MAS-BIT：用于端到端脑信号转文本解码的多智能体系统

Xinjie Zhang[1, *] **Linyang He**[1, *, †] **Yizi Zhang**[2] **Yuesheng Ma**[1] **Nima Mesgarani**[1]1哥伦比亚大学 2斯坦福大学
*共同贡献者 †项目负责人{xinjie.zhang, linyang.he}@columbia.edu nima@ee.columbia.edu

arXiv:2611.29457v1 [q-bio.NC] 22 Nov 2026

摘要

语音脑机接口旨在通过将神经活动转换为文本来恢复通信能力，但长期以来，端到端解码器的准确率一直落后于级联系统。我们提出脑信号转文本多智能体系统（MAS-BIT），这是首个基于皮层内记录进行端到端脑信号转文本解码的多智能体框架。MAS-BIT 协调三个专门的智能体：用于解读连续神经表征的神经智能体、用于解码软音素后验概率的语音智能体，以及仲裁智能体；后者可直接访问神经表征，整合互补证据以生成最终转录文本。专家智能体通过基于词元嵌入的潜在通信与仲裁器进行通信，将消息构建为词元嵌入的概率加权组合。这些消息保留备选词元假设，并提供可微接口，通过该接口，句子级监督可在不同参与者之间联合训练神经编码器和专用智能体。MAS-BIT 在 Brain-to-Text '24 和 '25 基准测试中实现了最先进的端到端解码性能，在 '24 上的词错误率为 6.929%，在 '25 公开测试集上的词错误率为 3.606%（私有测试集上为 3.892%）。这些结果超过了已发表的 BIT 级联在 '25 公开集上的表现，并缩小了在 '24 上的差距，且无需集成独立训练的模型，也无需使用外部语言模型重打分。[1]

1 引言

语音脑机接口（BCI）旨在通过将与尝试发声相关的神经活动转换为文本来恢复通信（Akbari 等人，2019年；Moses 等人，2021年；Metzger 等人，2023年；Willett等，2023年；Card等，2024年；Wairagkar 等人，2025年）。准确的解码需要从噪声录音和有限的配对训练数据中恢复语言内容。级联系统将神经信号到音素解码器与外部 n-gram/WFST 解码（以及可选的重打分）相结合，以重建句子。相比之下，端到端系统将神经表征直接连接到预训练语言模型，并针对目标转录文本进行优化（Feng 等人，2024年；Zhang 等人（2026））。

在 Zhang 等人（2026）报告的 Brain-to-Text基准测试比较中，端到端解码的准确率仍低于级联解码。

尽管如此，从建模和部署的角度来看，端到端解码仍然颇具吸引力。通过将神经活动直接映射到文本，它避免了独立的音素到词阶段所带来的信息瓶颈和错误传播，直接将神经表征与句子级目标对齐，并且无需固定的发音词典或经过大量调优的搜索启发式方法，即可利用预训练语言模型。此外，传统的基于 WFST 的级联系统依赖于独立的解码堆栈，其中包含大型 n 元语法/WFST 图以及外部语言模型重评分，这可能带来数百 GB 的主机内存开销（表 2）；相比之下，端到端解码器提供了一个统一且显著更轻量的推理流水线。因此，我们希望探究：在保留这些优势的同时，端到端解码能否达到或超越级联系统准确率。

[1]代码可在此处获取，模型权重可在此处获取。

神经编码器为同一段录音提供了双视图：连续隐藏表征，以及通过监督式音素读出获得的音素后验概率。隐藏表征将学习到的神经特征呈现给语言解码器，而后验概率则围绕音素标识及其相对概率来组织这些证据。语音识别中的基于后验概率的接口展示了如何利用此类结构化分布来为语言模型提供条件（Ma et al., 2025）。对于脑信号转文本解码而言，这两种视图体现出一种分工：一名专家解读连续神经表征，另一名专家解读明确的语音学证据。尽管二者都源自同一神经观测，但不同的输入可能导致不同的转录假设。我们将研究这些专门化模块能否学会为共同的最终转录文本提供互补证据。

通信接口决定专家判断中的哪些部分能够传递给最终解码器。传递单个生成的文本序列虽然能表达选定的词元，却会丢弃每个位置上这些词元相对于其他备选词元的支持度。离散选择还会阻断消息向生成该消息的专家概率进行常规反向传播。关于可微通信以及语言模型智能体之间连续交换的先前研究，为这种文本接口提供了替代方案（Foerster 等人，2016年；Pham 等人，2024年；Zou 等人，2026年）。对于神经解码，我们希望设计一种接口，既能传达有关相互竞争的词元假设的信息，又能让最终转录目标通过消息影响专家。

在此，我们提出一种端到端**多**智能体**系**统，用于**脑**到**文本**解码（**MAS-BIT**），将专家解码与仲裁分开。神经专家 读取连续的编码器表征，语音专家 读取软音素后验概率，而仲裁器 则将它们的消息与直接神经证据整合起来。各个角色共用一个语言模型主干网络，但使用不同的适配器。专家并非只传递选定文本，而是通过按概率加权的词元嵌入进行通信，从而保留备选假设，并使最终转录损失能够通过消息监督专家。参与者特定输入投影使两个以 BIT 初始化的编码器能够处理任一参与者的录音；随后通过分阶段专门化和联合优化，在单一端到端系统中共同训练编码器与智能体。
我们在 Brain-to-Text ’24 和 ’25 基准上评估了一个联合训练的模型，在 ’24 留出集上取得 6.929% 的 WER，在 ’25 公开／私有子集上分别取得 3.606%/3.892% 的 WER。在 ’24 留出集和 ’25 公开子集上，MAS-BIT 达到最先进的端到端性能，超越 BIT 的单模型解码器和端到端集成，且无需集成独立训练的模型或进行外部语言模型重评分（Zhang 等，2026）。它在 ’25 公开集上的表现也优于已发表的 BIT 级联，并缩小了在 ’24 上的差距；与复现的级联和 BIT + LightBeam 相比，它还大幅降低了内存占用和解码时间（Feghhi 等，2026）。受控比较表明，专门的专家模型优于同时接收两种视图的单个解码器（包括扩展容量的变体），也优于参数量和解码轮次相匹配的顺序细化对照；在验证集上，即使两份草稿都出错，仲裁器也能恢复出正确转录文本。这些结果表明，与将两种视图都输入单个解码器或在专家草稿之间进行选择相比，联合训练的专门模型能更有效地利用神经与语音证据。

2 相关工作

语音神经假体与端到端神经解码。 皮层内语音神经假体通常会先解码中间语音单位，再通过基于语言模型的句子重构进行还原（Willett等，2023年；Card等，2024年）。近期研究则将神经表征直接连接到预训练语言模型（Feng 等人，2024年；Zhang 等，2026）。具体而言，BIT 将跨任务、跨物种神经预训练与音素适配编码器和音频-大语言模型解码器相结合（Zhang 等，2026）。MoDAl 通过对齐损失和去相关损失结合并行神经编码器（Chen 和 Chin，2026），而 BrainWhisperer 则将预训练的 Whisper 编码器–解码器适配到神经输入（Boccato 等人，2026）。其他直接解码器包括使用 BART 词解码器（Olak 等人，2026）、将 Wav2Vec2 表征迁移到字符级 CTC 解码（Fiedler 等人，2025），或训练字符级 CTC Conformer（Khanday 等人，2026）。Fogg 等（2026）使用参与者特定输入投影，在六名参与者之间训练共享神经音素解码器。

基于后验概率的接口与基于信号的精化。 CTC后验概率为低层序列表征与语言模型提供了概率接口。LegoSLM 使用后验加权词元嵌入，将语音编码器连接到大语言模型，而无需确定

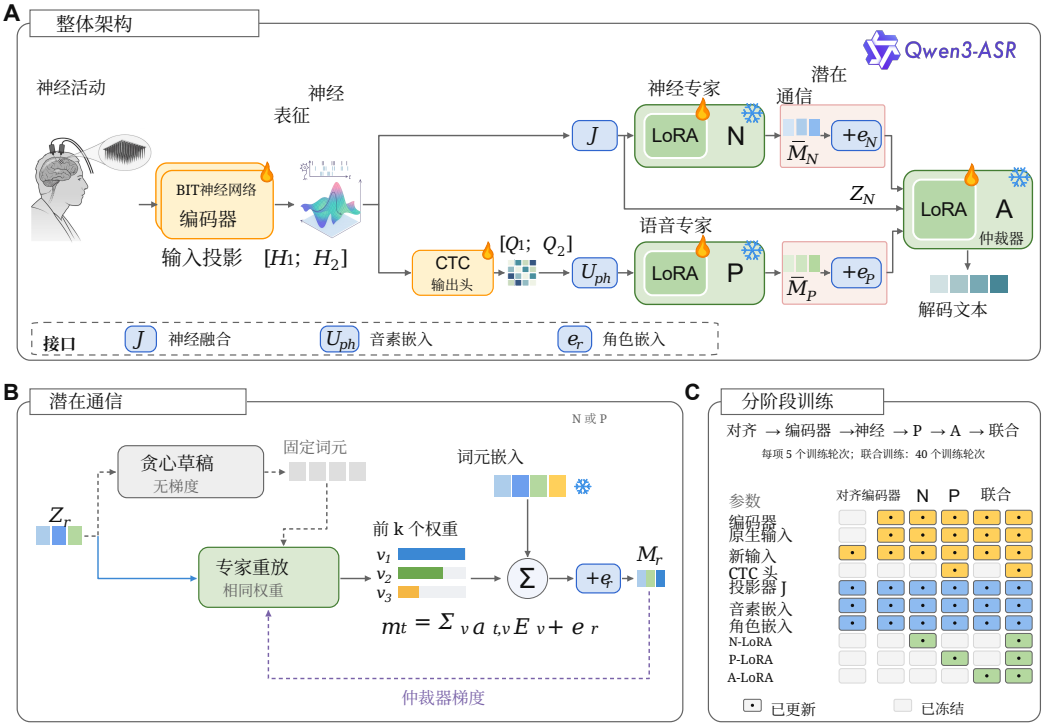

图1：*MAS-BIT* 通过专业化专家和仲裁器融合神经与语音证据。（**A**）两个以 BIT 初始化的编码器通过参与者特定输入投影处理每段录音。融合后的隐藏状态为神经专家和仲裁器提供条件；两个音素后验概率则为语音专家提供条件。三个语言智能体共用一个冻结的文本骨干网络，并各自使用独立的适配器。（**B**）每个专家都会重放其贪心草稿，根据概率加权词元嵌入构建潜在消息。 Z_r 表示其输入， Z_N 或Z_P，而 e_r 标识消息来源。仲裁器损失通过这些潜在消息训练处于激活状态的上游组件。（**C**）在联合优化的40个训练轮次之前，先进行五个训练轮次的对齐、编码器适配和专门化阶段。填充的单元格表示参数更新；冻结的下游权重仍保留输入梯度。

一优中间序列（Ma 等人，2025年）。审议系统在保留对原始信号访问权限的同时，对首轮假设进行精化（Hu 等人，2020年；Dekel 等人，2026年）。MAS-BIT 结合了这些思路：神经专家读取连续编码器状态，语音专家读取 CTC 后验概率，而仲裁器则将两者的消息与直接神经证据整合起来。

多智能体与可微语言模型通信。多智能体语言模型系统通过将提案生成与下游聚合分开来提升生成效果（Du 等人，2024年；Wang 等人，2025年）。连续通信进一步使信息能够通过可微表示传递，而不仅限于离散文本（Foerster 等人，2016年；Pham 等人，2024年；Ramesh 和 Li，2025 年；Zou 等人，2026 年）。MAS-BIT 采用固定的双专家单仲裁器拓扑，共用一个冻结的语言模型骨干网络，并配备角色专属的 LoRA 适配器（Hu 等人，2022年）。每个专家都专门处理其输入证据，并将生成的草稿重放为潜在消息，通过下游仲裁器损失进行优化。

3 方法

MAS-BIT 通过三个语言智能体将神经录音映射为句子：两个专门的专家——神经专家 N 和语音专家 P——以及一个仲裁器 A （图1）。专家读取同一录音的互补表示，并向仲裁器发送潜在消息。仲裁器将这些消息与直接神经证据相结合，生成最终转录。我们介绍共享骨干网络架构及其分阶段到联合训练流程。

3.1 架构

神经与语音证据。令 $X \in \mathbb{R}^{T\times C}$ 表示皮层内特征的话语，$y=(y_1, \ldots\ldots, y_U)$表示其包含序列结束词元的文本词元序列，$\pi$ 表示其音素目标。输入结合了阈值交叉活动和尖峰频带功率（Nason 等人，2020年）；我们在附录 A.1中介绍预处理。我们使用从已发布的 T12 和 T15 BIT 编码器及其匹配的音素读出初始化的两个神经 Transformer，处理每个录音（Zhang 等，2026年）。对于输入参与者 s，输入投影 $R_{j,\,s}$ 将神经补丁映射到编码器 $j \in \{1,2\}$中，生成 $H_j = f_{\theta_j}(R_{j,\,s}(X)) \in \mathbb{R}^{S\times d_n}$，其中 $S=\lceil T/5\rceil$和 $d_n = 384$。每个编码器都保留其为原始参与者提供的已发布输入投影，并为另一位参与者学习新的输入投影。两个编码器处理同一录音；三个语言智能体共享其输出。

两层投影器 J 将层归一化隐藏状态融合到语言模型的嵌入空间中。每个编码器的音素读出都会生成一个独立的后验概率，两个期望音素嵌入相加：

$$\begin{aligned} Z_N &= J([\mathrm{LN}(H_1);\mathrm{LN}(H_2)]),\\ Q_j &= \mathrm{softmax}(H_j W_{C,j} + b_{C,j}),\\ Z_P &= Q_1 U_{\mathrm{ph},1} + Q_2 U_{\mathrm{ph},2}. \end{aligned} \tag{1}$$

其中，$d=2048$表示语言模型的宽度，每个 $Q_j \in \mathbb{R}^{S\times 41}$ 都包含 39 个语音音位、静音和 CTC 空白符的音素后验概率。学习得到的嵌入表 $U_{\mathrm{ph},\,j} \in \mathbb{R}^{41\times d}$ 将两个独立归一化的后验概率转换为期望音素嵌入。因此，Z_N 保留连续的神经特征，而 Z_P 则以受监督的音素清单来表示这些证据。语音接口会在每个有效补丁位置保留完整分布，包括空白符概率。

语言智能体与角色。三个智能体均使用 Qwen3-ASR-1.7B（Shi 等人，2026年）文本解码器。我们共享并冻结其权重和词元嵌入表，同时为每个角色配置独立的 LoRA 适配器（Hu 等人，2022年）。附录 A.1 列出了适配器设置。我们为 MAS-BIT 初始化神经投影器、音素嵌入表、消息角色嵌入和角色适配器。

神经专家 N 以 Z_N 为条件，直接从神经表示预测英语转录文本。语音专家 P 以 Z_P 为条件，同样预测英语文本词元；其语音专门化来自基于后验概率的输入。仲裁器 A 以 Z_N以及下文定义的两条专家消息为条件。每种角色都以其连续输入序列为条件，自回归地生成文本。角色专用提示描述输入证据，并要求生成英语转录文本（附录 A.1；表 3）。输入证据、提示和独立适配器共同定义每个解码角色。填充位置会被屏蔽。

