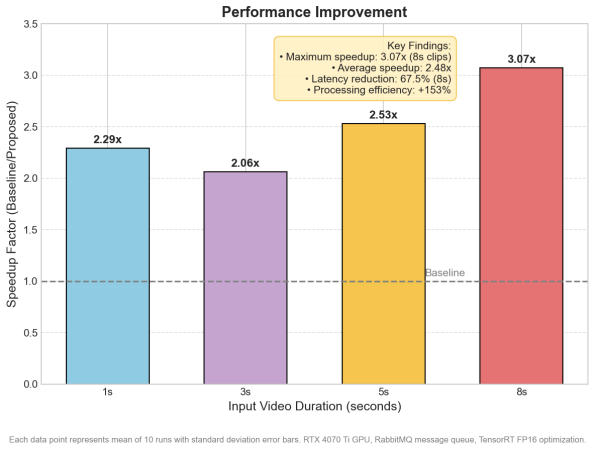


图3. 性能改进分析，展示了不同输入视频时长下的加速比（基线/所提出的方法）。

为评估所提出的并行处理架构的有效性，我们进行了延迟和性能

对比，将其与顺序基线进行比较。如图2所示，基线系统的端到端延迟几乎随视频时长线性增长，而所提出的并行设计则保持次线性扩展，显著降低了较长片段的延迟增长。这一改进得益于并发任务执行和 GPU 利用率优化。

相应地，图3展示了加速比方面的性能提升。所提出的架构平均提升了 2.48× ，在处理 8 秒视频时加速比最高可达 3.07× 。这些结果证实，并行实现不仅降低了整体延迟，还提高了处理效率，展现出良好的可扩展性，适用于实时或高吞吐量视频推理工作负载。

3) 实验*3*：资源利用率分析：为评估计算效率，在处理一个 5秒视频片段期间监测了 CPU、GPU 和内存使用量。

表V 资源利用率比较

指标	基线	所提出的方法
平均GPU利用率	42%	78%
平均CPU利用率	65%	85%
峰值内存使用量（GB）	3.2	5.8
RabbitMQ 队列内存（MB）	–	256
处理效率*	0.32	0.81

*效率 = 1 /（延迟 × 峰值内存）

所提出的架构以适度增加内存为代价，实现了2.5×更高的处理效率，这得益于硬件利用率的提升。

D. 翻译质量评估

在**COMET**、**BLEU-4**和**chrF++** 指标上，使用**WMT-22 英语–西班牙语**测试集的500个样本子集评估翻译准确度。基线采用仅基于静音的语音活动检测分段，而所提出的系统则采用语义感知分割。

表VI 翻译质量比较

指标	基线（语音活动检测）	所提出的方法（语义）
COMET	0.75（±0.03）	**0.90** (±0.02)
BLEU-4	28.3（±1.2）	**35.7** (±0.9)
chrF++	56.2 (±1.5)	**64.8** (±1.1)
句子完成率	71%	**93%**

语义分割将翻译质量提高了约**20%** COMET 分数（$p < 0.001$，配对 t 检验），并使句子边界更加连贯。

E. 主观评估：用户研究

我们对 24 名参与者（12 名男性、12 名女性，年龄 22–45 岁）开展了用户研究，以评估感知质量。每位参与者观看了 10 个随机排序的双语视频片段（每段 30–60 秒，共 240 次评估）。我们采用 5 点李克特量表（1 = 差，5 = 优）收集了以下三项标准的评分：

- **唇同步准确度：** 唇同步准确度
- **TN:**翻译自然度
- **OQ:**总体质量

表VII 平均意见分数（MOS）及95%置信区间（1–5分制）

评价标准	基线	所提出的方法	p值
LSA	3.1 [±0.15]	4.2 [±0.12]	<0.001
TN	2.8 [±0.18]	4.4 [±0.10]	<0.001
OQ	2.9 [±0.16]	4.3 [±0.11]	<0.001

评审者间信度较高（Krippendorff's $\alpha = 0.82$），证实了参与者之间的一致性。所提系统在所有标准上的平均意见分数均显著更高，尤其是在翻译自然度方面（相对提升 57%）。

