FlashLabs Chroma 1.0：具备个性化语音克隆功能的实时端到端语音对话模型

Tanyu Chen[*], **Tairan Chen**[*], **Kai Shen**[*], **Zhenghua Bao**[*,†], **Zhihui Zhang**,**Man Yuan**,**Yi Shi**[†]FlashLabs

联系邮箱：zhenghua.bao@flashlabs.ai，chroma@flashlabs.ai

arXiv:2601.11141v1 [cs.SD] 16 Jan 2026

摘要

近期的端到端语音对话系统借助语音分词器和神经音频编解码器，使大语言模型能够直接处理离散语音表示。然而，这些模型往往难以充分保持说话人身份特征，阻碍了个性化语音交互的发展。在本研究中，我们推出了**Chroma1.0**，这是首个开源、实时的端到端语音对话模型，兼具低延迟交互和高保真个性化语音克隆能力。Chroma 采用交错的文本—音频令牌调度方案（1:2），支持流式生成，实现了低于一秒的端到端延迟，同时在多轮对话中保持高质量的个性化语音合成。实验结果表明，与人类基线相比，Chroma 的说话人相似度相对提升了**10.96%**，实时因子为**0.43**，同时保持了出色的推理和对话能力。我们的代码和模型已在以下平台公开发布：GitHub和HuggingFace。

1 引言

构建实时交互式语音对话系统仍然颇具挑战。大多数已部署的系统都采用级联流水线：首先通过自动语音识别（ASR）将输入语音转录为文本，然后由下游大语言模型（LLM）处理转录文本，最后通过文本转语音（TTS）模块将生成的回复转换为语音。虽然这类流水线适用于离线或对延迟不敏感的场景，但端到端延迟较高，容易出现错误传播，而且往往会丢失说话人身份、语速、音色、情感和韵律等副语言信息。这些限制在实时对话场景中尤为突出，因为自然度、响应速度和说话人保真度至关重要。

语音标记化和神经编解码器的最新进展（Hsu 等，2021年；Défossez 等，2022年；Ji 等，2024年）使大语言模型能够直接处理离散语音表示，由此催生了语音到语音（S2S）系统，绕过了显式的中间转录。GPT-4o（Hurst 等，2024年）的成功推动了端到端大音频语言模型（LALMs）的出现；GPT-4o 展示了无需显式文本中间表示即可进行端到端语音处理的可行性。后续研究（Défossez 等，2024年；Zeng 等，2024年；Xu 等，2025a,b；Nguyen 等，2025年；Huang 等，2025a,b；Wu 等，2025年）沿用了这一原则，通过多种架构策略直接处理和生成语音：交错排列的语音—文本标记序列（Nguyen 等，2025年）、统一音频文本表示（Huang 等，2025a,b；Wu 等，2025年；Zeng 等，2024年），或将语义推理与声学生成解耦的并行双流架构（Xu 等，2025a,b）。

然而，现有的端到端语言音频语言模型在语音理解和语音生成方面仍面临若干关键局限。早期模型，如 Spirit LM（Nguyen 等，2025年）和 GLM-4-Voice（Zeng 等，2024年），主要关注将语音中的语义内容对齐到文本，往往难以捕捉自然交互所必需的副语言线索。尽管 Qwen2-Audio（Chu 等人，2024年）和 Kimi-Audio（Ding 等，2025年）等较新的音频理解模型能够理解这类副语言信息，但它们只生成文本输出，未能利用这种理解来生成富有表现力的语音。此外，当前的语音到语音系统通常更重视对话质量，而非个性化语音保真度。尽管 VALL-E（Wang 等，2023年）和 CosyVoice 系列（

[*]贡献相同。
[†]通讯作者。

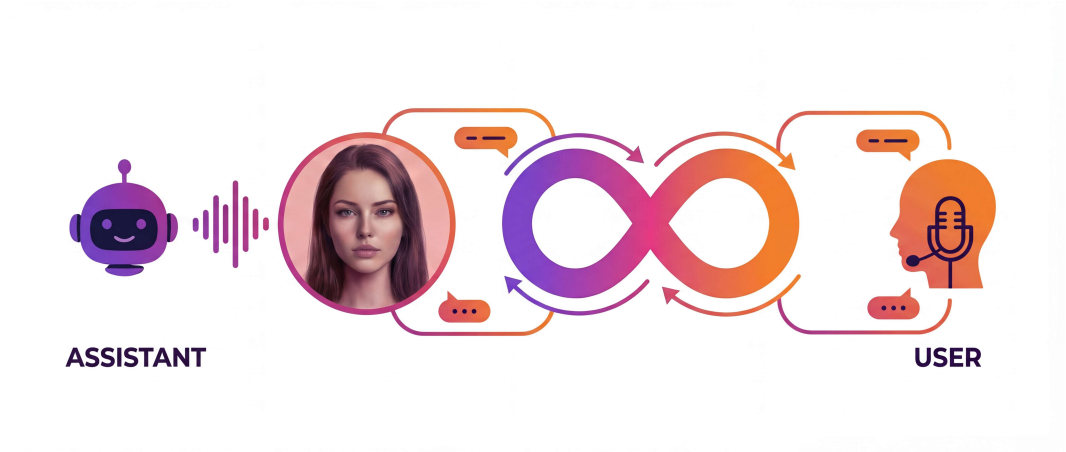

图 1：**Chroma 1.0** 的系统工作流程。Chroma 以语音为输入并生成语音作为输出，在整个对话过程中保持一致的说话人身份。

（Du 等，2024a，b，2025年）以及工业解决方案（如 Qwen-TTS[1] 和 ElevenLabs[2]）等具备语音克隆能力的语音系统，在说话人适配方面表现出色，却缺乏实时流式处理能力，无法在多轮对话中持续一致地克隆声音。相反，实时对话模型（Défossez 等，2024年；Xie 和 Wu，2024a，b；Fang 等，2024年）则牺牲了精细的说话人控制，以实现低延迟。

为解决这些局限性，我们提出了 **Chroma 1.0**，这是首个同时实现低延迟交互和高保真个性化语音克隆的开源实时端到端口语对话模型。我们的主要贡献如下：

- 一种流式架构，通过语义状态表示将语音理解与语音生成紧密耦合，实现端到端延迟低于一秒。
- 高保真语音克隆：利用仅数秒参考音频中的音频嵌入对生成模型进行条件控制，相较于人类基线，使说话人相似度提升 10.96%。
- 一种交错文本-音频词元调度方案（1:2），通过增量式文本输出实现声学编码的同步生成，从而实现实时流式合成。
- 仅用 4B 个参数便具备出色的推理和对话能力，在理解、推理和口语对话任务中均展现出有竞争力的表现。