3.2 潜在通信

潜在消息是由概率加权词元嵌入组成的序列，并通过学习得到的角色嵌入标明其来源。每个专家 $r \in \{N,P\}$ 接收连续输入 Z_r：神经专家接收投影后的神经状态 Z_N，语音专家则接收后验加权音素嵌入 Z_P。专家以该输入为条件，生成贪心草稿 $\hat{y}^r$。草稿生成使用专家先前生成的词元，并在遇到序列结束词元或达到固定长度上限时停止。训练期间，此轮生成过程不进行梯度跟踪，并禁用语言模型 dropout。随后，我们将生成的草稿重新输入同一专家，以计算下一词元 logit

$$\ell_t^r = g_r(Z_r, \hat{y}^r_{<t}). \tag{2}$$

该重放过程具有可微性，并使用专家自身生成的草稿作为文本上下文。设 $\mathcal{K}_t^r$ 为位置 t处 16 个最大 logit 的索引。在温度 $\tau=1$下，消息内容是按概率加权的词元嵌入（Pham 等人，2024年）：

$$a_{t,v}^r = \frac{\exp(\ell_{t,v}^r/\tau)}{\sum_{u\in\mathcal{K}_t^r}\exp(\ell_{t,u}^r/\tau)}, \qquad \bar{m}_t^r = \sum_{v\in\mathcal{K}_t^r} a_{t,v}^r E_v, \tag{3}$$

其中， $E_v \in \mathbb{R}^d$ 是文本词元 v的共享冻结嵌入。权重会在所选支持集上重新归一化，从而允许多个候选词元为每个消息位置作出贡献。我们将学习得到的消息角色嵌入 e_r 添加到每个消息词元中，以表明其源自神经专家还是语音专家。在推理过程中，这将得到 $m_t^r = \bar{m}_t^r + e_r$。训练时，会在添加角色嵌入之前对消息内容应用 dropout（附录 A.2）。仲裁器的连续前缀是以下内容的拼接

$$Z_A = [Z_N; M_N; M_P], \qquad p_A(y \mid X) = \prod_{t=1}^{U} p_A(y_t \mid Z_A, y_{<t}). \tag{4}$$

消息掩码保留有效的草稿位置，包括第一个序列结束词元。这定义了一轮通过直接神经条件输入进行的专家到仲裁器通信。仲裁器损失通过选定的 softmax 权重和重放计算反向传播至活跃的专家适配器、输入接口和神经编码器。在此反向传播过程中，离散草稿词元和 top-k 成员关系保持固定。

推理。 在推理过程中，模型对录音进行一次编码，生成并重放两个专家草稿，然后贪婪解码仲裁器的最终句子。神经增强和 dropout 均被禁用。仲裁器的转录文本即为系统的最终输出。

3.3 训练目标与分阶段训练

我们使用目标句子上的教师强制交叉熵监督每个智能体，同时根据生成的草稿构建专家消息。令 $\mathcal{L}_{\mathrm{CE}}^{r}$ 表示角色 r的词元损失总和，该损失按目标文本词元数量进行归一化；令 $\mathcal{L}^{(l)}{}_{\mathrm{CTC}}$ 表示两个编码器在第 l层的平均音素 CTC 损失（Graves 等，2006），并按目标音素数量进行归一化。我们将辅助 CTC 损失应用于中间编码器层（Lee 与 Watanabe，2021）。训练目标为

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\mathrm{CE}}^{N} + \mathcal{L}_{\mathrm{CE}}^{P} + \mathcal{L}_{\mathrm{CE}}^{A} + 0.3\,\mathcal{L}_{\mathrm{CTC}}^{(7)} \\ + 0.05\,\rho\,\frac{\mathcal{L}_{\mathrm{CTC}}^{(3)} + \mathcal{L}_{\mathrm{CTC}}^{(5)}}{2}. \end{aligned} \tag{5}$$

系数 ρ在编码器首次更新的训练轮次内从零线性增加至一，此后保持为一。每个训练阶段都会计算全部三种句子损失；阶段决定更新哪些参数。文本损失和音素损失保留各自独立的分母，分布式归一化方式详见附录 A.4。

受 Fogg 等（2026）的神经增强方法启发，在专家生成、重放和句子监督期间，三个智能体均使用相同的增强录音。数据增强包括时间扰动和电极扰动，以及块掩码（附录 A.3）。

分阶段训练。 训练首先进行五轮对齐训练，更新输入接口和新初始化的参与者特定输入投影。随后进行五轮编码器适配，再依次进行五轮神经专家专门化、五轮语音专家专门化和五轮仲裁器专门化。之后将所有适配参数共同优化 40 轮：

$$\mathrm{Align}_5 \;\rightarrow\; \mathrm{Enc}_5 \;\rightarrow\; N_5 \;\rightarrow\; P_5 \;\rightarrow\; A_5 \;\rightarrow\; \mathrm{Joint}_{40}. \tag{6}$$

下标表示训练轮次数，因此最多训练 65 轮。图1C总结了各阶段特定的更新；附录 A.4 中的表 4 列出了启用的参数组。输入接口在整个训练过程中始终保持启用；对齐阶段之后，两个编码器主干和参与者特定输入投影也始终保持启用。 N、 P和 A阶段还会更新各自的角色适配器，而在 P阶段，两个音素读出层都会更新。最后的联合阶段会更新所有适配参数。

Froz下游模块传递输入梯度，使仲裁器目标能够监督
这游组件，但它们不会接收参数更新或优化器状态更新。

我们使用 Muon 优化编码器权重矩阵，并使用 AdamW 优化其余部分（Jordan 等，2024；Loshchilov 与 Hutter，2019），采用块局部预热—余弦调度（附录 A.4）。

4 评估

4.1 任务

我们使用分别来自参与者 T12 和 T15 的 Brain-to-Text ’24 和 ’25 数据集，评估 MAS-BIT 对尝试发声的解码效果（Willett等，2023年；2024；Card等，2024年）。在这项任务中，我们记录参与者尝试说出句子时的神经活动，并由解码器预测该句子的转录文本。我们以语料库词错误率（WER）衡量性能：词替换、删除和插入的总数除以参考词的数量；WER 越低，解码越准确。

我们使用 8,800 条 T12 话语和 8,072 条 T15 话语联合训练一个模型，采用参与者特定输入投影以及共享的编码器、接口和语言智能体。我们使用提供的全部 880 条和 1,426 条带标签验证话语来选择检查点。基准比较涵盖全部五种对照，并报告模型在分别包含 1,200 条和 1,450 条话语的隐藏留出集上的表现。T15 公有集和私有留出集分数分别报告。检查点选择以两个提供的带标签验证集上的合并词编辑数最小为准，且在留出集评估前固定解码设置。所有预测结果均在不访问参考转录文本的情况下生成，且声明划分中的每条话语都纳入评估。附录 A.5 详述相关内容。

4.2 基线

我们将 MAS-BIT 与已发表的端到端解码器和级联解码器进行基准比较，考察级联复现中的候选句子数量，并评估多智能体框架的五项核心消融实验，以及单智能体对照的适配器容量和更大骨干网络变体，同时评估顺序双视图细化、分离潜在消息和仅消息仲裁器对照。

已发表的基线。 我们的主要比较采用参与者和隐藏留出子集上的报告语料库词错误率（表 1）。不使用集成的端到端解码，会直接从神经输入生成文本，不使用外部加权有限状态转换器、 n元语法模型、发音词典或独立的语言模型重排序器；音素预训练、辅助 CTC 损失和分阶段训练均符合此定义。BIT 端到端解码 + 集成单独列出，已发表的级联参考结果见附录D。附录D还单独总结了其他评估集上的已发表结果。

候选句分析。 我们使用已发布的 BIT 编码器，为每位参与者复现级联解码。MAS-BIT 也从同一编码器初始化。按照 Zhang 等人（2026年）的方法，我们将 CTC音素逻辑值输入 OpenWebText2 五元语法语言模型，并使用 OPT-6.7B 对候选句子进行重打分（Zhang 等人，2022年）。我们调整保留候选项的数量 $K \in \{1, 2, 3, 4, 6, 9, 12, 18, 25, 37, 50, 75, 100\}$，同时固定编码器、语言模型和搜索设置。为区分编码器适配与多智能体解码，我们还使用 MAS-BIT 中每位参与者经微调的编码器分支和音素读出，对该级联系统进行评估。附录 C 介绍格重评分和候选项选择。

多智能体消融实验。 我们将 MAS-BIT 与 单智能体、双视图、双神经专家、硬文本通信、参与者独立训练和 无分阶段训练进行比较。所有对照组均采用相同的双编码器前端、骨干网络、初始化、数据增强和检查点选择协议；相应组件均从相同权重开始。附录 A.6 对各对照组作了定义。

5 实验与结果

5.1 MAS-BIT 实现最先进的端到端解码

MAS-BIT 在 Brain-to-Text ’24 留出集和 ’25 公共基准上实现了新的端到端最先进水平，词错误率分别为 6.929% 和 3.606%（图 2A；表 1）。同一模型在 ’25 私有子集上的词错误率为 3.892%。这些结果由一个统一的跨参与者模型和贪心解码实现，优于两个基准中匹配子集上的所有已发表端到端语音脑机接口，无论是否采用集成。

表 1：已发表的端到端基准测试结果。数值为语料库词错误率（%）；Feng 等人（2024）和 BIT 的分数遵循 Zhang 等人（2026）。MoDAl 报告五个随机种子上的均值 ± 95% CI；MoDAl-Full 也使用区域44。BrainWhisperer-跨模态使用了额外的神经数据集。 *未注明 T15 公开集/私有集子集。– 表示未报告结果。

	T12	T15	
方法	**留出集**	**公开集**	**私有集**
未使用集成的端到端解码			
Feng 等人（2024年）	24.69	–	–
循环神经网络编码器 + 端到端大语言模型解码器（Zhang 等，2026）	19.9	16.0	–
MoDAl-1（Chen 和 Chin，2026年）	18.8 ± 0.5	–	–
MoDAl-Full（Chen 和 Chin，2026年）	17.7 ± 0.4	–	–
BIT-TFS 端到端（Zhang 等，2026）	16.5	12.2	–
BIT-Human 端到端（Zhang 等，2026）	15.8	11.4	–
BIT 端到端（Zhang 等，2026）	15.67	11.06	–
BrainWhisperer（Boccato 等人，2026）	–	10.2*	
BrainWhisperer-跨模态（Boccato 等人，2026）	–	8.7*	
BEST（无外部语言模型）（Feng 等人，2026）	12.61	5.17*	
MAS-BIT（我们的）	**6.929**	**3.606**	**3.892**
集成式端到端解码			
BIT 端到端 + 集成（Zhang 等，2026年）	10.22	7.76	–

与单模型 BIT 相比，T12 和 T15 公开集上的词错误率分别降低 8.74 和 7.45 个百分点（相对降幅分别为 55.8% 和 67.4%）；MAS-BIT 还分别以 3.29 和 4.15 个百分点（相对降幅分别为 32.2% 和 53.5%）进一步超越了 BIT 的端到端集成（Zhang 等，2026）。在 T12 上，MAS-BIT 比 MoDAl-Full 报告的均值低 10.8 个百分点（Chen 和 Chin，2026），比同期无需外部语言模型的 BEST 解码器低 5.68 个百分点（Feng 等，2026）。在 T15 上，BrainWhisperer 和 BrainWhisperer-cross 报告的竞赛词错误率分别为 10.2% 和 8.7%（Boccato 等人，2026），BEST 报告的词错误率为 5.17%（Feng 等，2026）；这些结果均未说明使用的是公开集还是私有集。MAS-BIT 在 T15 的任一子集上的词错误率都低于上述每项结果，且不到 BrainWhisperer 两项结果的一半。

5.2 MAS-BIT超越复现的级联系统，在两个基准测试上均表现更优

我们将 MAS-BIT 与基于已发布 BIT 编码器的参与者专属级联解码进行比较。在基线级联系统中，这些编码器保持固定；而 MAS-BIT 则通过

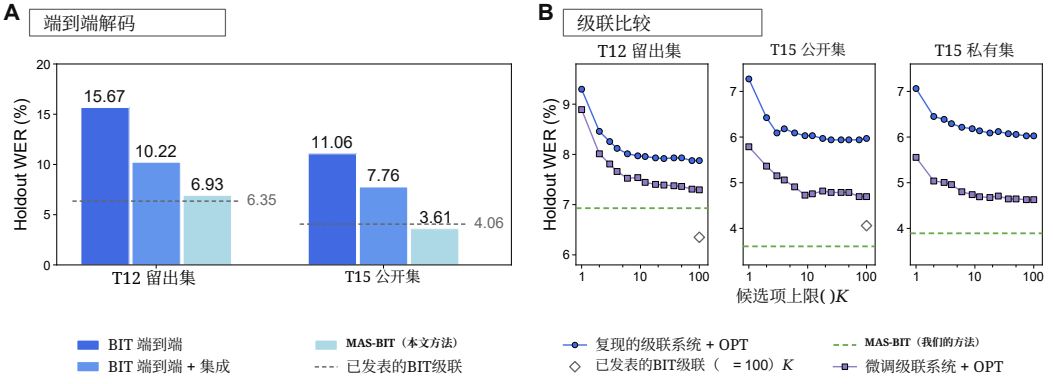

图2：联合训练的多智能体解码实现了最先进的端到端性能。（**A**）已发表的 BIT 端到端结果（Zhang 等，2026年）以及 MAS-BIT 在 T12 留出集和 T15 公开集上的表现。（**B**）蓝色曲线采用已发布的参与者专用BIT编码器；紫色曲线采用 MAS-BIT 中相应的微调编码器分支和最终音素读出。两个级联系统使用相同的加权有限状态转换器、未剪枝的五元语法，以及在 13 个测量候选项上限下的 OPT重评分设置；绿色线条表示 MAS-BIT。灰色菱形在 $K = 100$ 处标示已发表的BIT级联在 T12 和 T15 公开集上的结果。所有面板均报告语料库留出集词错误率（%，越低越好）；候选项上限采用对数坐标轴。

Tab表 2： 单张 *NVIDIA A40 GPU* 上的峰值主机内存、*GPU* 显存和解码时间。内存包括 s
load编程和审计；时间不包括两者（附录 C）。

	主机内存（吉比字节）		GPU 峰值显存（吉比字节）		解码时间（秒）	
方法	**T12**	**T15**	**T12**	**T15**	**T12**	**T15**
复现的级联系统（$K = 100$）（Zhang 等，2026）	321.8	334.8	25.211	25.274	2789.6	3761.6
BIT +LightBeam（Feghhi 等，2026年）	17.2	31.8	12.553	14.750	3563.5	6008.0
MAS-BIT（我们的方法）	**12.7**	**31.1**	**8.559**	**8.565**	**1538.8**	

