arXiv:2611.33117v1 [cs.GR] 24 Nov 2026

Read My Lips：基于上下文的唇读与语言和视听模型

Amir Razlighi[1,2,3] Zhuo Ning[1] Naomi Ken Korem[4] Nisarg Patel[5]
Yiorgos Chrysanthou[2,3] Daniel Cohen-Or[6] Ali Mahdavi Amiri[1]

[1]西蒙弗雷泽大学 [2]塞浦路斯大学 [3]CYENS卓越中心 [4]Lightricks [5]MARZ VFX [6]特拉维夫大学

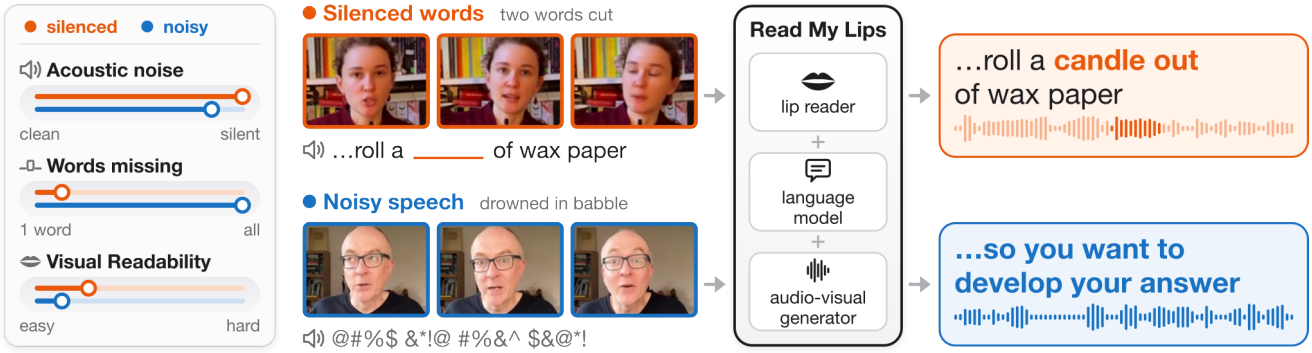

图1。 **Read My Lips 通过结合唇读器、语言模型和视听生成器，** 还原音频中的各种损坏。 左： 输入内容各不相同，包括音频的损坏程度、受影响（缺失）的词数，以及嘴唇的可读难度。 右： 两个具有不同损坏情况的示例。当音频中有两个词被剪辑片段截去（橙色）时，Read My Lips 会将其补全；当整个片段被多人嘈杂声淹没（蓝色）时，它会根据嘴唇动作还原句子。

摘要

现实世界中的语音通常只能被部分观测：音频可能缺失，或受到噪声和竞争说话者的干扰而损坏，而说话人的嘴唇仍清晰可见。恢复语音需要结合能够听到、看到以及从上下文中推断出的信息。我们提出 *RML*（ReadMy Lips），一种结合唇读器、语言模型和视听生成模型的协同方法。唇读器提供发音动作假设，语言模型利用上下文解决含糊不清或缺失的词语，视听生成器则在保留说话者身份的同时重建同步语音。由于源可靠性各不相同，*RML* 会生成多个重建结果，并利用可用证据 *favor* 最可靠的结果。这使 *RML* 能够在音频缺失、强干扰和视觉证据不足的情况下运行，而无需手动设定阈值。在这些条件下开展的定量和定性实验表明，结合视觉、声学和语言信号能够带来显著益处。

1. 引言

想象一下，你正在拥挤的招待会上努力听清一段对话。几个人同时说话，音乐和背景噪声淹没了部分语音，但你面前的人仍清晰可见。即使他们的声音变得难以听清，他们的唇部运动以及对话的上下文仍能为其所说内容提供线索。类似情况还会出现在音频中断的视频通话、在嘈杂环境中录制的录音、声音受损的档案影像，以及需要修复受损语音的电影或内容制作中。在

在所有这些情况下，要还原语音，就需要利用音频、视频和周围对话中仍然可用的各种信息。

更普遍地说，现实世界中的语音往往只能被部分观测到。背景噪声或竞争说话者可能会掩盖音频；录音或传输故障可能导致词语消失；说话人的嘴部也可能被遮挡，或从难以辨认的视角呈现。我们从三个维度对这些挑战进行刻画，如图 1 所示： 声学噪声，从干净语音到严重噪声或竞争说话者； 缺失词语，从孤立的词到

话语的大段内容缺失；以及视觉可读性，从清晰的正面人脸到难以辨认的视角、遮挡，或动画和生成的说话人。这些维度可以彼此独立地变化，从而形成没有任何单一模态始终可靠的条件。

在这些条件下恢复语音，需要结合三种互补的证据来源。视觉发音动作提供有关语音的信息，但其本质上存在歧义，因为不同音素可能产生相似的唇部运动。声学证据即使受到严重干扰，也能提供有关音素、时间和说话者身份的线索。最后，语言上下文可以利用周围可靠的语音消除歧义。例如，在“我午餐点了一份——”这句话中，上下文对缺失的词有很强的限制作用。然而，单靠上下文无法可靠地推断姓名、数字或意外表达。因此，有效的语音修复必须结合所见、所闻和所推断的信息。

现有模型针对这些信息来源提供了互补能力。唇读模型能够捕捉细粒度的视觉发音动作，但难以应对含糊不清的视素、遮挡和困难的视角。大语言模型利用语言上下文消除歧义，但可能会偏向合理的词，而非实际发出的内容。音视频生成模型能够合成同步且与说话人特征一致的语音，但并非专为精准唇读而设计。它们各自互补的优势表明，与其依赖任何单一模型，不如将它们结合起来。

我们介绍 RML（*Read My Lips*），这是一个用于视听语音修复的协作框架。给定一段视频和可能受损的音频，唇读器会对发音内容提出假设，语言模型会根据前后语音推断哪些内容更合理，而视听生成模型则结合这些预测、视频和剩余的声学证据，重建最终语音。它们的相对重要性会随输入自然变化：音频缺失时，视觉证据占主导；发音动作含糊不清时，上下文有助于消除歧义；剩余音频则约束发音、时序和说话者身份。

因此，RML 将语音修复建模为一个多模态推理问题（见图1）：唇读器估计发音者似乎在说什么，语言模型则推断在上下文中什么内容合乎情理，而音视频模型确定恢复的语音应当听起来如何，以及如何与视频对齐。为适应不同的输入条件，RML 使用不同的组件和音频条件生成多个重建结果，根据观测到的音频和唇部运动评估这些结果获得的支持，并通过基于证据加权的最小贝叶斯风险方法将其结合起来，从而优先选择最可靠的重建结果。我们在不同程度的声学损坏、语音缺失和视觉可读性条件下评估 RML，展示了结合这些互补信息来源的益处。

2. 背景：音素与视素

音素是口语中区分词语的基本声音单位。例如，/p/、/b/ 和 /m/ 是不同的音素。视素是它们在视觉上的对应形式：一个人说话时呈现的嘴部和唇形。然而，音素与视素之间并非一一对应关系。不同的声音可能产生几乎相同的嘴部运动。例如，/p/、/b/ 和 /m/ 都涉及闭唇，因此*pat*、*bat*和*mat* 这样的词很难仅凭视觉区分。即使嘴部清晰可见，这种歧义也让唇读充满挑战。

额外信息有助于消解这些歧义。语言上下文会限制句子中可能出现的词，而即使是微弱的声学线索，也可能区分嘴唇外观相同的声音。这解释了为什么结合视觉、声学和语言证据，可能比单独依赖唇读更有效。

3. 相关工作

视觉语音识别。唇读通过音视频自监督预训练 [50] 以及在大规模自动标注或伪标注语料库 [22, 23, 36]上的训练取得了进展。如今，英语以外的其他语言也有了原生唇读器 [35]。近期的一项研究方向是将视觉特征连接到大语言模型。一些方法将视觉语音单元或特征直接输入 LLM [4, 57]；另一些方法则预测音素，再由 LLM 将其组合成句子 [54]。这些模型利用了 LLM 的语言知识。然而，当它们作为唇读器使用时，只能看到视频，而忽略了围绕着所需读取词语、仍然可以听到的语音。我们将唇读器作为一个黑箱，使其返回一个 N最佳列表，并在下一阶段引入可听上下文。

用于识别错误纠正的语言模型。 生成式错误纠正训练 LLM 将语音识别器的 N 最佳列表映射为纠正后的转录文本 [7]。音频-视觉变体基于唇部特征 [19]对这一纠正过程进行条件化， N或者让 LLM 结合独立的音频和视觉 N最佳列表 [27]。我们的语言模型也将唇读器的假设与音频的转录文本相结合，但它会根据静音片段前后的语音以及该片段的时长，判断静音片段中说了什么。

视频到语音合成。视频到语音方法从无声视频生成语音。早期工作为每位说话人训练一个模型 [43]；后来的方法可在真实环境中使用，通过中间语音单元 [12]，或根据音视频表征进行重新合成 [25]。近来有几种系统以唇读器预测的文本为条件。它们将文本输入

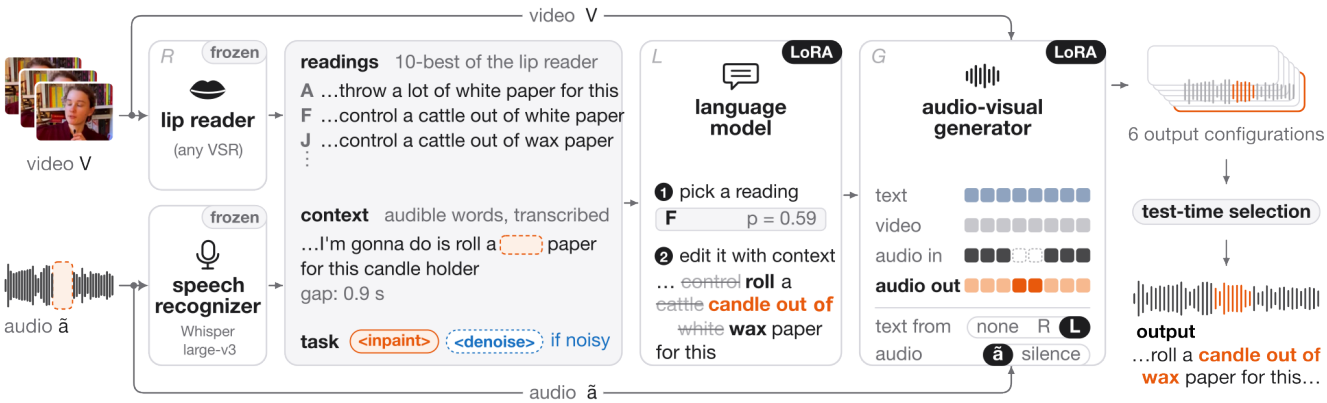

图 2. **Read My Lips 概述**在一个包含静音词的真实测试片段上。冻结的唇读器 R 返回整个片段的 N个最佳识读结果，而冻结的语音识别器则转录可听上下文（如果存在）。语言模型 L 选出最接近的识读结果，并对其进行编辑以符合上下文。视听生成器 G 接收视频、音轨和作为干净上下文词元的文本，并生成语音。切换文本来源（无、 R或 L）以及音轨（$\tilde{a}$ 或静音），即可得到 G的六种可能配置；测试时选择（第4.3节）会选出最佳配置。

视觉文本转语音模型 [24], , 利用它通过语音识别器引导扩散 [56], , 或将其与视频以及说话人的声音或面部参考 [13, 33]共同纳入注意力机制。研究还表明，在信噪比极低时，唇动到语音生成的效果优于音视频增强 [47]。所有这些方法都会为整个片段生成语音。它们都不是为保留受损录音中仍完好的原始语音而设计的，也无法通过推断其中的词语来恢复缺失内容。

语音修复与编辑。视听语音修复可填补说话人面部录音中最长 1.6 秒的间隙，无需转录文本 [38, 39]。文本引导的修复与编辑模型会根据转录文本重新生成录音中标记的区间，同时保留说话人的语音特征 [3, 32, 41, 55]。零样本文本转语音模型可根据一段简短提示克隆语音 [5, 11]。这些模型需要以正确的词作为输入，而且看不到嘴唇。因此，它们既无法还原说过的内容，也无法使重新生成的语音与视频对齐。我们的处理流程从嘴唇和上下文中获取词语，并由生成器在说出这些词语时关注视频。

视听语音增强。视听增强与分离利用目标说话人的唇部运动和面部信息，从噪声和其他说话人的干扰中分离出其语音 [1, 16, 18, 45]。生成式变体会重新合成语音，而不是对其进行滤波。它们使用神经声码器 [37] 或扩散模型 [14], , 基于声码器的方法还能在低信噪比条件下进行增强。仅音频增强器 [49] 以及还会使用转录文本的修复模型 [28] 都无法利用视频。在我们的实验中，基于提示的分离模型 [51] 几乎无法改变严重受损的语音。增强方法旨在恢复仍然存在于干扰中的语音。我们的生成器也从受损音轨开始，但还会以转录文本为条件；当可听证据消失时，转录文本承载着相应的内容。

音视频联合生成。 如今，开放的基础模型可在同一个扩散变换器中联合生成视频和音频，并分别处理视频流和音频流 [21]。Just-Dub-It [6] 基于相同的骨干网络，重新生成嘴唇动作和语音，从而将视频配音为另一种语言。我们的任务与此互补：视频保持固定并作为关注对象，只生成与之匹配的语音。

4. 方法

给定一段说话人视频 V 及其受损音轨 $\tilde{a}$，目标是恢复说话内容和语音本身。我们将损坏表示为

$$\tilde{a} \;=\; (1 - \mathbf{1}_\Omega) \odot a \;+\; \mathbf{1}_\Omega \odot (\gamma\, a + n)\,, \qquad (1)$$

其中， a 是干净语音，Ω 是由指示函数 $\mathbf{1}_\Omega$ 标记的受损时间区间集合， n 是干扰信号。静音（$\gamma = 0$， $n = 0$）会移除 Ω 内的语音。噪声（$\gamma = 1$）则会在其上叠加干扰，例如多人嘈杂声、竞争说话人或环境声音，且信噪比未知。Ω 可以涵盖单个词、多个词或整个片段。输出包括转录文本 $\hat{}y$ 和与 V 同步、并保留说话人声音的语音信号 $\hat{a}$。这些视频均来自真实场景，因此头部姿态、光照条件以及嘴唇的清晰可见程度各不相同。

4.1. 概述

RML 通过三个互补模型融合视觉、语言和声学信息（图 2）。唇读器 R



图 3. **测试时选择** 在一个含多人嘈杂声的测试片段上，噪声为 −20 分贝。每张卡片对应一种配置（行：提供给 G的文本；列：起始音频），并显示其转录文本（红色：错误词语）、证据 e （公式（2））和风险（公式（3））。$\hat{y}_L$ 仅进行投票。在此情况下，所有保留含噪音频的配置都失败了，因此选择过程将其关闭。

（视觉模态）生成一个包含 N 个最佳转录文本假设的列表 $\mathcal{H} = (h_1, \ldots\ldots, h_N)$。语言模型 L（上下文模态）将这些假设与 Whisper 的转录文本结合起来 c，其中，通过简单的能量阈值定位的静音区间被标记为间隙。最后，视听生成器 G（视听模态）根据 V、$\tilde{a}$ 和恢复的文本这一条件，重建语音。模型之间通过纯文本进行通信，因此每个模型都可以替换；尤其是，R 可以替换为任何视觉语音识别器，而无需重新训练 L 或 G。我们首先介绍 RML 的各个组件，然后引入一个自适应选择过程，根据可用音频和视觉证据的可靠性，确定如何使用这些组件。

4.2. RML 框架

唇读器作为视觉模态。当音频缺失或受损时，唇读器可提供互补的视觉证据，帮助判断说了什么。我们对英语使用 Auto-AVSR [36]，对西班牙语和法语使用 Ma 等人的 CMU-MOSEAS 识别器。[35] 识别器的选择并不重要——换用 USR 20 [23] 而无需重新训练其他任何组件，也能获得相同的提升（第 5.5节）。我们使用预训练视觉语音识别器，该识别器以 V 的嘴部区域为输入，并通过束搜索返回排名前 N 的转录文本。我们处理完整片段，而非仅处理 Ω，使唇读器能够利用周围的视觉线索，并在后续将其假设与可听语音转录文本对齐。我们保留排名前 N 的候选列表，为语言模型提供备选视觉假设，从而提高正确词语出现在候选项中的概率。

上下文推理（语言模型）。语言模型（Gemma-4 12B [52]）结合视觉假设与可听上下文，以恢复转录文本。我们通过 Whisper large-v3 [44], 转录 $\tilde{a}$ 来获取上下文 c，并用占位符标记检测到的静音区间，同时提供其时长；对于噪声输入，转录文本中也可能包含错误或缺失的词语。模型接收可听上下文 c、任务标签（检测到静音间隙时为 <inpaint>，否则为 <denoise>）以及候选转录文本列表。对于静默输入，候选项是唇读器排名前 N 的假设 $\mathcal{H}$。对于音频仍包含信息的含噪输入，我们将$\mathcal{H}$ 与视听 Auto-AVSR 模型 [36]（或任何视听识别器）排名前 N 的假设交错排列，并加入 Whisper 对含噪音频生成的转录文本，同时删除重复项。模型首先选出最可能的候选项，然后利用周围的上下文对其进行修正，生成最终转录文本。我们使用已知词语时间戳的静默和含噪片段对模型进行微调，并以最佳候选项和参考转录文本作为监督信号。为避免依赖特定唇读器，训练假设从不同架构和质量的识别器中采样，包括劣化视频变体；假设数量则在一到 N 之间均匀采样；我们始终使用N=10 。