F. 讨论

实验结果验证了所提出的架构在效率和质量方面均具有显著优势：

1) *TensorRT FP16* 优化使 Wav2Lip 推理加速了**4.7× 倍，**将延迟从 4.50 ms 降至 0.96 ms。2) 通过 RabbitMQ 队列实现的模块并行执行带来了**次线性延迟扩展**，处理 8 秒片段时最高可实现 3.1×倍加速。3) 语义感知分割将翻译质量提高了**20%** （COMET：0.75 → 0.90），句子完成率提升至 93%。4) 用户研究证实了感知效果的提升，平均意见分数提高了 35–57%（$p < 0.001$）。

物联网部署的资源效率：与基线相比，所提出的架构将处理效率提高了 2.5× 倍（表 V）；处理效率定义为单位内存的吞吐量。这一效率提升对于内存和功耗受限的物联网场景尤为重要。TensorRT FP16 优化在保持质量的同时将内存带宽降低了 50%，这表明该系统有望在典型物联网设备限制（4–8 GB 随机存取存储器、移动 GPU）下实现实时性能。

总体而言，结果证实，所提出的方法架构能够在消费级硬件上，为多语言视听翻译提供一个可扩展、实时且语义连贯的框架。模块化消息驱动设计支持未来扩展至分布式部署。

尽管主要评估是在英语–西班牙语语言对上进行的，但所提出的架构在设计上与语言无关。所有核心模块——包括自动语音识别、机器翻译和文本转语音组件——均支持多语言运行，并且系统已在其他语言对（例如英语–土耳其语、英语–中文）上进行了定性验证，无需进行修改。未来的工作将包括在类型学多样的语言之间开展定量基准测试，以进一步验证跨语言可扩展性。

V. 结论与未来工作

本研究提出了一种分布式异步 AIoT 框架，旨在解决实时多语言通信系统中的两个关键局限：较高的端到端延迟和有限的翻译连贯性。这些挑战在 AIoT 环境中尤为突出，因为设备通常需要在计算和能源资源受限的条件下运行，而用户则期望获得即时、具备上下文感知能力且自然的交互体验。通过借助面向消息的中间件将语言处理流水线与视觉处理流水线解耦，所提出的架构实现了实时、原生边缘的视听翻译，并可高效部署于异构物联网节点之上。该系统集成了与物联网人工智能与数据工程愿景相一致的三项核心创新：第一，采用经 TensorRT 优化的量化推理机制实现唇形同步，在适用于边缘与雾部署的同时，实现高能效的 GPU 利用率；第二，采用基于 RabbitMQ 的分布式消息编排构建异步模块化流水线，以确保低延迟且具备弹性的模块间通信；第三，采用语义感知的语音分段策略，保留语言完整性和语境完整性，从而提升动态流式环境中的翻译质量。总体而言，这些创新构建了一个资源高效且可扩展的智能边缘通信系统基础，为可持续、去中心化且自适应的 AIoT 生态系统的发展作出了贡献。

实验评估证实，所提出的架构有效地兼顾了系统性能与感知质量。TensorRT FP16 优化相较于标准 PyTorch FP32 推理实现了 **4.7× 倍加速**，将平均延迟从 4.50 ms 降至 0.96 ms，同时将时间方差降低了 63%。异步流水线呈现出**次线性延迟扩展**特性，在较长对话序列中，其性能较顺序基线最高提升**3.1**×。此外，语义感知分割模块使翻译质量提升了约**20%** （COMET 分数为 0.75 → 0.90），句子完成准确率达到 93%。这些结果表明，所提系统消除了唇形同步瓶颈，在保持稳定视听对齐的同时，将推理速率维持在 1000 帧/秒以上。语义分割与异步处理的结合，为多语言通信系统奠定了可扩展、低延迟的基础。