跨参与者的语言智能体对两个编码器进行联合适配。对于所有测试的候选项上限（最高为 100 个句子），MAS-BIT 在 T12 和 T15 的两个子集上均取得更低的留出集词错误率（图 2B）；与复现结果中的最低词错误率相比，T12 提高了 0.947 个百分点（从 7.877% 降至 6.929%），T15 公开集和私有集分别提高了 2.33 和 2.13 个百分点。为区分编码器适配与多智能体解码的作用，使用 MAS-BIT 中各参与者微调后的编码器分支和音素读出运行相同的级联系统，可将 $K = 100$ 词错误率降至 T12 上的 7.295%，以及 T15 公开集/私有集上的 4.696%/4.629%（表15）。但在每种候选项上限下，MAS-BIT 仍优于这一适配编码器级联系统（T12 上高出 0.37 个百分点，T15 上在 $K = 100$时高出 1.09/0.74 个百分点），表明编码器适配和多智能体解码均有贡献。已发表的 BIT 级联解码在 $K = 100$ 时提供了另一项参考结果：T12 上为 6.35%，T15 公开集上为 4.06%（Zhang 等，2026）。已发表的级联系统采用了未公开发布的音素级编码器；我们的复现则以已发布的句子级检查点为起点（附录 C）。MAS-BIT 在 T15 公开集上的结果比这一已发表结果提高了 0.454 个百分点；在 T12 上则仅差 0.58 个百分点，而 BIT 的端到端解码器落后 9.32 个百分点，且 MAS-BIT 无需使用加权有限状态转换器解码或外部候选重评分。

5.3 MAS-BIT L 增强推理内存和解码能力 TIME

共享一个冻结的 17 亿参数骨干网络，无需外部图搜索或候选重打分，也能降低推理成本（表 2）。在 A40 GPU 上，复现的 $K = 100$ 级联系统需要 10.8–25.4×MAS-BIT 的峰值主机内存和 $2.95\times$其 GPU 峰值显存，而 MAS-BIT 的解码速度约快 1.8× 。词典约束的 LightBeam 解码（Feghhi 等，2026）可减少内存占用，但仍需要 MAS-BIT GPU 显存的 1.47–1.72× 倍，且解码速度慢 2.3–2.9× 倍。在消费级 RTX 5090 笔记本 GPU 和 RTX 4090 台式机 GPU 上，由于 $K = 100$ 级联系统超出主机内存，MAS-BIT 在 8.7 吉比字节 GPU 显存下复现了完全相同的 FP32 预测结果，其解码速度比 BIT +LightBeam 快 2.6–3.6× 倍和 2.7–3.0× 倍（表 11–12）。

5.4 专业化 ，通信 ，与分阶段训练改善解码

五项核心控制中的每一项都会降低三个留出集划分上的准确率，从而分别揭示专家专门化、语音学证据、潜在消息、跨参与者联合训练和分阶段训练的作用（图3B–C；表16）。

专业化能够更充分地利用相同的输入证据。 与单智能体相比，专门化专家和仲裁机制在同时接收神经表征与音素后验概率时可降低词错误率；后者在 T12 上的留出集词错误率比 MAS-BIT 高 3.06 个百分点，在 T15 公开集/私有集上分别高 2.52/2.78 个百分点。为检验这一差距是否源于参数容量，将单智能体适配器扩展至秩 24，使其与三个秩 8 角色适配器合计的 LoRA 参数量相当，得到 T12 为 10.231%、T15 公开集/私有集为 5.787%/6.889% 的词错误率；而将单智能体骨干网络扩展至 Borealis-5B 的 4B 音频训练解码器（VikhrModels，2025年）（秩 13 LoRA）后，词错误率分别为 12.478% 和 8.515%/8.553% （图 3B–C ；附录 A.6 ）。即使通过顺序双视图细化匹配 MAS-BIT 的三个秩 8 适配器、中间监督和多轮潜在消息解码，留出集词错误率在 T12 上仍高出 0.23 个百分点，在 T15 上高出 0.33/0.85 个百分点（表16）；因此，在参数容量和解码轮次相当的情况下，专门的神经专家和语音专家仍能进一步提升性能。

潜在消息优于专家离散草稿。 用困难文本替代潜在消息后，T12 的留出集词错误率上升 0.23 个百分点，T15 公开集/私有集分别上升 0.58/0.52 个百分点。潜在接口保留了替代词元假设，并允许仲裁器损失通过重放更新专家

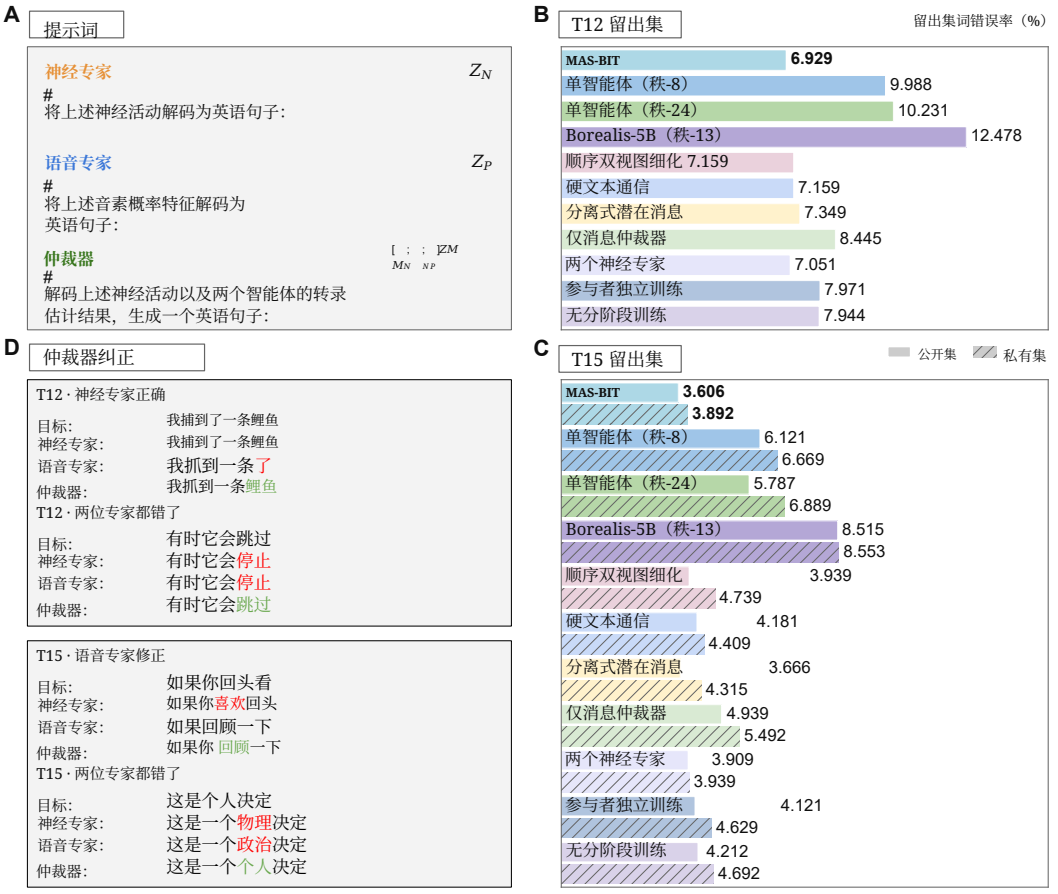

图 3：专家专门化与潜在通信可提升端到端解码。 **(A)** 角色专用提示；# 标记连续输入槽位。 **(B–C)** T12 和 T15 留出集词错误率；实心柱和斜线柱分别表示 T15 公开集和私有集结果。八种对照设置和秩为 24 的单智能体变体采用相同的骨干网络和双编码器初始化。Borealis-5B 单智能体变体使用带秩为 13 的 LoRA 的 4B 解码器；所有单智能体变体均接收两种输入视图。柱状图的零点起始刻度一致。**(D)** 选取的验证集转录文本显示，仲裁器能够匹配正确的专家结果，并在两份草稿都出错时恢复正确句子。专家错误以红色标出；仲裁器中对应的正确词语以绿色标出。

概率，而困难文本对照仅传递选定的草稿词元，并阻断这条梯度路径。为区分梯度传递与替代词元内容，分离潜在消息对照在前向传递中保留前16个概率加权词元嵌入，同时在消息处停止仲裁器梯度（附录 A.6）；这使 T12 的留出集词错误率上升 0.42 个百分点，T15 公开集/私有集分别上升 0.06/0.42 个百分点，证实端到端消息梯度能够改善专家协同。相反，仅消息仲裁器仅从$[M_N;\ M_P]$解码，而没有直接神经输入 Z_N，导致 T12 的词错误率下降 1.52 个百分点，T15 公开集/私有集分别下降 1.33/1.60 个百分点，在每个划分上都超过了困难文本惩罚，表明仲裁依赖直接神经依据，而非仅仅完善专家草稿。

语音学证据补充了神经解释。 用第二个神经专家替换语音专家，会使留出集词错误率在 T12 上增加 0.12 个百分点，在 T15 公开集/私有集上分别增加 0.30/0.05 个百分点。双神经专家系统仍比单智能体在 T12 上提升 2.94 个百分点，在 T15 上提升 2.21/2.73 个百分点，因此大部分收益来自专家草稿和基于神经信息的仲裁，而音素后验视图则带来较小但稳定的增益。这两种接口所呈现的句子级几何结构也比双神经专家对照更为不同（附录 E）。

跨参与者联合训练可提升留出集泛化。 在每位参与者的数据上分别训练 MAS-BIT 模型，在 T12 上的留出集词错误率为 7.971%，在 T15 公开集/私有集上分别为 4.121%/4.629%

。在相同的分阶段训练轮次计划下，联合训练使 T12 的词错误率降低 1.042 个百分点，使 T15 公开集/私有集的词错误率分别降低 0.515/0.737 个百分点，对应的相对降幅分别为 13.1% 和 12.5%/15.9%。因此，联合训练的编码器、接口和语言智能体能够提升留出集准确率，同时以单个模型服务两位参与者。

专业化为系统的联合优化做好准备。从第一个训练轮次开始训练所有适配组，会使 T12 的留出集词错误率增加 1.015 个百分点，并使 T15 公有集/私有集的留出集词错误率分别增加 0.606/0.800 个百分点。分阶段前缀会在最终联合阶段之前建立输入接口和专家角色。

5.5 仲裁整合互补的专家判断

在完整验证集上，相较于神经专家和语音专家，仲裁使 T12 上的词编辑数分别减少了 28.4% 和 30.5%，使 T15 上的词编辑数分别减少了 33.0% 和 40.3%（附录 B）。关键在于，当两位专家都出错时，仲裁器在 T12 上恢复了 31 个完整句子，在 T15 上恢复了 49 个：其中 26 个和 38 个无法通过沿任意最小编辑草稿对齐重组词语得到；在这两组中，分别有 26 个和 36 个句子包含两份草稿中都不存在的正确词语（图4；示例见图3D）。因此，模型将专家消息与神经证据相结合，产生了超出选择或拼接专家草稿能力的修正（附录 B）。

在固定权重和专家草稿的情况下，同时掩蔽两条消息会使 T12 和 T15 上的验证集词错误率分别增加 4.130 和 2.668 个百分点，而 Top-16 消息相较于前1可减少 41 和 29 个词编辑数（附录 B.2–B.3）。将直接神经输入 Z_N 置零，会移除全部 26 个和 38 个超出草稿重组能力的恢复结果；而将其替换为来自同一 partic-

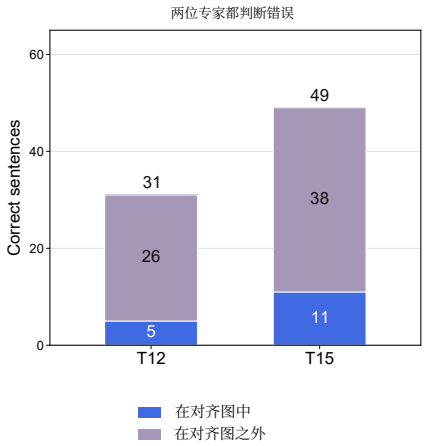

图4：仲裁可恢复超出草稿重组能力的句子。在完整验证集上，当两位专家均出错时的精确恢复结果。

参与者移除了配对子集中全部 25 个和 36 个此类恢复结果（附录 B.2）。

6 讨论

总结。我们提出了 MAS-BIT，这是一种端到端语音解码器，通过潜在通信和基于信号的仲裁，结合神经专家与语音专家。它并未将解释和生成工作都交给单个解码器，而是由专门的专家分别解读连续神经状态和软音素后验概率，再由仲裁器直接访问录音来解决双方的不一致，从而恢复两位专家均未能识别的句子。在分阶段专门化后，该模型通过跨参与者联合训练，在 Brain-to-Text ’24 和 ’25 上创下端到端新纪录（词错误率分别为 6.929% 和 3.606%/3.892%）；在 ’25 公开集上超越了 BIT 已发表的级联系统，在 T12 上与之仅差 0.58 个百分点；在所有候选项上限下均优于复现的级联系统；并在保留不同视角的同时，使跨参与者接口与语音及语言几何相契合（附录 E）。

局限性与未来方向。我们的评估是在离线条件下进行的，涵盖两名参与者。对于在线流式脑机接口，我们的编码器对完整话语采用双向注意力，并在仲裁之前进行顺序草稿生成；下一步的重要工作是将多智能体协调适配到因果编码器和更低延迟的消息传递，同时将共享接口扩展到更大的队列、内言以及持续漂移适应。将解码扩展到内言还需要明确的用户同意和隐私保护措施（Kunz 等，2025年）。

大语言模型使用声明

我们使用了生成式人工智能工具，包括 ChatGPT、Codex 和 Claude，以协助进行语法检查、文本润色和软件实现（例如编程、调试和可视化脚本）。所有研究构想、假设、实验设计以及科学解释均源自作者；作者审查并验证了本文中的所有代码、实验结果和文本，并对此承担全部责任。

可复现性声明

方法、评估和补充章节中提供了详细的架构、优化、预处理、评估和分析流程。代码和模型权重可在以下平台公开获取：GitHub和HuggingFace。

参考文本

Hassan Akbari、Bahar Khalighinejad、Jose L Herrero、Ashesh D Mehta 和 Nima Mesgarani。致力于从人类听觉皮层重建可理解语音。《科学报告》，9(1)：874，2019年。

Tyler Benster、Guy Wilson、ReshefElisha、FrancisRWillett 和 ShaulDruckmann。采用跨模态方法实现无声语音，并结合大语言模型增强识别，2024年。网址 https://arxiv.org/abs/2403.05583。

Tommaso Boccato、Michal Olak 和 Matteo Ferrante。BrainWhisperer：利用大规模自动语音识别模型进行神经语音解码，2026年。arXiv:2603.13321v1。

Nicholas S. Card、Maitreyee Wairagkar、Carrina Iacobacci、Xianda Hou、Tyler Singer-Clark、Fran- cis R. Willett、Erin M. Kunz、Chaofei Fan、Maryam Vahdati Nia、Darrel R. Deo、Aparna Srinivasan、Eun Young Choi、Matthew F. Glasser、Leigh R. Hochberg、Jaimie M. Henderson、Kiarash Shahlaie、Sergey D. Stavisky 和 David M. Brandman。精准且可快速校准的语音神经假体。《新英格兰医学杂志》，391(7):609–618，2024年。ISSN 1533-4406。doi: 10.1056/nejmoa2314132。网址 https://doi.org/10.1056/NEJMoa2314132。

