EEG-Jev：脑电图的统一单次决策模型

Zhifei Zhu[1,2,3,6,#]；Hongtao Zhang[1,2,3,6,#]；Yu Ren[1,2,4,6]；Wei Zhou[1,2]；Zeping Xie[1,2,5,6]；Yi He[1,2,5,6,*]

[1]生物医学成像科学与系统国家重点实验室，中国科学院深圳先进技术研究院，中国深圳518055[2]Paul C. Lauterbur生物医学成像研究中心，中国科学院深圳先进技术研究院，中国深圳518055[3]东南大学，中国江苏南京211189[4]哈尔滨工业大学，中国黑龙江哈尔滨150001[5]先进医疗器械研究所，深圳理工大学，中国深圳518107[6]国家先进医疗器械创新中心，中国深圳518100**#这些作者贡献相同：Zhifei Zhu和Hongtao Zhang。*通讯作者：y.he3@siat.ac.cn（Y.H.）。**

地址：生物医学成像科学与系统国家重点实验室，中国科学院深圳先进技术研究院，中国科学院，中国广东省深圳市学苑大道1068号，邮编518055

摘要。脑电图分析涵盖选择候选项、判断命题以及评定有序属性，因此需要具有明确含义且可校准的概率。我们提出 EEG-Jev，一种用于脑电图的统一单次决策模型。给定一个脑电图窗口、一个问题和答案规范，一次前向传播即可返回类型化读出：CHOICE提供候选项分布，NOUL提供命题概率，SCORE提供有序等级分布及其期望。在35个公开数据源的留出划分上联合训练一个检查点并进行一次评估。核心原始输出在 Stieger2021 运动想象任务上达到54.08%的平衡准确率，在 Sleep-EDF-SC 睡眠分期任务上达到70.45%，在阈值为0.5时于 CHB-MIT 癫痫发作检测任务中达到每窗口72.11%的准确率和0.825的曲线下面积，并在0–10分的 ds005292 疼痛评分任务上取得1.700的平均绝对误差。先验校正和温度缩放将 Sleep-EDF-SC 上的负对数似然从0.766降至0.606，并将 CHB-MIT 上的负对数似然从0.166降至0.019；与此同时，直接读出将批次大小为1时的延迟从122.43毫秒降至63.11毫秒（1。94×），相较于匹配骨干生成式比较器。交互式演示和汇总结果：https://neuroimglab.github.io/EEG-Jev/。

1 引言

脑电图分析服务于多种决策。有些决策需要在多个选项中进行选择，例如某个时段的睡眠阶段，或参与者想象的动作。有些决策需要判断某个陈述是否成立，例如某个窗口是否包含癫痫发作。还有一些决策需要评估有序属性，例如参与者报告的疼痛程度。这三种形式广泛存在于临床监测、脑机接口和认知研究中。在每种情况下，用户需要的不仅是一个标签，还需要一个含义明确、可直接使用且经过校准的概率。

如今，这些决策大多仍是一次服务一个任务。LaBraM、CBraMod 和 REVE 等预训练脑电图编码器提供了可跨数据集迁移的表示[1–3]，但每个下游任务通常仍需配备独立的任务头并进行单独的训练运行，而且输出的含义由数据集决定，而不是由模型明确说明。NeuroLM 和 UniMind 等脑电图语言模型将 EEG 编码器与语言模型连接起来，使一个模型能够覆盖多种任务，并通过生成文本作答[4,5]。生成方式具有灵活性，但必须从文本中解析答案，而且每个答案词元都需要额外的一步解码。通过对每个候选项的词元进行评分，可以恢复一个概率，但此类分数取决于表面形式和答案先验[6,7]，并且属于词元序列，而不是属于某个明确说明的事件。

采用统一形式进行决策，可以应对这些限制。如果将每项任务都表述为自然语言问题，并明确规定允许的答案，共享模型便能处理形式各异的任务。答案规范还规定了每个输出概率的含义，因此所有任务的输出都可以用相同的方式评分和校准。Jev 最初作为 System One 模型推出，定义了这种形式：它针对输入回答类型化问题，并返回带有概率的结构化答案，而不是生成文本[8,9]。

EEG-Jev 基于这种形式构建了脑电图统一决策模型。任务由一个脑电图窗口、一个问题和一份答案规范组成，类型化读出可直接返回结果：候选项上的分布（CHOICE）、某个命题成立的概率（NOUL），或有序等级上的分布及其期望（SCORE）。冻结的 REVE-L 编码器和可训练连接器将脑电图输入经 LoRA 适配的通义千问 Qwen3.5-2B[1, 10–12]。一组权重在全部三种类型上进行训练，一次前向传播即可回答问题，无需生成文本。

任务定义具有普适性，而我们的验证涵盖其中一部分：运动想象（MI）、睡眠分期、癫痫发作检测和疼痛评分，数据来自 35 个公开数据源。我们仅使用一个检查点，在留出测试数据上进行一次评估，考察任务性能以及其概率在简单后处理前后的质量；此外，还在开发窗口上测量其推理延迟，并与一个基于相同骨干网络构建的自回归对照模型进行比较。我们的贡献如下：

1. **单一模型中的脑电图类型化决策形式**。 遵循公开的 Jev 接口，EEG-Jev 将选择、判断和评分表述为一种统一的输入形式（脑电图窗口、问题、答案规范），并提供三个类型化读出。所有三种类型均使用同一组权重进行训练，并在一次前向传播中回答问题，无需生成或解析文本。在本研究中，每种应用使用一种读出类型。

2. **可评分且可校正的概率**。 不同于生成文本的脑电图语言模型，每个输出都是指定事件的概率，因此可以直接应用适当评分规则。针对类别均衡训练抽样的先验校正，以及每种应用各自的一个温度参数，均在测试评估前固定；它们在不改变模型的情况下，降低了睡眠和癫痫发作核心数据源上的负对数似然；随后必须根据预期用途设定决策阈值。