我们公开了完整的代码库、训练流程和预训练模型权重，以支持结果复现和进一步研究[3] [4]。

2 相关工作

2.1 级联语音-文本流水线与端到端语音到语音对话

级联方法仍被广泛采用，这得益于自动语音识别、大语言模型和文本转语音模块的成熟，以及既有的工程实践（Zhang 等人，2023a；Radford 等人，2023；Wang 等人，2023；Chu 等人，2023；Tan 等人，2024；Huang 等人，2024；Lin 等人，2024；Fang 等人，2024）。这类方法灵活性高，但会累积延迟与错误；而且，语音一旦被转换为文本，副语言线索往往也会丢失，从而限制下游的说话人保真度和情感表现力。相比之下，端到端语音到语音对话系统通过学习直接在离散语音单元之间进行映射，避免了显式转录。SeamlessM4T（Duquenne 等人，2023）确立了多语言语音到语音基线，而 GPT-4o（Hurst 等人，2024）则展示了无需中间文本转换的原生音频处理在商业上的可行性。近期研究探索了端到端语音建模的多种架构策略。Moshi（Défossez 等，2024）引入了用于实时对话的全双工流式架构；Spirit LM（Nguyen 等，2025）则采用交错语音-文本词元序列，以保持多模态对齐。GLM-4-Voice（Zeng 等人，2024）

[1] https://qwenlm.github.io/blog/qwen-tts/ [2] https://elevenlabs.io/voice-cloning

[3] https://github.com/FlashLabs-AI-Corp/FlashLabs-Chroma [4] https://huggingface.co/FlashLabs/Chroma-4B

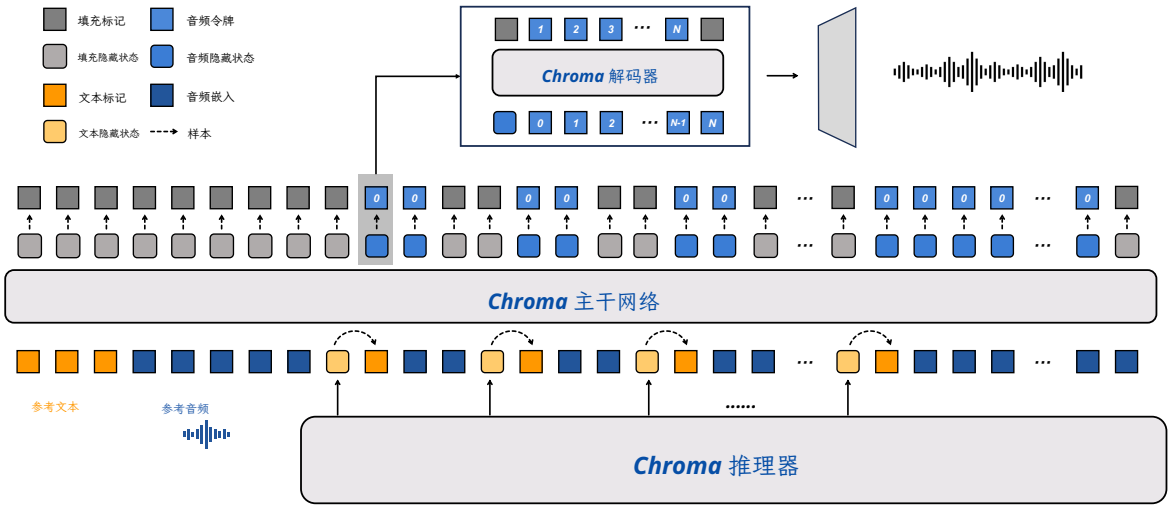

图 2：**Chroma 1.0**的总体架构。**推理器**输出文本标记和隐藏状态。这些构成一个交错的文本—音频嵌入序列（1:2），供**骨干网络**使用，以生成粗粒度声学编码 c_t^0和隐藏状态 h_t。**解码器**预测其余的残差向量量化级别 $c_t^{1:N-1}$，而**编解码器解码器**则将完整的码本序列重建为连续波形，从而实现高保真语音到语音生成。

以及早期的 Step-Audio 系列（Huang 等，2025a，b）采用统一令牌表示，将语音和文本映射到共享词汇空间；Step-Audio-2（Wu 等，2025年）则进一步运用检索增强生成（RAG）和强化学习技术提升生成质量。Qwen2.5-Omni（Xu 等，2025a）和 Qwen3-Omni（Xu 等，2025b）提出了并行双流（Thinker-Talker）架构，将语义推理与声学生成解耦。具体而言，Qwen3-Omni 在 Talker 中采用多码本词元预测（MTP）模块，按帧预测残差码本。主要专注于音频理解的模型——例如 Qwen2-Audio（Chu 等人，2024）和 Kimi-Audio（Ding 等人，2025）——会生成文本回复，展现了从语音输入中理解副语言信息的能力。

2.2 语音克隆与个性化语音合成

语音克隆领域的一项重大进展来自神经编解码器语言模型（NCLM），这类模型将文本转语音视为离散声学令牌生成。VALL-E（Wang 等人，2023年）证明，利用大型条件语言模型对 EnCodec 编码进行建模，可以生成自然且富有表现力的零样本文本转语音，仅需约 3 秒的参考音频，同时保留副语言线索。VALL-E X（Zhang 等人，2023b）将这一能力扩展到了多种语言。扩散和流匹配方法进一步提升了保真度和可控性。StyleTTS-2（Li 等人，2023年）引入了风格扩散，实现了出色的自然度和零样本适配；NaturalSpeech 3（Ju 等人，2024年）将内容、韵律和音色解耦，实现了最先进的相似度；Voicebox（Le 等人，2023年）则展示了高速非自回归流匹配。CosyVoice 系列（Du 等人，2024a，b，2025年）和 OpenVoice（Qin 等人，2023年）等开源系统展示了实用的克隆能力，在提升内容一致性和精细控制方面有所进步，后续版本还加入了流式支持。商业平台（如 ElevenLabs）也进一步证明，几秒钟的参考音频即可实现可用于生产环境的克隆。

3 模型架构

Chroma 1.0 构建于端到端的语音对话架构之上，旨在通过紧密耦合语音感知与语言理解，实现高质量的实时语音交互。如图 2所示，该系统由两个集成子系统组成：(i) **Chroma 推理器**，负责多模态输入理解和文本响应生成；以及 (ii) 一个语音

合成流水线，由用于**声学建模的 Chroma 主干网络**、用于**音频码本解码的 Chroma 解码器**以及用于**波形重建的 Chroma 编解码器解码器**组成。

3.1 Chroma 推理器

Chroma 推理器基于 Thinker 模块（Xu 等，2025a）构建，能够执行多模态理解和语义表示构建。它通过标准的 Qwen2-Audio 编码流程（Chu 等人，2024年）处理文本和音频输入，生成同时捕捉语言内容与声学特征的高层语义表征。