视听模态与语音重建。虽然可以使用任何视听生成模型，但我们采用 LTX-2。5 [21]，因为它是开源且公开可用的。LTX-2.5 是一种扩散变换器，通过双向注意力和文本条件控制联合建模视频和音频潜变量。我们使用双画布低秩适配器 [2]对生成器进行适配 V，并仅用它来生成语音： V 的视频潜变量提供视觉线索，而音频流则包含输入轨道，先是作为干净上下文，随后是待生成的目标音频。任务标签 <修复> 或<降噪>，后面还可选择附上转录文本，以提供文本条件控制。这些信号共同使生成器能够让语音与唇部运动对齐，保留说话人的声音和可听语音，并重建静音或被掩码的内容。生成语音与 V同步。更多训练和实现细节请参见附录第B节。

4.3. 自适应组件选择

每种模型的实用性取决于输入：有可听语音时，音频更有优势；缺失词语时，语境能带来帮助；噪声很大时，视觉证据更有优势；而嘴唇不可见时，唇读器就不可靠。我们不手动挑选这些情况，而是运行多个具有不同组件和音频条件的配置，并选择最受可用证据支持的输出。具体而言，每个配置 θ 都是为生成器 G 选择两种输入：文本和音轨。文本可以是

省略、唇读器的最优假设 h_1，或语言模型的转录文本 $\hat{}y_L$。音轨可以是受损输入 $\tilde{a}$，这样 G便可以利用可听语音；也可以是静音，这样 G便会忽略音轨。我们根据生成文本的组件为配置命名： G(无文本)、 $R \to G(h_1)$，以及 $R \to L \to G(\hat{}y_L)$，每种配置均可选择是否使用 $\tilde{a}$。组合这两种选择后，共有六种配置（图 3）。整段片段静音时， $\tilde{a}$ 本身也处于静音状态，因此只剩下三种文本选项。每种配置都会生成一个候选语音信号 $\hat{a}_\theta$ 及其转录文本 y_θ。

基于证据的评分。 为选择最佳配置，我们需要确定哪个候选项与可用证据最为一致。测试时的挑战在于，我们没有真实标签可供比较或据此进行选择。因此，我们为每个候选转录文本（y）定义一个证据分数（e（y）），用于衡量其获得实际音频和唇部运动支持的程度，同时消除识别器对其自身看来更有可能出现的句子的先验偏好。音频和视觉证据分数（e_A（y）和e_R（y））分别计算，然后合并：

$$\begin{aligned} e_A(y) &= \log p_A(y \mid \tilde{a}) - \log p_A(y \mid \varnothing_a), \\ e_R(y) &= \log p_R(y \mid V) - \log p_R(y \mid \bar{V}), \\ e(y) &= e_A(y) + e_R(y), \end{aligned} \tag{2}$$

其中， $\varnothing_a$ 表示静音， V 重复第一帧口部画面，分别提供没有声学证据或视觉证据时的基线。减去这些基线分数，即可分离实际音频和唇部运动对e_A（y）和e_R（y）的贡献，而二者之和e（y）则给出对y的总体证据。因此，当某种模态缺乏信息时，例如音频受到严重干扰或嘴唇不可见，其证据分数就接近于零，对最终选择的影响也很小。**最终选择。** 我们使用最小贝叶斯风险 [17, 20, 29]来选择输出。投票者为各配置生成的转录文本以及$\hat{}y_L$，但不包括在没有音频的情况下由 h_1生成的输出（$R \to G$）；$\hat{}y_L$ 已将其纳入考虑，因为 L 会读取整个 N最佳列表：

$$\theta^\star = \arg\min_{\theta \in \Theta} \sum_{y' \in \mathcal{Y}} e^{\frac{e(y')}{T}}\, d(y_\theta, y'), \tag{3}$$

其中， $\mathcal{Y}$是投票者集合， d 是双向词错误率（WER），T 是在验证集上设定的。$\hat{}y_L$ 参与投票，但无法被选中。选定的配置同时提供$\hat{a}_{\theta^\star}$ 和$\hat{}y = y_{\theta^\star}$。

选择会根据可用音频和视频的可靠性进行调整。当音频受到严重干扰时，忽略音频的配置更有可能被选中（见图 3）。当音频干净时，则更倾向于选择使用音频的配置。证据分数还会降低那些缺乏音频或唇部运动支持的预测所产生的影响，使系统无需使用固定规则即可选择最可靠的信息。

表1. **测试集。**每个4.8秒窗口通过改变缺失跨度长度 k生成多个静音示例，并生成五个采用不同噪声条件的含噪示例。未能通过干净音频上的自动语音识别验证的静音示例将被舍弃。

Set	视频 时间窗		k	静音 含噪	
TalkVid（英语）	23	114	1–10	875	570
VoxMM（英语）	28	159	1–10	1,217	795
困难（英语）	23	44	1, 3, 5	116	220
TalkVid（西班牙语、法语）	76	431	1–4	401	–

5. 实验

我们首先介绍实验配置，然后评估我们的方法及其各个组件的影响。

数据。 我们在 TalkVid 的一个小型子集上进行训练 [9]: 使用 6,509 个时长为 4.8 秒的时间窗（英语、西班牙语和法语）训练生成器，并使用 6,792 个英语时间窗训练语言模型，总计不足 1,430 个完整视频和 12 小时语音。我们在未见过的 TalkVid 视频、域外 VoxMM [30, 31], 西班牙语和法语视频，以及一个包含侧面视角、遮挡、动画角色和生成的说话人的挑战性集合上进行测试（表1）。我们通过移除 1–10 个连续词语，或用多人嘈杂声、音效或另一位说话人的语音遮盖完整音频，来扰乱每个时间窗。超参数在 99 个验证视频上调优，并在所有测试集上保持固定；完整细节见附录第A 节。

指标。 我们根据输出的可听内容进行评估。我们使用 Whisper large-v 转写生成音频3 [44] ，并将其与参考内容进行比较。对于静音语音，我们报告精确片段恢复和词准确率，仅对 Whisper 在损坏前能够正确转写的样例进行评分。对于噪声语音，我们报告窗口级词错误率。我们使用源视频上的 95% 配对自助法置信区间来确定统计显著性。置信区间见附录第D.1节；附录第D.6节提供了使用第二个识别器进行的检查；附录第E节评估生成语音的自然度、说话者身份和唇部同步。

基线方法。据我们所知，现有方法中没有一种能结合唇读、语言上下文和残留音频来修复静音语音与噪声语音，并针对每个输入采用最可靠的信息来源。因此，我们针对这两项任务，使用词语恢复准确率与最接近的基线方法进行比较。对于静音语音，我们将方法与唇读/视听识别器（Auto-AVSR [36], USR2.0 [23], Llama- AVSR [4]）、获得缺失词语及其位置的语音编辑器（F5-TTS [11], VoiceCraft [41], AlignDiT [13]）以及 LipVoicer [56]进行比较。对于噪声语音，我们将方法与音频修复方法（DeepFilterNet3 [49], Sidon [40]）和视听分离方法（VisualVoice [18],SAM Audio [51]）进行比较。

表2. **恢复被静音的语音。** 精确片段恢复和词准确率（Acc.），单位为%。文本是方法从唇读器（R）或我们的语言模型（L）接收的转录文本；区间表示已告知区间位置的方法。

方法	Text	区间	精确恢复 TalkVid	精确恢复 VoxMM	精确恢复 合并	Acc. 合并
文本输出						
Auto-AVSR [36]	–	–	51.0	16.1	30.7	59.7
USR 2.0 [23]	–	–	53.5	23.0	35.8	64.9
Llama-AVSR [4]	–	–	47.7	18.0	30.4	58.5
语音输出						
LipVoicer [56]	own	–	43.4	17.3	28.3	58.6
VoiceCraft [41]	L	✓	51.8	18.3	32.3	59.8
F5-TTS [11]	L	✓	57.4	20.2	35.8	62.2
AlignDiT [13]	L	✓	56.3	20.9	35.7	62.8
我们的方法（将 Auto-AVSR 用作 R）	L/R	–	<u>60.3</u>	<u>24.2</u>	<u>39.3</u>	<u>65.7</u>
我们的方法（将 USR 2.0 用作 R）	L/R	–	**64.0**	**29.3**	**43.8**	**69.2**

5.1. 恢复被静音的语音

表2 显示，无论使用 Auto-AVSR 还是更强的 USR 2.0 作为唇读器，我们的方法在所有词语准确率和片段恢复指标上都取得了最佳结果。使用 Auto-AVSR 时，该方法准确恢复了39.3%的静音片段，而 Auto-AVSR 单独使用时为30.7%；这一结果已与更强的 USR 2.0 持平，甚至更胜一筹。使用 USR 2.0 后，准确恢复率升至43.8%，比单独使用 USR 2.0 高出 8 个百分点。值得注意的是，F5-TTS 和 AlignDiT 等语音编辑器既获得了我们语言模型生成的转录文本，也获得了片段位置，而我们的生成器必须自行推断片段位置，但表现仍然更好。这些提升表明，我们的方法优于唇读器，并能从联合利用可见唇部和周围可听语音中获益。

5.2. 恢复噪声与重叠语音

表3 报告了完整 4.8 秒语音窗口上的词错误率（WER），该数值是不同噪声条件下的平均值；各测试集的详细结果见附录第D.3节。平均而言，我们的处理流程将词错误率从69.0%（噪声输入）降至25.5%，优于所有基线方法。仅音频方法在强噪声下表现不佳，而仅视频方法则会丢弃有用的语音信息。音频噪声较小且仍能提供大部分信息时，USR 2.0 (AV) 等视听识别器能够从中受益：在 -5 分贝多人嘈杂声条件下，USR 2.0 (AV) 的表现最接近我们的方法，因为相对较轻的噪声仍使音频能够揭示大部分语音内容。随着损坏加剧，这一优势便会消失；在 -10 分贝多人嘈杂声条件下，我们的方法显著优于所有视听识别器。在竞争说话人条件下，专为说话人分离而构建的模型 VisualVoice 表现与我们的方法不相上下。

这些结果表明，对更可靠的信号赋予更高权重是有益的，这也促使我们采用下文所述的自适应选择方法。第 5.6节进一步展示了每个信号对最终性能的贡献。

表3。**恢复含噪语音。**在 TalkVid 和 VoxMM 上合并统计的完整 4.8 秒时间窗的词错误率（%，越低越好）。其他说话人：音量为 -5 分贝的竞争说话人。文本输出直接进行评分。语音输出首先进行转录。V：仅嘴唇；AV：嘴唇和含噪音频。 †仅统计预处理成功的时间窗。

方法	多人嘈杂声 -10	音效 -5	音效 已见说话人	音效 未见说话人	其他	Avg
噪声输入	91.8	52.9	47.7	48.0	104.8	69.0
文本输出						
Auto-AVSR [36] (V)	49.4	49.4	49.4	49.4	49.4	49.4
Auto-AVSR [36] (AV)	82.1	47.7	68.7	61.0	93.6	70.6
USR 2.0 [23] (V)	47.4	47.4	47.4	47.4	47.4	47.4
USR 2.0 [23] (AV)	62.4	28.8	35.4	29.0	104.3	52.0
Llama-AVSR [4](V)	58.8	58.8	58.8	58.8	58.8	58.8
Llama-AVSR [4](AV)	57.8	34.3	30.7	26.0	90.4	47.9
语音输出						
DeepFilterNet3 [49]	92.5	70.6	51.6	43.3	102.8	72.2
Sidon [40]	99.4	93.0	61.5	64.8	103.1	84.4
VisualVoice [18]†	80.5	41.8	28.7	29.1	**20.7**	40.1
LipVoicer [56]	52.9	52.9	52.9	52.9	52.9	52.9
AlignDiT [13](R)†	50.5	50.5	50.5	50.5	50.5	50.5
我们的方法 (Auto-AVSR as R)	44.0	28.1	22.0	<u>19.9</u>	22.1	27.2
我们的方法（USR 2.0 作为 R）	**40.1**	**26.6**	**20.6**	**19.0**	<u>21.1</u>	**25.5**

5.3. 适应音频损坏

不同类型和程度的损坏会使不同信号变得更加可靠，但在测试时，我们并不知道哪种配置最适合给定输入。因此，我们的方法会考虑全部六种配置，并利用它们的证据分数选择最可靠的输出。为检验这一点，我们使用相同的测试窗口，并改变损坏类型及其持续时间：多人嘈杂声以 -5、-10 或 -20 分贝的强度，或静音，影响一个词、四个词或整个片段。我们也纳入干净片段。这样一来，共有 1,421 个 TalkVid 样本和 1,998 个 VoxMM 样本，并在整个窗口上使用词错误率进行评估。

图4（底部）展示了随着音频变得不那么可靠，我们的方法如何改变其选择。对于干净片段，它几乎总是保留音频。随着损坏变得更加严重并覆盖更多片段，它会越来越多地选择用静音替换受损音频的配置。例如，当整个片段都受到-20分贝多人嘈杂声（右图）损坏时，它会移除88%片段中的音频，此时语音是根据说话人的外貌推断出来的，而不是取自录音（附录第E节）。重要的是，方法并未获得噪声级别或受损区域的信息；它依据证据分数作出这一决定。音频证据从 $\sim$ 80干净片段中的 $\sim$ 3 降至 -20分贝时的水平，从而更加重视转录文本和视觉证据。

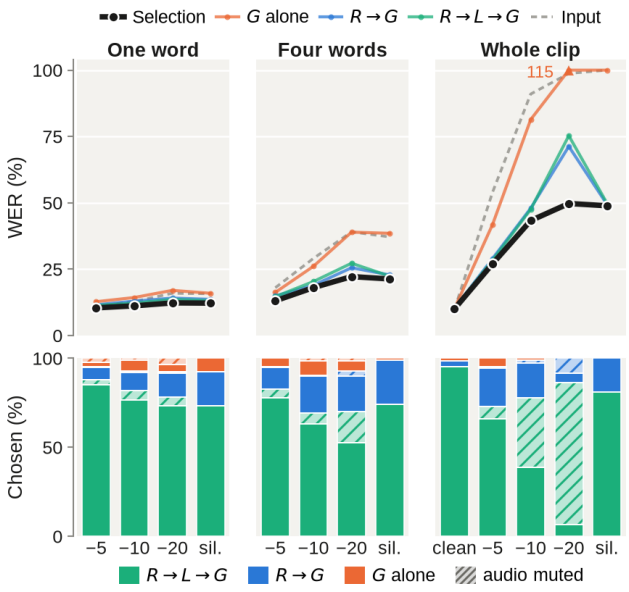

图4。**适应输入**。不同损坏级别和持续时间下的词错误率。上图：黑线表示我们的自适应选择，其性能与最佳固定配置相当或更优，而没有任何固定配置能在所有场景中都取得最佳表现。下图：随着损坏情况变化，所选配置也随之变化；条纹表示输入音频被静音替代。

图4（顶部）展示了这一选择的优势。黑色曲线表示我们的自适应选择方法，它在不同损坏条件下均达到或优于最佳固定配置。没有任何单一固定配置能在所有损坏级别下都取得最佳效果。噪声较轻时，单独使用生成器效果良好；但在严重损坏下会变得不可靠，当整个片段在 -20分贝下受到损坏时，词错误率甚至超过100%。使用转录文本有所帮助，但保留受损音频的配置仍会受到噪声影响。我们根据音频的可靠性选择保留或移除音频，在这一困难情况下将词错误率降至49。8%。在所有条件下，自适应选择达到23。1%的词错误率，相比之下，始终使用 $R \to G$时为26。8%，使用 $R \to L \to G$时为40。8%，单独使用生成器时为49。1%，而唇读器为该

5.4. 难以辨读的嘴唇

在这里，我们通过检测微弱的视觉证据并转向更可靠的信息源（例如语言和音频），测试唇读不可靠时我们的方法能否保持稳健。我们从 23 个场景（例如电影、动画和生成的视频）中收集了 44 个英语窗口；由于侧面视角、遮挡或动画角色的影响，这些窗口中的嘴部难以或无法读取。在这些窗口中，我们将 $k = 1$、3 和 5 个词的区间静音（116 个已验证片段），并

表4。**难以辨读的嘴唇**。在116个静音片段上的精确片段恢复（%），以及在220个噪声样本上的词错误率（%）。标记为 * 的方法仅在其人脸预处理成功的情况下进行评分（附录第D.1节中的匹配比较）。

方法	静音 ↑	含噪 ↓
受损输入	–	54.1
Auto-AVSR [36] （V/ AV）	8.6	72.5*
USR 2.0 [23]	12.0*	56.3*
Llama-AVSR [4]	13.3*	60.3*
SAM Audio [51],文本提示	–	53.7
VisualVoice [18]	–	41.6*
LipVoicer [56]	10.8*	92.8*
F5-TTS [11](L)	23.3	–
AlignDiT [13](L/ R)	23.6*	97.0*
我们的方法（Auto-AVSR 作为 R）	**25.9**	**29.6**

采用表 3 中的噪声条件（220 个噪声样本），并在两项任务上与基线方法进行比较（表 4）。

我们的方法在两项任务上的得分都是最高的，恢复的静默片段数量是其自身唇读器的三倍。当嘴唇无法提供可靠信息时，它会更多地依赖其他组件：语言模型根据上下文预测缺失词语，生成器则用说话人的声音将这些词语说出来。这也解释了为何使用我们的方法的语言模型生成的转录文本的语音编辑器，在静默语音任务上的表现最接近我们的方法。重要的是，我们的方法不会获知唇读何时失效；语言模型和剩余的音频会弥补嘴唇无法呈现的信息。唇读基线方法在静默语音任务中仅使用嘴唇（V），在含噪语音任务中则同时使用嘴唇和音频（AV），但表现仍差得多。一些基线方法（标记为 *）需要检测到人脸；若嘴部始终不可见，它们便会失效，因此我们只在这些方法能够处理的时间窗上为其评分。

5.5 泛化

由于各组件通过文本进行交互，R 可以替换而无需重新训练。使用 USR 2.0 作为 R，并采用相同的选择（表2 和3），我们的方法在两项任务上的表现均有所提升：更强的唇读器能直接带来更强的处理流程，而 L和 G还能进一步提升性能。有趣的是，即使使用 Auto-AVSR 作为 R，我们的方法整体上仍位居第二，仅次于使用 USR 2.0 的版本。这表明性能提升源自框架本身，并且在不同唇读器上都能保持显著效果。不同语言下也有相同的表现（表5b）：使用西班牙语或法语唇读器作为 R，且 L 和 G保持不变时，该处理流程的精确恢复效果比唇读器提升了三分之一。