3. **在 35 个数据源上进行一次测试评估，并考察延迟**。一个检查点未经过检查点选择，直接在来自 35 个公开

个数据源的留出数据上评估，涵盖四种应用，并与使用相同编码器的逐来源基线进行比较。在这 35 个数据源中，点估计在 26 个数据源上高于冻结特征基线，在 19 个数据源上高于微调基线；在四个核心数据源上，共享模型在睡眠分期方面与微调基线表现相当，在疼痛评分方面与两种基线表现相当，但在四分类运动想象和癫痫发作 AUROC 方面表现较差。在一个包含四个数据源的开发阶段计时面板中，对于同一个请求，直接读出的速度约为同一骨干网络上的自回归对照模型的 1.94 倍（批次大小为 32 时快 4–6%）。

2 相关工作

脑电图编码器和脑电图语言模型。预训练脑电图编码器包括 BIOT、LaBraM、EEGPT、CBraMod 和 REVE[1–3,13,14]。REVE 将每个信号片段与其电极位置和时间一同编码，因此同一个编码器可接受不同的电极布局和窗口长度。鉴于这一特性以及其公开权重，我们使用其公开发布的大型检查点 REVE-L[10]，冻结版；但这并不意味着它是每项任务中最强的编码器：在 NeuralBench-EEG 中，REVE 条目（端到端微调的基础模型）在核心运动想象任务中排名第一，在睡眠分期中排名第四，在癫痫发作检测中排名第十二[15,16]。在脑电图语言模型中，NeuroLM 和 UniMind 在多个数据集上经过指令微调，并以生成文本的形式输出答案[4,5]；UniMind 的作者指出，其输出不包含概率分数。其他系统则会生成解释或临床笔记[17,18]，或将脑电图表征与标签文本嵌入或原型对齐[19–21]。EEG-Jev 的 Choice 读出在形式上与这些对齐方法相近，因为它将投影状态与候选项的文本嵌入键进行匹配。EEG-Jev 在相同任务定义下增添了命题概率和评分分布，为这三种类型训练一组权重，并将所有输出作为概率进行评估。

类型化决策与概率质量。 三种答案类型及其输出契约遵循 Jev 的公共接口[8,9]。Jev 的架构和训练过程尚未公开，因此本文所述模型、读出层和训练均为我们自行实现；EEG-Jev 是一项独立的研究实现，与 TypeSafe 无关联。TypeSafe 表示，Jev 使用其称为校准决策强化学习（RLCD）的方法进行训练[8,9]，，并且已有该方法的独立开放实现方案[22]；EEG-Jev 仅使用监督目标。由于语言模型的候选项得分取决于表面形式和答案先验[6,7]，，我们使用适当评分规则进行评估[23–25]，并在模型外应用两种标准校正：先验偏移校正[26]（下文简称先验校正）和温度缩放[27]。

3 方法

3.1 统一任务定义

我们定义任务时不依赖任何应用。一个任务实例是一个三元组(x, q, a)：包含电极位置的脑电图窗口x、以自然语言表述的问题q，以及答案规范a。该规范说明答案类型$\tau(a) \in \{\text{Choice}, \text{Noul}, \text{Score}\}$，并列出允许的答案。答案为候选项集合、命题或有序量表的任何任务都符合这种形式，且每个问题都会附上候选项、命题和等级文本。每个问题恰好选择一种类型，模型则返回其允许答案上的分布，

$$p_\theta(y \mid x, q, a),\ y \in \begin{cases} \{1, \ldots, K\}, & \tau = \text{Choice}, a = (a_1, \ldots, a_K) \text{ candidate texts}, \\ \{\text{no}, \text{yes}\}, & \tau = \text{Noul}, a = \phi \text{ a proposition}, \\ \{0, \ldots, J\}, & \tau = \text{Score}, a = (t_0, \ldots, t_J) \text{ levels, values } v_0 < \cdots < v_J. \end{cases} \tag{1}$$

该类型确定每个概率所指向的事件。对于 CHOICE，p^k 表示参考答案为候选项 a_k 的概率。对于 NOUL，$p(\text{yes})$表示命题 ϕ 在该窗口内成立的概率。对于 SCORE，p_j表示该

属性处于级别 j 的概率，而点答案是期望值 $\hat{y} = \sum_j v_j p_j$。该类型还决定训练损失和概率质量的度量，而任务指标则依据类型和应用协议确定，而不是针对每个数据集临时选择。

3.2 架构与类型化读出

图1展示了从脑电图窗口和问题到类型化输出的处理路径。

脑电图编码与连接器。 冻结的 REVE-L 编码器接收带有电极位置的预处理窗口，并为每个通道和时间片段返回一个特征向量，$f \in R^{C \times}{}_{P \times 1216}$。这些特征在不进行池化的情况下展平为 $M = CP$ 个词元。可训练的多层感知机连接器将每个词元映射到语言模型的宽度 $d = 2048$ ，

$$e_j = \gamma \ \mathrm{RMSNorm}\left(W_2 \mathrm{GELU}\left(W_1 f_j + b_1\right) + b_2\right), \tag{2}$$

其中，学习得到的标量 γ 初始化为文本嵌入的均方根幅值。

输入行与骨干网络。 每个问题在语言模型的聊天格式中构成一个输入行：

[系统提示词] $Ee_1 \cdots e_M$/E[问题 q][答案规范 a][助手前缀]R .