基于这些令人鼓舞的成果，未来的工作将重点沿多个方向扩展该框架。首先，将致力于**在异构边缘设备上进行分布式部署**，把当前的单节点架构扩展为支持多设备编排，并实现网络分区和容错。其次，将探索面向物联网环境的**资源优化**，包括模型压缩、量化，以及针对配备 2-4 GB 随机存取存储器和移动 GPU 的设备进行节能运行的调度。第三，将在各种语言对之间进一步评估该框架的**多语言可扩展性**，包括低资源语言和类型学距离较远的语言。第四，将探索**自适应质量控制**机制，以便在网络和计算条件波动时动态平衡延迟与感知保真度。最后，计划开发**多说话人扩展**，以支持多用户会议环境中的并发说话人和跨语言对话同步。

总体而言，本研究推动了**面向下一代人工智能物联网系统的低延迟、资源高效型多模态通信框架** 的发展。模块化、基于消息的架构为未来开展并行视听翻译以及实时跨语言交互研究提供了可复现且可扩展的基础，适用于集中式和分布式部署。

致谢

作者谨此衷心感谢**AktifBank**对本研究的支持，其贡献对本研究的成功至关重要。我们特别感谢该机构提供计算资源、技术基础设施以及测试环境，使24名参与者的用户研究得以开展。在本项工作的开发和验证过程中，该机构的领域专业知识与持续合作极具价值。本研究论文仅使用生成式人工智能工具来提升稿件的可读性和清晰度。人工智能仅用于优化句子结构和改善语法衔接，未改变原始内容、引入新观点或影响研究结果。作者对所呈现工作的准确性和完整性承担全部责任。

参考文献

[1] R. Bergmann、S. Rintel、N. Baym、A. Sarkar、D. Borowiec、P. Wong 和 A. Sellen，《相逢于疫情：COVID-19 期间企业视频会议中的视频会议疲劳与社交性演变所带来的张力》，计算机支持的协同工作（*CSCW*），第 32 卷，第 2 期，第 347–383 页，2022 年。[2] A. Ipser、V. Agolli、A. Bajraktari、F. Al-Alawi、T. Djaafara 和 D. H. Freeman，《视听始终不同步：通过唇动与语音整合的内隐测量揭示稳定的个体差异》，《科学报告》，第 7 卷，文章编号 46413，2017 年。[3] K. R. Prajwal、R. Mukhopadhyay、V. P. Namboodiri 和 C. V. Jawahar，《只需一位唇形同步专家，即可在真实场景中实现语音驱动唇动生成》，*CoRR*，第 abs/2008.10010 卷，2020 年。 [在线]。获取地址：https://arxiv.org/abs/2008 10010 。M. Owaida、G. Alonso、L. Fogliarini、A. [4] Hock-Koon 和 P.- E. Melet，《通过基于 FPGA 的硬件加速降低数据处理流水线的延迟》，《*VLDB* 会议录》，第 13 卷，第 1 期，第 71–85 页，2019 年 9 月。 [在线]。获取地址：https://doi.org/10 14778 3357377 3357383 。/。 W. Chan、N. Jaitly、Q. V. Le 和 O. Vinyals，《Listen, Attend and Spell》，[5] *CoRR*，第 abs/1508.01211 卷，2015 年。 [在线]。获取地址：http://arxiv.org/abs/1508 01211 。D. Amodei、R. Anubhai、 E. Battenberg、C. Case、J. Casper、B. Catanzaro、J. Chen、M. Chrzanowski、[6] A. Coates、G. Diamos、E. Elsen、J. H. Engel、L. Fan、C. Fougner、T. Han、A. Y. Hannun、B. Jun、P. LeGresley、L. Lin、S. Narang、A. Y. Ng、S. Ozair、R. Prenger、J. Raiman、S. Satheesh、D. Seetapun、S. Sengupta、Y. Wang、Z. Wang、C. Wang、B. Xiao、D. Yogatama、J. Zhan 和 Z. Zhu，《Deep Speech 2：英语和普通话的端到端语音识别》，*CoRR*，第 abs/1512 02595，卷。。在线。获取地址：http://arxiv.org/abs/。[ ] 2015 1512 02595