5.6 消融实验

组件。表。 5 对我们的处理流程中的组件进行消融实验，以展示每个模型的贡献。首先，将唇读器的 N个最佳假设提供给语

言模型，并结合可听上下文，得到的转录文本远胜于仅使用唇读器的结果（$R \rightarrow L$相比于 R）。其次，生成器 G 不仅仅是渲染器。如果 G 只是像文本转语音模型一样朗读输入文本，那么它所看到的嘴唇就无关紧要。相反，在转录文本相同但唇部动作不匹配的情况下，其表现会显著变差（附录第D.5节）。第三，语言模型对于较长的间隙最为重要：没有它（$R \rightarrow G$），处理流程会在所有长度下丢失区间；对于缺失七到十个词的情况，完整流水线比唇读器多恢复42%的区间。

没有一种固定配置在所有情况下都表现最佳。语言模型有助于处理静音语音，但对于含噪语音，音频仍承载着大部分词语，而 $R \rightarrow G$在不使用它时表现更好。我们的选择机制无法判断当前属于哪种情况，但在含噪语音上取得了最低的错误率，在静音语音上也接近最佳表现。唯一的例外是静音困难集合，在该集合上差异并不显著。在这里，嘴唇无法被读懂，因此选择机制所依据的视觉证据并不可靠，有时还会导致它选择遵循唇读器错误预测的输出。

供语言模型使用的唇部特征。 人们可能会认为，如果语言模型能直接看到嘴唇，而不是通过唇读器的假设来获取信息，它的识别效果会更好。我们用一个小型预测头对此进行测试，该预测头读取 AV-HuBERT [50]的嘴部特征，并将其视素后验概率传递给 L 作为软输入标记，与 L联合训练。模型确实依赖该预测头，因为使用另一个片段的特征会使验证集上的精确恢复率降低 22.8 个百分点；然而，该预测头在任何测试集上都没有带来提升，并且在唇部特征最不可靠的区间上表现更差（表 5c）。因此，排序后的假设包含了 L 所需的视觉信息。

选择规则。 为检验选择规则（式（3））中的每一项是否都必不可少，我们每次移除一项。每一项至少在一种设置下有帮助，而移除任何一项都不会在任何设置下带来显著改善（附录第C节）。

剩余错误。 给定真实转录文本，G在 TalkVid 和 VoxMM 上恢复了 96.6% 的静音片段，优于任何使用相同文本的语音编辑器。在噪声条件下，它将词错误率大致减半，使其接近干净录音的水平；困难集也呈现出相同的规律。这表明，大多数剩余错误源于转录文本，而非生成器，因此唇读器和语言模型是需要改进的主要组件。

6. 结论

我们介绍了 Read My Lips，这是一种通过结合唇读、语言理解和视听生成来恢复语音的方法。这项工作揭示了语言与视听模型近期进展所带来的一个引人注目的机遇：我们现在可以结合这些模型互补的能力，恢复

(*a*) 组件

	静默 ↑		含噪 ↓	
	合并	Hard	合并	Hard
R (文本)	30.7	8.6	49.4	91.4
$R \rightarrow L$(文本)	36.2	22.4	42.8	74.5
G 无文本	–	–	45.2	37.6
$R \rightarrow G$	33.1	14.7	<u>29.0</u>	<u>30.6</u>
$R \rightarrow L \rightarrow G$	<u>38.6</u>	**27.6**	29.5	35.2
选择（我们的方法）	**39.3**	<u>25.9</u>	**27.2**	**29.6**

(b) 西班牙语和法语，精确片段恢复（%）↑

	西班牙语	法语	合并
R = *Ma*等人。 *[35]*			
R （文本）	34.0	17.5	30.7
$R \rightarrow L$(文本)	<u>36.8</u>	<u>21.2</u>	<u>33.7</u>
$R \rightarrow L \rightarrow G$	**43.9**	**28.8**	**40.9**

(c) 的输入，精确片段恢复率（%）

	TalkVid	VoxMM	Hard	es, fr
N最佳列表（我们的方法）	60.5	22.9	27.6	40.9
N最佳列表 +唇部特征	60.6	22.9	25.9	38.4

表5。**消融实验。**（a）我们的处理流程以 Auto-AVSR 作为 R 时的每种配置：在静音片段上的精确片段恢复（%），以及在噪声窗口上的词错误率（%）；结果汇总了 TalkVid 和 VoxMM，以及困难集上的数据。（b）西班牙语和法语：仅替换 R，使用 Ma 等人的唇读器。 [35]； L 和 G 保持不变。（c） $R \rightarrow L \rightarrow G$ 在语言模型中有无唇部特征头的情况。

单个模型可能遗漏的信息。仅凭嘴唇难以辨认的词，可能会因对话内容而变得清晰；被噪声淹没的语音，也可能通过说话人的嘴部运动得以揭示。这类似于人类感知，会根据具体情境着重关注最清晰的感官线索。综合这些线索，原本难以理解的录音也能重新变得清晰。

要使这种协作发挥作用，必须知道何时应当信任每个来源。更多信息并不总是更好：严重损坏的音频可能会误导模型，而忽略音频反而可能改善恢复效果。同样，语言模型可以生成一句令人信服的句子，却无法确定这句话是否真的被说出；当嘴部画面不清晰或被遮挡时，唇读也会变得不可靠。挑战在于，推理时并不知道对于给定输入，这些来源中的哪一个是可靠的。因此，必须根据可用证据判断它们的可靠性。这为结合语言模型推理语音的能力与音频和唇部运动中的证据提供了机会：当某个来源能够提供有用信息时依赖它，在其信息不足时则减少依赖。

尚未解决的问题同样令人振奋。一旦词语正确，生成器就能

以很高的准确度重建语音。随着语音生成能力日益增强，恢复正确词语成为核心挑战。强大的语言与视听模型的出现，使这一方向正值探索时机：通过结合能够听到、看到和推断出的信息，我们究竟能在多大程度上恢复所说的内容?

参考文献

[1] Triantafyllos Afouras、Joon Son Chung 和 Andrew Zisserman。《The conversation：深度视听语音增强》。arXiv 预印本 arXiv:1804.04121，2018。3[2] Matan Ben-Yosef、Tavi Halperin、Naomi Ken Korem、Mohammad Salama、Harel Cain、Asaf Joseph、Anthony Chen、Urska Jelercic 和 Ofir Bibi。《Avcontrol：用于训练视听控制的高效框架》。arXiv 预印本 arXiv:2603.24793，2026。4[3] Zalan Borsos、Matthew Sharifi 和 Marco Tagliasacchi。《SpeechPainter：文本条件语音修复》。载于 Interspeech 2022，第431–435页，2022。3[4]Umberto Cappellazzo、Minsu Kim、Honglie Chen、Pingchuan Ma、Stavros Petridis、Daniele Falavigna、Alessio Brutti 和 Maja Pantic。《大语言模型是强大的视听语音识别学习器》。载于 ICASSP 2025——2025 IEEE国际声学、语音与信号处理会议（ICASSP），第1–5页，2025。2、5、6、7[5]Edresson Casanova、Kelly Davis、Eren G¨olge、G¨orkem G¨oknar、Iulian Gulea、Logan Hart、Aya Aljafari、Joshua Meyer、Reuben Morais、Samuel Olayemi 和 Julian Weber。《XTTS：大规模多语言零样本文本转语音模型》。载于 Interspeech 2024，第4978–4982页，2024。3[6] Anthony Chen、Naomi Ken Korem、Tavi Halperin、Matan Ben Yosef、Urska Jelercic、Ofir Bibi、Or Patashnik 和 Daniel Cohen-Or。《Just-dub-it：通过联合视听扩散实现视频配音》。载于计算机协会图形学与交互技术特别兴趣组会议论文集，美国纽约州纽约市，2026。计算机协会。3、14[7]CHEN CHEN、Yuchen Hu、Chao-Han Huck Yang、Sabato Marco Siniscalchi、Pin-Yu Chen 和 Eng-Siong Chng。《Hyporadise：基于大语言模型的生成式语音识别开放基线》。载于《神经信息处理系统进展》，第31665–31688页。Curran Associates, Inc.，2023。2[8] Honglie Chen、Weidi Xie、Andrea Vedaldi 和 Andrew Zisserman。《VGGSound：大规模视听数据集》。载于 ICASSP 2020——2020 IEEE国际声学、语音与信号处理会议（ICASSP），第721–725页。IEEE，2020。1[9]Shunian Chen、Hejin Huang、Yexin Liu、Zihan Ye、Pengcheng Chen、Chenghao Zhu、Michael Guan、Rongsheng Wang、Junying Chen、Jianye Hou 等。《TalkVid：用于音频驱动说话头合成的大规模多样化数据集》。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第3492 3500页，2026 5 –。[10] Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen 和 Jiajun Qi。用于说话人验证的局部与全局特征融合增强型 Res2Net。arXiv 预印本 arXiv:2305.12838，2023。14[11] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chun- hui Wang, Jian Zhao, Kai Yu 和 Xie Chen。F5-TTS：一种通过流匹配生成流畅且忠实语音的“说书人”。载于第63届计算语言学协会年会论文集（第1卷：长篇论文），第6255–6271页，奥地利维也纳，2025。计算语言学协会。3, 5, 6, 7, 2, 10, 14 [12]

Jeongsoo Choi、Minsu Kim 和 Yong Man Ro。使用语音单元实现清晰易懂的唇动到语音合成。载于 Interspeech 2023，第4349–4353页，2023。2[13] Jeongsoo Choi、Ji-Hoon Kim、Kim Sung-Bin、Tae-Hyun Oh 和 Joon Son Chung。AlignDiT：用于同步语音生成的多模态对齐扩散 Transformer。载于第33届ACM国际多媒体会议论文集，第10758–10767页，美国纽约州纽约市，2025。计算机协会。3, 5, 6, 7, 2, 10, 14[14] Ju-Chieh Chou、Chung-Ming Chien 和 Karen Livescu。Av2wav：基于连续自监督特征的扩散再合成，用于视听语音增强。载于 ICASSP 2024——2024年 IEEE国际声学、语音与信号处理会议（ICASSP），第10806–10810页，2024。3[15] Joon Son Chung 和 Andrew Zisserman。Out of Time：野外自动唇语同步。载于亚洲计算机视觉会议，第251–263页。Springer，2016。14[16] Ariel Ephrat、Inbar Mosseri、Oran Lang、Tali Dekel、Kevin Wilson、Avinatan Hassidim、William T. Freeman 和 Michael Rubinstein。鸡尾酒会中的听音辨人：用于语音分离的说话人无关视听模型。《ACM图形学汇刊》，37(4)，2018。3[17] Jonathan G Fiscus。用于降低词错误率的后处理系统：识别器输出投票错误率降低（ROVER）。载于1997年IEEE自动语音识别与理解研讨会论文集，第347–354页。IEEE，1997。5[18] Ruohan Gao 和 Kristen Grauman。VisualVoice：基于跨模态一致性的视听语音分离。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第15495–15505页，2021。3, 5, 6, 7, 2, 14 – 。[19] Sreyan Ghosh、Sonal Kumar、Ashish Seth、Purva Chiniya、Utkarsh Tyagi、Ramani Duraiswami 和 Dinesh Manocha。LipGER：用于鲁棒自动语音识别的视觉条件生成式错误纠正。载于 Interspeech 2024，第1920–1924页，2024。2[20] Vaibhava Goel。最小贝叶斯风险自动语音识别。2002。5[21] Yoav HaCohen、Benny Brazowski、Nisan Chiprut、Yaki Bitterman、Andrew Kvochko、Avishai Berkowitz、Daniel Shalem、Daphna Lifschitz、Dudu Moshe、Eitan Porat 等。Ltx-2：高效的联合视听基础模型。arXiv 预印本 arXiv:2601.03233，2026。3, 4[22] Alexandros Haliassos、Rodrigo Mira、Honglie Chen、Zoe Landgraf、Stavros Petridis 和 Maja Pantic。统一语音

识别：一种适用于听觉、视觉和视听输入的单一模型。载于《神经信息处理系统进展》，第139673–139699页。Curran Associates, Inc.，2024。2[23] Alexandros Haliassos、Rodrigo Mira 和 Stavros Petridis。关注连接时序分类（CTC）：用于统一语音识别的快速鲁棒伪标签生成。载于第十四届国际学习表征会议，2026。2, 4, 5, 6, 7[24] Sindhu Hegde、Rudrabha Mukhopadhyay、C.V Jawahar 和 Vinay Namboodiri。迈向野外精准的唇动到语音合成。载于第31届ACM国际多媒体会议论文集，第5523–5531页，美国纽约州纽约市，2023。计算机协会。3[25] Wei-Ning Hsu、Tal Remez、Bowen Shi、Jacob Donley 和 Yossi Adi。Revise：利用视觉输入进行自监督语音重合成，实现通用且泛化的语音再生成。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第18795–18805页，2023。2[26] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen- Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适配。arXiv 预印本 arXiv:2106.09685，2021。3[27] Sungnyun Kim、Kangwook Jang、Sungwoo Cho、Joon Son Chung、Hoi-Rin Kim 和 Se-Young Yun。两个预测头胜过一个：基于双假设的视听语音错误纠正。载于计算语言学协会：ACL 2026 论文集，第544–564页，美国加利福尼亚州圣迭戈，2026。计算语言学协会。2[28] Yuma Koizumi、Heiga Zen、Shigeki Karita、Yifan Ding、Ko- hei Yatabe、Nobuyuki Morioka、Yu Zhang、Wei Han、Ankur Bapna 和 Michiel Bacchiani。Miipher：融合自监督语音与文本表征的鲁棒语音修复模型。载于2023年 IEEE音频与声学信号处理应用研讨会（WASPAA），第1–5页，2023。3[29] Shankar Kumar 和 Bill Byrne。用于统计机器翻译的最小贝叶斯风险解码。载于北美计算语言学分会人类语言技术会议论文集：HLT-NAACL 2004，第169–176页，2004。5[30] Doyeop Kwak、Jaemin Jung、Kihyun Nam、Youngjoon Jang、Jee-Weon Jung、Shinji Watanabe 和 Joon Son Chung。VoxMM：野外对话的丰富转录。载于 ICASSP 2024——2024年 IEEE国际声学、语音与信号处理会议（ICASSP），第12551–12555页，2024。5, 1[31] Doyeop Kwak、Jeongsoo Choi、Suyeon Lee 和 Joon Son Chung。LRS-VoxMM：野外视听语音识别基准。arXiv 预印本 arXiv:2604 27866，2026 5, 1 。 。[32] Matthew Le、Apoorv Vyas、Bowen Shi、Brian Karrer、Leda Sari、Rashel Moritz、Mary Williamson、Vimal Manohar、Yossi Adi、Jay Mahadeokar 和 Wei-Ning Hsu。Voicebox：大规模文本引导的多语言通用语音生成。载于《神经信息处理系统进展》，第14005–14034页。Curran Associates, Inc.，2023。3[33] Gunwoo Lee、Yoori Oh 和 Yoseob Han。《留意你的语音：通过文本条件控制实现文本感知的视频转语音合成》，2026。3[34] Ilya Loshchilov 和 Frank Hutter。《解耦权重衰减正则化》。arXiv预印本 arXiv:1711.05101，2017。3[35] Pingchuan Ma、Stavros Petridis 和 Maja Pantic。《真实环境下的多语言视觉语音识别》。Nature Machine Intelligence，4(11):930–939，2022。2、4、8[36] Pingchuan Ma、Alexandros Haliassos、Adriana Fernandez- Lopez、Honglie Chen、Stavros Petridis 和 Maja Pantic。《Auto-AVSR：带自动标签的视听语音识别》。载于 ICASSP 2023——2023年 IEEE国际声学、语音与信号处理会议（ICASSP），第1–5页。IEEE，2023。2、4、5、6、7[37] Rodrigo Mira、Buye Xu、Jacob Donley、Anurag Kumar、Stavros Petridis、Vamsi Krishna Ithapu 和 Maja Pantic。《La-Voce：使用神经声码器进行低信噪比视听语音增强》。载于 ICASSP 2023——2023年 IEEE国际声学、语音与信号处理会议（ICASSP），第1–5页，2023。3[38] Juan Felipe Montesinos、Daniel Michelsanti、Gloria Haro、Zheng-Hua Tan 和 Jesper Jensen。《语音修复：视频引导的基于上下文的语音合成》。载于 Interspeech 2023，第4459–4463页，2023。3[39] Giovanni Morrone、Daniel Michelsanti、Zheng-Hua Tan 和 Jesper Jensen。《基于深度学习的音视频语音修复》。载于 ICASSP 2021——2021年 IEEE国际声学、语音与信号处理会议（ICASSP），第6653–6657页，2021。3[40] Wataru Nakata、Yuki Saito、Yota Ueda 和 Hiroshi Saruwatari。《Sidon：用于大规模数据集清洗的快速、鲁棒开源多语言语音修复》。载于 ICASSP 2026——2026年 IEEE国际声学、语音与信号处理会议（ICASSP），第17617–17621页。IEEE，2026。5、6、2、14[41] Puyuan Peng、Po-Yao Huang、Shang-Wen Li、Abdelrahman Mohamed 和 David Harwath。《VoiceCraft：真实环境中的零样本语音编辑与文本转语音》。载于第62届计算语言学协会年会论文集（第1卷：长篇论文），第12442–12462页，泰国曼谷，2024。计算语言学协会。3、5、6、2、7、10[42] KR Prajwal、Rudrabha Mukhopadhyay、Vinay P Namboodiri 和 CV Jawahar。《只需一位唇同步专家，即可在真实环境中实现语音到唇动生成》。载于第28届 ACM国际多媒体会议论文集，第484–492页，2020。14[43] K R Prajwal、Rudrabha Mukhopadhyay、Vinay P. Namboodiri 和 C.V. Jawahar。《学习个体说话风格，实现准确的唇动到语音合成》。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2020。2[44] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。《通过大规模弱监督实现鲁棒语音识别》。载于国际机器学习会议，第28492–28518页。PMLR，2023。4、5、1