Yuanhao Chen 和 Peter Chin。MoDAl：通过去相关实现用于语音神经假体的自监督神经模态发现，2026年。网址 https://arxiv.org/abs/2605.00025。

Avihu Dekel、Samuel Thomas、Takashi Fukada 和 George Saon。NLE：基于大语言模型的非自回归 ASR 转录文本编辑，2026年。网址 https://arxiv.org/abs/2603.08397。

Yilun Du、Shuang Li、Antonio Torralba、Joshua B.Tenenbaum 和 Igor Mordatch。通过多智能体辩论提升语言模型的事实准确性和推理能力。载于 Ruslan Salakhutdinov、Zico Kolter、Katherine Heller、Adrian Weller、Nuria Oliver、Jonathan Scarlett 和 Felix Berkenkamp（编），第*41*届国际机器学习会议论文集，第235卷，机器学习研究论文集，第11733–11763页。PMLR，2024年7月21日至27日。网址https://proceedings.mlr.press/v235/du24e.html。

Ebrahim Feghhi、Shreyas Kaasyap、Nima Hadidi 和 Jonathan C. Kao。用于神经语音解码的带掩码 Transformer，结合轻量级测试时自适应。在 神经信息处理系统进展，第 38 卷，2025年。网址：https://papers.nips.cc/paper_files/paper/2025/hash/85e351d89eee3f33d3d1356f38590bb6-Abstract-Conference.html。

Ebrahim Feghhi、Junlin Hu、Nima Hadidi 和 Jonathan C. Kao。LightBeam：一种适用于语音神经假体的高精度、内存高效 CTC解码器，2026年。网址：https://arxiv.org/abs/2603.14002。

Sheng Feng、Heyang Liu、Yu Wang 和 Yanfeng Wang。《借助大语言模型实现侵入式脑信号解码的端到端框架》。载于*Interspeech 2024*，第 1495–1499 页，2024年。doi：10.21437/Interspeech.2024-382。

Zehui Feng、Weichuan Wang、Xiaohan Chen、Cuntai Guan 和 Ting Han。语音脑机接口是否需要更大的模型？重新思考规模扩展之外的神经解码。载于神经信息处理系统进展，2026年。网址：https://arxiv.org/abs/2610.23175。

Tobias Fiedler、Leon Hermann、Florian M¨uller、Sarel Cohen、Peter Chin、TobiasFriedrich 和 Eilon Vaadia。让 Wav2Vec2 掌握大脑的语言，2025年。URL https://arxiv.org/abs/2501.09459v1。

Jakob Foerster、Ioannis Alexandros Assael、Nando de Freitas 和 Shimon Whiteson。通过深度多智能体强化学习学习通信。载于 D. Lee、M. Sugiyama、U. Luxburg、I. Guyon 和 R. Garnett（编），神经信息处理系统进展，第29卷。Curran Associates, Inc.，2016年。URL https://proceedings.neurips.cc/paper_files/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf。

Zachery M. Fogg、Nicholas S. Card、Maitreyee Wairagkar、Aparna Srinivasan、Tyler Singer-Clark、Xianda Hou、Elizaveta Okorokova、Hamza Peracha、Carrina Iacobacci、Tiffany Brailow、Justin Jude、Hadar Levi-Aharoni、Trung Le、Domenick Mifsud、Pranav I. Deevi、Samuel R. Nason- Tomaszewski、Anna L. Pritchard、Yunnuo Zhang、Brice Richards、Payton Bechefsky、Leigh R. Hochberg、Ziv Williams、Kiarash Shahlaie、Nicholas AuYong、Daniel B. Rubin、Chethan Pan- darinath、David M. Brandman 和 Sergey D. Stavisky。一种可泛化的语音神经假体。*bioRxiv*，2026年。doi：10.64898/2026.07.23.739430。URLhttps://www.biorxiv.org/content/early/2026/07/27/2026.07.23.739430。

Alex Graves、Santiago Fern´andez、Faustino Gomez 和 J¨urgen Schmidhuber。连接时序分类：使用循环神经网络标注未分段序列数据。载于第*23*届国际机器学习会议论文集——*ICML '06*，ICML '06，第369–376页。ACM出版社，2006年。doi：10.1145/1143844.1143891。网址：http://dx.doi.org/10.1145/1143844.1143891。

Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、LuWang 和 Weizhu Chen。LoRA：大型语言模型的低秩适配。发表于*ICLR2022*，2022年4月。网址：https://www.microsoft.com/en-us/research/publication/lora-low-rank-adaptation-of-large-language-models/。

Ke Hu、Tara N. Sainath、Ruoming Pang 和 Rohit Prabhavalkar。基于审议模型的两遍端到端语音识别。载于 *ICASSP 2020 - 2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第7799–7803 页。IEEE，2020年。doi: 10.1109/ icassp40776.2020.9053606。网址 https://doi.org/10.1109/ICASSP40776.2020.9053606。

Keller Jordan、Yuchen Jin、Vlado Boza、Jiacheng You、Franz Cesista、Laker Newhouse 和 Jeremy Bernstein。Muon：一种用于神经网络隐藏层的优化器，2024年。网址 https://kellerjordan.github.io/posts/muon/。

Owais M. Khanday、Jose A. Gonzalez-Lopez、Marc Ouellet、Alberto Gald´on 和 Gonzalo Olivares Granado。基于神经活动的端到端皮层内语音解码。载于 *Odyssey 2026*，第 165–172 页，2026年。doi: 10.21437/Odyssey.2026-24。网址https://www.isca-archive.org/odyssey_2026/khanday26_odyssey.html。

Nikolaus Kriegeskorte、Marieke Mur 和 Peter A. Bandettini。表征相似性分析——连接系统神经科学的各个分支。系统神经科学前沿，第 2 卷，2008年，2008年。ISSN 1662-5137。doi：10.3389/neuro.06.004.2008。网址：https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008。

Erin M Kunz、Benyamin Abramovich Krasa、Foram Kamdar、Donald T Avansino、Nick Hahn、SeonghyunYoon、Akansha Singh、Samuel RNason-Tomaszewski、NicholasSCard、JustinJ Jude 等。《运动皮层中的内言及其对语音神经假体的启示》。*Cell*，188(17)：4658–4673.e17，2025年。doi：10.1016/j.cell.2025.06.015。

Jaesong Lee 和 Shinji Watanabe。用于基于 CTC 的语音识别的中间损失正则化。载于*ICASSP 2021 - 2021*年*IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 6224–6228 页。IEEE，2021年6月。doi: 10.1109/ICASSP39728.2021.9414594。网址: https://doi.org/10.1109/ICASSP39728.2021.9414594。

Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。发表于 国际学习表征会议，2019年。网址 https://mlanthology.org/iclr/2019/loshchilov2019iclr-decoupled/。

Rao Ma、Tongzhou Chen、Kartik Audhkhasi 和 Bhuvana Ramabhadran。LegoSLM：使用 CTC 后验概率将大语言模型与语音编码器连接起来。载于 Christos Christodoulopoulos、Tanmoy Chakraborty、Carolyn Rose 和 Violet Peng（编），计算语言学协会论文集：*EMNLP 2025*，第 18171–18186 页，中国苏州，2025年11月。计算语言学协会。ISBN 979-8-89176-335-7。doi：10.18653/v1/2025.findings-emnlp. 985。网址 https://aclanthology.org/2025.findings-emnlp.985/。

Sean L Metzger, Kaylo T Littlejohn, AlexanderB Silva, David AMoses, MargaretPSeaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger 等。用于语音解码和虚拟形象控制的高性能神经假体。《自然》，620(7976):1037– 1046，2023年。

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams 等。针对一名患有构音不能的瘫痪患者的语音解码神经假体。《新英格兰医学杂志》，385(3):217–227，2021年。

Samuel R. Nason, Alex K. Vaskov, Matthew S. Willsey, Elissa J. Welle, Hyochan An, Philip P. Vu, Autumn J. Bullard, Chrono S. Nu, Jonathan C. Kao, Krishna V. Shenoy, Taekwang Jang, Hun- Seok Kim, David Blaauw, Parag G. Patil, and Cynthia A. Chestek。一种低功耗神经元尖峰活动带，其中局部单神经元占主导，可提升脑机接口的性能。《自然·生物医学工程》，4(10):973–983，2020年。doi: 10.1038/s41551-020-0591-0。URL https://doi.org/10.1038/s41551-020-0591-0。

Michał Olak、Tommaso Boccato 和 Matteo Ferrante。Brain Translation：用于皮层内尝试性语音解码的多任务序列到序列建模。在 计算语言学协会年会论文集：*EMNLP 2026*，第 43901–43916 页，2026年。网址：https://aclanthology.org/2026.findings-emnlp.2062/。

Daniel S. Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D. Cubuk 和 Quoc V. Le。SpecAugment：一种用于自动语音识别的简单数据增强方法。在 *Interspeech 2019*，第 2613–2617 页，2019年。doi：10.21437/Interspeech.2019-2680。

Kyubyong Park 和 Jongseok Kim。g2pE。https://github.com/Kyubyong/g2p，2019年。

Chau Pham、Boyi Liu、Yingxiang Yang、Zhengyu Chen、Tianyi Liu、Jianbo Yuan、Bryan Plum- mer、Zhaoran Wang 和 Hongxia Yang。让模型使用密码交流：通过嵌入进行多智能体辩论。载于 B. Kim、Y. Yue、S. Chaudhuri、K. Fragkiadaki、M. Khan 和 Y. Sun（编），国际学习表征会议，第 2024 卷，第 51899–51928 页，2024年。网址：https://proceedings.iclr.cc/paper_files/paper/2024/file/e444859b2a22df6b56af9381ad1e9480-Paper-Conference.pdf。

Vignav Ramesh 和 Kenneth Li。语言模型智能体之间的激活值通信。载于 Aarti Singh、Maryam Fazel、Daniel Hsu、Simon Lacoste-Julien、Felix Berkenkamp、Tegan Maharaj、Kiri Wagstaff 和 Jerry Zhu（编），第 *42* 届国际机器学习会议论文集，第 267 卷，机器学习研究论文集，第 51094–51116 页。PMLR，2025年7月13–19日。网址: https://proceedings.mlr.press/v267/ramesh25a.html。

Xian Shi、Xiong Wang、Zhifang Guo、YongqiWang、Pei Zhang、Xinyu Zhang、ZishanGuo、Hongkun Hao、Yu Xi、Baosong Yang、Jin Xu、Jingren Zhou 和 Junyang Lin。Qwen3-ASR 技术报告，2026年。网址https://arxiv.org/abs/2601.21337。

Jianlin Su、Murtadha H. M. Ahmed、Yu Lu、Shengfeng Pan、Wen Bo 和 Yunfeng Liu。RoFormer：采用旋转位置嵌入的增强型 Transformer。*Neurocomputing*，568:127063，2024年。doi：10.1016/J.NEUCOM.2023.127063。网址https://doi.org/10.1016/j.neucom.2023.127063。

VikhrModels. Borealis：用于语音理解的音频语言模型，2025年。网址 https://huggingface.co/Vikhrmodels/Borealis-5b-it。

Maitreyee Wairagkar、Nicholas S Card、Tyler Singer-Clark、Xianda Hou、Carrina Iacobacci、Lee M Miller、Leigh R Hochberg、David M Brandman 和 Sergey D Stavisky。即时语音合成神经假体。《自然》, 644(8075):145–152, 2025年。

Junlin Wang、Jue Wang、Ben Athiwaratkun、Ce Zhang 和 James Y Zou。智能体混合增强大语言模型的能力。载于 Y. Yue、A. Garg、N. Peng、F. Sha 和 R. Yu（编），国际学习表征会议，第 2025 卷，第 33944–33963 页，2025年。网址https://proceedings.iclr.cc/paper_files/paper/2025/file/5434be94e82c54327bb9dcaf7fca52b6-Paper-Conference.pdf。

Francis R. Willett、Erin M. Kunz、Chaofei Fan、Donald T. Avansino、Guy H. Wilson、Eun Young Choi、Foram Kamdar、Matthew F. Glasser、Leigh R. Hochberg、Shaul Druckmann、Krishna V. Shenoy 和 Jaimie M. Henderson。高性能语音神经假体。《自然》，620 (7976):1031–1036，2023年。ISSN 1476-4687。doi: 10.1038/s41586-023-06377-x。网址 https://doi.org/10.1038/s41586-023-06377-x。

Francis R. Willett、Jingyuan Li、Trung Le、Chaofei Fan、Mingfei Chen、Eli Shlizerman、Yue Chen、Xin Zheng、Tatsuo S. Okubo、Tyler Benster、Hyun Dong Lee、Maxwell Kounga、E. Kelly Buchanan、David Zoltowski、Scott W. Linderman 和 Jaimie M. Henderson。大脑到文本基准 '24：经验教训，2024年。网址https://arxiv.org/abs/2412.17227。

Seonghyun Yoon、Donald T. Avansino、Sasidhar Madugula、Alisa D. Levin、Chaofei Fan、Benyamin Abramovich Krasa、Akansha Singh、Christina Vo、Nick V. Hahn、Nicholas S. Card、Zachery Fogg、Maitreyee Wairagkar、Samuel R. Nason-Tomaszewski、Brandon G. Jacques、Payton H. Bechefsky、Carrina Iacobacci、Darrel R. Deo、Leigh R. Hochberg、David M. Brandman、Sergey D. Stavisky、Nicholas Au Yong、Chethan Pandarinath、Jaimie M. Henderson 和 Francis R. Willett。利用深度神经集成模型对语音进行神经解码。*bioRxiv*，2026年。doi：10.64898/2026.06.02.729705。网址：https://www.biorxiv.org/content/10.64898/2026.06.02.729705v1。预印本。

Susan Zhang、Stephen Roller、Naman Goyal、Myle Ott、Moya Chen、Shuohui Chen、Christo- pher Dewan、Mona Diab、Xian Li、Xi Victoria Lin、Todor Mihaylov、Myle Ott、Sam Shleifer、Kurt Shuster、Daniel Simig、Punit Singh Koura、Anjali Sridhar、Tianlu Wang 和 Luke Zettle- moyer。OPT：开放式预训练 Transformer 语言模型，2022年。网址 https://arxiv.org/abs/2205.01068。

Yizi Zhang、Linyang He、Chaofei Fan、Tingkai Liu、Han Yu、Trung Le、Jingyuan Li、Scott Linderman、Lea Duncker、Francis Willett、Nima Mesgarani 和 Liam Panin- ski。用于端到端语音解码的跨物种神经基础模型。载于 C. Vondrick、B. Hariharan、C. Raffel、L. Pinto、D. Yang 和 A. Faust（编），国际学习表征会议，第 2026 卷，第 30773–30797 页，2026年。网址 https://proceedings.iclr.cc/paper_files/paper/2026/file/34307cd20546fa1c6de90f303973ccf7-Paper-Conference.pdf。

Jiaru Zou、Ruizhong Qiu、Gaotang Li、Xiyuan Yang、Katherine Tieu、Pan Lu、Ke Shen、Hanghang Tong、Yejin Choi、Jingrui He、James Zou、Mengdi Wang 和 Ling Yang。多智能体系统中的潜在协作。载于第四十三届国际机器学习会议，2026年。