文本通过冻结的嵌入表进行嵌入。标记器〈 E 〉 和〈 /E〉， 以及读出标记〈 R 〉， 都是经过学习的向量。答案规范以文本形式写出：每个候选项（“Candidate i: a_i.”）、命题或量表的每个等级。该行通过 Qwen3.5-2B[12,28], 因果语言模型的文本模型进行一次处理；该模型有 24 层、宽度为 2048，并结合了 Gated DeltaNet 和注意力层；其视觉编码器未被使用。其基础权重保持冻结状态，并对所有层的线性映射训练秩为 16 的 LoRA 适配器[11]。读出标记是该行的最后一个位置，因此其最终层状态为

row，$u \in R^d$ 已经看到了脑电图、问题和答案规范。

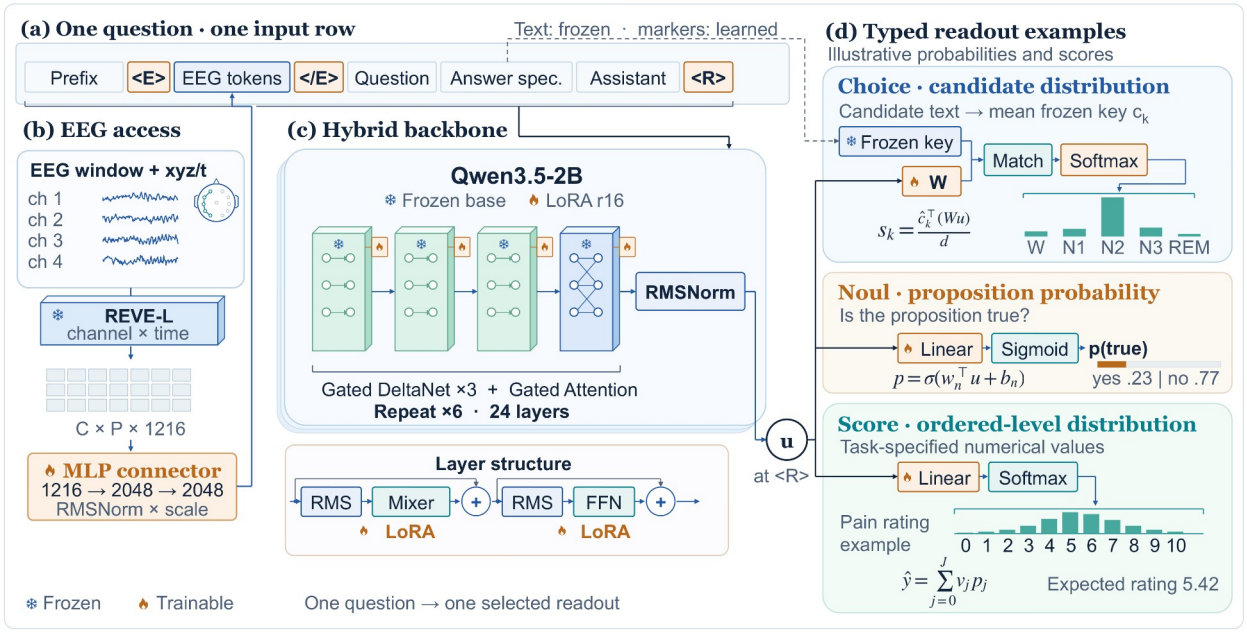

图1. EEG-Jev 架构与类型化读出。 **(a)** 系统提示词、脑电嵌入、自然语言问题、答案规范和读出标记〈 R 〉构成一个输入行。**(b)** 冻结的 REVE-L 编码器结合电极位置和时间信息处理多通道脑电图；可训练连接器将其特征映射到语言模型维度。**(c)** 通义千问3.5-2B 混合骨干网络由冻结的基础模型和可训练的 LoRA 适配器组成；u是〈 R 〉处的最终归一化状态。**(d)** 每个问题都会选择一个读出层。CHOICE 将u的可训练投影与候选项键进行匹配，$\hat{c}_k = c_k / \mathrm{rms}(c_k)$，其中$c_k$ 是其冻结文本嵌入的均值，d是状态维度；NOUL 给出命题成立的概率$p(\text{yes})$，并显示为 p(true)；SCORE 则给出任务指定数值v_j 的分布及其概率加权期望。0–10 刻度条和 5.42 的期望评分用于说明疼痛评分示例。概率、分数、脑电图轨迹和头皮示意图均为示意内容。

类型化读出。问题的答案类型决定选用三个读出层中的哪一个；其余两个不参与评估。加上连接器、标记器和 LoRA 适配器，读出层共有 2770 万个可训练参数；REVE-L 和语言模型的 18.8 亿基础权重均被冻结（附录B）。

选择。 候选项 k 的键为 $c_k \in R^d$，即其文本 a_k 中各词元的冻结输入嵌入的均值。该键不会经过骨干网络，也不会接收

梯度。使用归一化键 $\hat{c}_k = c_k / \mathrm{rms}(c^k)$ 和可训练投影 $W \in R^{d \times d}$，读出层会将投影状态与每个键进行匹配：

$$s_k = \frac{\hat{c}_k^{\top}(Wu)}{d}, \quad p_k = \frac{\exp(s_k)}{\sum_{j=1}^{K} \exp(s_j)}. \tag{3}$$

W初始化为零，因此训练从均匀分布开始。该读出没有类别专用输出单元，因此其形式不取决于候选项的数量或措辞；我们使用两个到五个候选项对其进行了训练和评估。

NOUL. 命题成立的概率为 $p(\text{是}) = \sigma(w_n^{\top} u + b_n)$。

SCORE. 任务指定一组有限的有序数值支持值 $v_0 < \cdots < v_J$。J读出会计算这些值对应的概率 p_j，并返回 $\hat{y} = \sum_{j=0} v_j p_j$。我们的支持值工具接受 2–11 个严格递增的实数值，在 Softmax 之前屏蔽未使用的输出槽位，并以任务所用单位计算期望。在当前实验中，所有疼痛数据源都使用同一个 11 槽位读出，$p = \mathrm{Softmax}(W_s u + b_s)$, 其中$v_j = j$ 对应 $j = 0, \ldots, 10$。因此，0–10 指定了这些实验中使用的疼痛评分量表。三种答案类型遵循公开的 Jev 接口[9]。

3.3 共享训练和类型化监督

问题采样与候选项重排序。每次更新使用 32 个问题，每个问题将一个训练窗口与一种措辞及答案规范配对（图 2）。每个数据源都有一个问题族，其中包含五种训练措辞；每个问题都会以相同概率抽取其中一种。对于 CHOICE，每个问题的候选项都会重新排列，并将目标索引映射到标注候选项的新位置。每个数据源、受试者和类别内的目标位置均保持平衡，因此位置不携带任何关于答案的信息，并且每个键都会随其候选项文本一起移动。NOUL 和 SCORE 没有需要重排序的内容。