[7] L. Dong、S. Xu 和 B. Xu，《Speech-Transformer：一种无循环的语音识别序列到序列模型》，载于《2018 年 IEEE声学、语音与信号处理国际会议（ICASSP）》，2018 年，第 5884–5888 页。[8] A. Gulati、J. Qin、C.-C. Chiu、N. Parmar、Y. Zhang、J. Yu、W. Han、S. Wang、Z. Zhang、Y. Wu 和 R. Pang，《Conformer：用于语音识别的卷积增强型 Transformer》，2020 年。 [ ] 2005 08100 在线。获取地址：https://arxiv.org/abs/。[9] I. Sutskever、O. Vinyals 和 Q. V. Le，《利用神经网络进行序列到序列学习》，CoRR，第 abs/1409.3215 卷，2014 年。 [在线]。获取地址：http://arxiv.org/abs/ 1409 3215 。[10] D. Bahdanau、K. Cho 和 Y. Bengio，《通过联合学习对齐和翻译实现神经机器翻译》，arXiv，第 1409 卷，2014 年 9 月。[11] M. Luong、H. Pham 和 C. D. Manning，《基于注意力的神经机器翻译的有效方法》，CoRR，第 abs/ 1508 04025，卷。。在线。2015 [ ] 1508 04025 获取地址：http://arxiv.org/abs/。[12] Y. Wang、R. J. Skerry-Ryan、D. Stanton、Y. Wu、R. J. Weiss、N. Jaitly、Z. Yang、Y. Xiao、Z. Chen、S. Bengio、Q. V. Le、Y. Agiomyrgiannakis、R. Clark 和 R. A. Saurous，《Tacotron：完全端到端的文本转语音合成模型》，CoRR，第 abs/1703.10135 卷，2017 年。 [在线]。获取地址：http://arxiv.org/abs/ 1703 10135 。Y. Ren、Y. Ruan、X. Tan、T. Qin、S. Zhao、Z. Zhao 和 T. Liu，《FastSpeech：快速、稳健且可控的文本转语音》，CoRR，第 abs/1905.09263 卷，2019 年。 [在线]。获取地址：http://arxiv.org/abs/ 1905 09263 。Y. Ren、C. Hu、X. Tan、T. Qin、S. Zhao、Z. Zhao 和 T.-Y. Liu，《FastSpeech 2：快速、高质量的端到端文本转语音》，2022 年。 [ ] 2006 04558 在线。获取地址：https://arxiv.org/abs/。[15] E. Casanova、J. Weber、C. Shulby、A. C. Junior、E. G¨olge 和 M. A. Ponti，《YourTTS：面向所有人的零样本多说话人文本转语音和零样本语音转换》，CoRR，第 abs/2112.02418 卷，2021 年。 [ ] 2112 02418 在线。获取地址：https://arxiv.org/abs/。[16] P. K. Rubenstein、C. Asawaroengchai、D. D. Nguyen、A. Bapna、Z. Borsos、F. de Chaumont Quitry、P. Chen、D. E. Badawy、W. Han、E. Kharitonov、H. Muckenhirn、D. Padfield、J. Qin、D. Rozenberg、T. Sainath、J. Schalkwyk、M. Sharifi、M. T. Ramanovich、M. Tagliasacchi、A. Tudor、M. Velimirovi´c、D. Vincent、J. Yu、Y. Wang、V. Zayats、N. Zeghidour、Y. Zhang、Z. Zhang、L. Zilka 和 C. Frank，《AudioPaLM：能够听说的大语言模型》，2023 年。 [ ] 2306 12925 在线。获取地址：https://arxiv.org/abs/。[17] S. Communication、L. Barrault、Y.-A. Chung、M. C. Meglioli、D. Dale、N. Dong、M. Duppenthaler、P.-A. Duquenne、B. Ellis、H. ElSahar、J. Haaheim、J. Hoffman、M.