A 实现细节

A.1 数据与初始化

第 4.1 节介绍了数据集划分和检查点选择。

输入特征按阈值穿越、尖峰频带功率的顺序拼接。每个输入时间 bin 的跨度为 20 毫秒，因此 T12 有 $C = 256$ 个通道，T15 有 $C = 512$ 个通道。编码器采用高斯平滑、不重叠的五帧分块，以及七个双向 Transformer 模块。T12 使用 128 个电极、按块进行 z 分数标准化，并根据该日的训练试次估算后续的日级归一化。T15 使用全部 256 个电极，以及发布的 HDF5 文件中提供的归一化特征。针对参与者的预处理保留了已发布的 T12/T15 帧数上限（分别为 1,250/2,500 帧）；两个编码器主干均支持 T15 输入使用的 2,500 帧时长。转录文本均转换为小写，并删除除撇号以外的标点符号。使用 g2p en（Park & Kim，2019年）根据转录文本生成音素目标，删除重音标记，将空格映射为静音，并在末尾附加静音。41 类清单包含 39 个 ARPAbet 音素、静音和空白符。

发布的两个 BIT 模型提供了各自的神经编码器、原生输入投影和相应的音素读出层。MAS-BIT 在预训练的通义千问文本解码器之上，增加了两个跨参与者输入投影、一个神经融合投影器、两张音素嵌入表、两个消息角色向量，以及三个 LoRA 适配器。融合投影器分别对两个 384 维编码器状态进行层归一化，将它们拼接起来，并在两个线性层之间应用带有 GELU 的 $768 \to 1024 \to 2048$ 多层感知器。两张包含 41 行的音素嵌入表以标准差为 $1/\sqrt{2}$ 的值进行初始化，随后将相应的嵌入求和。这些模块在一次训练中完成训练，训练从对齐阶段开始。神经编码器的宽度为 384，包含七层、六个注意力头，前馈层宽度为 512，采用 GELU 激活函数和旋转位置嵌入（Su 等人，2024年）。其参与者专属的分块输入层依次应用层归一化、从 $5C$ 到 384 的线性投影，以及另一层层归一化。高斯平滑的标准差为两个输入时间 bin（40 毫秒）；嵌入 dropout 和 Transformer dropout 分别为 0。2 和 0。4。文本解码器的宽度为 2048，包含 28 层。每个角色的 LoRA 都应用于查询、键、值、输出、门控、上投影和下投影，秩为 8，缩放参数为 32（缩放因子为 $32/8$），dropout 为 0。2。

角色专用提示词。 文本分隔符用于标识各角色的连续输入，后接生成英语句子的指令。神经专家的提示词提及神经活动，语音专家的提示词提及音素概率特征，而仲裁器的提示词则提及神经活动及两位专家的转录估计。图 3A 展示了完整提示词，其中 # 标记连续输入的插入位置。表 3 列出了训练和推理期间 MAS-BIT 及所有消融对照所用的准确提示词文本。硬文本通信、分离式潜在消息、参与者独立训练和无分阶段训练均保留相同的三个角色提示词；单智能体提示词则提及两种输入视图。在双神经专家对照中，两位专家均使用神经专家模板，仲裁器提示词保持不变。秩为 24 和 Borealis-5B 的单智能体变体使用相同的单智能体模板。仅消息对照保留两位专家的提示词，并使用仅消息仲裁器模板。顺序双视图细化使用单智能体文本及 $[Z_N;\ Z_P]$ 作为第一位智能体的提示词，使用细化智能体模板及 $[Z_N;\ Z_P;\ M_1]$作为第二位智能体的提示词，并使用仲裁器模板及 $[Z_N;\ M_1;\ M_2]$ 生成最终转录。分隔符和指令使用预训练分词器现有的词汇表。我们使用对应的解码器分词器对 # 两侧的字面文本进行分词，并禁用自动特殊标记插入，且不添加其他系统消息或聊天消息。对于单智能体对照，v_N 和 v_P 分别是为神经输入和语音输入学习的视图嵌入。

A.2 消息构造

神经专家、语音专家和仲裁器的连续输入分别为 Z_N、 Z_P和$[Z_N;\ M_N;\ M_P]$。硬文本对照根据离散草稿词元嵌入构建 M_N和 M_P。单智能体对照在添加学习得到的视图标识符后，将 Z_N 和 Z_P 拼接起来。交叉熵监督答案词元（包括EOS）；条件词元和填充不接收标签。EOS词元ID为151643和151645。

表 3：*MAS-BIT* 和所有消融对照的提示词。占位符 # 将替换为中间列中的连续输入。成对的文本分隔符和指令文本均按原样显示。每个模板均为单行；换行仅为排版需要。

Role	**输入**	**提示**
神经专家	Z_N	<|神经_活动|># <|/神经_活动|> 将上述神经活动解码为一个英文句子：
语音专家	Z_P	<|音素_概率_特征|># <|/音素_概率_特征|> 将上述音素概率特征解码为一个英语句子：
仲裁器	$[Z_N;\ M_N;\ M_P]$	<|神经_和_专家_消息|># <|/神经_和_专家_消息|> 解码上述神经活动和两个智能体的转录估计，生成一个英语句子：
单个智能体，两个视图	$[Z_N + v_N;\ \ Z_P + v_P]$	<|神经_和_音素_特征|># <|/神经_和_音素_特征|> 将上述神经活动和音素概率特征解码为一个英语句子：
仅消息仲裁器	$[M_N;\ M_P]$	<|专家_消息|># <|/专家_消息|> 将上述两个智能体的转录估计解码为一个英语句子：
精炼智能体	$[Z_N;\ Z_P;\ M_1]$	<|神经_音素_和_草稿|># <|/神经_音素_和_草稿|> 使用上述神经活动完善第一版草稿，并将音素概率特征转化为英语句子：

对于潜在消息重放，保留的草稿包括第一个EOS标记；生成的填充通过其有效性掩码排除。每个专家的完整消息以0.15的概率独立置零，不对保留的消息进行重新缩放；随后逐元素嵌入丢弃的概率为0.1，并采用标准的逆保留率缩放。角色向量在这两项操作之后添加。

A.3 神经输入增强

每个训练样例都会进入数据增强流水线。时间扭曲和掩蔽遵循 SpecAugment（Park等，2019年）所采用的原则。首先，我们选取两个内部切点并去除重复点，以定义时间片段；然后，将每个片段按从 $\mathcal{N}(1, 0.05_2)$中抽取的因子进行线性重采样，并将其裁剪至 $[0.9, 1.1]$。短于九帧的试次保留原始时间网格。如果提出的扭曲超出输入上限，或导致编码器补丁数量少于目标音素数量加相邻重复音素所需的数量，我们就会拒绝该扭曲。即使扭曲被拒绝，其余变换仍会执行。其次，我们以0.05的概率独立掩蔽每个电极，在整个试次中同时将其阈值交叉特征和尖峰功率特征置零。第三，我们抽取长度为10–20帧的原始时间跨度，累计跨度长度预算为 $\lfloor 0.1 T_{\mathrm{aug}} \rfloor$，其中 T_{aug} 是扭曲后的序列长度；最后一个跨度会缩短至剩余预算长度，并允许跨度重叠。第四，我们添加独立高斯噪声，其标准差为归一化特征单位下的0.2。这些操作在编码器的高斯平滑之前执行。

令 $S_{\mathrm{aug}} = \lceil T_{\mathrm{aug}}/5 \rceil$ 表示有效增强补丁的数量。在补丁输入投影之后，从 1 到 $\lfloor 0.2 S_{\mathrm{aug}} \rfloor$ 均匀采样一个连续跨度长度，并将其均匀放置在有效起始位置中。将该跨度内的补丁替换为冻结的预训练掩码嵌入，同时保留序列长度和注意力有效性。当 $\lfloor 0.2 S_{\mathrm{aug}} \rfloor = 0$ 时不应用跨度。对于每个样本，三个智能体共享同一次数据增强结果，包括在对齐阶段。评估时保留确定性平滑，并禁用这些随机扰动。

A.4 优化与可复现性

Tab第 4 表：各阶段更新的参数组。 预训练语言模型骨干网络始终保持冻结 n thro状态。

阶段	更新的参数
对齐（1–5）	New 参与者特定输入投影， 神经融合 投影器， 音素嵌入表、消息角色嵌入
编码器适配（6–10）	两个编码器和所有输入投影；所有输入接口
神经网络专门化（11–15）	编码器适配组和 N 适配器
语音专门化（16–20）	编码器适配组、两个音素读出层和 P 适配器
仲裁器专门化（21–25）	编码器适配组和 A 适配器
联合优化（26–65）	所有适配参数，包括全部三个角色适配器和两个音素读出层

方程式 6 中的计划包括五个为期五个训练轮次的对齐、适应和专业化模块，随后是一个为期 40 个训练轮次的联合模块。中间 CTC 递增计划在编码器激活训练期间推进，从第 6 个训练轮次开始。三个交叉熵项的权重均为 1，且不使用标签平滑。中间 CTC 使用与最终状态相同的输出归一化和音素读出，在最终层权重 0.3 之外，对第 3 层和第 5 层的总权重为 0.05。CTC 实现使用空白索引 0，并抑制不可行对齐所导致的无穷损失。

我们对两个编码器主干中的线性权重矩阵使用 Muon，动量设为 0.95，采用 Nesterov 更新、五次 Newton–Schulz 迭代、权重衰减 0.01，以及峰值学习率 4×10^{-5}（Jordan 等，2024）。对于形状为 $m \times n$ 的权重，更新采用矩阵形状缩放 $0.2\sqrt{\max(m, n)}$。输入投影、接口、音素读出层、角色适配器，以及编码器偏置和归一化参数使用 AdamW，其参数为（β_1, β_2） = （0.9,0.999）、 $\epsilon = 10^{-8}$; 矩阵参数使用权重衰减 10^{-5}，偏置和归一化参数的权重衰减则为零。AdamW 的峰值学习率分别为: 编码器偏置和归一化参数、所有输入投影以及音素读出层使用 2×10^{-5}，接口和适配器使用 10^{-4}; 对齐期间，新输入投影使用 10^{-4}。在梯度累积和分布式归约后，我们对激活参数的并集统一进行一次梯度范数裁剪，阈值为 1。学习率在每个区块内分别遵循预热–余弦曲线。对于包含 K 次优化器更新的区块，令 $w = \max(1, \text{round}(\omega K))$，并令 t 表示该区块中已完成的更新次数。某个激活组的学习率为

$$\eta_g(t) = \eta_g^{\text{peak}} \, \mu_c \, s(t), \qquad s(t) = \begin{cases} 0.1 + 0.9 t / w, & t < w, \\ 0.1 + 0.45 \left[1 + \cos\!\left(\pi \, \frac{t - w}{K - w} \right) \right], & t \geq w, \end{cases} \tag{7}$$

五轮训练的阶段使用峰值乘数 1；最终的 40 轮联合阶段使用 0.125。整个训练过程中，AdamW 均采用预热比例 $\omega = 0.05$；在每个编码器激活的五轮阶段中，Muon 采用 0.10，而在联合阶段采用 0.05。每次阶段转换时，阶段内的预热都会重新开始，但训练期间始终保留优化器动量。学习率计划是固定的，不会因验证集平台期而降低。

我们使用八个数据并行秩进行训练，有效批量大小为 32。每个秩累积两个微批次，每个微批次包含两个话语；每个微批次中的录音均来自同一参与者。在每个秩内，交叉熵总和除以该窗口实际的目标词元数量（包括 EOS），CTC 总和则除以该窗口的目标音素数量。更新时对八个由此得到的秩本地目标取平均值。最后一个部分窗口使用其实际计数，填充或未激活的重复槽位不产生损失。偏置和一维参数不进行权重衰减。训练使用 BF16 自动混合精度和随机种子 42。验证集生成和损失评估的批量大小均为 1。

A.5 检查点选择与评估

每个训练轮次都会在所有提供的带标签验证话语上，通过贪心生成句子进行评估。检查点选择从第 25 个训练轮次开始，此时完整的对齐和专家–仲裁器前缀已完成。更早的训练轮次会进行评估，但不具备选择资格。我们保留在 2,306 个验证话语（14,941 个参考词

）上汇总词编辑数最少的合格检查点；如果连续 15 个训练轮次没有严格改进，则停止训练，或最迟在第 65 个训练轮次停止。参考词分母是固定的，因此最小化编辑次数等价于最小化语料库验证集词错误率。若出现并列，则保留较早的检查点；阶段转换不会重置耐心计数，训练可能在一个阶段块内停止。

T15 训练集包含已发布的三个“ 不做任何处理”对照试验。所有声明的行均保留在训练和评估中。带标签验证集用于检查点选择和专家预测分析。基准比较和全部五个训练后对照均报告留出集词错误率；T15 公开分数和私有分数保持分开。对于最终的留出集解码，选定检查点使用 FP32 和批量大小 1。专家和仲裁器生成均采用贪心解码，T12 和 T15 分别固定上限为 24 个和 29 个新词元，并在遇到序列结束词元时提前终止。参考转录文本不用于生成、停止和候选项选择。

A.6 消融对照的实现

所有对照共享数据划分、数据增强、有效批量大小 32 和选择规则。对应组件使用共同的初始权重；后续的独立训练可能会改变其表征。

单个智能体，两个视图。 输入拼接 Z_N 和 Z_P，并包含学习得到的视图标识符。其提示要求根据神经活动和音素概率特征生成英语转录文本。单个适配器从未经训练的仲裁器适配器开始，并在对齐与编码器适配之后安排的全部 55 个训练轮次中保持激活状态。其三个为期五个训练轮次的专业化模块每次都会更新解码器，其中中间模块还进行音素读出更新，之后进入最终的 40 个训练轮次联合模块。解码器使用一个角色适配器和单个句子损失。另一项容量对照将该适配器的秩提高到 24，缩放参数设为 96，与相同目标模块上的三个秩为 8 的角色适配器相匹配，同时保留缩放因子 4。其初始 LoRA 因子沿秩维度拼接三个未经训练的角色适配器；初始有效更新仍为零，其余单智能体训练和评估设置保持不变。我们还评估了一个更大骨干网络的单智能体对照：将冻结的 Qwen3-ASR-1.7B 文本解码器（宽度 2048）替换为 Borealis-5B（VikhrModels，2025年）的 4B 经音频训练的文本解码器和分词器（宽度 2560，36 层），并使用秩为 13、缩放参数为 52 的 LoRA 适配器，共使用 26,836,992 个 LoRA 参数，比三个秩为 8 的角色适配器合计多 2.63%，同时在相同的分阶段训练计划下，将神经与语音视图投影到宽度 2560。