[45] Akam Rahimi、Triantafyllos Afouras 和 Andrew Zisser- man。《在鸡尾酒会中读唇以聆听：多模态语音分离》。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第10493–10502页，2022。3[46] Chandan KA Reddy、Vishak Gopal 和 Ross Cutler。《DNS-MOS：用于评估噪声抑制器的非侵入式感知客观语音质量指标》。载于 ICASSP 2021——2021年 IEEE国际声学、语音与信号处理会议（ICASSP），第6493–6497页。IEEE，2021。14[47] Julius Richter、Danilo de Oliveira、Tal Peer 和 Timo Gerk- mann。《LipDiffuser：利用条件扩散模型实现唇动到语音生成》，2025。3[48] Takaaki Saeki、Detai Xin、Wataru Nakata、Tomoki Ko- riyama、Shinnosuke Takamichi 和 Hiroshi Saruwatari。《UT-MOS：东京大学SaruLab的VoiceMOS挑战赛2022系统》。arXiv预印本 arXiv:2204.02152，2022。14[49]Hendrik Schr¨oter、Alberto N. Escalante-B.、Tobias Rosenkranz 和 Andreas Maier。《DeepFilterNet：以感知为导向的实时语音增强》。载于 Interspeech 2023，第2008–2009页，2023。3、5、6、2、14[50] Bowen Shi、Wei-Ning Hsu、Kushal Lakhotia 和 Abdelrahman Mohamed。《通过掩码多模态聚类预测学习音视频语音表征》。arXiv预印本 arXiv:2201.02184，2022。2、8、3[51] Bowen Shi、Andros Tjandra、John Hoffman、Helin Wang、Yi-Chiao Wu、Luya Gao、Julius Richter、Matt Le、Apoorv Vyas、Sanyuan Chen 等。《SAM Audio：音频中的任意分割》。arXiv预印本 arXiv:2512.18099，2025。3、5、7、2[52] Gemma团队、Sherif El Abd、Vaibhav Aggarwal、Robin Al- gayres、Alek Andreev、Olivier Bachem、Ian Ballantyne、Cor- mac Brick、Victor C˘arbune、Michelle Casbon 等。《Gemma 4技术报告》。arXiv预印本 arXiv:2607.02770，2026。4、3[53]Joachim Thiemann、Nobutaka Ito 和 Emmanuel Vin- cent。《多样环境多通道声学噪声数据库（DEMAND）：多通道环境噪声录音数据库》。载于声学会议论文集，第035081页。美国声学学会，2013。1[54] Marshall Thomas、Edward Fish 和 Richard Bowden。《VALLR：用于唇读的视觉ASR语言模型》。载于 IEEE/CVF国际计算机视觉会议（ICCV）论文集，第2846–2856页，2025。2[55] Yuancheng Wang、Haoyue Zhan、Liwei Liu、Ruihong Zeng、Haotian Guo、Jiachen Zheng、Qiang Zhang、Xueyao Zhang、Shunsi Zhang 和 Zhizheng Wu。《MaskGCT：使用掩码生成式编解码器变换器进行零样本文本转语音》。载于第十三届国际学习表征会议，2025。3[56] Yochai Yemini、Aviv Shamsian、Lior Bracha、Sharon Gannot 和 Ethan Fetaya。《LipVoicer：在唇读引导下从无声视频生成语音》。载于国际学习表征会议，第32147 32166页，2024 3、5、6、7、2、14 – 。[57] Jeonghun Yeo、Seunghee Han、Minsu Kim 和 Yong Man Ro。《视觉语音与语言的交汇：VSP-LLM框架-用于高效且具备上下文感知能力的视觉语音处理。在 计算语言学协会会议论文集：*EMNLP2024*，第 11391–11406 页，2024 年。 2

ReadMyLips：基于上下文的语言与视听模型唇读

补充材料

A. 设置

本节详细介绍评估集、损坏、评分和基线实现。成对比较和置信区间见第D节。

A.1. 测试集

表 6. **评估集**。 静音片段会从一个窗口中移除 k 个连续词语。只有经过验证的 剪辑片段才会参与评分；这些片段中的词语均能被 Whisper 从原始音频中准确转录。每个窗口在每种干扰条件下各提供一个噪声样本。 †场景。

Set	视频时间窗		k	静音片段 所有已验证样本		含噪
TalkVid（英语）	23	114	1–10	958	875	570
VoxMM（英语）	28	159	1–10	1,424	1,217	795
困难（英语）	23†	44	1, 3, 5	119	116	220
TalkVid（西班牙语、法语）	76	431	1–4	431	401	–
验证集（英语）	99	99	1–10	373	360	495

所有测试窗口的时长均为 4.8 秒，且均来自未用于任何组件训练的视频（表 6）。VoxMM [30, 31]窗口来自其测试集。这些对话录音的来源与 TalkVid 不同，而且语言模型或生成器的训练都未使用 VoxMM 数据。西班牙语和法语数据集包含 334 个西班牙语窗口和 97 个法语窗口，分别来自 69 个和 7 个 TalkVid 视频，且每个窗口都包含一个剪辑片段。

困难集包含英语窗口，其中由于说话人从侧面被拍摄、嘴巴被手或物体遮挡，或者说话人是动画或生成的，嘴唇难以或无法读取。28 个窗口来自生成的视频，其余 16 个来自电影和动画。作为视觉可用性的一项衡量指标，我们在每一帧上运行面部关键点检测器。对于 32 个窗口，它在每一帧中都找到了嘴巴；对于 5 个窗口，它在部分帧中找到了嘴巴；而对于剩余 7 个窗口，它在任何帧中都没有找到嘴巴。

所有超参数，包括选择的温度 T，均在英语 TalkVid 窗口的验证集上确定；这些窗口分别来自不同的视频，且与训练集或测试集不共享任何视频。

A.2. 损坏

静音语音。 对于每个窗口以及从 1 到 10 的每个 k，我们使用强制对齐的时间戳，寻找一个包含 k 个连续词语的区间。我们排除以停用词或填充词开头或结尾的区间，因为对齐器会将它们的边界放置得最不可靠。区间两侧至少各需保留一个仍可听见的词语，使该区间远离窗口边缘，因为边缘处的词语可能会被截断。我们将从区间第一个词语之前 80 毫秒到最后一个词语之后 80 毫秒的音频置零，并在相邻词语之前 20 毫秒处停止。如果该区间使区间内某个词语仍有部分可听见，或者区间边缘的某个词语持续时间少于 80 毫秒（表明存在错误对齐），则拒绝该位置。在所有有效位置中，我们优先选择包含窗口剪辑片段所围绕的词语的位置，其次选择最接近窗口中心的位置。如果没有位置符合这些规则（见表 k），窗口对于给定的 6 不产生剪辑片段。对于 TalkVid，以及西班牙语和法语，转录文本和词语时间戳来自 WhisperX。对于 VoxMM，我们原样使用 LRS-VoxMM 版本中的转录文本和词语时间戳。如果人工复核更正了参考文本，更正后的词语沿用其所替换词语的时间区间；而仅覆盖更正词语序列部分内容的剪辑片段则会被移除。

含噪语音。 对于多人嘈杂声，训练使用 DEMAND 自助餐厅噪声 [53]，来源于麦克风通道 1 至 12。对于测试，我们将每个完整窗口分别与多种干扰混合，以考察该方法在不同条件下的表现。主要评估包含五种条件。其中两种使用来自保留通道 13 至 16 的 DEMAND 自助餐厅多人嘈杂声，信噪比为 -10 和 -5分贝。另两种使用 VGGSound 环境声音 [8]，信噪比为 -5分贝；这些声音来自训练中使用的九个类别的测试片段，或来自三个保留类别（教堂钟声、打喷嚏声、火车鸣笛声）的测试片段。第五种是信噪比为-5分贝的竞争说话人语音，取自同一测试集中的另一个视频。第 D.2 节使用两个独立来源重复了多人嘈杂声评估：DEMAND 餐厅场景和六位其他说话人的混合语音；我们的词错误率没有显著变化。我们使用干净窗口和噪声的均方根幅度来设置信噪比。

条件网格。 对于第5.3节，我们同时改变损坏的强度和范围。在 -5、 -10 或 -20分贝下的多人嘈杂声，或静音，会影响一个词、四个词或整个时间窗，同时也包含干净的时间窗。由此得到的网格包含 1,421 个 TalkVid 样本和 1,998 个 VoxMM 样本，表 12 报告了每种条件。我们以相同方式，利用验证窗口构建验证网格。

A.3. 评分与统计

使用 Whisperlarge-v3 [44]，对整个窗口中的每个输出进行转录，并在文本规范化后将转录文本与参考进行对齐。静音切分计为

完全恢复，当且仅当每个静音词都在原位匹配，且区间内没有插入任何词。词准确率是指原位匹配的静音词所占的百分比。对于唇读器等仅生成文本的方法，则使用相同的对齐方式对其文本进行评分。

Whisper 本身也可能误识别言语不流畅、含糊书写的数字，或在时间窗边缘被截断的词语。为避免将此类错误计为恢复失败，我们仅在 Whisper 根据原始、未损坏的音频准确转写出缺失词语时，才对静音切分进行评分。这些经过验证的剪辑片段占 TalkVid 剪辑片段的 91%，占 VoxMM 剪辑片段的 85%（表6）。对于噪声语音，我们报告按时间窗平均的整窗词错误率。当输出插入大量词语时（例如包含了竞争说话人的话语），词错误率可能超过 100%。

由于同一视频中的剪辑片段和时间窗彼此相关，我们以源视频为单位进行配对自助法，或在困难集上以场景为单位进行配对自助法，至少重复抽样 2,000 次，计算 95% 百分位置信区间。当置信区间不包含零时，我们认为差异具有统计显著性。

A.4. 基线方法

每种基线方法均使用其官方代码和检查点，在与我们的方法相同的时间窗上运行，并由同一评分器进行评分。

识别器。 Auto-AVSR [36], USR 2.0 [23] 和 Llama-AVSR [4]仅生成文本。对于静默语音，它们只通过读取嘴唇进行识别。对于噪声语音，我们分别仅基于嘴唇（V），以及结合嘴唇和含噪音频（AV）运行每个模型。视觉 Auto-AVSR 模型也是我们的唇读器R。

语音编辑器。 F5-TTS [11]和 VoiceCraft [41] 使用说话人的声音，将提供的词填入标记的音频区间，而 AlignDiT [13] 还以嘴唇信息作为条件。每个编辑器都会接收我们的语言模型生成的转录文本（即生成器可用的相同文本），以及静音片段的位置；我们的方法必须从输入中推断出该位置。我们根据每个编辑器官方推理代码返回的输出进行评分。仅将 AlignDiT 生成的区间粘贴回输入后，其精确恢复率便从 35.7% 提升至 36.6%，且不会影响任何比较。用唇读器的最佳假设替代语言模型的转录文本后，F5-TTS 和 AlignDiT 分别能够精确恢复 30.3% 和 31.4% 的区间，比使用 L的转录文本（表13a）少约五个百分点。

视频到语音。 LipVoicer [56]根据视频生成整个窗口的语音，并由其自身的唇读器提供指导，同时忽略输入音频。对于含噪语音，Align- DiT 接收唇读器给出的最佳假设，并以另一个窗口中同一说话人的干净录音作为语音参考。它也会忽略噪声输入，因此其错误不受噪声条件影响。

增强与分离。 DeepFilterNet3 [49] 和 Sidon [40]仅处理含噪音频。VisualVoice [18]利用目标说话人的面部分离其语音。SAM Au- dio [51] 的提示可以是文本“语音”，也可以是目标说话人的遮罩视频。

表 7。**每种方法接收的内容。**区间：静音区间的位置。转录文本：从方法外部提供的转录文本，来自我们的唇读器（R）或语言模型（L）。语音：来自另一个窗口的说话人干净录音。 *对目标声音的文本描述，而非转录文本。

方法	视频	音频	区间	转录文本	语音
识别器 (V)	✓	–	–	–	–
识别器 (AV)	✓	✓	–	–	–
VoiceCraft ，F5-TTS	–	✓	✓	L	–
AlignDiT，静默	✓	✓	✓	L	–
AlignDiT，含噪	✓	–	–	R	✓
LipVoicer	✓	–	–	–	–
DeepFilterNet3，Sidon	–	✓	–	–	–
VisualVoice	✓	✓	–	–	–
SAM Audio，视频提示	✓	✓	–	–	–
SAM Audio，文本提示	–	✓	–	–*	–
Ours	✓	✓	–	–	–

输入。表 7 列出了各方法可用的输入。我们的方法接收视频和损坏音频。语音编辑器还会接收静音片段的位置以及我们的语言模型生成的转录文本。对于噪声语音，AlignDiT 还会接收该说话人的干净录音。解读比较结果时，应考虑这些额外输入。

覆盖率。 依赖面部信息的方法无法处理所有窗口。在 TalkVid 和 VoxMM 上，VisualVoice 覆盖了 1,365 个噪声窗口中的 1,245 个，而 AlignDiT 覆盖了 1,315 个。在困难集上，覆盖率较低：USR2.0 和 Llama-AVSR 只能处理约四分之三的剪辑片段和窗口，而 VisualVoice 处理的噪声窗口不到一半（表 14）。对于每个此类基线，我们都在其覆盖的剪辑片段或窗口上计算两项分数。

B. 实现

B.1. 模型与训练

唇读器与可听上下文。唇读器 R 将视觉 Auto-AVSR [36] 应用于整个窗口的嘴部裁剪图像。束搜索会返回其 $N = 10$ 个最佳且互不相同的转录结果。对于西班牙语和法语，我们改用 Ma等人的唇读器。 [35] Whisper large-v3 对 $\tilde{a}$ 进行转录，以提供可听上下文 c。为了标记该上下文中的间隙，我们使用 5 毫秒的帧、比音轨峰值低 70 分贝的阈值以及 50 毫秒的最短持续时间，直接在 $\tilde{a}$ 中检测静音区间。如果有多个区间符合条件，我们会选择两侧均有音频且持续时间最长的区间。在测试集上，该检测器



图5. **语言模型的提示及其回答**（粗体显示），针对一个 TalkVid 测试剪辑片段，其中被静音词是“candle out”。没有任何假设包含这些词。模型指出最接近的假设，并结合上下文对其进行改写。

在 99.9% 的剪辑片段中检测到静音区间；对于 TalkVid 剪辑片段，99.2% 的片段中两个边界均与真实边界相差不超过 10 毫秒，对于 VoxMM 剪辑片段，这一比例为 99.6%。第D.8 节考察了并非数字静音、而是由室内底噪或舒适噪声填充的间隙。正如图 c 所示，该区间在5中由一个占位符及其持续时间表示。

语言模型。 L为 Gemma-4 12B [52]（经指令微调），并使用 LoRA [26] 秩为 32 的适配器（$\alpha = 64$，dropout 为 0.1），应用于每个注意力层的查询、键、值和输出投影。模型仅使用英语 TalkVid 时间窗进行训练，有效批量大小为 16（四个样本，累积四步），并采用余弦学习率调度，分三个阶段进行。第一阶段（3,000 步，学习率 10^{-4}）中，模型学习选出最佳候选项或撰写转录文本。第二阶段（3,000 步，$5 \cdot 10^{-5}$）中，模型学习第4.2 节所述的回答，即先选出最佳候选项，再对其进行编辑。第三阶段（1,000 步，$5 \cdot 10^{-5}$）中，向静音示例中加入含噪示例。训练期间展示的假设来自四种识别器：对原始视频运行的 Auto-AVSR（占示例的 50%）、对嘴部裁剪图像进行缩小（从 96 像素缩至 24 像素）再放大的 Auto-AVSR（15%）、对一半帧替换为前一帧的视频运行的 Auto-AVSR（15%），以及 AV-HuBERT [50]（20%）。展示的假设数量在 1 到 N之间均匀抽取。测试时采用贪心解码。

唇部特征头。 对于表5c 的消融实验，一个小型预测头读取 AV-HuBERT [50] 嘴部特征，并将其视素后验概率传递给 L作为软输入词元。该预测头与 L联合训练。为确认模型确实使用了这些特征，我们在测试时用另一个片段的特征替换它们。在 360 个经验证的验证剪辑上，$R \to L \to G$的精确恢复率从 64.4% 降至 41.7%（-22.8，置信区间$[-28.4, -17.2]$）。

生成器。 G 这是一个拥有 220 亿参数的 LTX-2.5 Transformer [21], ，通过秩为 64 的 LoRA 进行微调（$\alpha = 64$），作用于音频自注意力、音频文本交叉注意力、音频前馈层和视频到音频注意力。视频路径保持不变。模型使用 AdamW [34]进行训练，学习率为 10^{-4}，批量大小为 2，在一块 H100 GPU 上，使用英语、西班牙语和法语的 TalkVid 数据，分三个阶段进行。第一阶段（5,000 步）使用约 15,800 个由一至十个词组成的静音片段，每个片段都配有唇读器的转录文本。第二阶段（3,000 步）使用 13,018 个样本，即来自约 1,350 个视频的 6,509 个时间窗，每个时间窗各有一个静音版本和一个含噪版本。此阶段中，含噪样本会接收任务标签，但不附带转录文本。含噪训练样本会混入多人嘈杂声、已见过的音效、有色噪声或竞争说话人，信噪比在 -15 至 5分贝之间均匀抽取；有时还会加入第二个声源，其音量低 5 分贝。第三阶段（2,000 步）使用相同的样本，但会改变含噪样本所配的文本：一半使用唇读器的转录文本，四分之一使用参考转录文本，其余不提供转录文本。因此，生成器学会将转录文本视为可能有误的提示，并在没有转录文本的情况下工作。所有样本中有 10% 会去掉文本，由此训练出用于无分类器引导的无条件分支。静音区域内的损失权重为 $\lambda = 12$，并延伸至其边缘之外最多 60 毫秒。推理时，我们使用固定随机种子的 30 个去噪步骤，并将文本引导设为 5（<修复> 时），设为 1（ <去噪>时）。生成器会生成整个时间窗：受损区域之外的音频也会与修复部分一同生成，而不是从输入中复制；所有评分都在生成的时间窗上计算。第D.6 节表明，这不会改变完整语音的可懂度。