推理器采用跨模态注意力机制融合文本和音频特征。编码后的表征通过时间对齐的多模态旋转位置嵌入（TM-RoPE）（Xu 等，2025a）统一为隐藏状态序列。这种融合使模型能够将语音中的韵律和节奏线索与文本语义结合起来，从而增强后续语音合成中的对话理解和上下文建模能力。

3.2 Chroma 主干网络

Chroma 主干网络采用 1B 参数的 LLaMA（Touvron 等人，2023年）架构变体，旨在生成音色与给定参考音频相匹配的语音。为实现高保真语音克隆，我们使用 CSM-1B 将参考音频及其对应的转录文本编码为嵌入提示[5]，并将其前置于输入序列之前，使模型明确以目标说话人的声学特征为条件。在推理过程中，为确保音频输出与推理器的文本模态严格对齐，同时保持较低的参数量，我们采用共享词元嵌入策略：将推理器的词元嵌入和隐藏状态作为统一文本上下文输入主干网络。

为支持高效的流式生成，我们以固定的 1 : 2 比例交错排列文本标记和音频编码词元 c^0，即每个文本标记对应两个音频编码。借助这一机制，骨干网络能够与推理器的增量文本生成并行，自回归地生成音频序列，使系统无需等待完整的文本序列即可生成输出。因此，首令牌时间（TTFT）显著缩短，从而提升实时交互体验。

[5] https://github.com/SesameAILabs/csm

3.3 Chroma 解码器

为了在保持生成质量的同时大幅加快推理速度，我们引入了一个轻量级模块——**Chroma 解码器**——负责生成其余的声学编码（c_t^1，……，c_t^{N-1}），而不是让骨干网络直接生成所有码本。Chroma 解码器采用 LLaMA 架构的一个变体实现，参数量约为 1 亿。与骨干网络不同，该模块不依赖完整的文本输入历史或参考音频上下文；相反，它仅以当前时间步的骨干网络输出为条件，执行帧同步推理，从而大幅降低长上下文处理所带来的计算开销。

具体而言，在每个时间步 t，Chroma 解码器以骨干网络生成的隐藏状态特征 h_t 和初始音频码本 c_t^0 为输入，并在每帧内基于先前生成的层级，通过针对不同层级的投影头自回归地生成其余的残差向量量化码本 c_t^i（$i = 1$，……，$N-1$）。这种解耦设计不仅降低了推理延迟，还使 Chroma 解码器能够在骨干网络提供的粗粒度语义表征基础上，丰富韵律和发音细节等精细的声学属性。

3.4 Chroma 编解码器解码器

作为最终的声学重建模块，**Chroma 编解码器解码器** 将离散编码序列映射为连续的高保真语音波形。在每个时间步，模块将粗码本（c^0）与 Chroma 解码器生成的精细声学码本（c^1、……、c^{N-1}）拼接起来，形成完整的离散声学表征。在架构上，该模块遵循 Mimi 声码器的解码器设计（Défossez 等，2024年），并采用因果卷积神经网络（Causal CNN），确保波形重建过程中严格遵循时间因果性，从而支持流式生成。

为满足实时交互需求，我们采用 8 个码本（$N = 8$）。这一配置大幅减少了 Chroma 解码器所需的自回归细化步骤，从而提高了推理效率。

3.5 训练数据集

公开可用的数据集缺少符合我们模型对语义理解和推理能力要求的高质量语音对话数据。为解决这一局限，我们提出了一种数据生成流程，利用大语言模型与文本转语音系统之间的协同合作。

流程概述。 生成过程由两个阶段组成：(1)文本生成：将用户问题输入类似推理器的大语言模型模块，以生成相应的文本回复。(2)语音合成：使用 TTS 系统将文本回复合成为语音，其音色特征与参考音频相匹配。该合成语音作为训练目标，使骨干网络和解码器模块能够学习语音克隆和声学建模。

3.6 训练目标

我们的训练策略优化两个主要组件：骨干网络和解码器，同时将推理器冻结为特征提取器。对于每个音频文本对，推理器提供固定的文本嵌入和多模态隐藏状态，作为语义和韵律条件。骨干网络经过训练，以自回归方式预测第一层粗粒度声学编码（c^0）。为确保因果对齐，其仅关注声学编码的前缀以及相应的推理器表征。该目标使模型能够捕捉长期时间结构，并使声学生成与文本进程对齐。解码器通过预测剩余的残差向量量化（RVQ）层级来细化粗粒度声学表征（$c^{1:N-1}$）。在骨干网络的粗编码和隐藏状态的条件下，解码器通过帧内自回归过程运行。这种分解使其能够在保持与骨干网络所建立的粗粒度轨迹一致性的同时，逐步提升声学保真度。更多详情请参见附录 C。

4 实验

4.1 实验设置

数据集。我们在多个基准数据集上评估模型，以衡量语音生成质量的不同方面。对于一般性能评估，我们使用 **CommonVoice 数据集**（Ardila 等人，2020年），该数据集包含多样的说话人和录音条件，可用于测量说话人相似度。为评估声音克隆能力，我们开展主观实验，测量生成音频与参考音频之间的自然度和声音保真度。为评估模型理解并推理输入语音的能力，我们使用 **URO-Bench**（Yan 等人，2025年），该基准提供基于语音的问答任务，要求具备语义理解能力。除非另有说明，所有实验均在 NVIDIA H200 GPU 上进行。

训练配置。 我们的实现基于 PyTorch 2.7.1。我们采用 AdamW 优化器（Loshchilov 和 Hutter，2017），学习率为 5×10^{-5}，每台设备的批次大小为 4。模型在 8 块 NVIDIA H200 GPU（每块显存 141GB）上训练 100K 步，约 6 小时后收敛。我们采用最大范数为 1.0 的梯度裁剪，以确保训练稳定性。

评估指标。 我们采用客观和主观指标进行全面评估，涵盖语音质量、自然度、说话人相似度和系统效率。客观评估采用说话人相似度（SIM）、首令牌时间（TTFT）和实时因子（RTF）。主观评估则通过人工听众进行比较平均意见分（CMOS）测试。我们的 SIM 评估基于 SEED-TTS-EVAL 框架[6]。

说话人相似度（SIM）通过衡量生成语音与目标说话人的声音特征有多接近，来评估语音克隆质量。我们使用 WavLM-Large（Chen 等人，2022a），这是一种针对说话人验证任务进行微调的自监督语音表征模型（Chen 等人，2022b），从生成音频和参考音频中提取 192 维说话人嵌入。相似度通过这些嵌入之间的余弦相似度计算。