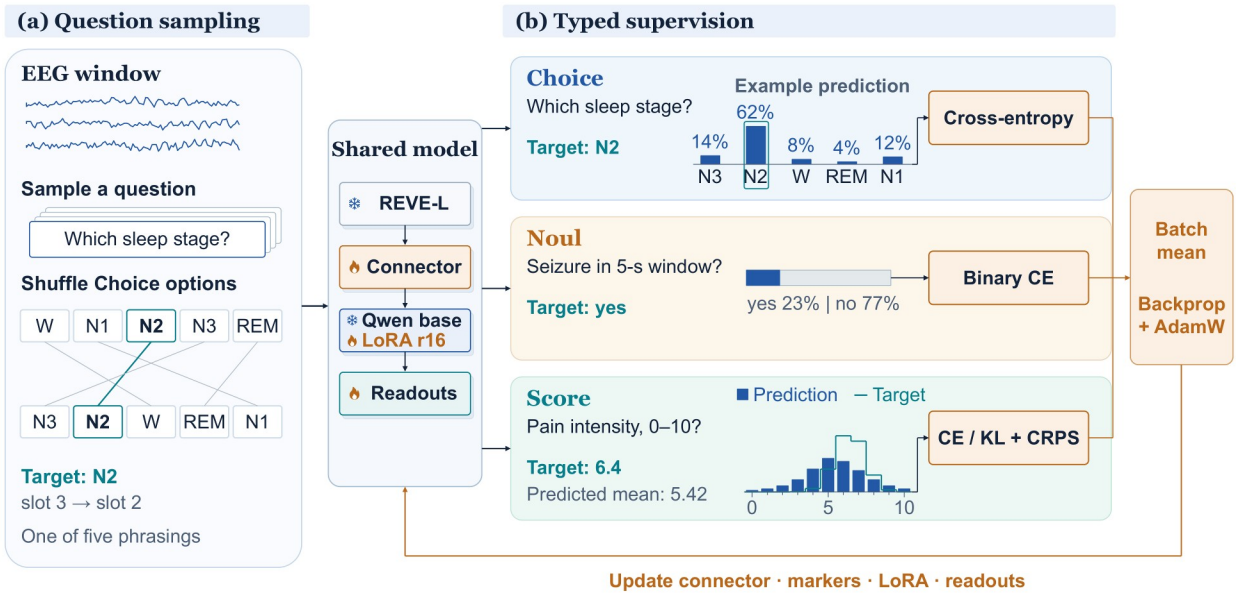

图 2. 问题采样与类型化监督。 **（a）** 每个训练样本都将一个脑电图窗口与五种问题措辞中的一种配对。对于选择题，选项会被打乱，目标索引也会随标注选项移动；此处即从第 3 个槽位移至第 2 个槽位。 **（b）** 每个训练问题仅通过一种类型进行监督：选择题使用交叉熵，NOUL 使用二元交叉熵，而 SCORE 则根据训练阶段使用软目标交叉熵或 KL散度，此外还加上 CRPS 乘以 0.2 再除以 10 分制范围所得的项。SCORE 目标轮廓是以 6.4 为中心的离散高斯分布；预测均值为 5.42。批次均值损失会更新连接器、标记器、LoRA 和读出层；REVE-L 与骨干网络基础部分保持冻结。概率和脑电图轨迹仅用于示意，并非模型输出。

类型化监督。一个问题仅由其自身类型的目标函数进行监督。CHOICE 使用标注候选项所在位置的交叉熵，NOUL 使用p(是) 的二元交叉熵。SCORE 将等级分布与以报告评分为中心的软目标r进行比较，并加上 ^{0}.2 CRPS / $(v_{J} - v_{0})$，其中 CRPS 是连续排序概率评分。在更新步 3,600 之前，目标r会将其质量分配到包含该评分的两个等级之间（交叉熵）；之后，它会变为标准差为一个等级的离散化高斯分布（ KL ($r \parallel p$) ）。32 个问题的损失会先求和再除以 32，因此无论问题类型和候选项数量如何，每个问题的权重都相同，并且通过对该均值执行一步 AdamW 更新来更新所有可训练参数。

四个阶段。一次包含 13,400 个更新步的运行（随机种子 8）分为四个阶段（附录B）。预热阶段（更新步 1 至 600）仅训练连接器、标记器和读出层，并冻结 LoRA，使脑电嵌入和读出层首先适配保持不变的语言模型。混合训练（601 至 3,600）加入 LoRA 适配器，并按固定比例抽取四种应用的数据。困难任务强化（3,601 至 11,000）采用一份抽样计划，该计划在开发面板显示运动想象仍处于随机水平后，于更新步 3,600 重新生成了一次。新计划根据各数据源训练数据的数量设定每个数据源的问题数，使运动想象在抽样中的占比从 0.30 提高到 0.46；对更新步 3,600 的检查点在固定开发子集上表现较弱的睡眠数据源和阶段增加抽样次数，并将部分癫痫发作抽样分配给该检查点评分较低的训练窗口。开发数据仅用于调整这些抽样统计；没有任何开发窗口参与梯度计算。最终阶段（11,001 至 13,400）将峰值学习率降低十倍，并将各项占比恢复到接近前两个阶段的水平。该次运行原计划进行 8,000 个更新步，并在抽样计划重新生成时延长至 13,400 个更新步。训练在这一固定预算下结束：未采用停止准则，最终阶段的训练损失仍在缓慢下降，并且直接使用最后一个检查点，不进行检查点选择。