生成器从不会被告知受损的位置。在推理时， G 会接收视频、受损音轨、整个窗口的转录文本以及任务标签。它既不会通过掩码、时间戳，也不会通过文本中的标记获知受损的位置或长度，因此必须自行判断哪些地方缺失了语音，哪些地方必须保留。噪声情况下也是如此：没有任何组件会被告知窗口的哪一部分遭到破坏或破坏程度如何，选择过程完全依据证据来决定（第C节）。只有在语言模型的可听上下文中，才会用到间隙的位置。在那里，静音区间会被

标记为占位符并注明其时长，因为 Whisper 之类的识别器不会报告转录文本中哪些位置缺少词。同一检测步骤还会设置任务标签： <修复>检测到静音区间时使用 <降噪> 否则使用。这个检测步骤是一个前端，可以替换而不影响处理流程的其余部分。第D.8 节已经用一个同时利用音频和嘴唇的检测器替换了它，其他部分均未更改。能够标记未听到语音的位置，或为每个词报告置信度的识别器，可以直接提供占位符。

B.2 资源与运行时间

所有训练和推理任务均使用一块显存为 80 GB 的 NVIDIA H100 GPU。生成器的 22B 参数主干保持冻结，并以 bfloat16 格式运行，只有其适配器参与训练。我们提前对训练时间窗的音频和视频潜变量进行一次编码。随后，批次大小为 2 的生成器训练步骤耗时约 2 秒，其三个阶段总计约需 6 GPU 小时。语言模型在一块 GPU 上每步耗时约 2.5 秒，三个阶段总计约需 5 GPU小时。推理时，对一个 4.8 秒语音窗口进行一次 30 步生成约需 11 秒。固定配置只需运行一次。选择过程则每种配置需运行一次：受损窗口需运行六次，而没有可听语音时需运行三次。之后，使用真实输入和无信息输入分别通过 Whisper 和唇读器对每个不同的转录文本进行评分（第C节）。

C. 选择

我们详细介绍第 4.3 节提出的测试时选择，检查其选项、消融其中的各项，并完整报告第 5.3 节的条件网格。

证据。 公式（2）中的两项都是对每个候选转录文本进行教师强制后得到的对数似然，并对各个词元求和。音频项 e_A 使用 Whisper large-v3 对转录文本进行评分：先以受损输入 $\tilde{a}$ 为条件，再以等长的静音音轨为条件。视觉项 e_R使用视觉 Auto-AVSR 的加权注意力和连接时序分类对数似然（0。9 和 0。1，即其解码器权重），先以窗口中的嘴部裁剪图像为条件，再以第一幅裁剪图像的重复版本为条件。重复的裁剪图像保留了说话人的外观，同时消除了发音动作。如果生成器未生成语音，其空转录文本不会获得证据、不参与投票，也不会被选中。当使用 USR2.0 替代 Auto-AVSR 作为唇读器时，只有 R 的假设会发生变化。证据项仍按上述方式计算，而噪声输入对应的音视频假设仍来自 Auto-AVSR，因此这种替换无需重新训练，也无需新增设置。

每个项都相对于其自身的基线进行归一化。减去同一识别器在无信息输入下对同一转录文本的对数似然值，在很大程度上会抵消识别器的语言先验以及转录文本长度带来的代价。剩下的就是对转录文本与输入之间逐点互信息的估计。因此，这两项都是以相同单位表示的对数似然比，我们将它们相加而不使用模态权重：不携带词信息的模态对每个候选项给出的值都接近零，因此不会产生影响。我们对词元求和而不是取平均，这样，逐词得到支持的转录文本就会胜过仅由一个高置信度词支持的转录文本。每个窗口的长度都相同，因此这一尺度在不同片段之间具有可比性。唯一的全局尺度是温度T。

表8报告了第5.3节条件网格中的两项指标。随着损坏程度增加，音频证据会减弱。其在候选项上的逐样本最大值（即第5.3节中引用的数值）在语音仍可听见时较高；随着多人嘈杂声变大并覆盖更多窗口，该数值会下降；当整个片段被静音时则降为零。由于视频从未受到破坏，各种条件下的视觉证据都维持在相近水平。

表8。**条件网格中的证据。**平均音频证据 e_A 和视觉证据 e_R（单位：奈特），对应于图4中的各个条件。*Sel.*：选定输出。*Max*：候选项上的逐样本最大值；候选项包括六种配置和两个文本输出 h_1 和$\hat{y}_L$。1周、4周、片段：受影响范围分别为一个词、四个词或整个窗口。

		e_A		e_R	
条件	n	Sel.	Max	Sel.	Max
干净	268	79.0	80.4	57.1	69.8
多人嘈杂声 −5分贝，1周	268	78.8	81.2	57.1	69.7
多人嘈杂声 −5分贝，4w	252	76.0	78.5	56.4	70.1
多人嘈杂声 −5分贝，片段	268	55.2	58.7	53.6	69.6
多人嘈杂声 −10分贝，1w	268	78.1	81.1	57.1	69.7
多人嘈杂声 −10分贝，4周	252	70.1	73.5	54.5	70.1
多人嘈杂声 −10分贝，片段	268	18.8	23.3	55.1	69.4
多人嘈杂声 −20分贝，1周	268	75.9	79.5	56.4	69.6
多人嘈杂声 −20分贝，4周	252	62.3	66.7	53.9	69.9
多人嘈杂声 −20分贝，片段	268	0.2	3.3	55.8	69.3
静音，1周	268	74.8	77.0	57.0	69.6
静音，4w	252	60.9	64.1	55.6	69.8
静音，片段	267	0.0	0.0	59.1	69.3

投票者。 语言模型会读取唇读器完整的 N最佳列表，因此其转录文本 $\hat{}y_L$ 已经包含了来自 h_1的信息。如果允许两者都进行投票，就会将该信息计算两次。因此，当 $\hat{}y_L$ 进行投票时，我们会保留将音频置于静音之上的配置作为候选项，但将其排除在投票者之外。距离 d 取两个方向上的平均词错误率。对于完全相同的风险平局（这种情况会在不同配置产生相同转录文本时出现），优先选择保留音频的配置，其次选择使用更多组件的配置。

我们没有对这一排除规则进行调优：这是根据上述重复计算论证得出的设计规则。选择过程唯一经过调优的参数是温度，我们在验证网格上将其设为 $T = 100$ 。

表9. **选择机制在静音片段上的配置选择**（合并 TalkVid 和 VoxMM）以及在噪声窗口上的配置选择，以样本百分比表示。柔和：生成器从静音开始，而不是从 $\tilde{a}$ 开始。

	保留音频			柔和	
	$R{\to}L{\to}G$	$R{\to}G$	G	$R{\to}L{\to}G$	$R{\to}G$
静音（2,092）	73.2	21.7	2.5	1.9	0.8
含噪（1,365）	62.5	21.1	3.8	11.9	0.7

选择机制会选择什么。 在静音片段上，选择机制几乎总是保留音频；对于四分之三的片段，它会向生成器提供语言模型的转写文本（表 9）。在噪声窗口上，它更常将输入静音，而且干扰越强，静音的情况越多：在 -10分贝的多人嘈杂声下，43% 的窗口会被静音；在 -5 分贝的多人嘈杂声下，这一比例为 7%；在音效或竞争说话人的干扰下，则为 4% 至 5%。在条件网格上，对于每个干净片段，它都会保留音频；但对于被 -20分贝多人嘈杂声完全覆盖的片段，它会将其中 88% 静音（图 4）：在这 268 个片段中，它选择 $R \to L \to G$而非静音的有 213 个，选择 $R \to G$ 而非静音的有 23 个，其余 32 个则保留音频。在静音片段或噪声窗口上，它从不选择 G 而非静音；该配置既没有文本，也没有音频。

表10。 **选择规则的消融实验。** 网格 和含噪：图 4 中的3,419个网格样本和1,365个含噪窗口上的词错误率。 静默：在2,092个已验证片段上的精确片段恢复。括号内为变体减去我们的方法；当其95%区间不包含零时以粗体显示。预言机为每个样本选取最佳配置。

	网格 ↓	含噪 ↓	静默 ↑
Ours	23.1	27.2	39.3
所有输出进行投票	23.6 (+0.4)	32.2 (+**5.0**)	36.3 (−**3.0**)
无证据 ($T \to \infty$)	23.6(+**0.4**)	27.9 (+**0.7**)	39.0 (−0.3)
仅使用证据 ($T \to 0$)	23.9(+**0.8**)	27.9 (+**0.7**)	34.3 (−**5.0**)
无 e_A	23.3 (+0.1)	28.5 (+**1.3**)	39.3(+0.0)
不含 e_R	23.5 (+**0.3**)	27.4(+0.3)	38.9 (−0.4)
预言机选择	20.0	23.2	44.3

消融。 表 10 评估了第 5.3节条件网格、含噪窗口以及移除一个部分后的静音片段上的选择规则。每个部分至少在一种设置下都有显著贡献，而没有任何消融变体能显著优于完整规则。投票者排除的影响最大，在含噪语音（置信区间 $[3.7, 6.4]$ 个百分点）和静音语音（置信区间 $[1.2, 4.9]$）上均是如此。选择不经投票而选择证据最高的输出，在三种设置下表现都更差。等权重（$T \to \infty$）平均只造成不到一个百分点的损失，因为证据仅在大约十分之一的输入中改变选择。在这些输入上，其影响很大（表 11b）。音频证据对含噪语音很重要，视觉证据对网格很重要；而在静音语音上，移除任一项都没有显著影响。预言机行展示了更优规则仍有提升空间。

表11。**Whisper 评分下的选择诊断。**(a) 针对按条件进行事后选择的最佳固定配置进行选择。(b) 在两者选择不同的样本上，将证据加权与等权重进行比较，并列出其词错误率（E / eq.）。 Δ：选择减去比较基准。B/W：(a) 中显著更好或更差的条件，(b) 中更好或更差的样本。

(a) 按条件的事后规则

Set	选择。固定	Δ [95% 置信区间]	B/W
含噪（1,365）	27.2	28.7 −1.5 [−1.9, −1.0]	4/0
网格（3,419）	23.1	24.2 −1.0 [−1.3, −0.7]	8/0

（b）不一致子集

Set	变更的 E/等式	Δ [95% 置信区间]	B/W
含噪	145 (10.6%) 40.6/47.3	−6.7 [−9.3, −3.9]	80/29
Grid	304（8.9%）35.0/39.8	−4.8 [−6.2, −3.5]	155/77

完整的条件网格。表12 给出了图4 中选择、每种固定配置以及两种文本输出对应的词错误率。第5.3 节中引用的平均值位于最后一列，其中固定配置使用了输入音频。没有一种固定配置在所有条件下都表现最佳。语音仍可听见时，保留音频的配置表现更好；但在整段片段都被 -20分贝多人嘈杂声覆盖时，这些配置的表现不如基于静音的相同配置。选择方法能适应这两种情况：语音可听见时保留音频；多人嘈杂声覆盖整个片段时，则切换到基于静音的配置。在整段片段都被−20分贝多人嘈杂声覆盖时，即使没有获知噪声级别，其表现与柔和配置及唇读器自身的文本相比也仅差约一个点。表12 中的柔和行是参考行，它们会获得受损区域，并仅将该区域替换为静音，因此会重复静音列中的结果。选择方法不会接收受损区域。其柔和候选项会丢弃整个输入音轨，而它仅依据证据，在这些候选项与保留音频的候选项之间作出选择。即便如此，除整段片段都被 -20分贝多人嘈杂声覆盖的情况外，它在所有条件下都能达到或优于柔和行；在该情况下，它也仅差半个点。

与按条件固定选择相比。 表格中的固定配置12 对每个输入都使用同一种选择。作为更强的回顾性比较基准，我们根据每个条件的

表12. **完整条件网格。** 图中 3,419 个样本的词错误率（%）。 4。多人嘈杂声或静音会影响一个词（1周）、四个词（4周）或整个窗口（片段）。 保留：生成器从 $\tilde{a}$ 开始。 柔和：将 $\tilde{a}$ 中受损部分替换为静音的参考。这些行会获知受损部分的位置，不同于选择，其静音候选项从完全静音的窗口开始。 R 和 $R \to L$ 根据其文本进行评分。

		多人嘈杂声 −5分贝			多人嘈杂声 −10分贝			多人嘈杂声 −20分贝			静音			
方法	干净	1w	4w	clip	1w	4w	clip	1w	4w	clip	1w	4w	clip	Avg
样本	268	268	252	268	268	252	268	268	252	268	268	252	267 3,419	
选择（我们的方法）	10.1	10.4	13.1	27.1	11.2	18.1	43.3	12.3	22.2	49.8	12.2	21.3	48.9	**23.1**
$R \to L \to G$，保留	10.3	11.4	14.6	28.4	12.2	20.3	47.5	13.7	27.3	75.4	13.2	22.4	50.0	26.8
$R \to L \to G$，柔和	50.0	13.2	22.4	50.0	13.2	22.4	50.0	13.2	22.4	50.0	13.2	22.4	50.0	30.3
$R \to G$，保留	10.2	11.6	14.9	29.0	12.9	19.0	47.9	14.1	25.5	71.3	13.6	22.9	49.3	26.4
$R \to G$，柔和	49.5	13.6	22.9	49.5	13.6	22.9	49.5	13.6	22.9	49.5	13.6	22.9	49.3	30.4
G 无文本，保留	10.2	12.7	16.2	42.0	14.4	26.1	81.4	17.1	39.0	114.9	15.9	38.6	100.1	40.8
G 无文本，柔和	100.1	15.9	38.6	100.1	15.9	38.6	100.1	15.9	38.6	100.1	15.9	38.6	100.1	55.6
R(文本)	48.8	48.8	49.9	48.8	48.8	49.9	48.8	48.8	49.9	48.8	48.8	49.9	48.8	49.1
$R \to L$(文本)	16.9	18.4	24.6	38.7	20.5	32.6	46.4	22.8	38.9	47.7	17.5	30.5	47.9	31.0

测试结果，分别为每个条件选择词错误率最低的配置。因此，它知道条件及其结果；其柔和配置也知道受损区域，但仍对该条件下的每个样本使用同一种配置。表格中的逐样本预言机10则是另一种比较基准。按照 Whisper 评分，选择在噪声集和条件网格上的表现都优于这条按条件划分的规则（表格11a）。在网格的 13 个条件中，共有五种配置表现领先，包括 G 在干净输入上单独使用，以及在 -20分贝多人嘈杂声下使用柔和配置。

证据改变时的选择。 证据只会改变少数输入的所选配置。在这些不一致子集上，证据所选方案的词错误率显著低于等权投票（表 11b）。这种条件分析解释了为何证据对个别输入可能很重要，而表 10 中的总体差异却不大。

D. 补充结果

我们首先考察与基线方法的配对比较，然后检验噪声语音结果是否适用于不同干扰源和测试集。其余分析将探讨生成器、评分协议、恢复语音示例、并非数字静音的间隙，以及语言模型从唇读器获取的信息。

D.1. 与基线方法的比较

正文报告了各方法的得分。这里给出这些比较背后的配对差异和95%区间（第A.3节）。我们先考察合并的静默和含噪评估结果，然后检查各基线都能处理的输入上的困难集。

静音语音。使用 Auto-AVSR 作为唇读器时，我们的方法显著优于除 USR2.0 以外的所有基线；在合并的剪辑片段上，与 USR2.0 无显著差异（表13a）。各测试集的比较结果有所不同：在 TalkVid 上，我们领先 USR 2.0 6.9 个百分点（置信区间 $[0.5, 13.2]$），而在 VoxMM 上领先 1.2 个百分点（置信区间 $[-4.6, 7.0]$）。在 VoxMM 上，与 Llama-AVSR 的差异也包含零（$+6.2$，置信区间 $[-0.4, 11.9]$）。

USR 2.0 是比 Auto-AVSR 更强的唇读器。用它替换 Auto-AVSR 后，我们的精确恢复率从 39.3% 提升至 43.8%（$+4.5$，置信区间 $[1.1, 7.7]$），比单独使用 USR2.0 高 8.0 个百分点（置信区间 $[5.3, 11.0]$）。

含噪语音。在五种条件下取平均后，我们的方法的词错误率显著低于所有基线（表13b）。在任何合并条件下，都没有基线显著优于我们的方法，但有三个基线分别在一种条件下与我们的方法无显著差异。

哪种基线最接近我们的方法，取决于干扰类型。VisualVoice 在其覆盖的时间窗上与我们的竞争说话人结果相当（$+1.6$，置信区间 $[-0.8, 4.0]$），而在仅使用 VoxMM 时表现更好（$+3.9$，置信区间 $[0.8, 6.9]$）。在 -10 分贝的多人嘈杂声下，仅使用嘴唇信息的 USR 2.0 与我们的方法无显著差异（-3.5，置信区间 $[-7.5, 0.6]$），此时音频提供的信息最少。在 -5 分贝的多人嘈杂声下，同时使用嘴唇和音频信息的 USR 2.0 与我们的方法无显著差异（-0.6，置信区间 $[-4.5, 2.6]$），此时仍有大量语音清晰可闻。

将我们的唇读器替换为 USR2.0，可将平均词错误率从 27.2% 降至 25.5%（-1.7，置信区间 $[-3.1, -0.2]$）。第 D.2 节使用其他录音重复了多人嘈杂声对比。

我们还在合并后的 TalkVid 和 VoxMM 噪声样本上运行 SAM Audio，使用“语音”作为文本提示，或使用屏蔽的目标说话人视频作为视觉提示。这两种设置的平均词错误率都明显高于我们的方法（无论采用哪种输出转录器）（表 13b 和表 20）。表 14 还评估了 SAM Audio 在困难集上的表现。

表13. **与基线方法比较**，在 TalkVid 和 VoxMM 上合并统计。 Δ：我们的方法减去基线方法，单位为百分点，并附上其 95%区间；区间不包含零时以粗体显示。在 (a) 中，正值 Δ 有利于我们的方法；在 (b) 中，负值 Δ 有利于我们的方法。无显著差异：差异不显著的噪声条件。 †适用于方法所覆盖的时间窗。