比较平均意见分（CMOS）通过成对比较来衡量自然度（NCMOS）和说话人相似度（SCMOS）。在 NCMOS 测试中，参与者比较两个系统，并从四个选项中选择：“A 听起来更自然”“B 听起来更自然”“差不多”或“难以判断”。在 SCMOS 测试中，参与者先听参考音频，再比较两个合成样本中哪一个与参考音频中的说话人特征更匹配

[6] https://github.com/BytedanceSpeech/seed-tts-eval

模型	SIM ↑
人类基线	0.73
F5-TTS	0.64
Seed-TTS	0.76
FireRedTTS-2	0.66
Step-Audio-TTS	0.66
CosyVoice 3	0.72
Chroma 1.0	**0.81**[7]

表 1：语音模型在零样本语音克隆任务上的性能对比。SIM 越高，说明说话人相似度越高。

并使用相同的四选项量表。样本 A 和 B 的顺序随机排列，以避免位置偏差。CMOS 分数按平均偏好差值计算，正值表示参与者偏好我们的系统。

延迟指标。 我们使用两个互补指标来衡量系统效率：首个令牌生成时间（TTFT）和实时因子（RTF）。TTFT 通过记录从接收输入到生成首个音频令牌所用的时间来衡量系统响应速度，这对保持自然的对话节奏至关重要。RTF 则以生成时间与音频时长之比衡量计算效率。RTF 低于 1.0 表示系统生成语音的速度快于实时播放。评估实时交互性能离不开这两个指标，数值越低，效率越高。

4.2 客观语音克隆评估

我们通过测量 SIM 来评估 Chroma 的声音克隆能力，这是本研究的主要关注点。鉴于 Chroma 目前仅生成英语音频，我们遵循 Seed-TTS（Anastassiou 等人，2024年）制定的方案，使用 Common-Voice 数据集中的英语样本，在零样本设置下评估其性能。我们的模型以其原生采样率 24kHz 进行评估。表 1展示了与当前最先进的语音模型及人类基线的对比结果。

如表 1所示，大多数当代文本转语音模型的说话人相似度分数相近，通常比人类基线低 1–9%。一个显著的例外是 Seed-TTS，其表现略高于人类基线。另一方面，Chroma 的表现优于所有竞品模型，相较于人类基线提升了 10.96%。这一结果表明，Chroma 能够有效捕捉细粒度的副语言特征，从而生成高保真、个性化的语音，并出色地保持说话人身份。

[7]Chroma以24kHz的采样率运行，与其他模型使用的16kHz相比，这一采样率能更好地保留说话人特征。

指标	Chroma	ElevenLabs	Deuce
NCMOS	24.4%	**57.2%**	18.3%
SCMOS	40.6%	42.4%	17.0%

表2：Chroma 与 ElevenLabs 的对比评估。

偏好	参考音频	ElevenLabs
总计	4 (8.0%)	**46 (92.0%)**

表3：ElevenLabs 与参考音频（真实标注）的比较。

4.3 主体语音克隆评估

我们对 Chroma 和 ElevenLabs 进行了比较实验。ElevenLabs 是最先进的商业语音克隆系统。实验测量了 NCMOS 和 SCMOS。我们使用其 eleven_multilingual_v2 API[8] 生成输出。对于每个维度，我们选取了 15 个样本，并使用两个模型分别生成对应输出，共得到 30 个比较样本，在 12 轮独立评估中进行评估。“大致相同”或“难以判断”的回答被归入“平局（Deuce）”类别，表示无法明确判定偏好的情况。

表 2展示了对比结果。在 NC-MOS 方面，ElevenLabs 表现更胜一筹，获得了 57.2% 的偏好，Chroma 获得 24.4%（Deuce 为 18.3%），这表明其语音听起来明显更加自然。然而，在 SCMOS 评估中，参与者会先听参考音频，再比较合成输出，结果却非常接近：ElevenLabs 获得 42.4%，Chroma 获得 40.6%（Deuce 为 17.0%）。双方几乎打成平手，仅相差 1.8 个百分点，这表明两者在捕捉说话人特征方面的能力相当。

这一现象反映了系统设计上的根本差异。ElevenLabs 采用两阶

[8]https://elevenlabs.io/docs/overview/models

组件	首个令牌生成时间（毫秒）	每帧平均延迟（毫秒）	总时长（秒）
推理器	119.12	26.03	3.74
骨干网络	8.48	8.75	4.27
解码器	19.27	17.56	8.57
编解码器解码器	–	3.08	2.99
整体生成延迟	**146.87**	52.34	**16.58**
生成音频长度	–	–	38.80
生成实时因子		**0.43**	

表4： 语音生成的延迟细分。系统的首个令牌生成时间为 **146.87ms**，实时因子为 **0.43** .

段方法：首先从参考音频创建语音配置，然后将其用于文本转语音生成。相比之下，Chroma 采用端到端方法，一次性直接处理参考音频。ElevenLabs 优化了自然度和清晰度，但其两阶段流水线在语音配置提取过程中可能会丢失说话人的细微特征。Chroma 的端到端架构在整个生成过程中始终能够直接访问参考音频，因此保留了这些细腻的特征。

为探究这种权衡，我们开展了另一项实验，将 ElevenLabs 的输出直接与参考音频进行比较。表3 展示了 5 轮实验的结果，每轮包含 10 个样本，评估者需选择听起来更自然、更像真人的音频。值得注意的是，评估者压倒性地偏好 ElevenLabs生成的音频（92.0%），而非真实的人类录音（8.0%）。这揭示了一个关键见解：听众的主观偏好不一定与说话人相似度一致。

这一发现重新诠释了我们对 SCMOS 的理解。虽然 ElevenLabs 的 SCMOS 偏好略高（42.4% 对 40.6%），但这种优势很可能反映了听众对自然度的固有偏好，而非更高的说话人保真度。合成音频受到的压倒性偏好（92%）远高于真实录音，这表明在主观评估中，自然度比说话人相似度更具决定性。因此，Chroma 的 SCMOS 表现具有竞争力，说明其说话人保真度可能比原始百分比所显示的更强，因为尽管 Chroma 优先忠实复现说话人特征——包括自然的不完美之处、语速变化和细微的韵律特征——而非 ElevenLabs 所优化的感知自然度，仍然保持了竞争力。

4.4 实际生成延迟

当前的 Chroma架构不支持批处理，因此我们测量了其延迟——并发数为 1 时的延迟。表4 展示了响应生成过程中各个组件的延迟细分。总生成延迟是所有单个组件延迟之和。我们通过一个实际示例展示测得的延迟：生成一段时长为 38.80 秒的音频响应。

预填充策略。在生成语音之前，我们会对提示文本和提示音频进行预填充，以缩短首个令牌生成时间（TTFT）。具体而言，我们将提示文本和提示音频编码并拼接，得到提示嵌入，然后将其输入骨干网络以执行预填充计算并生成相应的键值缓存（Pope等人，2023年）。通过在生成阶段开始前构建键值缓存，模型无需在推理期间重新处理提示内容，因此接收到用户输入后即可立即进行自回归生成。这一策略有效降低了响应延迟，并显著提升了实时交互性能。