3.4 概率后处理

类型化输出支持通过先验校正和温度缩放进行事后概率校准。训练抽样遵循抽样计划，其类别占比为 $\tilde{q}_k$，可能不同于数据源频率 π_k。先验校正用于调整这一由抽样引起的差异。在类别条件分布保持稳定的标准先验偏移模型下，校正形式如下[26]。

$$p'_k \propto p_k \frac{\pi_k}{\tilde{q}_k}. \tag{4}$$

$\tilde{q}_k$ 和 π_k均在每个数据源的训练集上统计，因此不涉及开发集或测试标签。SCORE没有类别先验，会跳过此步骤。这里先验偏移假设仅近似成立：不同训练阶段的抽样类别占比发生了变化，并且从第3,601次更新开始，部分癫痫发作抽样由模型自身的分数选出（附录B）。因此，我们将该校正视为一项固定规则，并通过测试得分进行评判。第二步是在先验校正后，针对每个应用在开发集划分上拟合温度 $p''_k \propto (p'_k)^{1/T}$。仅当在开发集受试者上进行交叉验证能够降低对数损失时，才采用该温度。

所得映射在测试评估之前被冻结：对于运动想象，仅进行先验校正（$T=1$）；对于睡眠分期，先进行先验校正，再应用温度（$T=1.078$）；对于癫痫发作检测，同样先进行先验校正，再应用温度（$T=1.252$）；对于疼痛评分，仅应用温度（$T=1.309$）。对于 NOUL，这两个步骤都是单调的，无法改变排序。对于 CHOICE，温度从不改变决策，但先验校正可能会改变决策。

4 实验设置

实验围绕三个问题展开：一个共享模型能否在不同应用中实现有用的任务性能，其概率是否具有所述含义并能通过简单的后处理得到改善，以及直接读出在推理时间上节省了什么。

4.1 数据与评估范围

我们在四种应用中使用 35 个公开数据源进行训练和评估（表1；各数据源见附录 A）。运动想象（17 个数据源）被设定为在两到四种提示条件中进行选择，部分数据源包含静息条件；睡眠分期（7 个数据源）则被设定为在一个 30 秒时段的五个阶段中进行选择。每个时段单独分期：模型不会查看相邻时段，而人工评分规则和基于序列的自动分期器会利用这些信息。癫痫发作检测（2 个数据源）被设定为 NOUL，命题为“此窗口内的某个时间点会发生一次标注的癫痫发作。”，窗口长度为 5 秒。疼痛评分（9 个数据源）被设定为在 0–10 量表的 11 个等级中评分，评估

激光刺激后的自我报告疼痛。任务是预测该数据源中未参与训练的受试者的单试次评分；这并非对临床疼痛的评估。

21 个数据源遵循锁定的 NeuralBench 协议，包括其窗口、标签和分区[15,29]：所有运动想象数据源、3 个睡眠数据源以及 CHB-MIT。其余 14 个数据源——4 个睡眠数据源、锡耶纳和 9 个疼痛数据源——采用各自的窗口、标签和受试者级划分。所有数据源都采用 NeuralBench 为 REVE 使用的信号处理方案（带通滤波器为 0.5–99.5 Hz，采样率为 200 Hz，按通道标准化，在 ± 15 处截断）。与 NeuralBench 流程一样，零相位带通滤波器和按通道标准化均基于整个记录计算，因此这些结果描述的是离线记录级预处理，而非仅能看到当前窗口的模型。每个数据源都设有训练集、开发集（验证集）和测试集。除 5 个运动想象数据源外，各分区在受试者层面互斥；这 5 个数据源的协议会将同一受试者的不同会话或试次划分到不同分区。

表1。数据与评估范围。 每个应用使用一种读出类型；应用的核心来源在测试评估前已固定，并承载主要结果。NB：保留 NeuralBench 代码 0.4.0版的窗口、标签和受试者划分的数据来源（核心任务或完整基准测试）。补充材料：采用相同信号处理方式以及我们自定义的窗口、标签和受试者划分的补充数据来源；九个疼痛数据来源均来自同一研究组的激光疼痛研究，不属于 NeuralBench。测试集：留出受试者（癫痫发作任务中为患者）和窗口，按数据来源汇总；五个运动想象数据来源在受试者内部进行划分，且同一应用的数据来源可能共享参与者。各数据来源的详细信息见：表A1。

应用	数据来源	读出	答案空间	核心来源	方案范围	测试受试者 / 窗口
运动想象	17	选择	*K*=2–4 个类别	Stieger2021	17 名新生儿（1 名核心，16 名完整）	128 / 44,200
睡眠分期	7	选择	*K*=5 阶段	Sleep-EDF-SC	3 名新生儿（1 个核心，2 个完整）；4 份补充材料	291 / 293,681
癫痫发作检测	2	Noul	是 / 否	CHB-MIT	1 NB（核心）；1 补充材料	9 / 155,753
疼痛评分	9	评分	0–10、11 个等级	ds005292	9 份补充材料	109 / 5,080
总计	35				21 新生儿；14 补充材料	537 / 498,714

每个应用的一个核心来源在读取任何测试结果之前就已固定：NeuralBench 核心任务中的 Stieger2021[30]，Sleep-EDF 的睡眠记录（Sleep-EDF-SC）[31] 和 CHB-MIT[32,33]，，以及疼痛任务中的 ds005292[34]，，后者依据元数据规则选定（附录 A）。我们评估了一个检查点（随机种子 8，更新步数 13,400）。该检查点在每个数据来源的锁定测试集划分上各评分一次，总计 498,714 个窗口，使用每个问题族的首次训练措辞和规范候选顺序。没有任何测试结果用于选择模型、阈值或问题；评估后唯一做出的选择——所展示的癫痫发作指标——见第 4.3 节。

4.2 比较基线

B1 和 *B2* 是针对每个数据源的基线，使用同一个 REVE-L 编码器，由我们训练，并在与 EEG-Jev 相同的测试窗口上评分。B1 是基于冻结特征的读出：对每个 REVE-L 词元使用一个小型多层感知机，进行均值池化，再接一个线性头（0.38 百万个可训练参数）。B2 微调 22 个编码器块中的最后四个，并使用相同的线性头（71.2 百万个可训练参数）。每个基线都针对每个数据源分别训练和选择，随机种子为 7 和 8；我们报告这两个随机种子的均值。对于三个 NeuralBench 核心任务中的每一个，*NeuralBench* 最佳结果指公开的 NeuralBench-EEG 结果[15,16]条目中均值最高的测试结果。这些是外部报告的结果，来自针对各任务进行端到端训练或微调的模型。表2列出了每个数值对应的模型，并以粗体标出本研究的三个模型中表现最佳者。