顺序双视图细化。在此对照实验中，我们保留 MAS-BIT 完全相同的可训练参数清单、三个秩为 8 的角色适配器、三个权重为 1 的句子交叉熵损失，以及最终和中间 CTC 损失，同时以顺序双视图起草与细化取代视图专属专家专门化。第一个智能体接收拼接后的连续视图 $[Z_N;\ Z_P]$ 以及单智能体双视图提示，生成贪心草稿，再通过其适配器重放该草稿，以构造前16概率加权潜在消息 M_1。第二个智能体接收 $[Z_N;\ Z_P;\ M_1]$ 以及成对的 <|神经_音素_和_草稿|> 分隔符，并附有指令“利用上述神经活动和音素概率特征，将第一份草稿润色成一个英语句子：”，随后生成润色后的贪心草稿，并重放该草稿以构造 M_2；仲裁器接收 $[Z_N;\ M_1;\ M_2]$ 以及 MAS-BIT 的仲裁器提示，复用 M_1 及其计算图，不再重新生成或重放它。该设置在推理时与 MAS-BIT 一样，进行三次贪心自回归生成和两次消息重放（训练时保留消息角色嵌入和消息丢弃），并采用相同的模型初始化、六阶段 65 训练轮次计划（在 40 轮联合优化之前，分别在三个为期五轮的专门化模块中更新第一草稿、细化和仲裁器适配器）、优化器超参数以及合并验证集检查点选择。

硬文本通信。 每条消息都使用专家贪心草稿中的冻结词元嵌入，而不是概率加权嵌入。保留向仲裁器提供的直接神经输入和消息角色嵌入。离散消息会阻断传向发送专家的仲裁梯度，而各专家仍保留自身的句子监督。因此，该对照实验同时改变了保留的词元信息和梯度路径。

分离式潜在消息。 在该对照实验中，每个专家重放其贪心草稿，以构建相同的前16个概率加权词元嵌入，同时在消息构建之前分离重放逻辑值，使仲裁器交叉熵损失不会通过消息反向传播至发送专家、投影器或神经编码器。这样既保留了前向传播中的备选词元假设，又移除了消息梯度路径。在与 MAS-BIT 相同的模型初始化、分阶段训练计划、优化器超参数和验证检查点选择条件下，向仲裁器提供的直接神经输入、消息丢弃、消息角色嵌入以及各专家自身的句子监督均予以保留。

仅消息仲裁器。 在此对照中，仲裁器仅接收拼接后的潜在消息$[M_N;\ M_P]$，在训练、验证和留出集解码期间均省略直接神经输入 Z_N。为引导解码，仲裁器提示指示模型将两个智能体的转录估计翻译成一个英语句子。两条消息仍为可微的前16概率加权词元嵌入，并带有消息角色嵌入和消息丢弃，因此仲裁器的交叉熵损失能够通过消息反向传播至发送专家、投影器和神经编码器。我们沿用与 MAS-BIT 相同的模型初始化、分阶段训练计划、优化器超参数和验证集检查点选择。其留出集词错误率比硬文本对照在 T12 上高 1.29 个百分点，在 T15 公开集/私有集上分别高 0.76/1.08 个百分点（表16）。

两个神经专家。 第二个专家接收与第一个专家相同的融合神经视图，以第二个独立适配的神经专家替代音素后验视图。两个专家均保留各自的适配器和消息角色标识符；相应的专家模块会更新第二个神经适配器。两个编码器主干、直接神经输入、由分阶段转为联合训练的计划以及辅助音素监督均予以保留。

参与者独立训练。 我们分别在 T12 的 8,800 条训练话语和 T15 的 8,072 条训练话语上训练两个独立的 MAS-BIT 模型，同时保留双编码器前端、共同初始化、65轮分阶段转联合训练计划、32 的有效批量大小以及优化器设置。分块学习率计划根据各参与者的训练轮次长度设定，因此 T12/T15 每轮分别进行 275/253 次优化器更新，而联合训练每轮为 528 次。从第 25 轮起，依据相应参与者的完整带标签验证集（T12 为 880 条话语，T15 为 1,426 条话语）选择检查点，并采用相同的整数词编辑准则、较早检查点平局规则以及 15轮耐心值，然后进行 FP32留出集评估。

无分阶段训练。 所有适配组（包括两个编码器、输入投影、读出层、接口和角色适配器）均从第 1 轮开始更新。我们采用单一的 65 轮预热—余弦学习率计划，其中 Muon 的预热比例为 10%，AdamW 为 5%，并保留相同的优化器、最大训练轮次预算和汇总检查点选择规则。

B 专家预测分析

我们分析所选联合检查点在完整 FP32、批量大小为一条件下的验证预测：880 条 T12 话语和 1,426 条 T15 话语。这些带标签验证集支持对专家错误的分析。三个智能体均基于同一录音和检查点进行贪心生成，生成完成后才使用参考文本。仲裁、草稿重组、消息干预和输入掩码均使用这些专家输出以及相同的完整输入仲裁器基线。我们首先比较同一模型基于相同录音生成的三种不参考文本的输出。

表5：仲裁优于两种专家输出。词编辑数和完全匹配数涵盖所有验证集话语。所有三种输出均在不使用参考文本的情况下生成。

	T12		T15	
输出	**词编辑数**	**完全匹配**	**词编辑数**	**完全匹配**
神经专家	778	477	627	1078
语音专家	801	492	703	1059
仲裁器	557	570	420	1189

在 T12 上，仲裁将神经/语音专家的 778/801 词编辑数降至 557，分别减少 28.4%/30.5%；完全匹配数则从 477/492 增至 570。在 T15 上，相应的编辑次数从 627/703 降至 420（分别减少 33.0%/40.3%），完全匹配数则从 1078/1059 增至 1189。

B.1 超越对齐草稿重组的仲裁

在查看参考文本之前，我们根据神经草稿和语音草稿构建词语对齐图。对于替换、删除和插入，我们均采用单位代价，并保留所有最小代价对齐。在每条对齐边上，图中都提供任一专家的词语；若对齐处存在空缺，则提供空选项。该图保留了词语顺序，并包含两份原始草稿。对于其候选句子集合 $\mathcal{L}_i$，与参考文本之间的最小编辑距离为

$$e_i^{\mathrm{comb}} = \min_{z \in \mathcal{L}_i} \mathrm{ED}(z, y_i).$$

构建图后，我们通过对图和参考文本位置进行动态规划来计算该距离。当 $e_i^{\mathrm{comb}} > 0$ 时，精确的仲裁器恢复超出了对齐草稿重组的范围。

在两份草稿都出错时恢复的31/49个句子中，5/11个可通过对齐图获得，另有26/38个不在图中（图4）。后者中有26/36个包含两份草稿中都没有的正确词语。这些恢复结果表明，仲裁器能够生成超越专家草稿对齐组合的正确答案。

词级互补性。 我们使用相同的单位编辑代价，将每个草稿和仲裁器输出与参考文本对齐，并按对角线、删除、插入的顺序解决平局。在每个参考词位置，我们记录哪些专家是正确的，以及仲裁器是否保留了正确词语。在 T12/T15 上，神经专家分别在 259/264 个位置上唯一正确；仲裁器在其中分别有 222/233 个位置正确。语音专家分别在 241/187 个位置上唯一正确，其中分别有 190/150 个位置被仲裁器保留。

表6：恰好在一个专家草稿中正确的词位置。 各列标识正确的专家；替换/删除计数描述另一位专家的错误。训练频率根据汇总的神经微调转录文本计算。音素距离取词典覆盖的替换的平均值，并明确显示覆盖率。

	T12		T15	
数量	**神经**	**语音**	**神经**	**语音**
词位置	259	241	264	187
仲裁者正确	222	190	233	150
其他专家：替换	254	237	261	184
其他专家：删除	5	4	3	3
实词	204	185	215	153
虚词	50	53	42	29
其他标签	5	3	7	5
训练中未见	31	22	31	21
训练计数 1–5	71	51	86	60
训练计数 6–20	48	44	45	41
训练计数 > 20	109	124	102	65
平均音素距离	0.597	0.534	0.594	0.528
词典覆盖率	245/254	223/237	251/261	173/184

替换占专家不一致的大多数（表6）。我们统计汇总的16,872份神经微调转录文本（8,800份T12和8,072份T15）中的目标词出现次数，并将其分为未出现、出现1–5次、6–20次或超过20次。我们使用NLTK的上下文词性标签对词语进行分组：封闭类标签和固定的代词/助动词列表定义虚词，其余的名词、动词、形容词和副词标签定义实词。对于替换，我们使用去除重音的CMU 发音，并在所有发音对中取编辑距离除以较长发音长度所得值的最小值。词典中不存在的词语保留缺失的音素距离值。

仲裁器在T12/T15上分别保留了神经专家唯一正确词位置的85.7%/88.3%，以及语音专家唯一正确词位置的78.8%/80.2%，从而汇总保留率分别为87.0%（455/523）和79.4%（340/428）。在汇总神经微调转录文本中未出现的445个参考词位置中，神经专家和语音专家分别在62个（13.9%）和43个（9.7%）位置上唯一正确；仲裁器分别保留了其中58/62（93.5%）和36/43（83.7%）个正确词语。在语音专家唯一正确的位置上，另一位专家的词典覆盖的替换具有更小的平均归一化音素距离（T12/T15上为0.534/0.528），而在神经专家唯一正确的位置上则为0.597/0.594；汇总平均值分别为0.532和0.596。这些互补的词级贡献通过仲裁得以保留，包括那些在神经微调文本中未出现的词语。

B.2 专家消息补充直接神经条件输入

我们冻结选定检查点，并在五种输入条件下对完整验证集进行解码（表7）。对于每个话语，专家各自生成一次草稿和潜在消息；随后，仲裁器分别在完整输入、单独掩蔽每条消息、同时掩蔽两条消息，或将直接神经输入置零的情况下进行解码。消息掩蔽遵循训练时的 dropout：在添加其角色嵌入之前将消息内容置零，同时保留序列长度、位置和有效性掩码。将投影后的神经输入置零不会改变两条专家消息。

表7：同时使用两条专家消息和直接神经条件输入时，仲裁准确度最高。词语编辑数、语料库验证集词错误率（%）和句子完全匹配数涵盖全部 880/1,426 个 T12/T15 话语以及 5,496/9,445 个参考词。所有条件均使用固定权重、FP32批量为一的解码，以及相同的缓存专家输出。完整输入下的预测与前述仲裁分析及 Top-16消息条件相符。

	T12			T15		
仲裁器输入	**编辑数**	**词错误率**	**完全匹配**	**编辑数**	**词错误率**	**完全匹配**
完整输入	557	10.135	570	420	4.447	1189
神经消息已掩蔽	605	11.008	559	488	5.167	1158
语音消息已掩蔽	631	11.481	545	488	5.167	1153
两条消息均已屏蔽	784	14.265	465	672	7.115	1043
直接神经输入已置零	808	14.702	518	5092	53.912	544

完整输入使验证集词错误率最低，并使两位参与者的完全匹配数最多。掩蔽任一消息、两条消息或直接神经输入都会增加词错误并减少完全匹配数。掩蔽两条消息会使词错误率增加 4.130/2.668 个百分点，并使完全匹配数从 T12/T15 的 570/1189 降至 465/1043。在权重和专家草稿保持不变的情况下，这些干预表明，训练后的仲裁器会利用两条专家消息和直接神经证据来提高全数据集解码准确率。

针对具体试次的神经输入支持草稿之外的恢复。我们追踪了在两份专家草稿都错误的话语中，全输入仲裁器的精确恢复结果的保留情况。将 Z_N 置零，同时保留两条完整的潜在消息后，T12/T15 上原有的 31/49 项恢复中仍保留了 2/4 项。原有的 26/38 项图外恢复均未能保留，其中包括 26/36 项包含了两份草稿中都没有的正确单词的恢复。

为使用非零输入测试神经内容，我们仅将 Z_N 替换为同一参与者另一验证试次的表征，且该试次具有相同的有效神经标记数量，同时固定接收试次的潜在消息和掩码。三种确定性供体映射在不参考参考文本或假设文本的情况下进行选择，优先选择同一录制日期的供体，其次选择日期最近的供体；当可用供体少于三个时，则循环重复使用。配对子集包含 T12 的 856/880 条话语和 T15 的 1,363/1,426 条话语，参考词共 5,258/8,902 个；其他行没有长度匹配的供体。自我替换会在每条话语上重现基线标记序列。在三种供体映射下，平均词编辑数从 494/398 上升至 5,611/9,589，配对集词错误率从 9.395%/4.471% 上升至 106.714%/107.717%。每种映射都会消除这些子集中的全部 30/43 个原始双方均错误的恢复结果，包括全部 25/36 个图外恢复结果。这些固定权重干预表明，仲裁器会结合专家消息使用试次特定的神经内容。

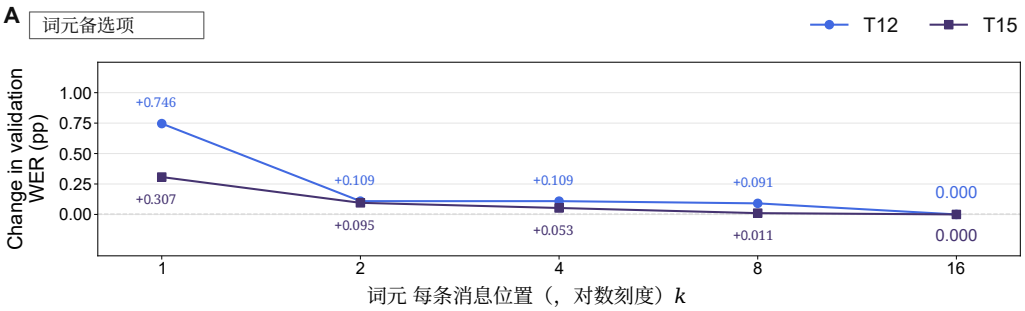

图 5：词元备选项可改善两位参与者的解码效果。语料库验证集词错误率的变化是相对于 $\tau = 1$ 处的 Top-16 消息计算的，涵盖全部 880 条 T12 验证话语和 1,426 条 T15 验证话语。正值表示词错误率更高。词元支持集在对数坐标轴上从 1 到 16 变化。数值标注到小数点后三位。模型权重、专家草稿、直接神经输入以及贪心解码仲裁器解码均保持不变。

B.3 冻结专家消息中的词元备选项

在固定选定的模型权重、专家贪心草稿以及仲裁器的直接神经输入时，我们改变专家消息。对于每条话语，我们缓存神经表征，以及沿每位专家自身草稿回放得到的前 16 个词元对数几率。在每个消息位置保留前 $k \in \{1, 2, 4, 8, 16\}$个候选项，在 $\tau = 1$处重新归一化其概率，并根据公式 3 构造加权嵌入。两位专家使用相同的 k；消息长度、掩码和学习得到的角色嵌入保持固定。

仲裁器始终采用相同的贪心解码规则。

图 5 展示了相较于原始前16消息，语料库验证集词错误率的变化。表 8 和表 9 报告了完整的提供验证集以及各专家草稿组内的绝对词错误率。专家一致、不一致和正确性均根据原始固定草稿在统一文本规范化规则下确定。这些比较衡量的是固定权重下的词元支持敏感性；附录 A.6 中的硬文本对照则使用离散消息接口进行训练。

表 8：固定模型权重和专家草稿下的词元支持集。在 $\tau = 1$，报告全部 880 条 T12 和 1,426 条 T15 提供验证话语上的语料库验证集词错误率（%）和词编辑数。对应的参考词分母分别为 5,496 和 9,445。前1会改变训练后的 MAS-BIT 模型的消息；这不同于单独训练的硬文本对照。