组件级延迟分析。推理器的首个令牌生成时间（TTFT，119.12 毫秒）表示从处理输入数据到生成首个音频令牌所需的时间。骨干网络的首个令牌生成时间（TTFT，8.48 毫秒）表示骨干网络接收到推理器生成的音频令牌后，生成首个音频隐藏状态所需的时间。随后，解码器会生成剩余的 7 个码本（$c^{1:7}$），以捕获精细的声学特征，平均每帧耗时 17.56 毫秒。最后，编解码器解码器根据完整的码本序列重建波形。为提高批处理效率，我们会先将每 4 帧拼接起来，再输入编解码器解码器，因此此组件不适用首个令牌生成时间（TTFT）。

总体系统的首个令牌生成时间（TTFT）为 146.87ms，展现出低于一秒的响应能力，适用于实时交互。平均每帧延迟为 52.34ms，实时因子（RTF）为 0.43，系统生成语音的速度显著快于实时播放，从而实现流畅

模型	大语言模型规模	理解			推理			口语对话				总体
		重复↑	摘要↑	高考↑	Storal↑	真实性↑	Gsm8k↑	MLC↑	Alpaca↑	常见问题↑	开放域↑	
GLM-4-Voice	9B	90.95	91.07	64.47	73.80	59.28	30.93	57.82	80.77	63.07	78.76	69.09
LLaMA-Omni	8B	45.62	80.68	16.06	50.65	45.13	3.89	44.44	64.36	58.40	72.19	48.14
Freeze-Omni	7B	70.89	78.87	26.29	57.74	46.95	2.81	42.56	52.23	48.70	55.80	48.28
Mini-Omni	0.5B	5.07	32.20	0	23.25	25.06	0	2.82	30.99	29.80	31.42	18.06
Mini-Omni2	0.5B	8.10	40.06	0.66	28.49	26.92	0	6.97	34.81	30.70	36.43	21.31
SLAM-Omni	0.5B	12.26	66.21	1.32	36.95	34.65	0	21.85	48.98	41.03	52.61	31.59
Chroma（我们的方法）	**4B**	**69.05**	**74.12**	**38.61**	**71.14**	**51.69**	**22.74**	**60.26**	**60.47**	**62.07**	**64.24**	**57.44**

表 5：端到端语音对话模型在理解、推理和口语对话能力方面的任务完成得分。粗体数值表示 Chroma 的表现；尽管仅使用 40 亿参数，它在各个维度上仍具备竞争力。

流式生成。

实时因子。 生成实时因子通过将总生成延迟除以生成音频的时长来计算：

$$\text{RTF} = \frac{T_{\text{generation}}}{T_{\text{audio}}} = \frac{16.58\text{s}}{38.80\text{s}} \approx 0.43. \quad (1)$$

实时因子（RTF）为 0.43，表明 Chroma 生成语音的速度比实时播放快 2.3×，展现出强大的流式应用性能。即使在持续进行多轮对话时，这种高效率也能让系统保持低延迟。

4.5 推理与对话能力

虽然 Chroma 的主要侧重点是高保真语音克隆，但我们也评估了它的一般对话能力，以证明个性化语音生成不会损害其认知和对话能力。表 5使用 **URO-Bench** (Yan 等人，2025年）的基础赛道，从三个维度将 Chroma 与现有端到端语音对话模型进行了比较。

尽管针对语音克隆进行了优化，Chroma 在所有能力方面都展现出出色的性能。在**推理任务**中，Chroma 始终取得第二优的表现：在 Storal 上为 71.14%（GLM-4-Voice 为 73.80%），在 TruthfulQA 上为 51.69%（GLM-4-Voice 为 59.28%），在 GSM8K 上为 22.74%（GLM-4-Voice 为 30.93%）。值得注意的是，GLM-4-Voice 使用了 9B 参数，规模超过 Chroma 的 4B 的两倍。在**口语对话**中，Chroma 在 MLC（60.26%）和 Common-Voice（62.07%）上取得最高分，展现出自然的对话流畅性。在**理解任务**中，Chroma 在复述和摘要方面分别取得 69.05% 和 74.12% 的竞争力分数，在所有模型中排名第二。

更重要的是，Chroma 是此次比较中唯一具备个性化语音克隆能力的模型。其他所有模型都专注于对话和推理，不具备克隆特定说话人特征的能力。这使得 Chroma 的竞争力表现尤为值得关注：它在保持强大认知能力和对话能力的同时，还支持高保真度的声音个性化，而这一能力在所有参与比较的系统中都不存在。此外，Chroma 仅使用 4B 参数，较大型模型（7B-9B）具有效率优势，并且在所有评测任务中，相较于较小模型（0.5B）都取得了更优的性能。

5 结论

我们推出 Chroma 1.0，这是首个兼具低延迟交互与高保真个性化语音克隆能力的开源实时端到端语音对话模型。Chroma 完全使用合成语音到语音数据进行训练，仅凭数秒参考音频便显著提升了说话人相似度，同时保持实时性能。对比评估表明，在语音克隆方面，Chroma 的表现可与最先进的商业系统相媲美，同时具备出色的推理和对话能力。该系统高效的架构支持流畅的流式生成，适用于自然的多轮对话。通过结合低延迟流式处理与高保真语音克隆，Chroma 为各类领域中易于使用的个性化语音 AI 应用开辟了新的可能。我们将发布代码和模型，以推动个性化语音对话系统的进一步研究。

6 致谢

我们感谢 Yusen Lin、Kaiwen Zhou 和 Liujie Zhen 在本项目早期阶段作出的贡献。我们也感谢参与人工评估实验的志愿者。

参考音频

Philip Anastassiou、Jiawei Chen、Jitong Chen、Yuanzhe Chen、Zhuo Chen、Ziyi Chen、Jian Cong、Lelai Deng、Chuang Ding、Lu Gao 等。2024年。Seed-TTS：一系列高质量、多用途的语音生成模型。*arXiv* 预印本 *arXiv:2406.02430*。

Junyi Ao、Rui Wang、Long Zhou、Chengyi Wang、Shuo Ren、Yu Wu、Shujie Liu、Tom Ko、Qing Li、Yu Zhang 等。2022年。SpeechT5：用于口语处理的统一多模态编码器-解码器预训练。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第5723–5738页。

Rosana Ardila、Megan Branson、Kelly Davis、Michael Kohler、Josh Meyer、Michael Henretty、Reuben Morais、Lindsay Saunders、Francis Tyers 和 Gregor Weber。2020年。Common Voice：一个大规模多语言语音语料库。载于第十二届语言资源与评测会议论文集，第4218–4222页。

Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao 等 1 位作者。2022a。Wavlm：面向全栈语音处理的大规模自监督预训练。*IEEE*《信号处理精选主题期刊》，16(6):1505–1518。

Zhengyang Chen、Sanyuan Chen、Yu Wu、Yao Qian、Chengyi Wang、Shujie Liu、Yanmin Qian 和 Michael Zeng。2022b。面向自动说话人验证的大规模自监督语音表征学习。载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第6147–6151 页。IEEE。

Yunfei Chu、Jin Xu、Qian Yang、Haojie Wei、Xipin Wei、Zhifang Guo、Yichong Leng、Yuanjun Lv、Jinzheng He、Junyang Lin 等 1 位作者。2024年。Qwen2-audio 技术报告。*arXiv* 预印本 *arXiv:2407.10759*。

Yunfei Chu、JinXu、XiaohuanZhou、Qian Yang、Shil- iang Zhang、Zhijie Yan、Chang Zhou 和 Jingren Zhou。2023年。Qwen-audio：通过统一的大规模音频语言模型推进通用音频理解。*arXiv* 预印本 *arXiv:2311.07919*。

Raffel Colin。2020年。探索使用统一的文本到文本转换器进行迁移学习的极限。《机器学习研究杂志》，21。

Alexandre Défossez、Jade Copet、Gabriel Synnaeve 和 Yossi Adi。2022年。高保真神经音频压缩。*arXiv* 预印本 *arXiv:2210.13438*。

Alexandre Défossez、Laurent Mazaré、Manu Orsini、Amélie Royer、Patrick Pérez、Hervé Jégou、Edouard Grave 和 Neil Zeghidour。2024年。Moshi：用于实时对话的语音-文本基础模型。*arXiv* 预印本 *arXiv:2410.00037*。Ding Ding、Zeqian Ju、Yichong Leng、Songxiang Liu、Tong Liu、Zeyu Shang、Kai Shen、Wei Song、Xu Tan、Heyi Tang 等人。2025年。Kimi-Audio 技术报告。*arXiv* 预印本 *arXiv:2504.18425*。Zhihao Du、Qian Chen、Shiliang Zhang、Kai Hu、Heng Lu、Yexin Yang、Hangrui Hu、Siqi Zheng、Yue Gu、Ziyang Ma 等人。2024a。CosyVoice：一种基于监督语义令牌的可扩展多语言零样本文本转语音合成器。*arXiv* 预印本 *arXiv:2407.05407*。

Zhihao Du、Changfeng Gao、Yuxuan Wang、Fan Yu、Tianyu Zhao、Hao Wang、Xiang Lv、Hui Wang、Chongjia Ni、Xian Shi 等共 12 人。2025年。Cosyvoice 3：通过扩规模和后训练，实现真实场景中的语音生成。*arXiv* 预印本 *arXiv:2505.17589*。

Zhihao Du、Yuxuan Wang、Qian Chen、Xian Shi、Xiang Lv、Tianyu Zhao、Zhifu Gao、Yexin Yang、Changfeng Gao、Hui Wang 等共 12 人。2024b。Cosyvoice 2：基于大型语言模型的可扩展流式语音合成。*arXiv* 预印本 *arXiv:2412.10117*。

Paul-Ambroise Duquenne、H Elsahar、H Gong、K Hef- fernan、J Hoffman、C Klaiber 等。2023年。SeamlessM4T——大规模多语言、多模态机器翻译。

Qingkai Fang、Shoutao Guo、Yan Zhou、Zhengrui Ma、Shaolei Zhang 和 Yang Feng。2024年。LLaMA-Omni：借助大型语言模型实现无缝语音交互。*arXiv* 预印本 *arXiv:2409.06666*。

Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdel- rahman Mohamed。2021年。Hubert：通过对隐藏单元进行掩码预测实现自监督语音表征学习。*IEEE/ACM* 音频、语音与语言处理汇刊，29：3451–3460。

Ailin Huang、Bingxin Li、Bruce Wang、Boyong Wu、Chao Yan、Chengli Feng、Heng Wang、Hongyu Zhou、Hongyuan Wang、Jingbei Li 等。2025a。Step-audio-aqaa：一个完全端到端、富有表现力的大型音频语言模型。*arXiv* 预印本 *arXiv:2506.08967*。

Ailin Huang、Boyong Wu、Bruce Wang、Chao Yan、Chen Hu、Chengli Feng、Fei Tian、Feiyu Shen、Jingbei Li、Mingrui Chen 等 1 位作者。2025b。Step-Audio：智能语音交互中的统一理解与生成。*arXiv* 预印本 *arXiv:2502.11946*。

Rongjie Huang、Mingze Li、Dongchao Yang、Jia- tong Shi、Xuankai Chang、Zhenhui Ye、Yuning Wu、Zhiqing Hong、Jiawei Huang、Jinglin Liu，以及另外 1 位作者。2024年。Audiogpt：理解与生成

语音、音乐、声音和说话头像。发表于 人工智能促进协会人工智能会议论文集，第 38 卷，第 23802–23804 页。

Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、AdityaRamesh、AidanClark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford及另外1位作者。2024年。GPT-4o 系统卡。*arXiv* 预印本 *arXiv:2410.21276*。

Shengpeng Ji、Ziyue Jiang、Wen Wang、Yifu Chen、Minghui Fang、Jialong Zuo、Qian Yang、Xize Cheng、Zehan Wang、Ruiqi Li及另外1位作者。2024年。Wav-tokenizer：一种用于音频语言建模的高效声学离散编解码器分词器。*arXiv* 预印本 *arXiv:2408.16532*。

Zeqian Ju、Yuancheng Wang、Kai Shen、Xu Tan、Detai Xin、Dongchao Yang、Yanqing Liu、Yichong Leng、Kaitao Song、Siliang Tang及另外1位作者。2024年。NaturalSpeech 3：利用因子化编解码器和扩散模型实现零样本语音合成。*arXiv* 预印本 *arXiv:2403.03100*。

Matthew Le、Apoorv Vyas、Bowen Shi、Brian Karrer、Leda Sari、Rashel Moritz、Mary Williamson、Vimal Manohar、Yossi Adi、Jay Mahadeokar 等另外 1 位作者。2023年。Voicebox：大规模文本引导的多语言通用语音生成。神经信息处理系统进展，36：14005–14034。