4.3 指标

正文针对每种应用展示一个指标；对于存在相应任务的应用，采用 NeuralBench 指标：运动想象、睡眠分期和癫痫发作检测采用平衡准确率（BA）。疼痛评分没有对应的 NeuralBench 任务，因此我们报告 $\hat{y}$ 的平均绝对误差（MAE），单位为评分单位。癫痫平衡准确率按窗口计算，依据自然患病率下原始输出的决策 $p(\text{yes}) \geq 0.5$ 得出。协议将 ROC 曲线下面积

（AUROC）指定为主要癫痫发作指标。评估结束后，在已知 AUROC 结果的情况下，我们将阈值为0.5时的平衡准确率作为展示指标，使三种 NeuralBench 应用均采用 NeuralBench 指标；同时报告 AUROC。所有其他计算所得指标均见已发布的结果文件（附录 E）。

任务结果使用原始输出。对于 CHOICE 和 NOUL，概率质量通过负对数似然（NLL）和期望校准误差（ECE，十个等质量分箱）衡量；对于 SCORE，则通过 CRPS 除以量表范围，以及中心 80% 区间的覆盖率衡量。区间是针对某一数据源的测试受试者（癫痫发作任务中为患者）进行 2,000 次自助法重采样后得到的配对差异 95%区间。这些区间仅描述测试群体抽样带来的变化；训练随机种子之间的变化则是另一回事，单次 EEG-Jev 运行和两个基线随机种子无法对其进行量化。区间包含零并不能证明模型之间存在差异；研究未设定等效性界值。

4.4 效率实验

为衡量直接读出所节省的开销（图3），我们将 EEG-Jev 与一个基于相同骨干网络的生成模型进行比较；该生成模型以短文本回答相同的问题。该比较模型从 EEG-Jev 检查点初始化，随后使用 LoRA 适配器针对 16 个问题训练 50 个更新步，以生成答案词元，后接序列结束标记；与此同时，REVE-L、连接器和 EEG 标记保持不变。其输入行包含相同的脑电嵌入、问题与答案规范；系统提示词被替换为“给出一个简短答案”的指令，在问题后附加一句说明答案格式的句子，并省略读出标记。因此，这是一个针对短答案进行适配的、骨干网络匹配的自回归（AR）比较模型。它不与 EEG-Jev 共享权重，其任务质量也有所不同。

两条路径的计时均从主机内存中的预处理后的脑电图窗口及其电极位置开始，经过传输至图形处理器、REVE-L 编码、连接器和语言模型。该预处理需要完整记录（第 4.1 节），不计入计时。EEG-Jev 路径在类型化概率和决策位于中央处理器上时结束。AR 路径利用缓存状态进行贪心解码，直到模型生成序列结束标记（最多 32 个词元）；当完整文本在中央处理器上被解析为有效答案时结束。测试面板包含来自四个数据源的各 32 个开发窗口，每个应用一个，并经过选择以平衡类别和受试者：Lee2019（运动想象；不是核心来源）、Sleep-EDF-SC、CHB-MIT 和 ds005292。两条路径在同一个图形处理器上使用 32 位算术，交替针对相同窗口进行计时；硬件、重复次数、同步方式以及有效答案的判定规则见附录 D。

请求延迟是批次为 1 时单个请求所用的时间；我们对所有请求取均值，并对四个数据来源进行等权重处理。批次为 4 时的每个窗口的摊销延迟，是一个批次的总耗时除以 4 后的结果，并采用相同方式取均值；它并不是单个请求的响应时间。

5 结果

5.1 跨应用的任务性能

表2展示了四项核心测试结果。在 Sleep-EDF-SC 上，EEG-Jev 的平衡准确率达到 70.45%，比 B1 高 2.29 个百分点（95% 配对区间为 0.46 至 3.94），且与 B2（70.48%）相近。在 ds005292 上，其 SCORE 读出值的平均绝对误差为 1.700，而 B1 为 1.701，B2 为 1.717；两个配对区间均包含零。预测训练集均值的结果为 1.771，因此这三个模型相较于这一常数基准，误差均降低了 0.05 至 0.07。由此可见，共享模型在有序评分任务上的表现与特定来源基线相当。疼痛任务结果表明

SCORE 读出值可以与分类和二元判断一起训练，但这并不意味着模型能够以具有临床实用价值的准确率预测新受试者的单次试验疼痛评分。

表2。核心源测试结果（原始模型输出，后处理：表3）。*NeuralBench* 最佳结果：外部报告的结果，即 NeuralBench结果文件中针对该任务所列模型的最高测试集平均平衡准确率（BA）（提交版本 7aaccf3，2026年10月10日，代码发布版本 0.4.0；均值 ± s.d.，三个随机种子）[15,16]。B1 是基于冻结的 REVE-L特征、按数据源进行的读出；B2 针对每个数据源微调最后四个 REVE-L模块。两者均在本研究中训练，并在与 EEG-Jev 相同的测试窗口上进行评分（随机种子 7 和 8 的均值）。BA 以百分比表示；MAE 采用 0–10 量表。癫痫平衡准确率阈值原始为 p（是）0.5，并在评估后成为所显示的指标；在协议的主要指标——曲线下面积（AUROC）上，EEG-Jev 低于两个基线（0.825，而两者分别为 0.891 和 0.929；两个配对区间均不包含零）。**粗体**：在所显示的精度下，本研究三个模型（EEG-Jev、B1、B2）中每行的最佳值；NeuralBench 最佳结果不参与此比较；并列值一并加粗；粗体不表示存在显著差异。†：EEG-Jev 减去该基线的配对 95% 区间不包含零。所有计算指标：results/metrics-long.csv。