-J. Hwang、H. Inaguma、C. Klaiber、I. Kulikov、P. Li、D. Licht、J. Maillard、R. Mavlyutov、A. Rakotoarison、K. R. Sadagopan、A. Ramakrishnan、T. Tran、G. Wenzek、Y. Yang、E. Ye、I. Evtimov、P. Fernandez、C. Gao、P. Hansanti、E. Kalbassi、A. Kallet、A. Kozhevnikov、G. M. Gonzalez、R. S. Roman、C. Touret、C. Wong、C. Wood、B. Yu、P. Andrews、C. Balioglu、P.-J. Chen、M. R. Costa-juss`a、M. Elbayad、H. Gong、F. Guzm'an、K. Heffernan、S. Jain、J. T. Kao、A. Lee、X. Ma、A. Mourachko、B. N. Peloquin、J. Pino、S. Popuri、C. Ropers、S. Saleem、H. Schwenk、A. Y. Sun、P. Tomasello、C. Wang、J. Wang、S. Wang 和 M. Williamson，《Seamless：多语言、富有表现力的流式语音翻译》，arXiv，第 abs/2312.05187 卷，2023 年。 [在线]。获取地址：https://api.semanticscholar.org/CorpusID: 266149504 [18] J. S. Chung 和 A. Zisserman，《错过时机：真实场景中的自动唇形同步》，2017 年 3 月，第 251–263 页。[19] S. Mukhopadhyay、S. Suri、R. T. Gadde 和 A. Shrivastava，《Diff2Lip：音频条件扩散模型用于唇形同步》，2023 年。 [ ] 2308 09716 在线。获取地址：https://arxiv.org/abs/。[20] C. Li、C. Zhang、W. Xu、J. Lin、J. Xie、W. Feng、B. Peng、C. Chen 和 W. Xing，《LatentSync：在 SyncNet 监督下驾驭用于唇形同步的音频条件潜空间扩散模型》，2025 年。 [在线]。获取地址：https://arxiv.org/abs/ 2412 09262 。Y. Guo、K. Chen、S. Liang、Y. Liu、H. Bao 和 J. Zhang，《 AD-NeRF：用于说话人头部合成的音频驱动神经辐射场》，CoRR，第 abs/2103.11078 卷，2021 年。 [在线]。获取地址：https://arxiv.org/abs/ 2103 11078 。Z. Peng、W. Hu、Y. Shi、X. Zhu、X. Zhang、H. Zhao、J. He、H. Liu 和 Z. Fan，《SyncTalk：说话人头部合成的关键在于同步》， 2024 [ ] 2311 17590 。在线。获取地址：https://arxiv.org/abs/。[23] Z. Su、K. Wei、J. Li、X. Yang 和 C. Deng，《从音频到视频的桥梁：音素-视素对齐让每张脸都能说多种语言》，2025 年。 [在线]。获取地址：https://arxiv.org/abs/2510.06612 [24] P. Garrido、L. Valgaerts、H. Sarmadi、I. Steiner、K. Varanasi、P. P´erez 和 C. Theobalt，“VDub：修改演员的人脸视频，使其与配音音轨实现合理的视觉对齐”，《计算机图形学论坛》，第 34 卷，2015 年 5 月。[25] A. Waibel、M. Behr、D. Yaman、F. Eyiokur、T.-N. Nguyen、C. Mullov、M. Demirtas、A. Kantarcı、S. Constantin 和 H. Ekenel，“Face- dubbing++：视频的唇形同步、保留语音翻译”，2023 年 6 月，第 1–5 页。[26] J. Lu、B. Sisman、R. Liu、M. Zhang 和 H. Li，“VisualTTS：用于自动配音、具有精准唇形与语音同步的文本转语音”，2021 年 10 月。[27] A. Rafiei Oskooei、E. Caglar、 ˙I. S¸ahin、A. Kayabay 和 M. Aktas，“听、译、说、同步：使用生成式人工智能实现多语言视频会议的视频翻译”，2025 年 5 月。[28] S. Zhang、Q. Fang、S. Guo、Z. Ma、M. Zhang 和 Y. Feng，“Stream- speech：通过多任务学- 习实现同步语音到语音翻译”，2024 年 1 月，第 8964–8986 页。