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, O mer Levy, Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。载于计算语言学协会第*58*届年会论文集，第 7871–7880 页。Yinghao Aaron Li、Cong Han、Vinay Raghavan、Gavin Mischler 和 Nima Mesgarani。2023年。StyleTTS 2：通过风格扩散和与大型语音语言模型进行对抗训练，实现接近人类水平的文本转语音。神经信息处理系统进展，36：19594–19621。Guan-Ting Lin、Cheng-Han Chiang 和 Hung-yi Lee。2024年。推动大型语言模型掌握多样的说话风格，并在口语对话中作出恰当回应。*arXiv* 预印本 *arXiv:2402.12786*。Ilya Loshchilov 和 Frank Hutter。2017年。解耦权重衰减正则化。*arXiv* 预印本 *arXiv:1711.05101*。Tu Anh Nguyen、Benjamin Muller、Bokai Yu、Marta R Costa-Jussa、Maha Elbayad、Sravya Pop-uri、Christophe Ropers、Paul-Ambroise Duquenne、Robin Algayres、Ruslan Mavlyutov 等 1 人。2025年。Spirit-LM：交错式口语与书面语语言模型。计算语言学协会会刊，13：30–52。Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等 1 人。2022年。通过人类反馈训练语言模型遵循指令。神经信息处理系统进展，35：27730–27744。

Reiner Pope、Sholto Douglas、Aakanksha Chowdhery、Jacob Devlin、James Bradbury、JonathanHeek、Kefan Xiao、Shivani Agrawal 和 Jeff Dean。2023年。高效扩展 Transformer 推理。机器学习与系统会议论文集，5：606–624。

Zengyi Qin、Wenliang Zhao、Xumin Yu 和 Xin Sun。2023年。OpenVoice：多功能即时语音克隆。*arXiv* 预印本 *arXiv:2312.01479*。

Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。2023年。通过大规模弱监督实现鲁棒语音识别。发表于国际机器学习大会，第28492–28518页。PMLR。

Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher D Manning、Stefano Ermon 和 Chelsea Finn。2023年。直接偏好优化：你的语言模型暗中就是一个奖励模型。神经信息处理系统进展，36:53728–53741。

Xu Tan、Jiawei Chen、Haohe Liu、Jian Cong、Chen Zhang、Yanqing Liu、Xi Wang、Yichong Leng、Yuanhao Yi、Lei He 等。2024年。NaturalSpeech：实现人类级质量的端到端文本转语音合成。*IEEE*模式分析与机器智能汇刊，46(6):4234–4245。

Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar及另外1位作者。2023年。LLaMA：开放且高效的基础语言模型。*arXiv* 预印本 *arXiv:2302.13971*。Chengyi Wang、Sanyuan Chen、Yu Wu、Ziqiang Zhang、Long Zhou、Shujie Liu、Zhuo Chen、Yanqing Liu、Huaming Wang、Jinyu Li及另外1位作者。2023年。神经编解码器语言模型是零样本文本转语音合成器。*arXiv* 预印本 *arXiv:2301.02111*。Boyong Wu、Chao Yan、Chen Hu、Cheng Yi、Chengli Feng、Fei Tian、Feiyu Shen、Gang Yu、Haoyang Zhang、Jingbei Li及另外1位作者。2025年。Step-Audio 2 技术报告。*arXiv* 预印本 *arXiv:2507.16632*。Zhifei Xie和Changqiao Wu。2024a。Mini-Omni：语言模型能够在流式思考的同时听取和进行对话。*arXiv* 预印本 *arXiv:2408.16725*。Zhifei Xie和Changqiao Wu。2024b。Mini-Omni2：迈向具备视觉、语音和双工能力的开源GPT-4o。*arXiv* 预印本 *arXiv:2410.11190*。Jin Xu、Zhifang Guo、Jinzheng He、Hangrui Hu、Ting He、Shuai Bai、Keqin Chen、Jialin Wang、Yang Fan、Kai Dang及另外1位作者。2025a。Qwen2.5-Omni 技术报告。*arXiv* 预印本 *arXiv:2503.20215*。

Jin Xu、Zhifang Guo、Hangrui Hu、Yunfei Chu、Xiong Wang、Jinzheng He、Yuxuan Wang、Xian Shi、Ting He、Xinfa Zhu、Yuanjun Lv、Yongqi Wang、Dake Guo、He Wang、Linhan Ma、Pei Zhang、Xinyu Zhang、Hongkun Hao、Zishan Guo及另外19位作者。2025b。Qwen3-Omni 技术报告。预印本，arXiv:2509.17765。

Ruiqi Yan、Xiquan Li、Wenxi Chen、Zhikang Niu、Chen Yang、Ziyang Ma、Kai Yu 和 Xie Chen。2025年。Uro-bench：端到端语音对话模型的综合基准测试。*arXiv* 预印本 *arXiv:2502.17810*。

Jiahui Yu、Yuanzhong Xu、Jing Yu Koh、Thang Lu-ong、Gunjan Baid、Zirui Wang、Vijay Vasudevan、Alexander Ku、Yinfei Yang、Burcu Karagol Ayan 和其他 1 位作者。2022年。扩展自回归模型，实现内容丰富的文本到图像生成。*arXiv* 预印本 *arXiv:2206.10789*，2(3):5。

Aohan Zeng、Zhengxiao Du、Mingdao Liu、Kedong Wang、Shengmin Jiang、Lei Zhao、Yuxiao Dong 和 Jie Tang。2024年。Glm-4-voice：迈向智能、拟人化的端到端语音聊天机器人。*arXiv* 预印本 *arXiv:2412.02612*。

Dong Zhang、Shimin Li、Xin Zhang、Jun Zhan、Pengyu Wang、Yaqian Zhou 和 Xipeng Qiu。2023a。Speechgpt：赋予大型语言模型内在的跨模态对话能力。*arXiv* 预印本 *arXiv:2305.11000*。

Ziqiang Zhang、Long Zhou、Chengyi Wang、Sanyuan Chen、Yu Wu、Shujie Liu、Zhuo Chen、Yanqing Liu、Huaming Wang、Jinyu Li 和另外 1 位作者。2023b。用自己的声音说外语：跨语言神经编解码器语言建模。*arXiv* 预印本 *arXiv:2303.03926*。

附录

A 局限性与未来工作

当前系统未集成外部工具使用，也未采用特定任务的后训练技术，例如基于人类反馈的强化学习（RLHF）（Ouyang etal.，2022年）或直接偏好优化（DPO）（Rafailov et al.，2023）。集成此类方法或可进一步提升对话质量、指令遵循能力和用户偏好对齐，尤其能改善对话自然度和上下文适切性。

近期关于多码本令牌预测（MTP）的研究，例如 Qwen3-Omni（Xu 等，2025b），通过并行而非顺序预测残差码本，降低了首包延迟。研究如何在不损害语音克隆保真度的前提下，将 MTP 集成到 Chroma 的解码器阶段，是进一步降低延迟的一个有前景的方向。