	T12		T15	
消息	**编辑数**	**验证集词错误率**	**编辑数**	**验证集词错误率**
前1	598	10.881	449	4.754
前2	563	10.244	429	4.542
前4	563	10.244	425	4.500
前8	562	10.226	421	4.457
前16	557	10.135	420	4.447

增加词元支持集后，T12 上的词编辑数从 598 降至 557，T15 上则从 449 降至 420。保留 16 个词元备选项而非仅保留一个，可分别减少 41/29 个词编辑数，使验证集词错误率降低 0.746/0.307 个百分点。这些结果表明，多个词元假设所带来的贡献超出了专家排名最高的词元。

C 重现级联

神经编码器和语言模型。 我们使用已发布的BIT编码器，为每位参与者复现级联解码（Zhang 等，2026年）。这些参与者专属检查点还用于初始化在 MAS-BIT 中联合适配的两个编码器。它们是针对句子级解码微调的检查点——

表9：固定专家草稿组中的词元支持敏感性。数值为语料库验证集词错误率（%）。各组在干预前定义。一致与不一致分别划分每个验证集；双方错误语句与这些组存在重叠。

参与者	专家草稿	Rows	词语	前16	前1
T12	一致	460	2761	2.789	2.970
T12	不同意	420	2735	17.550	18.867
T12	两者都错	305	1975	25.924	28.101
T15	同意	1015	6565	0.716	0.716
T15	不同意	411	2880	12.951	13.958
T15	两者都错	262	1829	21.542	23.127

级解码；用于音素级解码并产生已发表级联结果的编码器并不属于公开发布版本，因此，复现的级联系统与已发表的级联系统预计不会完全一致（表15）。

我们使用基准测试的 OpenWebText2 五元语法加权有限状态转换器（WFST）解码音素逻辑值。在提取最多 100 个假设之前，我们通过用未剪枝模型替换经剪枝的五元语法贡献，对格进行重新评分。所有实验均固定此候选项上限； K 仅控制传递给 OPT-6.7B（Zhang 等人，2022年）的已排序假设数量。如果可用假设少于 K ，则保留全部假设。表10总结了解码器参数。

表10： 级联解码所用的超参数。

参数	值
最大五元语法假设数	100
剪枝波束 / 格点波束	17 / 8
最大 / 最小活跃状态数	7,000 / 200
声学分数缩放	0.325
空白惩罚 / 空白跳过阈值	90 / 1
OPT 重评分权重 (α)	0.55
长度惩罚	0

对于每个 K，输出会使 $s(h) = 0.325\, s_{\mathrm{acoustic}}(h) + 0.45\, s_{5\mathrm{gram}}(h) + 0.55\, s_{\mathrm{OPT}}(h)$ 最大化，其中 OPT 分数是词元对数概率之和，且未进行长度归一化。若分数相同，则保留第一个候选项。OPT 会对所有保留的候选项重新评分，包括 $K = 1$ 处的单个候选项。我们不会使用参考标签来生成或排序假设。

我们保留候选句子及其 OPT 分数，并在每个候选项上限下应用相同的选择规则。缓存的分数能够复现分别在原始七个候选项上限下解码得到的转写文本。在获取留出集分数之前，我们会预先指定所有额外的候选项上限。对于每个候选项上限，我们都会评估所有留出集句子的词错误率；T15 公开集和私有集的分数分别对应不同的隐藏子集。神经解码和 OPT 评分均使用 FP32，不使用 TF32，并且每次处理一个话语或候选句子。图 2B 展示了候选项上限对词错误率的影响。直线段连接各个实测设置；未拟合平滑曲线。为评估编码器适配后的级联解码，我们还使用每位参与者在 MAS-BIT 中微调后的编码器分支、输入投影和最终 CTC 读出层，在 13 个候选项上限下运行相同的 FP32 流水线（表15）。在 $K = 100$ 时，与发布的编码器相比，这使级联系统的留出集词错误率在 T12 上降低了 0.58 个百分点，在 T15 公开集/私有集上分别降低了 1.27/1.40 个百分点。

峰值主机内存和 GPU 显存。 我们比较完整推理运行过程中的峰值主机内存和 GPU 显存，包括模型与数据加载、解码和审计。测量使用单个 NVIDIA A40 GPU 和 AMD EPYC 7662 CPU 的 16 个核心，采用 FP32 计算和批量大小 1。每次运行都会加载解码器，并完成一名参与者的全部 1,200 条 T12 或 1,450 条 T15 留出语句。MAS-BIT 对两名参与者使用相同的联合训练模型；复现的级联使用发布的参与者专属检查点，其中包含 $K = 100$，包括 WFST 解码、未剪枝五元语法重评分和 OPT重评分。我们还在相同的 FP32 设置下评估 BIT +LightBeam（Feghhi 等，2026）（波束大小 1,000/900、100k 词典、剪枝后的 4-gram KenLM，以及

配备复现的下一词预测适配器的 Llama-3.2-1B），在 T12 上获得 8.066% 的留出集词错误率，在 T15 公开集/私有集上分别获得 6.151%/7.438%（表15）。我们每秒采样一次进程树的主机驻留内存和设备使用的 GPU 显存，并报告各自的峰值。解码时间不包括加载和审计，四舍五入到 0.1 秒；GPU 数值来自匹配的重复运行。

离线解码时间。一秒间隔的阶段时间戳估算，完整留出集解码耗时约为：MAS-BIT 在 T12/T15 上分别为 1538.8/2091.4 秒，而复现的 $K = 100$ 级联系统分别为 2789.6/3761.6 秒；在两位参与者身上，前者都快约 1.8× 倍。这些时长涵盖了在上述单张 A40、16 核、FP32、批量大小为一的设置下解码全部 1,200/1,450 条话语；不包括模型/数据加载和审计阶段。在相同设置下，BIT + LightBeam 在 T12/T15 上分别耗时 3563.5/6008.0 秒（分别是 MAS-BIT 解码时间的 2.3×/2.9× 倍，以及 $K = 100$级联系统解码时间的 1.3×/1.6× 倍）。

消费级笔记本电脑验证。我们还在一台消费级笔记本电脑上将 MAS-BIT 与 BIT + LightBeam（Feghhi 等，2026）进行了比较。该笔记本电脑配备 NVIDIA GeForce RTX 5090 笔记本电脑 GPU、Intel Core Ultra 9 275HX 处理器和 62.3 吉比字节主机内存，运行于 WSL2 下的 Ubuntu。两个系统采用相同的软件环境、FP32 计算、批量大小 1 和 16 个逻辑 CPU。每次运行均处理全部 1,200 条 T12 或 1,450 条 T15 留出话语，且所有预测结果均与对应的 A40 结果一致。我们每秒采样一次主机驻留内存和设备已用 GPU 显存，并报告加载、解码和审计期间的峰值；解码时间不包括加载和审计。由于测得的主机内存占用为 321.8/334.8 吉比字节，超过了笔记本电脑可用内存，因此未纳入复现的 $K = 100$级联系统。MAS-BIT 在 T12/T15 上完成留出集解码分别耗时 1210.2/1690.6 秒，而 BIT + LightBeam 耗时 3204.6/6021.1 秒；MAS-BIT 的主机内存和 GPU 显存峰值更低（表 11）。

表 11：消费级 *RTX 5090* 笔记本电脑 *GPU* 上的峰值主机内存、*GPU* 显存和解码时间。每个数值汇总了在 WSL2 下进行的一次完整留出集运行。内存数据包括加载和审计；解码时间不包括这两项，并四舍五入到 0.1 秒。GPU 数值为设备使用量的一秒峰值。

	主机内存（吉比字节）		GPU 峰值显存（吉比字节）		解码时间（秒）	
方法	**T12**	**T15**	**T12**	**T15**	**T12**	**T15**
BIT +LightBeam（Feghhi 等，2026）	17.3	32.0	12.629	14.785	3204.6	6021.1
MAS-BIT（我们的）	**12.6**	**31.0**	**8.662**	**8.648**	**1210.2**	

消费级台式机验证。 我们在一台配备 NVIDIA GeForce RTX 4090 GPU、Intel Core i9-14900KF 处理器（5.6 GHz P 核频率上限）和 61.5 吉比字节主机内存、运行原生 Ubuntu 的消费级台式机上重复了这一比较。两套系统均使用相同的 CUDA 13.2 软件环境、FP32 计算、批量大小 1，以及 16 个绑定的物理核心（八个 P 核和八个 E 核）。每次运行均处理全部 1,200 个 T12 或 1,450 个 T15 留出语句，且所有预测结果均与相应的 A40 结果一致。我们每秒采样一次主机驻留内存和每个进程的 GPU 显存，并报告加载、解码和审计期间的峰值；解码时间不包括加载和审计。由于测得的 321.8/334.8 吉比字节主机内存占用超过了该台式机的可用内存，因此省略了复现的 $K = 100$级联系统。MAS-BIT 在 T12/T15 上完成留出集解码分别耗时 550.5/775.1 秒，而 BIT+ LightBeam 为 1483.5/2332.7 秒，同时具有更低的主机内存和 GPU 显存峰值（表 12）。

表 12：消费级 *RTX 4090* 台式机 *GPU* 上的峰值主机内存、*GPU* 显存和解码时间。每个数值均汇总了在原生 Ubuntu 上完成的一次完整留出集运行。内存数据包括加载和审计；解码时间不包括这两项，并四舍五入至 0.1 秒。GPU 数值为进程级峰值，按一秒间隔测得。

	主机内存（吉比字节）		GPU 峰值显存（吉比字节）		解码时间（秒）	
方法	**T12**	**T15**	**T12**	**T15**	**T12**	**T15**
BIT +LightBeam（Feghhi 等，2026）	17.4	32.0	12.648	14.887	1483.5	2332.7
MAS-BIT（我们的方法）	**12.8**	**31.2**	**8.695**	**8.695**	**550.5**	**775.1**

D 基准测试与控制结果

表15和表16分别给出了图 2B 和图 3B–C 中汇总的确切留出集结果。表15还报告了涵盖全部 13 个候选项上限的微调编码器级联对照，以及 BIT + LightBeam（Feghhi 等，2026）的结果。表16还报告了秩为 24 的单智能体容量对照、Borealis-5B 单智能体骨干网络对照、顺序双视图细化对照、分离潜在消息对照和仅消息仲裁器对照。表 1 报告了已发表的基准比较结果。

已发布的端到端协议。 Feng 等人和 BIT 的分数遵循 Zhang 等人（2026）的基准测试表；BIT 使用核采样，而 MAS-BIT 使用贪心解码。MoDAl-1 使用 6v 区域编码器，而 MoDAl-Full 结合了三个神经编码器，并包含区域44信号（Chen 和 Chin，2026）。它们的 18.8% 和 17.7% 结果来自包含 1,200 个句子的竞赛留出集，置信区间基于五个随机种子计算；880 个句子的模型选择集上的结果则单独列于表 13 中。两种配置均使用 CTC 预训练，随后进行直接的自回归文本解码。BIT 端到端 +集成通过大语言模型合并器组合独立训练的解码器，并作为单独的集成参照保留。T12 系统使用经过微调的 GPT-3.5 合并十个模型，而 T15 系统使用经过微调的 GPT-4 合并 29 个模型；两个合并器均接收句子假设和音素输出（Zhang 等人，2026）。

其他已发表的端到端变体。表 1 中的 RNN、BIT-TFS 和 BIT-Human 端到端配置通过改变 BIT 中的编码器实现不同变体（Zhang 等，2026）。我们保留其图 2A 中列出的 T12 留出集和 T15 公开留出集词错误率数值，并保留一位小数。RNN 配置将其编码器与端到端大语言模型解码器配对。BrainWhisperer 具有直接的 Whisper 解码路径，以及独立的音素/WFST 路径（Boccato 等人，2026）；表 1 纳入了其直接解码竞赛结果。BrainWhisperer-cross 在 Card、Willett 和 Kunz 数据集上联合训练。BEST 在基于 Whisper 的解码器之前融合分层音素与字符表示（Feng 等人，2026）。表 1 使用该论文表 1 中无外部语言模型行的数据：T12 留出集上的词错误率为 12.61%，T15 留出集上为 5.17%；后者未说明采用公开集还是私有集子集。

其他评估集上的解码。 表 13 扩展了神经信号转文本解码器的覆盖范围，同时保留各项研究各自的评估集和推理流程。由于数据划分、训练数据和筛选流程各不相同，这些数值与匹配的留出集比较分开报告。Fiedler 等人（2025年）从每个训练日的首个数据块中划出数据，构成一个自定义的 880句测试集，并保留 7,920 个训练句子；原始带标签测试集用作验证集。表中列出的两个模型均经过完全微调；我们报告词错误率列中各设置下表现最佳的一次运行，且未使用外部语言模型解码。

Olak 等（2026年）排除了竞赛留出数据，并报告了基于 Transformer 和 BART 词解码器的原生贪心解码与束搜索解码。训练期间，梅尔频率倒谱系数目标用于提供辅助监督，而非在推理时作为额外输入。他们的论文将这组 880 句的带标签数据同时称为测试集和验证集，因此我们采用中性的“带标签集”称谓。内部音素选择器会利用神经条件音素似然，或字素到音素转换（G2P）后与预测音素的一致性，对词解码器候选项进行评分。这些配置不使用独立的语言模型分数，因此与原生词解码分开列出；候选项通过束搜索和核采样生成。

BrainWhisperer 的 Card测试结果与竞赛列中的结果不同；未注明测试集规模和划分标识符（Boccato 等人，2026年）。其跨数据集模型是一个联合训练解码器，而非集成。对于 Feng 等人（2024年），我们区分完整的 880句测试集及其 720句发声子集。我们列出在用于模型选择的 880句数据集上报告的额外 MoDAl 数值，并据此进行分组（Chen 和 Chin，2026年）。

其他解码协议。 Khanday 等人（2026年）报告了验证集字符错误率，但未报告词错误率，因此其字符级 CTC 模型在此处没有数值条目。Fogg 等（2026年）的多用户系统和 Benster 等人（2024年）、Yoon等人（2026年）的神经网络集成模型采用音素到单词的语言模型解码，仍属于级联系统。LightBeam 使用词典约束和外部

表 13：作者定义的评估集上的神经信号到文本解码结果。数值为报告的词错误率（%）。Fiedler 等人报告的是全量微调的最佳运行结果。Olak 等报告的是五个随机种子上的均值 ± 标准差；MoDAl 的数值为五个随机种子上的均值 ± 95% 置信区间。