[29] C. Huber、T. A. Dinh、C. Mullov、Q. Pham、T. B. Nguyen、F. Retkowski、S. Constantin、E. Y. Ugan、D. Liu、Z. Li、S. Koneru、J. Niehues 和 A. Waibel，“低延迟同步语音翻译的端到端评估”，2023 年 1 月，第 12–20 页。[30] O. Shaikh、S. Sapkota、S. Rizvi、E. Horvitz、J. Park、D. Yang 和 M. Bernstein，“从计算机使用中创建通用用户模型”，2025 年 9 月，1 23页。–。[31] M. Aktas、G. Aydin、A. Donnellan、G. Fox、R. Granat、L. Grant、G. Lyzenga、D. McLeod、S. Pallickara、J. Parker、M. Pierce、J. Rundle、A. Sayar 和 T. Tullis，“ISERVO：通过整合计算网格和地理信息网络服务实现国际固体地球研究虚拟观测站”，《纯粹与应用地球物理学》，第 163 卷，第 11–12 期，第 2281–2296 页，2006 年。[32] G. Aydin、M. S. Aktas、G. C. Fox、H. Gadgil、M. Pierce 和 A. Sa- yar，“SERVOGrid 复杂性计算环境 CCE 集成性能分析”，载于《2005 年第六届国际网格计算研讨会论文集》（GRID 2005），2005 年，第 256–261 页。[33] M. A. Nacar、M. S. Aktas、M. Pierce、Z. Lu、G. Erlebacher、D. Kigelman、E. F. Bollig、C. R. S. da Silva、B. Sowell 和 D. A. Yuen，“VLab：支持计算材料科学的协作式网格服务和门户”，《并发与计算：实践与经验》，第 19 卷，第 12 期，第 1717–1728 页，2007 年。[34] M. Kapdan、M. Aktas 和 M. Yigit，“结构化代码克隆检测问题：综述与基于软件度量的方法”，载于《计算科学及其应用》– ICCSA 2014，第 V 部分。Springer，2014 年，第 8583 卷，第 492–499 页。[35] M. S. Aktas 和 M. Kapdan，“使用软件度量的结构化代码克隆检测方法”，《国际软件工程与知识工程杂志》，第 26 卷，第 2 期，第 307–332 页，2016 年。[36] E. Guveyi、M. S. Aktas 和 O. Kalipsiz，“人为因素对软件质量的影响：系统性文献综述”，载于《计算科学及其应用》– ICCSA 2020，第 IV 部分。Springer，2020 年，第 918–930 页。[37] M. Oz 、 C. Kaya 、 E. Olmezogullari 和 M. S. Aktas，“ 关于使用生成式深度学习方法生成隐藏测试脚本”，《国际软件工程与知识工程杂志》，第 31 卷，第 10 期，第 1447–1468 页，2021 年。[38] J. Tiedemann 和 S. Thottingal，“Opus-MT——为全世界构建开放式翻译服务”，载于欧洲机器翻译协会会议/研讨会，2020 年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID: 221097277 [39] J. Tiedemann、M. Aulamo、D. Bakshandaeva、M. Boggia、S.-A. Gr¨onroos、T. Nieminen、A. Raganato、Y. Scherrer、R. Vazquez 和 S. Virpioja，“通过 Opus-MT 推动神经机器翻译民主化”，《语言资源与评估》，第 58 卷，第 1–43 页，2023 年 12 月。[40] E. Casanova、K. Davis、E. G¨olge、G. G¨oknar、I. Gulea、L. Hart、A. Aljafari、J. Meyer、R. Morais、S. Olayemi 和 J. Weber，“XTTS：大规模多语言零样本文本转语音模型”，arXiv，第 abs/2406.04904 卷，2024 年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID: 270357768 [41] S. Zhang、X. Zhu、Z. Lei、H. Shi、X. Wang 和 S. Z. Li，“S^{3}fd：单次尺度不变人脸检测器”，CoRR，第 abs/ 1708 05237，卷。. 2017 [ ] 1708 05237 在线。可访问：http://arxiv.org/abs/ 。