				报告结果	本研究		
应用	核心来源	指标	受试者 / 窗口	NeuralBench 最佳结果	EEG-Jev	B1	B2
运动想象	Stieger2021	平衡准确率 (%) ↑	13 / 25,340	68.04±0.73 (REVE-基础版)	54.08	61.95†	**65.19†**
睡眠分期	Sleep-EDF-SC	平衡准确率（%）↑	16 / 41,656	72.93±0.53 (CBraMod)	70.45	68.17†	**70.48**
癫痫发作检测	CHB-MIT	阈值为0.5时的平衡准确率 (%) ↑	5 / 121,650	85.54±0.54 (BDTCN)	**72.11**	68.81	68.65
疼痛评分	ds005292	平均绝对误差 ↓	26 / 779	未涵盖	**1.700**	1.701	1.717

在 CHB-MIT 上，EEG-Jev 在概率阈值为 0.5 时实现了 72.11% 的平衡准确率，而 B1 和 B2 分别为 68.81% 和 68.65%，且两组配对区间均包含零；由于 EEG-Jev 使用 1:1 的癫痫发作抽样，而基线模型在接近自然患病率的数据分布上训练，因此在两个原始量表上，0.5 对应不同的工作点。在该阈值下，EEG-Jev 将 121,650 个测试窗口中的 6,580 个标记为阳性，敏感度为 49.5%，精度为 3.5%，这是因为癫痫发作窗口仅占测试窗口的 0.38%。这些是来自五名患者的汇总窗口级结果，并非事件级检测结果。无阈值依赖的曲线下面积为 0.825，而 B1 和 B2 分别为 0.891 和 0.929，表明癫痫发作排序仍有提升空间。在四分类 Stieger2021 任务中，EEG-

Jev 达到 54.08% 的平衡准确率，低于 B1（61.95%）和 B2（65.19%）。B1 与 EEG-Jev 使用相同的冻结的REVE-L特征，因此与 B1 相差 7.87 个百分点不能归因于冻结编码器；差距产生于连接器、语言模型或联合训练环节。在 12 个跨受试者运动想象数据源中，有 10 个数据源上的 EEG-Jev 低于 B2；在其中 7 个具有配对区间的数据源中，有 5 个的区间不包含零。这些差异促使我们进一步研究源数据接触、任务平衡和跨受试者泛化。

在全部35个数据源中，EEG-Jev在26个数据源上优于B1，在19个数据源上优于B2，其中有16个数据源上同时优于两者。若以协议中的主要癫痫发作指标曲线下面积（AUROC）取代阈值为0.5时的平衡准确率，这些数量分别为24、17和14。上述统计采用展示的各应用指标（分类和癫痫发作检测采用平衡准确率，疼痛采用平均绝对误差），并将点估计与各基线的双随机种子均值进行比较；在有配对95%区间的26个数据源中，区间在有利于EEG-Jev的方向上不包含零的情况，B1为6个，对比方向为2个；B2为2个，对比方向为6个，其余18个区间均包含零。配对区间及所有数据源层面的结果见附录 E。图4a–c展示了各数据源的差异及其区间。综合来看，这些结果支持在异构脑电图应用中采用共享决策模型的可行性。该模型在大多数数据源上优于冻结特征读出B1；而在区间能够区分两者的数据源中，该模型落后于部分微调的B2的情况多于领先的情况（落后于6个数据源，领先于2个），其中最大的差距出现在跨受试者运动想象任务中。

5.2 概率质量与后处理

表3比较了 EEG-Jev 的原始输出与第3.4节中的冻结后处理结果。图4d展示了各阶段的 NLL。后处理降低了睡眠分期和癫痫发作检测中的概率预测误差与校准误差。在 Sleep-EDF-SC 上，NLL 从 0.766 降至 0.606，ECE 从 0.088 降至 0.044。在 CHB-MIT 上，NLL 从 0.166 降至 0.019，ECE 从 0.118 降至 0.001。这些结果表明，类型化输出可以通过标准

工具重新进行校准。大部分提升来自撤销类别均衡训练抽样的影响；对于任何使用重采样类别进行训练的分类器，这种校正都适用。作为参考，B1 和 B2 的原始输出（两个随机种子的均值）在 Sleep-EDF-SC 上的 NLL 分别为 0.778 和 0.794，在 CHB-MIT 上分别为 0.015 和 0.014；而以测试集患病率为常数进行预测时，NLL 为 0.025。因此，在 CHB-MIT 上，经过后处理的 EEG-Jev 输出优于常数基准，但仍高于两个基线，这与其较低的曲线下面积一致。在 Stieger2021 上，EEG-Jev 的原始 ECE（0.028）低于 B1（0.133）和 B2（0.183），而原始 NLL 为 1.093，相比之下，B1 和 B2 分别为 1.078 和 1.130。

逐步分析结果表明，NLL 的改善主要归功于先验校正。癫痫发作窗口占 CHB-MIT 训练抽样的一半，占其测试窗口的 0.38%。利用训练统计量校正采样先验后，Sleep-EDF-SC 上的 NLL 降至 0.614，CHB-MIT 上则降至 0.021。温度缩放进一步提升了概率质量；在 Sleep-EDF-SC 上，ECE 从先验校正后的 0.057 降至 0.044。在 Stieger2021 上，NLL 从 1.093 降至 1.082。在 ds005292 上，归一化范围后的 CRPS 从 0.121 降至 0.120，MAE 从 1.700 降至 1.687。

决策规则应考虑校准概率尺度和评估目标。在 Sleep-EDF-SC 上，准确率从 69.42% 提升至 75.57%，而类别平衡准确率从 70.45% 降至 66.79%，反映出按频率加权评估与类别平衡评估之间的权衡。在 CHB-MIT 上，后处理保留了窗口排序，曲线下面积为 0.825。在 0.5 的阈值下，阳性预测从 121,650 个窗口中的 6,580 个减少至 68 个，敏感度从 49.5% 降至 13.8%，特异度达到 99.997%，平衡准确率从 72.11% 降至 56.88%。因此，基于校准概率的二元决策需要在开发数据上针对目标敏感度或假阳性代价选择阈值。在 ds005292 上，温度使中心 80% 区间的覆盖率进一步偏离其名义值，从 82.7% 变为 85.6%。概率分数、点预测和区间覆盖率共同刻画了后处理后的输出。