(a) 静音语音，已验证片段

方法	精确恢复 %	精确恢复 Δ [95% 置信区间]	词准确率 %	词准确率 Δ [95% 置信区间]
我们的方法（将 Auto-AVSR 作为 R）	39.3		65.7	
Auto-AVSR [36]	30.7	**+8.6** [6.4, 11.0]	59.7	**+6.0** [4.7, 7.2]
USR 2.0 [23]	35.8	+3.5 [−0.6, 8.0]	64.9	+0.8 [−2.3, 4.0]
Llama-AVSR [4]	30.4	**+8.9** [3.5, 14.6]	58.5	**+7.1** [1.8, 11.9]
LipVoicer [56]	28.3	**+11.0** [7.0, 15.9]	58.6	**+7.0** [3.8, 10.1]
VoiceCraft [41] (L)	32.3	**+7.0** [5.5, 8.7]	59.8	**+5.8** [4.8, 6.9]
F5-TTS [11](L)	35.8	**+3.5** [2.4, 4.7]	62.2	**+3.5** [2.2, 4.7]
AlignDiT [13] (L)	35.7	**+3.6** [2.3, 4.9]	62.8	**+2.9** [1.7, 4.0]

(b) 噪声语音，五种条件的平均值

方法	词错误率	Δ [95% 置信区间]	Tied
我们的方法（Auto-AVSR 作为 R）	27.2		
噪声输入	69.0	**−41.8** [−46.1, −38.2]	–
Auto-AVSR（V）	49.4	**−22.2** [−26.4, −17.6]	–
Auto-AVSR（AV）	70.6	**−43.4** [−46.1, −40.6]	–
USR 2.0（V）	47.4	**−20.2** [−25.2, −15.5]	多人嘈杂声 −10
USR 2.0（AV）	52.0	**−24.8** [−28.2, −21.7]	多人嘈杂声 −5
Llama-AVSR（V）	58.8	**−31.6** [−39.4, −23.7]	–
Llama-AVSR（AV）	47.9	**−20.6** [−23.9, −17.4]	–
DeepFilterNet3	72.2	**−44.9** [−49.8, −41.1]	–
Sidon	84.4	**−57.1** [−61.6, −53.4]	–
SAM Audio，文本提示	67.1	**−39.9** [−44.7, −35.7]	–
SAM Audio，视觉提示	75.4	**−48.2** [−52.6, −43.7]	–
VisualVoice†	40.1	**−12.7** [−16.0, −9.8]	其他说话人
LipVoicer	52.9	**−25.7** [−30.2, −20.9]	–
AlignDiT (R)†	50.5	**−23.4** [−27.7, −18.9]	–

困难集。 表 14 列出了表 4 背后的差异，比较我们的方法与各基线在其覆盖的剪辑片段或时间窗上的表现。对于静音语音，在 Auto-AVSR、USR 2.0 覆盖的剪辑片段、LipVoicer 和 VoiceCraft 上，我们的优势都很显著。Llama-AVSR 在与 USR 2.0 相同的剪辑片段上表现较差，但其区间包含零。

F5-TTS 和 AlignDiT 都使用我们语言模型的转录文本和片段位置，在统计上与我们的方法无显著差异。本集合中的大部分提升来自语言模型（表 5a），这些编辑器也使用了该模型的转录文本。对于噪声语音，除 VisualVoice 外，我们的方法显著优于所有基线。其人脸检测器仅能在不到一半的时间窗中成功检测，在这部分数据上，两种方法无显著差异。

D.2. 其他录音中的多人交谈声

主要评估使用了训练期间未纳入的麦克风通道上的 DEMAND 自助餐厅多人嘈杂语音（第 A.2 节）。为考察其他多人嘈杂声，我们测试相同的语音窗口

表14. 困难集：覆盖率和差异。 n：某方法覆盖的剪辑片段或时间窗数。 Δ：在这些样本上，我们的方法与基线的差值，以及按场景计算的95%区间；若区间不包含零，则以粗体标示。对于静默语音，正值Δ 表示有利于我们的方法；对于噪声语音，负值 Δ 表示有利于我们的方法。

方法	静默（精确匹配） n	静默（精确匹配） Δ	含噪（词错误率） n	含噪（词错误率） Δ
受损输入	–	–	220	**−24.6** [−32.6, −17.6]
Auto-AVSR（V / AV）	116	**+17.2** [11.1, 22.9]	210	**−42.5** [−49.2, −34.7]
USR 2.0	83	**+10.8** [2.8, 17.6]	160	**−27.4** [−34.1, −20.5]
Llama-AVSR	83	+9.6 [−1.8, 18.1]	160	**−31.4** [−41.5, −22.6]
SAM Audio、文本	–	–	220	**−24.1** [−29.2, −18.3]
SAM Audio、视频	–	–	220	**−34.3** [−41.9, −26.4]
VisualVoice	–	–	90	−8.8 [−22.9, 2.0]
LipVoicer	111	**+16.2** [9.5, 22.5]	210	**−64.3** [−80.8, −46.0]
VoiceCraft (L)	116	**+6.0** [0.9, 11.8]	–	–
F5-TTS (L)	116	+2.6 [−1.1, 7.3]	–	–
AlignDiT (L / R)	110	+3.6 [−0.8, 8.8]	165	**−68.8** [−86.6, −47.7]

以及另外两个来源。其中一个是 DEMAND 的餐厅场景，录制于不同的房间和录音时段。另一个来源将六个留出的测试视频中六位说话人的语音相加，并排除每个窗口对应的源视频。这两个来源均未与训练数据共享录音、场景或说话人。我们将表 3 中全部 546 个样本在两个级别下的多人嘈杂声进行替换，同时固定视频、干净语音、生成器文本、采样种子和混音代码。只有干扰及其混音级别发生变化。

匹配输入词错误率。在固定信噪比下，不同的多人嘈杂声录音对语音的掩蔽程度不一定相同。自助餐厅录音中，频率在 300 至 3,400 Hz 之间的能量占 39%，而餐厅录音中这一比例为 69 %。在标称 −5分贝水平下，餐厅多人嘈杂语音和六位说话人的混合语音会使噪声输入词错误率从 52.9% 升至 90% 以上。因此，在相同比率下进行比较，会混淆多人嘈杂声来源和输入难度。为此，我们选择相应的比率，使 Whisper 在自助餐厅多人嘈杂语音条件下的噪声输入词错误率与之匹配：餐厅条件下为 −5.4 和 −1.4分贝，六位说话人条件下为−4.6 和 −0.4分贝。我们仅根据噪声输入确定这些比率，并在运行任何修复方法之前完成设定。在每个水平下，输入词错误率与自助餐厅条件相比最多相差 1.3 个百分点（表 15，首行）。

方法。 对于每种新的混合音频，我们都会重新生成使用输入音频的三种配置，并按照相同的规则和温度重新执行选择，仅重新计算依赖于输入的证据。我们也使用原始代码重新运行 DeepFilterNet3、Sidon 和 VisualVoice。LipVoicer 和 AlignDiT 不使用输入音频（表 7），因此它们在不同多人嘈杂声来源下的输出完全相同。作为可复现性检查，我们在第二个集群上重新运行自助餐厅多人嘈杂语音的输出，词错误率变化不超过 0.1 个百分点。

表15。 **其他录音中的多人交谈声**。 在表 3的546个多人嘈杂语音样本上，在主要实验的自助餐厅多人嘈杂语音以及两个与训练集不共享任何录音和说话人的多人嘈杂声源下，以相同的输入词错误率混合。 Δ：在相同时间窗上，选择与方法的差值，以及新来源的95%区间。每个区间均不包含零。 *不读取输入音频。 †在方法覆盖的时间窗上（VisualVoice为498个，AlignDiT为526个）。

	WER			Δ （我们的 −方法）		
方法	自助餐厅	餐厅	六位说话人	自助餐厅	餐厅	六位说话人
噪声输入	72.3	72.6	72.9	−36.3	−36.6 [−43.2, −31.0]	−35.7 [−42.4, −29.8]
DeepFilterNet3	81.6	81.1	89.4	−45.6	−45.1 [−51.0, −39.9]	−52.2 [−58.9, −46.2]
Sidon	96.2	93.9	99.7	−60.2	−57.9 [−63.9, −52.7]	−62.5 [−70.3, −55.6]
VisualVoice†	61.1	58.1	60.3	−24.7	−21.4 [−27.4, −16.1]	−22.5 [−28.6, −17.2]
LipVoicer*	52.9	52.9	52.9	−16.9	−16.9 [−20.4, −12.9]	−15.7 [−19.7, −11.3]
AlignDiT（R）*†	50.5	50.5	50.5	−14.7	−14.6 [−18.2, −11.4]	−13.4 [−16.8, −10.3]
Ours, G 单独	64.4	62.3	75.3			
Ours, $R \to G$	38.3	39.4	42.4			
Ours, $R \to L \to G$	39.1	38.1	41.0			
Ours, 选择	**36.0**	**36.0**	**37.2**			

结果。 我们的方法在自助餐厅和两个独立多人嘈杂声来源上的词错误率仍然相近（表 15）。相较于自助餐厅多人嘈杂语音，餐厅场景的差异为 $+0.0$ （置信区间 $[-1.7, 1.7]$），六人说话者场景的差异为$+1.2$ （置信区间 $[-0.1, 2.5]$）。两者均不显著，即使分别考察两个水平也是如此。对于这两个新来源，我们的方法仍显著优于此处评估的所有基线，优势幅度相近，基线排名顺序也相同。在餐厅多人嘈杂语音条件下，其固定配置也均未发生显著变化。

尽管输入词错误率相同，六人说话者混合音频对多种音频处理方法而言更为困难。DeepFilterNet3 和 Sidon 的性能均显著下降。在我们的固定配置中， G单独受到的影响最大，而提供转录文本的配置上升幅度要小得多。与自助餐厅多人嘈杂语音下的 25% 相比，该选择通过静音这些时间窗中 30% 的输入来限制影响，因此在音频更难恢复时更多地依赖嘴唇信息。结果，其相对于 DeepFilterNet3 的优势进一步扩大。

相同的名义比例。我们还在主实验采用的名义 -10 和 -5 分贝条件下测试了餐厅多人嘈杂声。这是更困难的输入条件：噪声输入词错误率从 72.3% 上升到 95.0%。所有方法的输出词错误率都上升，包括 $R \to L \to G$ （51.0%）和 DeepFilterNet3（95.0%）。它们之间的相对差异与自助餐厅多人嘈杂声下的情况（-44.0，置信区间 $[-52.4, -36.7]$，与 -42.4 相比）仍大致相当。

D.3. 各测试集上的噪声语音

表16 按测试集细分了表3 中的结果。即使在干净录音上，VoxMM 的词错误率也更高；单独使用唇读器时，其在 VoxMM 上的错误率约为 TalkVid 上的两倍。这些差异有助于解释为何所有方法在 VoxMM 上的表现都更差。

我们的方法在两个集合上的平均词错误率均为最低。两个集合在 -10分贝多人嘈杂声条件下的差异最大，该条件下的性能最依赖唇读器。在该条件下，在我们的处理流程中用 USR 2.0 替换 Auto-AVSR 几乎不会改变 TalkVid 上的结果，但会使 VoxMM 上的词错误率降低六个百分点。

D.4. 按片段长度统计的静音语音

图7按静音词数量细分了静音语音的结果。随着片段变长，每种配置恢复出的片段都更少。语言模型的表现优于唇读器，而生成器读出其转录文本后，在每种长度下表现都进一步提升：对于单词，唇读器的恢复率为 59.2%，生成器为 72.5%；对于七至十个词，恢复率则从 16.7% 提升至 23.7%。生成器对单个词的提升最大：在这种情况下，读出唇读器最佳假设（$R \to G$）就已经能恢复出明显多于该假设本身的片段；而对于七至十个词，两者相差不到一个百分点。对于长度不超过三个词的片段，选择机制使 $R \to L \to G$提升约两个百分点；对于更长的片段，则与其表现相当。

D.5. 生成器控制

以下控制项用于考察生成器对所提供的文本和视频的响应。

生成器会照给定内容生成。为了单独考察生成器的贡献，我们提供 G窗口的参考转录文本，而不是 L的转录文本。对于每个集合中经过验证的剪辑片段（包括困难集；表17a），Whisper 随后都能在超过 95% 的剪辑片段中准确还原静默片段；对于包含五到十个词的片段，这一比例同样成立（TalkVid 和 VoxMM 上为 95.8%）。在获得相同参考转录文本的情况下，尽管语音编辑器也会获得片段的位置，但在每个集合中，它们还原出的片段都更少。 G 在所有方法都能处理的 108 个困难集剪辑片段上，结果仍然最高。这些预言机文本结果表明，生成器通常能够准确定位并清晰地说出提供的词语，

表16. **各测试集上的噪声语音。** TalkVid（570个时间窗）和VoxMM（795个时间窗）上完整时间窗的词错误率（%，越低越好）。表 3 汇总了这两个集合。 †仅统计方法的人脸预处理成功的时间窗（VisualVoice为490和755个，AlignDiT为555和760个）。

	TalkVid						VoxMM					
方法	B−10	B−5	已见说话人	未见说话人	Avg		B−10	B−5	已见说话人	未见说话人	Avg	
噪声输入	87.2	49.5	42.1	46.0	102.9	65.5	95.0	55.3	51.7	49.4	106.2	71.5
Auto-AVSR（V）	29.3	29.3	29.3	29.3	29.3	29.3	63.8	63.8	63.8	63.8	63.8	63.8
Auto-AVSR（AV）	66.0	32.6	55.8	47.0	84.0	57.1	93.6	58.6	78.0	71.0	100.4	80.3
USR 2.0（V）	30.4	30.4	30.4	30.4	30.4	30.4	59.7	59.7	59.7	59.7	59.7	59.7
USR 2.0（AV）	46.6	22.9	28.2	21.6	104.7	44.8	73.8	33.0	40.6	34.3	104.0	57.1
Llama-AVSR（V）	35.2	35.2	35.2	35.2	35.2	35.2	75.8	75.8	75.8	75.8	75.8	75.8
Llama-AVSR（视听）	37.2	21.2	25.2	19.1	85.0	37.5	72.6	43.7	34.6	31.0	94.3	55.3
DeepFilterNet3	87.8	64.0	48.7	41.0	101.4	68.6	95.9	75.4	53.6	45.0	103.8	74.7
Sidon	98.9	88.6	57.2	61.6	102.0	81.7	99.8	96.1	64.6	67.0	103.9	86.3
VisualVoice†	69.8	31.3	24.9	24.8	15.7	33.3	87.4	48.6	31.1	32.0	24.0	44.6
LipVoicer	33.2	33.2	33.2	33.2	33.2	33.2	67.0	67.0	67.0	67.0	67.0	67.0
AlignDiT（R）†	30.5	30.5	30.5	30.5	30.5	30.5	65.0	65.0	65.0	65.0	65.0	65.0
我们的方法（Auto-AVSR 作为 R）	24.1	17.8	14.8	13.9	14.1	16.9	58.2	35.5	27.2	24.2	27.8	34.6
我们的方法（USR 2.0 作为 R）	23.4	16.3	14.6	13.5	13.7	16.3	52.0	34.0	24.8	23.0	26.5	32.1
干净录音	8.5	8.5	8.5	8.5	8.5	8.5	13.0	13.0	13.0	13.0	13.0	13.0

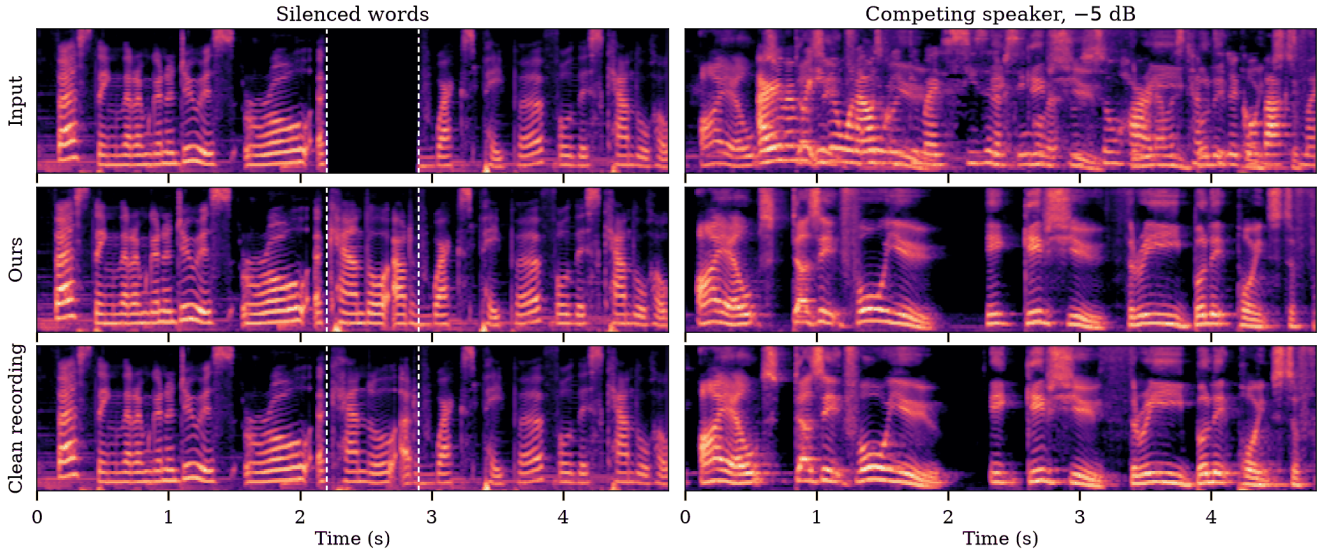

图 6. **两个测试示例** 的梅尔频谱图：输入、 的输出，以及 $R \to L \to G$干净录音。 左：图 5中被静音词“candle out”包围的剪辑片段，周围以虚线标示。 右：一个竞争说话人，音量为 −5分贝。Whisper 将输入转录为另一位说话人的句子，并将输出转录为参考内容：“雅思为你效劳。它会给你三个要点……”