模型	评估集	推理	词错误率 (%)
直接解码			
Wav2Vec2Conformer（Fiedler 等人，2025年）	T12定制测试集（880）	字符级CTC；无外部语言模型	26.67
Wav2Vec2（Fiedler 等人，2025年）	T12定制测试集（880）	字符级CTC；无外部语言模型	30.97
Transformer +梅尔频率倒谱系数 +BART（Olak 等人，2026年）	T12 标注集（880）	原生束搜索	24.8 ± 0.4
Transformer +梅尔频率倒谱系数 +BART（Olak 等人，2026）	T12 标注集（880）	贪心解码	25.6 ± 0.2
BrainWhisperer（Boccato 等人，2026年）	T15卡片测试	原生 Whisper 解码器	9.8
BrainWhisperer-跨模态（Boccato 等人，2026年）	T15卡片测试	原生 Whisper 解码器	8.9
GRU-BPE + Llama2-7B（Feng 等人，2024年）	T12 带标签测试集（880; all）	自回归	28.5
GRU-Phone + Llama2-7B（Feng 等人，2024年）	T12标注测试（880; all）	自回归	27.3
BGRU-Phone +Llama2-7B（Feng 等人，2024年）	T12标注测试 (880; all)	自回归	26.3
BGRU-Phone +Llama2-7B (Feng 等人，2024年)	T12发声子集 (720)	自回归	26.2
内部音素选择的词解码			
Transformer + 梅尔频率倒谱系数 +BART（Olak 等人，2026年）	T12 标注集（880）	音素头似然；G2P	22.6 ± 0.2
Transformer +梅尔频率倒谱系数 +BART（Olak 等人，2026年）	T12 标注集（880）	预测音素一致；G2P	23.2 ± 0.4
用于模型选择的标注评估			
MoDAl-1（Chen 和 Chin，2026年）	T12 选择集（880）	自回归	22.4 ± 0.2
MoDAl-Full（Chen 和 Chin，2026年）	T12 选择集（880）	自回归	21.6 ± 0.1
MoDAl，两个 area-6v 编码器（Chen 和 Chin，2026年）	T12 选择集（880）	自回归	22.5 ± 0.3

语言模型融合（Feghhi 等，2026年）；根据我们的定义，移除加权有限状态转换器并不会使其部署方式成为端到端，而表15报告了我们复现的 BIT + LightBeam 留出集词错误率。

表14：已发表的级联基线。数值为 Zhang 等人（2026）在基准测试表中报告的语料库留出集词错误率（%）。这些系统使用外部语言模型解码，并非单模型端到端基线。已发表的 BIT 级联使用 $K = 100$；未报告 T15 私有评分。

方法	T12 留出集	T15 公开集	T15 私有集
循环神经网络基线（Willett 等人，2023年；Card 等人，2024年）	9.76	6.67	–
时间掩码Transformer（Feghhi等，2025年）	7.98	–	–
已发表的BIT级联（Zhang 等，2026）	6.35	4.06	–

表15：候选句子数量与留出集解码。 数值为词错误率（%）；T15 公开分数和私有分数分别对应不同子集。 K 是我们使用已发布的 BIT 编码器或 MAS-BIT 微调后的参与者专属编码器分支和音素读出所复现的级联系统中的候选句子上限；BIT + LightBeam 使用已发布的 BIT 编码器。蓝色底纹标示 MAS-BIT。

		T12	T15	
方法	K	**留出集**	**公开集**	**私有集**
复现的级联系统 + OPT（Zhang 等，2026）	100	7.877	5.969	6.026
	75	7.877	5.939	6.026
	50	7.931	5.939	6.057
	37	7.931	5.939	6.073
	25	7.917	5.939	6.120
	18	7.931	5.969	6.089
	12	7.958	6.030	6.136
	9	7.971	6.030	6.183
	6	8.012	6.090	6.214
	4	8.120	6.181	6.293
	3	8.256	6.090	6.387
	2	8.459	6.424	6.450
	1	9.298	7.272	7.062
微调编码器级联系统 + OPT	100	7.295	4.696	4.629
	75	7.308	4.696	4.629
	50	7.362	4.787	4.645
	37	7.376	4.787	4.645
	25	7.389	4.787	4.708
	18	7.403	4.818	4.676
	12	7.443	4.757	4.692
	9	7.538	4.727	4.739
	6	7.525	4.909	4.802
	4	7.660	5.060	4.959
	3	7.809	5.151	5.006
	2	8.012	5.363	5.037
	1	8.892	5.787	5.555
BIT +LightBeam（Feghhi 等，2026）	–	8.066	6.151	7.438
MAS-BIT（我们的方法）	–	**6.929**	**3.606**	**3.892**

表16：*MAS-BIT* 及其消融实验的留出集性能。数值为在相同隐藏留出集上的词错误率（WER，%）。五个核心控制项、秩-24单智能体变体、顺序双视图细化对照、分离潜在消息对照和仅消息仲裁器对照均使用共享骨干网络、双编码器前端和检查点选择协议；Borealis-5B 单智能体变体则将 1.7B 解码器替换为 4B 音频训练解码器和秩-13适配器。

	T12	T15	
方法	**留出集**	**公开集**	**私有集**
MAS-BIT	6.929	3.606	3.892
单个智能体，两个视图	9.988	6.121	6.669
秩-24适配器	10.231	5.787	6.889
Borealis-5B 骨干网络（秩-13适配器）	12.478	8.515	8.553
顺序双视图细化	7.159	3.939	4.739
硬文本通信	7.159	4.181	4.409
分离式潜在消息	7.349	3.666	4.315
仅消息仲裁器	8.445	4.939	5.492
两个神经专家	7.051	3.909	3.939
参与者独立训练	7.971	4.121	4.629
无分阶段训练	7.944	4.212	4.692

E 表征相似性分析

我们使用句子级表征相似性分析（RSA）（Kriegeskorte 等，2008年），考察学习到的接口如何组织神经证据。分析围绕架构的三个部分展开：融合接口的对齐、两个编码器分支的适配，以及专家输入视图的几何结构。以下所有比较均使用选定检查点及其初始化时的句子池化表征。

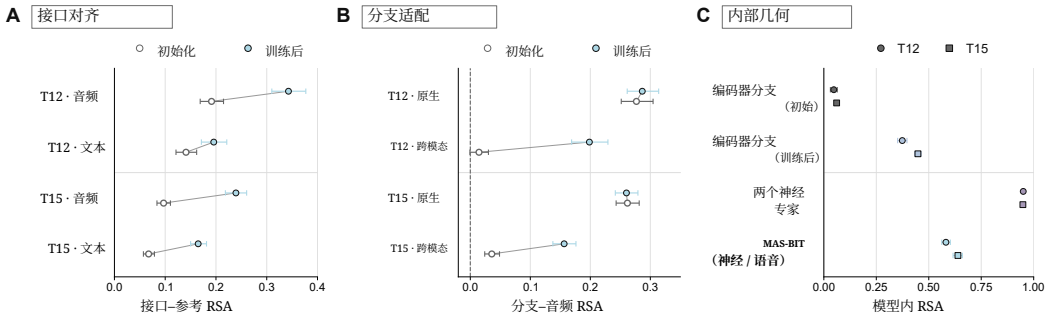

图 6：联合训练使神经接口实现对齐，同时保留不同的专家视图。（**A**）训练前后池化的 Z_N–音频/文本 RSA。（**B**）与音频参照的编码器分支对齐；原生分支和跨参与者分支根据输入参与者和编码器来源定义。（**C**）训练前后编码器分支之间的模型内 RSA，以及双神经专家对照和 MAS-BIT 中两个专家输入视图之间的 RSA。点和区间表示按词数分层的估计值及逐点 95% 句子自助法区间。音频参照使用固定音色文本转语音生成的语音，并通过通义千问 0.6B 音频塔提取；文本参照使用 1.7B 文本解码器。

E.1 联合训练使融合神经接口实现对齐

训练提升了 Z_N 与音频和文本句子几何结构的对齐程度（图6A；表17）。音频参照的表征相似性分析在T12上从0.191升至0.343，在T15上从0.097升至0.239。配对变化分别为 +0.152 [+0.123, +0.179] 和 +0.142 [+0.123, +0.163], ；对每位参与者而言，全部2,000个自助法变化值均为正。T12/T15上的文本参照对齐分别增加0.055/0.098，语音接口 Z_P 与音频参照的对齐程度也有所提升。这些结果表明，与训练相关的变化出现在专家和仲裁器读取的连续接口上。

表17：训练使神经接口和语音接口与参考句子几何结构对齐。数值为汇总后的按词数分层表征相似性分析结果。变化值采用配对自助抽样计算；括号给出逐点95%区间。

参与者	接口–参考文本	初始化	训练后	变化 [95% 间隔]
T12	Z_N–音频	0.191	0.343	+0.152 [+0.123, +0.179]
T12	Z_N–文本	0.141	0.196	+0.055 [+0.035, +0.074]
T12	Z_P–音频	0.199	0.233	+0.034 [+0.017, +0.051]
T15	Z_N–音频	0.097	0.239	+0.142 [+0.123, +0.163]
T15	Z_N–文本	0.067	0.165	+0.098 [+0.083, +0.113]
T15	Z_P–音频	0.050	0.133	+0.083 [+0.066, +0.101]

在所有五种长度处理方式下，Z_N–音频 RSA 的增幅均为正：词数分组、未分组 RSA，以及控制词数差异、有效神经位置数量差异或两者的偏相关（表 19）。在 T12 上，估计值介于 $+0.152$至 $+0.226$ 之间；在 T15 上，介于 $+0.125$至 $+0.142$ 之间。

E.2 跨参与者分支获取共享句子结构

每段录音都通过针对参与者定制的输入投影，由两个预训练编码器进行处理。对于 T12，以 T12 初始化的编码器是原生分支，而以 T15 初始化的编码器则是

跨参与者分支；对于 T15，这两个角色则互换。跨参与者输入投影经过全新初始化，而原生投影则保留发布时的初始化状态。融合接口将这些初始化方式不同的路径组合起来，并针对转录文本进行联合学习；其训练相关对齐增益是相对于初始化后的融合接口来衡量的。

原生分支在训练前后保留了相似的音频参照对齐：T12 上从 0.277 变为 0.287，T15 上从 0.262 变为 0.260。跨参与者分支的变化幅度更大，T12 上从 0.014 变为 0.198，T15 上从 0.036 变为 0.156（图 6B）。它们的配对增幅分别为$+0.184$ [0.150、0.219] 和 $+0.121$ [0.098、0.142], 。这两个分支还形成了更多共享句子结构： H_1–H_2 表征相似性分析在 T12/T15 上从 0.047/0.061 升至 0.374/0.448（图 6C）。因此，训练在调整预训练原生通路的同时，也调整了跨参与者通路；非线性融合则通过 Z_N呈现两者的联合证据。

表18：编码器分支的适应及其共享句子几何结构。原生/跨参与者表示相对于输入参与者的编码器来源。每位参与者对应的最后一行衡量两个分支之间的相似度。括号中给出逐点95%配对自助法区间。

参与者	比较	初始化	训练后	变化 [95% 区间]
T12	原生分支–音频	0.277	0.287	+0.010 [+0.005, +0.014]
T12	跨参与者分支–音频	0.014	0.198	+0.184 [+0.150, +0.219]
T12	H_1–H_2	0.047	0.374	+0.326 [+0.299, +0.354]
T15	原生分支–音频	0.262	0.260	-0.002 [-0.006, +0.003]
T15	跨参与者分支–音频	0.036	0.156	+0.121 [+0.098, +0.142]
T15	H_1–H_2	0.061	0.448	+0.387 [+0.369, +0.405]

E.3 神经与语音视图共享结构且保持区别

双神经专家对照为每位专家配备独立的融合投影器和角色适配器，但两个输入都是对同一连续编码器状态的变换。两种输入视图的句子几何结构非常接近：T12 上的表征相似性分析（RSA）为 0.950，T15 上为 0.949。在 MAS-BIT 中， Z_N–Z_P RSA 分别为 0.582 和 0.639（图6C）。因此，神经接口和语音接口在保留共享句子结构的同时，也向专家呈现出更具差异性的几何结构。结合第5.4节中语音专家留出比较的结果，这为神经与语音视图具有互补性提供了相互印证的证据。

通信接口也与学习得到的神经几何结构变化相关。MAS-BIT 的 Z_N–音频 RSA 为 0.343/0.239，而硬文本通信为 0.322/0.223。配对差值分别为 +0.021 [+0.011, +0.030]（T12）和 +0.016 [+0.010, +0.022] （T15），且在所有五种长度处理方式下，差值均为正（表 19）。这一比较表明，与硬文本对照相比，采用潜在消息的 MAS-BIT 与对齐度更高的句子级神经接口相关。

表 19：五种长度处理方式下神经接口对齐的变化。 条目表示汇总后的 Z_N–音频表征相似性分析差值。训练变化比较 MAS-BIT 与初始化；第二个对比比较 MAS-BIT 与硬文本通信。偏相关控制参考词数、有效神经位置数或两者的差异。

参与者	对比	分组	未分组	偏相关		
				词语	位置	Both
T12	训练变化	+0.152	+0.226	+0.196	+0.196	+0.189
T12	MAS-BIT − 困难文本	+0.021	+0.012	+0.014	+0.015	+0.015
T15	训练变化	+0.142	+0.132	+0.125	+0.138	+0.129
T15	MAS-BIT − 困难文本	+0.016	+0.008	+0.012	+0.010	+0.012

E.4 表征提取与统计分析

表征与参考文本。 我们使用冻结检查点的 FP32 推理路径，以批量大小为 1 提取所有提供的验证话语（880 条 T12 和 1,426 条 T15）。对于每个有效

时间位置，我们保留编码器状态 H_1，H_2、融合神经接口 Z_N，以及方程 1 中定义的语音接口 Z_P。包括非线性融合投影器在内的每项变换，均在时间平均之前应用。在双神经专家对照中，第二个输入视图是其独立神经投影器的输出。音频参照是对转录文本进行固定音色文本转语音后，由 Qwen3-ASR-0.6B 音频塔输出经均值池化得到的结果；文本参考是对转录文本进行处理后，由 Qwen3-ASR-1.7B 文本解码器最后一层状态经均值池化得到的结果。比较不同训练状态时，使用相同的参考特征和话语标识。

按词数分层的表征相似性分析。 我们对每位参与者的完整句子集合中的各项特征进行 z 分数标准化，并使用句子向量之间的 1 减皮尔逊相关来构建表征差异矩阵。我们保留参考词数相同且至少包含 20 个话语的分组，然后计算每个分组中上三角部分不同差异值之间的斯皮尔曼相关 ρ_g。总体得分为

$$\rho_{\mathrm{RSA}} = \frac{\sum_g \binom{n_g}{2} \rho_g}{\sum_g \binom{n_g}{2}}, \tag{8}$$

其中 n_g 是原始分组大小。T12 包含 848 个句子和 48,531 个组内配对（2–10 个词）；T15 包含 1,386 个句子和 134,018 个配对（3–11 个词）。未分组和偏相关估计量使用所有提供的验证集句子。对于偏相关表征相似性分析，我们对每个排序后的差异向量，以排序后的绝对长度差为自变量进行带截距回归，然后计算残差之间的相关性。

配对不确定性估计。 我们在每个词数分组内进行 2,000 次有放回的自助抽样；对于同一参与者，各模型、表征和参考文本共享相同的抽样结果（随机种子 0）。对于不同的原始句子 i，j，配对权重为 $c_i c_j$，其中 c_i 是该句子在抽样中的重复次数；排除同一原始句子的副本之间的配对。加权中秩保留这种重复配对分布。完整数据的标准化结果和原始分组权重保持不变。我们报告逐点 95% 句子自助法区间，取估计值或配对差异的第 2.5 和第 97.5 百分位数计算，并固定检查点权重。