尽管 Chroma 的语音推理器支持多语言输入（目前支持中文和英语），但该系统仅生成英语语音输出。扩展编解码器训练和解码器模块，以支持多语言输出生成，将扩大该系统在多样语言环境中的适用范围。跨语言语音克隆，即在输入语言与输出语言不同的情况下保留说话人身份，仍是一个重要的研究方向。

尽管 Chroma 的骨干网络采用了仅含解码器的架构，符合近期语音语言建模的发展趋势，但编码器-解码器架构已在多个领域展现出强劲的性能，包括机器翻译（Colin，2020年）、文本生成（Lewis 等人，2020年）、文本到图像生成（Yu 等人，2022年）以及口语处理（Ao 等人，2022年）。这些架构可能在可控性、跨模态对齐以及理解与生成过程的明确分离方面具有独特优势。探索用于语音到语音对话的编码器-解码器设计，或可为我们当前的方法带来互补益处，尤其适用于需要对语音克隆和流式生成中的语义内容与声学属性进行精细控制的场景。

B 伦理考量

虽然 Chroma 在语音克隆技术方面取得了重大进展，但它也引发了——

重要的伦理考量。只需少量参考音频即可生成高度逼真且个性化的语音，这种能力带来了若干风险，包括冒充他人、欺诈，以及未经个人同意创建误导性或有害音频内容的可能性。

为降低这些风险，我们建议采取有力的技术和基于政策的保障措施，例如：

- 要求任何形式的语音克隆都必须获得明确且可核实的同意，
- 开发并部署可靠的合成语音检测机制，
- 制定明确的使用政策和访问控制措施，以防止滥用，
- 探索为生成音频添加水印或采用可追溯技术。

如果负责任地开发和使用，像 Chroma 这样的语音克隆系统可以带来切实益处，例如为有言语障碍的人士提供语音修复支持、推动无障碍应用的发展，以及促进创意型或个性化通信工具的应用。我们鼓励研究界在语音合成系统不断进步的同时，并行推进伦理准则、监管框架和可靠的检测技术，以确保这些系统得到安全、负责任的部署。

C 训练目标

在训练期间，推理器被冻结，仅作为特征提供器使用。对于每个由音频样本 $\mathbf{x}_{\text{audio}}$ 及其对应转录 $\mathbf{x}_{\text{text}}$组成的训练对，推理器生成文本嵌入 $\mathbf{E}_{\text{text}} \in \mathbb{R}^{T\times d}$ 和多模态隐藏状态 $\mathbf{H}_{\text{reasoner}} \in \mathbb{R}^{T\times d}$，其中 T表示文本序列的长度，d 表示隐藏维度。这些表征在整个优化过程中保持固定，并作为下游声学模型的语义和韵律条件。

在参考音频 $\mathbf{x}_{\text{ref-audio}}$ 和参考文本 $\mathbf{x}_{\text{ref-text}}$的条件下，骨干网络以自回归方式预测粗粒度离散声学编码序列

$$\mathbf{c}^0 = \{c_t^0 \mid t = 1, \ldots, L\}, \qquad c_t^0 \in \{1, \ldots, V\},$$

其中，L 表示音频帧的数量，V表示码本大小。

Chroma 解码器通过生成 RVQ 表示中剩余的 $N-1$ 个残差量化级别，进一步细化每一帧。对于每个级别 $j \in \{1, \ldots, N-1\}$，解码器会预测

$$\mathbf{c}^j = \{c_t^j \mid t = 1, \ldots, L\}, \qquad c_t^j \in \{1, \ldots, V\}.$$

我们将完整的 N 级离散声学表示定义为

$$c_t^{0:N-1} = (c_t^0, c_t^1, \ldots, c_t^{N-1}).$$

主干网络损失。为保持文本与音频生成之间的因果对齐，骨干网络在时刻 t 只能关注前缀信息。我们定义因果条件集合

$$\mathbf{z}_{<t} = \big(c_{<t}^0,\ \mathbf{E}_{\text{text},<t},\ \mathbf{H}_{\text{reasoner},<t}\big),$$

其中，$c_{<t}^0$ 表示自回归生成的声学前缀，而 $\mathbf{E}_{\text{text},<t}$、$\mathbf{H}_{\text{reasoner},<t}$ 表示由推理器提供的前缀信号。

骨干网络预测粗粒度编码序列$\mathbf{c}^0 = \{c_t^0\}_{t=1}^L$，其损失定义为

$$\mathcal{L}_{\text{backbone}} = -\frac{1}{L}\sum_{t=1}^{L} \log p\big(c_t^0 \mid \mathbf{z}_{<t}\big). \qquad (2)$$

该目标训练骨干网络捕获帧间时序结构，并生成与文本进展对齐的粗粒度声学表征。

解码器损失。给定粗编码 c_t^0和骨干网络隐藏状态$\mathbf{h}_t^B$，Chroma 解码器通过帧内 自回归过程细化每个帧 t。具体而言，它预测残差量化级别

$$c_t^{1:N-1} = (c_t^1, c_t^2, \ldots, c_t^{N-1}), \qquad c_t^j \in \{1, \ldots, V\}.$$

细化级别上的条件分布分解为

$$p_\theta\Big(c_t^{1:N-1} \mid c_t^0,\, \mathbf{h}_t^B\Big) = \prod_{j=1}^{N-1} p_\theta\Big(c_t^j \mid c_t^0,\, \mathbf{h}_t^B,\, c_t^{1:j-1}\Big), \qquad (3)$$

其中，$c_t^{1:j-1}$ =(c_t^1，.....，c_t^{j-1})表示用于教师强制的前缀。

Chroma 解码器的训练目标是所有帧和细化级别上的负对数似然：

$$\mathcal{L}_{\text{decoder}} = -\frac{1}{L}\sum_{t=1}^{L}\sum_{j=1}^{N-1} \log p_\theta\Big(c_t^j \mid c_t^0,\, \mathbf{h}_t^B,\, c_t^{1:j-1}\Big). \qquad (4)$$

这一目标鼓励解码器逐步细化每一帧的声学表示，同时与骨干网络的粗编码 c_t^0 以及编码在$\mathbf{h}_t^D$中的上下文信息保持一致。

C.1 训练策略

我们采用两阶段训练策略来稳定优化，并逐步提升声学表征的精细程度。

在第一阶段，我们通过将损失权重设为 $\lambda = 0.5$，联合训练骨干网络和解码器。这种均衡的权重设置促使模型同时学习粗粒度声学编码分布和残差量化级别，帮助系统在训练早期建立一致的语义-声学对齐。

在第二阶段，我们冻结骨干网络参数并设置 $\lambda = 1$，从而将优化完全转移至解码器。此微调阶段专注于优化更高层级的量化层，使模型能够捕捉细粒度的语音特征，例如细微的音色差异、韵律变化和发音细节。因此，最终模型实现了更高的语音克隆保真度，并提升了整体语音自然度。