即使嘴部难以辨认也是如此，因此完整流水线中剩余的大多数错误都发生在生成之前。这也有助于解释为何更强的唇语识别器无需重新训练就能改善结果（第D.1节）。

生成器在噪声环境下也遵循所提供的转录文本（表 17b）。给定噪声输入和参考转录文本， G使我们在每个集合上的选择的词错误率均有所降低；合并后约降低一半，在 TalkVid 和 VoxMM 上也是如此

其结果与干净录音仅相差约三个百分点（表 16）。困难集呈现出相同的模式，词错误率降低了一半。

生成器会利用嘴唇信息。 如果 G它只是说出给定的文本，那么它看到的嘴唇就无关紧要。我们通过仅更改其视频输入来检验这一点。对于每个符合条件的样本，替换片段取自同一源视频中与原片段不重叠、且不包含任何目标内容词的最近片段，

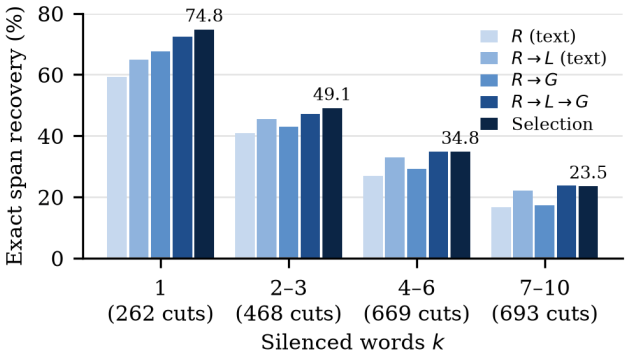

图7. **按静音词数量划分的精确片段恢复**，针对经核验的 TalkVid 和 VoxMM 剪辑片段。 R 和 $R \to L$ 根据其文本评分，另外三种方法则根据生成语音评分。Auto-AVSR 是唇读器。

表17. **给定真实转录文本**。(a) 每种方法均接收参考转录文本时，在经核验的静音片段上的精确片段恢复率（%）。(b) 含噪样本上的全窗口词错误率（%）：输入、我们的选择，以及给定参考转录文本的 G 。 †在该方法涵盖的110个剪辑片段上。

(a) 静音语音

方法	TalkVid	VoxMM	合并	Hard
Ours, G	**98.1**	**95.5**	**96.6**	**95.7**
F5-TTS [11]	96.2	92.8	94.2	90.5
AlignDiT [13]	94.3	88.4	90.9	92.7†
VoiceCraft [41]	86.7	82.5	84.3	87.9

(b) 含噪语音

Set	输入	选择	真实文本
TalkVid（570）	65.5	16.9	11.3
VoxMM（795）	71.5	34.6	16.5
合并（1,365）	69.0	27.2	14.3
困难（220）	54.1	29.6	15.2

停用词和填充词不计入这一排除条件。两个片段均来自同一源视频；此外，为排除其他说话人的干扰，我们要求 WavLM-Base-plus-SV 说话人余弦相似度至少为 0.86。音频、提供的转录文本、采样种子和所有推理设置均保持不变，替换片段中的嘴唇也仍在动，因此这项干预改变的是嘴唇所表达的内容，除此之外几乎没有别的变化。第三种输入则完全移除发音动作：将窗口的第一帧在整个窗口长度内重复播放，形成静止嘴部，正如视觉证据基线中的做法 e_R。通过这种第三种输入，我们可以区分正确的唇部动作带来的益处与唇部动作相互矛盾所造成的损害。实验涵盖了所有存在合适替换片段的样本：2,092 个经核验的静音片段中的 1,877 个，以及 914 个严重噪声样本；在这些样本中，多人嘈杂声以 -10 或 -20分贝的音量覆盖四个词或整个窗口。为了衡量变化在两种生成结果之间，我们还针对每项任务再次生成了 300 个样本，并为其配备各自的视频。

在转录文本保持不变的情况下，静止嘴部会在每种片段长度下都降低实际恢复率，并在每种严重噪声条件下提高词错误率；而不匹配的唇部动作带来的损失则更大（表 18）。在这些列中，转录文本仍来自正确的唇部动作，因此唇读器和语言模型对这两种损失都不起作用，二者均产生于生成器本身。静止嘴部下的损失表明，除了获得的词语之外，正确的唇部动作还能为生成器提供帮助。不匹配的唇部动作带来的额外损失表明，当嘴唇与词语相矛盾时，生成器对这些词语的遵循程度会进一步降低。在所有剪辑片段和所有严重噪声样本中，不匹配的唇部动作也显著劣于静止嘴部。嘴唇的影响在严重噪声下最大。当多人嘈杂声以 -20分贝覆盖整个片段时，生成器只有在视频正确的情况下才会根据提供的转录文本生成语音；在静止嘴部和不匹配的唇部动作下，它都会完全失败。使用同一样本自身的视频进行两次生成，结果最多相差 0.1 分，因此这些差异都不是抽样变异造成的。

没有转录文本时，只要有可恢复的内容，生成器就会依赖嘴唇动作。在较短的区间内，静止嘴部会使恢复效果减半，而不匹配的唇部动作造成的影响更大。即使视频正确，较长区间的恢复效果也接近于零；而整个窗口中的 -20 分贝多人嘈杂声已经超过单独使用 G 的情况，因此两者都无法再进一步下降。当唇读器也读取不匹配的唇部动作时，严重噪声下的损失会再次扩大（所有样本高出 $+40.9$ 个百分点，置信区间为 $[35.0, 48.6]$）。因此，视觉信号通过识别和生成两个环节进入处理流程。

D.6. 评分检查

我们检查比较结果是否受到静默剪辑片段验证筛选器或用于评估生成语音的自动语音识别模型的影响。第二次自动语音识别比较中的输出和选择结果保持不变。

所有剪辑片段。主要分析使用已验证片段，因为 Whisper 在干净音频上的错误会混淆恢复评分（第A.3节）。作为敏感性检查，表19 对除 54 个参考转录文本不正确的片段以外的所有剪辑片段进行评分。每种方法的得分都下降，但与基线方法的差异仍在已验证片段所得结果的 1.5 分以内，且显著性比较结果相同。

独立自动语音识别评分。 我们使用 Parakeet 替代 Whisper，对相同的输出重新评分，同时保持生成音频和选择决策不变。在含噪语音上，原始比较中的 15 种方法仍保持原有排名，我们的方法依然位居第一（表 20）。其中有五种方法是识别器，其文本直接接受评分。按照构造方式，这些方法在两种转录器下的得分完全相同，因此，这项检验对另外十种输出为

表18。**生成器从嘴唇中获取的信息**。仅提供给 G 的视频会发生变化：正确视频（正确）、静止嘴部或不匹配的唇部动作。相同转录文本：向 G 提供从正确的唇部动作获得的转录文本（(a) 中为 $R \to L \to G$，(b) 中为 $R \to G$）。静止嘴部和不匹配的唇部动作两列给出相对于正确视频的差值（以百分点计）及其对应的 95%区间。

		仅有 G			相同转录文本		
	n	正确	静止嘴部	不匹配的唇部动作	正确	静止嘴部	不匹配的唇部动作
(a)静音语音：精确片段恢复（%，越高越好），按静音词数量划分							
所有剪辑	1,877	7.1	-3.6 [$-4.9, -2.5$]	-4.6 [$-6.1, -3.4$]	38.8	-2.2 [$-3.2, -1.4$]	-6.3 [$-9.1, -4.3$]
1–3 个词	657	19.8	-9.9 [$-13.4, -6.8$]	-12.8 [$-16.6, -9.3$]	56.6	-3.2 [$-4.5, -1.8$]	-4.7 [$-6.3, -3.2$]
4–6 个词	598	0.5	-0.5 [$-1.2, 0.0$]	-0.5 [$-1.2, 0.0$]	34.4	-1.8 [$-3.6, -0.4$]	-7.7 [$-11.7, -4.8$]
7–10 个词	622	0.0	0.0	0.0	24.1	-1.6 [$-2.9, -0.5$]	-6.6 [$-11.5, -3.1$]
(b) 严重噪声：按受影响部分和多人言语噪声级别划分的全窗口词错误率（%，越低越好）							
所有样本	914	66.5	$+10.4$ [6.2, 15.0]	$+16.9$ [13.6, 21.1]	40.9	$+21.2$ [17.2, 26.0]	$+31.6$ [26.6, 38.0]
片段，-10分贝	234	81.7	$+24.4$ [16.2, 34.1]	$+34.0$ [26.1, 43.5]	47.1	$+24.3$ [19.1, 29.6]	$+55.3$ [46.5, 65.3]
片段，-20分贝	234	116.0	-2.3 [$-8.2, 4.3$]	$+1.1$ [$-3.5, 6.1$]	70.4	$+45.0$ [33.6, 58.4]	$+45.3$ [35.2, 58.6]
4个词，-10分贝	223	26.2	$+15.0$ [9.6, 22.7]	$+22.4$ [18.6, 27.1]	18.9	$+4.7$ [2.9, 6.5]	$+10.2$ [7.8, 12.9]
4个词，-20分贝	223	39.1	$+4.4$ [2.3, 6.7]	$+9.9$ [6.9, 13.4]	25.4	$+9.6$ [7.2, 12.4]	$+13.5$ [10.6, 17.3]

表19。 **未经过验证过滤器的静音语音**。 在全部2,328个TalkVid和VoxMM剪辑片段上的精确片段恢复（%），以及我们的方法与每个基线方法之差及其95%区间。粗体表示区间不包含零。表 13a给出了在2,092个已验证片段上的相同比较结果。

方法	精确	Δ [95% 置信区间]
我们的方法（Auto-AVSR 作为 R）	35.6	
Auto-AVSR	28.0	$+\mathbf{7.6}$ [5.5, 9.9]
USR 2.0	32.7	$+2.9$ [$-0.9, 7.0$]
Llama-AVSR	28.1	$+\mathbf{7.5}$ [2.3, 12.8]
LipVoicer	25.8	$+\mathbf{9.8}$ [6.0, 14.4]
VoiceCraft (L)	29.2	$+\mathbf{6.4}$ [5.0, 8.0]
F5-TTS (L)	32.3	$+\mathbf{3.3}$ [2.2, 4.3]
AlignDiT (L)	32.2	$+\mathbf{3.4}$ [2.2, 4.7]

表20。**噪声语音的独立自动语音识别检查**。在 Whisper 和 Parakeet 下，对 1,365 个时间窗的词错误率（%）进行评估，同时固定生成的输出和选择结果。 Δ：在 Parakeet 下，我们的方法减去该方法所得的差值及其 95%区间。 †在该方法覆盖的时间窗上进行比较。

方法	Whisper	Parakeet	Δ [95% 置信区间]
我们的方法，筛选	**27.2**	**28.8**	–
$R \to G$	29.0	29.3	-0.5 [$-1.0, 0.0$]
$R \to L \to G$	29.5	29.7	-0.9 [$-1.5, -0.4$]
VisualVoice†	40.1	39.3	-10.4 [$-13.4, -7.7$]
USR 2.0 (V)	47.4	47.4	-18.6 [$-23.7, -14.0$]
AlignDiT†	50.5	50.6	-21.9 [$-26.5, -17.2$]
LipVoicer	52.9	54.0	-25.1 [$-29.8, -20.3$]
噪声输入	69.0	62.5	-33.7 [$-37.7, -30.3$]
DeepFilterNet3	72.2	74.6	-45.8 [$-50.4, -41.8$]
Sidon	84.4	85.1	-56.3 [$-60.5, -52.8$]
SAM Audio，文本提示	67.1	67.2	-38.4 [$-43.0, -34.2$]
SAM Audio，视觉提示	75.4	75.2	-46.4 [$-50.9, -41.6$]

语音的方法具有参考意义。在这些方法中，我们的选择方法相较于外部基线方法的领先优势依然存在。在 Parakeet 下，其相较于固定的 $R \to G$并未显著领先，而相较于固定的$R \to L \to G$ 所取得的提升仍具有显著性。

在条件网格上，Parakeet 证实，相较于任务路由器，$R \to L \to G$对静音语音使用$R \to G$，并在其他情况下使用（表 21a；Whisper 下的网格见表 12）。在整段片段受到 -20分贝的多人嘈杂声干扰时，选择机制会在 268 个时间窗中的 236 个时间窗静音输入，并显著优于该任务路由器。在 Parakeet 下，该方法与第 C 节所述的回顾式条件逐项规则无显著差异。如该节所述，该规则使用每种条件下的测试结果，但它并非表 10 中的逐样本预言机。

对于静音语音，Parakeet 评分在两个已验证剪辑片段集合上保留了相对显著性模式（表 21b）。具体而言，USR2.0 与我们的选择仍无显著差异（差值 $+2.5$ 分，置信区间 $[-1.7, +7.0]$，基于 Parakeet 验证的剪辑片段）。文本基线方法的分数在不同列之间有所变化，这是因为已验证剪辑片段发生了变化，并非因为它们的输出经过了重新转录。

与听众的一致性。第E.1节的听力研究提供了一项无需依赖任何语音识别器的检验。听众对词语的偏好与此处报告的词错误率差异一致，在该研究的 120 个时间窗上，斯皮尔曼相关系数为 0.71。

干净输入。我们还测试了修复是否会改变完整语音。对于条件网格上的每个干净片段，选择都会保留输入音频（第C节）。在这 268 个窗口中，其输出的词错误率为 10.1%，而原始录音为 9.9%。这 0.2 个百分点的差异并不显著（置信区间 $[-0.6, 1.0]$），表明在这个干净集合上，可懂度没有可测量的损失。

D.7 示例说明

图 8 展示了四个测试示例：两个分别包含两个词和十个词的静音片段、一位竞争说话人，以及一位面部处于阴影中的困难集说话人。在每个示例中，



图 8. **四个测试示例**。 每张卡片展示一个视频帧、输入和我们的方法的输出的响度（$R \to L \to G$），以及两份转录文本。虚线框标出静音区间。仅嘴唇 表示唇读器给出的最佳假设，仅声音 表示 Whisper 对输入的转录文本，错误的词以红色标出。我们的方法 表示 Whisper 对我们的方法输出的转录文本，修复后的词语已高亮显示。

仅凭嘴唇或仅凭声音都会得出错误的词，而我们的输出与参考转录一致。图 6 以频谱图的形式展示了第一个和第三个示例。

在静默示例中，生成的区间与两侧可听见的语音自然衔接，没有明显接缝，同时保留了可听见的部分。对于同一剪辑片段，仅使用生成器，以及使用生成器和唇读器的假设，生成的语音分别被转录为“keletor of wax paper”和“gullet of wax paper”。只有使用语言模型的转录文本时，才会得到“candle out of wax paper”。在含噪示例中，竞争说话人的声音压过了输入，以至于 Whisper 转录了对方的句子，而不是目标说话人的句子。输出去除了该说话人的声音，并遵循干净录音的时间安排，包括其中的停顿。

D.8. 无数字静音的间隙

在主要评估中，被移除的区间是数字静音，而第 B.1 节的振幅阈值几乎能完美地定位它。此处我们测试处理流程是否依赖间隙为数字静音。我们使用相同的静音包络，并用三种音频之一填充该区间。室内底噪由区间外最安静的 5% 的 32 毫秒帧拼接而成，并采用 5 毫秒交叉淡化。舒适噪声是粉红噪声，其电平为区间外帧均方根电平的第 10 百分位数。衰减语音是将原始语音降低 40 分贝后的音频，这是最容易的情况，因为微弱语音按构造仍保留在输入中。这三种填充内容都是合成的对现有测试窗口进行的损坏，而非新的录音。

检测器。 我们将间隙定义为这样一个区间：音频中没有语音，而嘴部仍在进行发音动作。Silero VAD v5 提议使用阈值 0.7，并将至少持续 0.256 s 的内部非语音片段作为候选。来自 AV-HuBERT 特征上的视素头的视觉发音分数，$1 - P$(空白) $- P$(静音)，经过七帧平滑处理。检测器选择平均发音动作分数最高的候选，并且仅当该平均分数至少为 0.4 时才判定为间隙。作为消融实验，纯音频检测器选取最长的非语音片段（阈值 0.2，至少持续 0.448 s）。所有参数均根据 373 个验证剪辑和 99 个未损坏的验证窗口进行选择，以在三个填充词之间最大化平均检测率，同时将未损坏窗口上的误报率控制在不超过 5%；在任何测试运行前，这些参数均已固定。所选工作点的验证误报率为 4.0%，而没有任何纯音频设置满足该限制（其所选工作点的误报率为 9.1%）。真实区间仅用于构造损坏数据和进行评分，从不作为输入。

定位。 在经验证的测试剪辑上，检测器在两个集合中都能找出五分之四以上的这些间隙，中位边界误差不超过 70 毫秒（表 22）。在未受损的窗口中，TalkVid 的 113 个窗口里有 2.7%、VoxMM 的 153 个窗口里有 9.2% 被判定存在间隙。

表21。**针对可变损坏和静音语音的独立自动语音识别检查。**（a）Parakeet 下的网格词错误率。 Δ：选择结果减去比较基准。最后两行是整个片段的 −20分贝多人嘈杂声条件。（b）精确片段恢复（%）：所有语音输出均由 Parakeet 转录，所依据的剪辑片段则通过 Whisper 或 Parakeet 从干净音频中验证。 因此，语音输出的得分不同于表 13a。

（a）条件网格（3,419 个样本）

Rule	词错误率	Δ [95% 置信区间]
选择任务路由器	**24.2**	–
始终 $R \to G$	26.3	−2.1 [−2.5, −1.8]
始终 $R \to L \to G$	26.4	−2.2 [−2.6, −1.8]
按条件回溯	26.6	−2.5 [−2.8, −2.1]
选择任务路由器，−20分贝多人嘈杂声	24.3	−0.2 [−0.4, +0.1]
−20分贝多人嘈杂声	51.1	–
	69.5	−18.4 [−21.8, −15.3]

(b) 静音片段，Parakeet 评分

方法	Whisper 验证 (2,092)	Parakeet 验证 (2,066)
Ours, 选择	**38.3**	**38.9**
F5-TTS	35.8	36.1
USR 2.0	35.8	36.3
AlignDiT	35.4	35.9
VoiceCraft	31.8	32.0
Auto-AVSR	30.7	30.9
Llama-AVSR	30.4	30.9
LipVoicer	27.2	27.3

表22。**无数字静音时的间隙定位。**当声明的区间与真实区间的交并比至少为0.5时，即发现间隙。交并比和边界误差（边缘）是对声明区间计算的中位数。数字零是该检测器的参考标准，与主要评估中的振幅阈值相互独立。

Set	填充词	检出率（%）	中位交并比	边缘（毫秒）
TalkVid	数字零	91.2	0.966	13
	室内底噪	82.3	0.880	61
	舒适噪声	84.6	0.859	70
	衰减语音	89.6	0.953	20
VoxMM	数字零	90.0	0.966	16
	室内底噪	84.2	0.900	56
	舒适噪声	84.4	0.882	70
	衰减语音	85.1	0.945	20

恢复。 随后，我们根据检测到的区间构建可听上下文，并运行主要实验中的语言模型，其他设置保持不变。由于检测到的区间仅用于构建该上下文，因此我们对语言模型返回的转录文本进行评分。使用真实区间时，这会以 0.1 分以内的误差复现 $R \to L$ 表 5a 的结果。漏检间隙按未恢复计分，因此这些数字是下限。

每当判定存在一个间隙时，对于每种填充词和两种检测器，恢复值与使用真实区间时的结果相差不超过半个百分点（表 23，最后一列）。因此，相对于真实区间的损失来自漏检间隙，而非边界不精确。此

表23. **检测到间隙时的恢复。**在2,092个已验证片段中，语言模型转录文本的精确片段恢复率（%）。已判定。：判定出间隙的剪辑片段占比（%）。漏检间隙计为未恢复。 Δ：与真实区间的差值及其95%区间。已判定时。：对判定出间隙的剪辑片段，使用检测区间和真实区间计算恢复率。第一组将真实区间移动100毫秒。

Gap	检测器	声明。精确	Δ [95% 置信区间]	声明时。
数字零	真实区间	–	36.1 –	–
	边缘向外	–	35.6 −0.5 [−1.5, +0.5]	–
	边缘向内	–	36.1 +0.0 [−0.7, +0.8]	–
	向后移	–	36.6 +0.5 [−0.3, +1.2]	–
房间底噪	音频 + 嘴唇	88.1	30.4 −5.6 [−8.0, −3.8]	34.6 / 34.6
舒适噪声	音频 + 嘴唇	87.9	29.6 −6.5 [−8.9, −4.6]	33.7 / 34.2
衰减	音频 +嘴唇	92.0	33.2 −2.9 [−5.1, −1.1]	36.1 / 35.8
室内底噪	音频	76.7	23.8 −12.3 [−15.5, −9.6]	31.0 / 31.1
舒适噪声	音频	79.9	24.8 −11.3 [−13.9, −8.9]	31.1 / 31.2
衰减	音频	85.7	29.2 −6.9 [−8.7, −5.0]	34.0 / 33.6

外，将两个真实边界分别向外或向内移动 100 毫秒，或将区间平移 100 毫秒，恢复值的变化均不超过半个百分点；而检测器的边界误差中位数还小于这一数值。因此，定位间隙并不依赖数字静音。

嘴唇也有助于找到间隙。在相同的剪辑片段上，读取嘴唇的检测器比仅音频检测器多找回 4 到 7 个区间，且每个区间都不包含零。

D.9. 语言模型从唇读器获取了什么

语言模型会看到可听上下文和唇读器的 N-最佳列表。为区分两者，我们只更改它看到的列表，其余的 $R \to L$表 5a 行内容保持不变：剪辑片段、上下文、间隔时长和解码方式。我们移除该列表，将其替换为来自另一源视频中一个窗口的列表（该窗口的最优假设词数相近），或者将其缩短为前一个或前五个假设（表24）。

表24. **向语言模型展示的假设。** 在经核验的静音片段上，对 $R \to L$ 转录文本的精确恢复率（%）。Δ：与完整列表相比的差值，以百分点计，并给出其95%区间；若区间不包含零，则以粗体显示。

向 L展示	TalkVid + VoxMM %	Δ [95% 置信区间]	Hard %	Δ [95% 置信区间]
仅上下文	3.2	**−33.0** [−43.7, −25.0]	23.3	+0.9 [−14.0, 13.9]
其他视频的	2.3	**−33.9** [−44.4, −24.9]	17.2	−5.2 [−17.9, 5.2]
h_1 仅	32.0	**−4.2** [−6.5, −2.1]	22.4	0.0 [−5.2, 5.0]
前五个	35.0	−1.1 [−2.4, 0.0]	25.9	+3.4 [0.0, 7.9]
全部十个（我们的）	36.2		22.4	
唇读器	30.7	**−5.5** [−7.6, −3.3]	8.6	**−13.8** [−21.7, −5.1]

在 TalkVid 和 VoxMM 上，没有唇读器的帮助，语言模型无法补全区间。仅凭上下文，

或者借助另一段视频的假设，它几乎什么也无法恢复，效果远不及唇读器本身。它也不会照搬最优假设：仅有 h_1 时，它的结果仍与唇读器的结果相近；随着候选列表变长，效果提升也随之增加，其中大部分提升在提供五个假设时就已实现。在完整列表中恢复的区间里，有 13% 在排名较低的假设中正确、但在 h_1 中不正确；另有 9% 在十个假设中都没有出现。因此，语言模型会将列表视为候选项，并结合上下文从中选择和编辑。

正如其构建方式所预期的那样，困难集的表现有所不同。其中的嘴部画面难以辨认，甚至无法辨认，因此假设所含信息很少：仅凭上下文，语言模型恢复出的区间数量与使用完整列表时一样多；它恢复出的区间中，有 65% 未出现在十个假设中的任何一个。因此，同样是仅输入上下文，在 TalkVid 和 VoxMM 上会失效，但在上下文信息丰富的情况下却能奏效；这说明上述性能下降源于缺少唇部信息，而非输入形式所致。这与第5.4节报告的表现一致：嘴唇无法辨认时，上下文会起到补偿作用。

E. 语音质量、说话人相似度与唇部同步

表25 通过生成语音的质量补充了词级结果。我们使用 DNSMOS [46]和 UTMOS [48], 衡量自然度，使用 LSE-C 和 LSE-D [15, 42], 衡量唇部同步，并以 ERes2Net [10] 说话人嵌入的余弦相似度衡量说话人相似度，比较输出与同一窗口的干净录音。ERes2Net 是 Just-Dub-It 所使用的说话人模型 [6]。

相比 WavLM 的说话人嵌入，我们更倾向于使用它，因为在我们的数据上，它区分说话人的效果清晰得多。对于成对的干净测试录音，同一视频的两个时间窗的 ERes2Net 平均相似度为 0.61，不同视频的时间窗则为 0.05；而 WavLM 的结果分别为 0.89 和 0.59，且使用 WavLM 时，5% 的不同说话人配对得分高于 0.87。因此，WavLM 相似度接近 0.9，并不能说明两段录音具有相同的语音。

对于静音语音，各方法的差异主要集中在缺失区间附近，因此我们针对该区间测量说话人相似度，统计对象为区间持续时间至少为 1 秒的 1,586 个剪辑片段。在整个窗口上，F5-TTS 与我们的方法几乎相同（分别为 0.878 和 0.879），因为每个窗口的大部分内容都是两者都能复现的可听语音。在该区间上，我们的每一行结果都高于两种编辑器，而且差异具有显著性（相对于 F5-TTS 的选择的置信区间 $[0.009, 0.029]$ ，这是最接近的一对）。作为衡量该指标尺度的参考，干净行将干净录音的该区间与其自身窗口进行比较。在唇部同步方面，我们的输出也最接近干净录音，明显领先于两种编辑器；UTMOS 给出的评分高于两种编辑器和 LipVoicer，仅略低于干净录音。DNSMOS 对大多数

表25. **生成语音的质量**，在 TalkVid 和 VoxMM 上合并统计：自然度（DNSMOS 和 UTMOS，评分范围为 1 至 5）、与干净录音的说话人相似度（Spk.）以及唇部同步（LSE-C、LSE-D）。对于静音语音，Spk. 仅在静音片段上测量。粗体和下划线分别标示最佳和次佳方法。 †仅使用预处理成功的时间窗。

方法	DNSMOS↑	UTMOS↑	说话人↑	LSE-C↑	LSE-D↓
静默语音					
干净录音	2.54	2.77	0.839	6.38	8.07
输入（静音区间）	2.58	2.62	–	4.07	10.39
LipVoicer [56]	2.60	2.32	0.102	5.10	9.49
F5-TTS [11] (L)	**2.71**	2.54	0.597	4.70	9.70
AlignDiT [13] (L)	2.57	2.62	0.528	5.49	8.95
Ours, $R \to G$	2.69	2.68	0.624	**6.21**	**8.32**
我们的方法，$R \to L \to G$	2.70	2.69	**0.629**	6.16	8.36
我们的方法，筛选	**2.71**	**2.71**	0.616	6.20	8.33
含噪语音					
干净录音	2.57	2.78	–	6.38	8.05
噪声输入	1.33	1.38	0.492	2.36	11.02
DeepFilterNet3 [49]	2.14	1.54	0.382	3.07	10.93
Sidon [40]	2.47	2.60	0.185	2.26	12.37
VisualVoice [18]†	1.63	1.46	0.654	4.35	9.30
LipVoicer [56]	2.60	2.32	0.124	5.08	9.48
AlignDiT [13] (R)†	2.75	**3.10**	0.525	6.31	8.27
Ours, G 单独	**2.82**	2.70	0.664	**6.33**	8.17
Ours, $R \to G$	2.61	2.58	**0.698**	6.30	**8.15**
Ours, 选择	2.67	2.70	0.660	6.24	8.28

生成的输出在两项任务中甚至都高于干净录音，因此我们将其与 UTMOS 结合起来解读。

在噪声语音上，我们的固定配置 G单独 $R \to G$的 DNSMOS 和唇同步得分最高。选择方法和 AlignDiT 紧随其后，而 AlignDiT 是唯一一个 UTMOS 更高的方法。这些分数衡量语音听起来是否自然、同步是否良好，但它们都无法衡量其中的词是否正确。AlignDiT 用干净的参考语音读出唇读器的假设内容，其 UTMOS 评分高于干净录音本身，但在相同时间窗上，其词错误率比选择方法高 23.4 个百分点（表 13b）。它还会丢弃噪声输入，并且其说话人相似度低于我们每一行结果。选择方法的说话人相似度略低于 $R \to G$。这一差异集中在选择方法将输入静音的时间窗中：这些时间窗的相似度为 0.191，而保留输入音频时为 0.727。在这些时间窗中，输入里已没有可用的语音，生成器会根据说话人的外貌推断出一种语音。该结果高于不同说话人之间的相似度（0.05），也高于同样仅从视频生成语音的 LipVoicer，但远低于保留音频的时间窗。因此，静音是在录制的语音与可懂度之间进行权衡。

表 26。**含噪语音听力研究**。 在每个基线和问题的 64 次判断中，16 名评定者中偏好我们的输出（我们的方法）、认为二者相同或偏好基线（基线。）的判断比例（%）。得分：从 -1 （基线）到 $+1$ （我们的方法）的平均偏好及其跨评定者的 95%区间；区间不含零的结果以粗体标示。

	相同词语					相同说话人					更清晰、更自然				
我们的方法 vs.	我们的方法	相同	基线。	得分	[95% 置信区间]	我们的方法，	相同基线。		得分	[95% 置信区间]	我们的方法，	相同基线。		得分	[95% 置信区间]
DeepFilterNet3	81.3	15.6	3.1	$+\mathbf{0.78}$	$[0.66, 0.89]$	87.5	10.9	1.6	$+\mathbf{0.86}$	$[0.73, 0.97]$	92.2	6.3	1.6	$+\mathbf{0.91}$	$[0.81, 0.98]$
Sidon	84.4	12.5	3.1	$+\mathbf{0.81}$	$[0.72, 0.91]$	90.6	7.8	1.6	$+\mathbf{0.89}$	$[0.78, 0.97]$	93.8	4.7	1.6	$+\mathbf{0.92}$	$[0.84, 0.98]$
VisualVoice	48.4	35.9	15.6	$+\mathbf{0.33}$	$[0.16, 0.52]$	60.9	29.7	9.4	$+\mathbf{0.52}$	$[0.28, 0.73]$	82.8	14.1	3.1	$+\mathbf{0.80}$	$[0.69, 0.91]$
LipVoicer	78.1	17.2	4.7	$+\mathbf{0.73}$	$[0.50, 0.89]$	92.2	6.3	1.6	$+\mathbf{0.91}$	$[0.81, 0.98]$	90.6	7.8	1.6	$+\mathbf{0.89}$	$[0.80, 0.97]$
AlignDiT (R)	53.1	31.3	15.6	$+\mathbf{0.38}$	$[0.19, 0.56]$	46.9	23.4	29.7	$+0.17$	$[-0.02, 0.38]$	45.3	21.9	32.8	$+0.13$	$[-0.05, 0.28]$
所有基线方法	69.1	22.5	8.4	$+\mathbf{0.61}$	$[0.52, 0.68]$	75.6	15.6	8.8	$+\mathbf{0.67}$	$[0.58, 0.74]$	80.9	10.9	8.1	$+\mathbf{0.73}$	$[0.68, 0.78]$

E.1. 听力研究

到目前为止，所有结果都依赖自动指标。为了检验这些指标是否与人类感知一致，我们还针对噪声语音任务开展了一项听力研究，让评定者每次将我们选择方法的输出与一个基线方法的输出进行比较。

实验方案。 每个页面会呈现一个含噪测试窗口，其中包含使用相同视频的三个片段：作为参考的干净录音，以及两个修复音频，分别是我们的方法和一个基线方法的输出。它们在各页面中的顺序经过平衡处理，且不会向评定者透露。评定者需要回答三个问题：“哪一个说的是与原始录音相同的词语？”“哪一个听起来更像是同一个人？”以及“哪一个听起来更干净、更自然？”每个问题的答案可以是两个修复音频中的一个，或“差不多相同”。

该研究涵盖 TalkVid 和 VoxMM 的 120 个时间窗，五种基线方法各占 24 个，涉及表 3所列的噪声条件，以及第 D.2 节所述的多人 babble 噪声源。在这些时间窗上，我们的方法的输出词错误率为 30.7%，略高于完整含噪声集上的 27.2%，因此，与整体评估相比，该样本对我们的方法并不更容易。16 名评定者各自评估了 20 个时间窗，且全部通过了两项注意力检查。由此得到 960 条判断，表 26中的区间通过对评定者进行重采样获得。所有评定者都在研究首页表示同意，他们的匿名回答仅用于计算此处报告的统计数据。由于本研究不记录任何身份识别信息，且仅收集听觉偏好，因此无需进行伦理委员会审查。

结果。 听众在三个问题上都更偏好我们的方法的输出（表 26），而且16名评定者平均而言都更偏好该输出。相较于两个增强模型和 LipVoicer，这一偏好最为明显。在词语方面，VisualVoice 是最接近的基线，这一点与其在词错误率（表 13b）中的表现一致。AlignDiT 是唯一一个在语音和质量方面获得听众与我们的方法同等评价的基线，两者的区间都包含零。它获得了说话人的干净录音（表 7），并以该说话人的语音进行说话；表 25 中的自动评分也对其评价很高。在词语方面，听众同样更偏好我们的输出。

听众的判断也与自动评分一致。在 120 个时间窗中，对词的平均偏好与基线和我们的输出之间的词错误率差异相关（斯皮尔曼 $\rho = 0.71$）。由于这些判断不依赖任何语音识别器，因此这一一致性支持了全文采用的基于 Whisper 的评分方法。

F. 局限性与负责任使用

评估范围。 我们在每个包含一个受损区域的4.8秒时间窗上进行评估，并且提供给语言模型的上下文是同一时间窗中可听见的部分。较长的录音可以逐个时间窗进行处理，但我们没有研究如何将较早时间窗中的上下文向前传递。损坏是合成的，因此隐藏的词语已知晓并且可以进行评分。第 D.8节将其扩展到使用室内底噪或舒适噪声填充的间隙，但不涵盖因丢包、编解码器或混响造成损坏的录音。除英语之外，我们还在规模较小的集合上测试了西班牙语和法语（表 5b），而任何其他语言都需要使用该语言的唇读器。

对组件的依赖。 我们的方法继承了各组件的局限性。当音频和上下文都不包含可用信息时，例如片段被响亮的多人嘈杂声完全覆盖，结果便受唇读器的可读能力所限（表12）。当嘴唇难以辨认时，选择也可能跟随错误的唇读结果，例如在静音困难集上（第 5.6节）。由于唇读器被冻结且可以替换，更强的唇读器无需重新训练即可提升结果（表2）。

语音。 当选择过程舍弃录制的音频时，输入中便不再有可用语音，生成器会根据说话人的外貌推断语音（第 E节）。词语得以还原，但生成的语音并非录音中的那种语音。如果有相同说话人的干净录音，可以将其用作这些输入的语音参考；我们将此留待未来工作。

成本。 该选择过程每个窗口运行生成器六次（第 B.2节），因此我们的方法适用于离线恢复，而不适合实时使用。

单个固定配置只需一次处理，$R \to L \to G$在标准测试集的两个任务上都接近该选择过程（表 5a），从而提供了成本更低的工作点。

负责任地使用。 我们的方法会为真实人物生成语音，而修复后的词语可能有误：输出是重建内容，而非录音。因此，我们认为它是一种用于修复受损录音、档案影像和通话的工具，但须征得其中相关人员的同意。凡展示其输出之处，都应标注为修复内容；如果已知修复区间，还应注明该区间。输出不应被视为某人说过什么的证据。生成器所说的文本源自说话人自己的唇部运动和可听上下文，因为我们的目标是还原说过的话，而不是为某个个人生成新的语音。