5.3 推理效率

图3展示了直接类型化读出如何在一次前向传播后产生决策，而生成路径则先执行预填充，随后进行增量解码。在采用相同的骨干架构和优化执行的情况下，EEG-Jev 在批量大小为 1 时的平均请求延迟为 63.11 毫秒，而改编的自回归对比方法为 122.43 毫秒，延迟比为 1.94。EEG-Jev 在全部四个计时源上均更快。在批次大小为 4 时，其摊销处理时间为每个窗口 28.68 毫秒，而自回归方法为 44.01 毫秒，延迟比为 1.53。图4e展示了所有批大小。从绝对值来看，在批量大小为 1 时，直接读出使每次请求节省 59.3 毫秒。这些测量结果表明，与缓存自回归解码相比，即使生成式答案平均仅包含 2.0–3.0 个词元，直接读出在单次请求和小批次场景下仍具有延迟优势。该比较不包括答案词元对数几率的单次读出，也不包括逐源头部 B1 和 B2；后者不使用语言模型。

表3. EEG-Jev在四个核心测试集上的概率质量，以及冻结后处理后的原始输出 → 。后处理方法在测试评估前即已固定：先验校正使用训练抽样和训练集的类别占比；温度T则针对每种应用，利用开发数据拟合（运动想象保持$T=1$；SCORE没有类别先验）。NLL：负对数似然；ECE：采用十个等质量分箱计算的期望校准误差；CRPS/范围：连续排序概率评分除以量表范围；覆盖率：落在中央80%预测区间内的评分百分比（名义覆盖率为80%）。BA以百分比表示。表2报告原始输出；本表展示了后处理带来的变化。两种映射对于癫痫发作概率都是单调的，因此曲线下面积保持不变。后处理将癫痫发作概率调整到自然患病率的量级，因此当阈值为0.5时，敏感度从49.5%降至13.8%，平衡准确率也随之下降；应根据预期用途设定阈值。

核心来源	后处理	适当评分	校准	任务指标
Stieger2021	先验校正	负对数似然 1.093 → 1.082	期望校准误差 0.028 → 0.028	平衡准确率 54.08 → 54.03
Sleep-EDF-SC	先验校正，$T=1.078$	负对数似然 0.766 → 0.606	期望校准误差 0.088 → 0.044	平衡准确率 70.45 → 66.79
CHB-MIT	先验校正，$T=1.252$	负对数似然 0.166 → 0.019	期望校准误差 0.118 → 0.001	曲线下面积 0.825；阈值为 0.5时的平衡准确率 72. 11 → 56.88
ds005292	温度（$T=1.309$）	CRPS/范围 0.121 → 0.120	80%覆盖率 82.7 → 85.6	平均绝对误差 1.700 → 1.687

图3. 自回归生成与直接类型化决策（仅展示机制；实测延迟见第5.3节）。 (a)脑电嵌入、问题与答案规范经过预填充后输入混合骨干网络，语言模型头生成第一个词元。其余$L-1$ 个词元通过缓存的增量解码生成，解码过程经过完整骨干网络，并复用注意力键和值以及DeltaNet循环状态；完成后的序列会被解析为答案。L为实际生成的词元数。**(b)**对于每个问题，EEG-Jev从读出标记u的最终归一化状态中提取内容，并在每次骨干网络前向传播中应用所选的CHOICE、NOUL或SCORE读出。概率条和脑电图轨迹仅作示意。

计时使用融合式门控DeltaNet内核。在128窗口一致性面板上，分类决策与参考执行结果一致；在批次大小为1时，概率的最大绝对变化为0.00527，SCORE期望值的最大绝对变化为0.00485。任务性能和校准均基于冻结的参考执行进行评估。在不使用融合内核的情况下，批次大小为1时，参考执行耗时124.02毫秒，而另一者为180.84毫秒（比值为1.46）；批次大小为4时，每个窗口耗时分别为34.55毫秒和50.02毫秒。随着批量大小增加，优势逐渐缩小；在批次大小为32时，无论采用哪种执行方式，优势均为4–6%。此外，从批次大小为8开始，EEG-Jev的参考执行在每个窗口上的速度快于优化后的执行。AR比较器还进行了额外的简答适配，并且在任务质量上有所不同，因此该延迟比较刻画的是在匹配的骨干架构下实现的两条决策路径。附录 D提供了计时协议、一致性检查以及完整的批次测量结果。

6 讨论

统一决策模型所提供的价值。 EEG-Jev 将选择、二元判断和有序评分整合到一个共享模型中。问题和答案规范用于明确所需的决策，类型化读出层则返回具有明确含义的概率。35个来源评估表明，这一通用接口可通过一组权重支持四种脑电应用。在点估计方面，该模型在 26 个数据来源上优于冻结特征基线，在 19 个数据来源上优于微调基线；而在

核心运动想象数据来源和癫痫发作 AUROC 上，该模型仍低于两种基线。校准和直接推理进一步拓展了其价值，使其不仅能提供点预测，还能用于基于概率的决策系统。

提升跨任务性能。 异质源规模、类别频率和受试者分布使训练预算分配成为一项重要的设计选择。在 35 个数据来源之间共享模型容量也可能导致学习进度不均衡和任务干扰。这些因素可能是造成所观察差异的原因，值得通过受控实验进行研究。我们计划研究源感知采样、基于训练集和验证集表现的自适应曝光，以及能够改善特征对齐和跨受试者泛化的架构变更。当前结果为这些方向提供了动机，而因果归因则需要匹配消融实验加以验证。

可校准的概率输出。 明确的概率语义使得能够校正训练采样先验，并一致地应用温度缩放。睡眠和癫痫发作 NLL 的降低体现了这一实际益处。应结合预期用途，同时评估校准和决策阈值，因为先验校正可以改善概率质量，同时改变固定阈值下的平衡准确率和敏感度。

效率与更长形式的生成。匹配骨干网络比较表明，在批量大小为 1 和 4 时，直接类型化读出具有实现层面的延迟优势。作为比较对象的模型已经采用平均长度为两个至三个词元的短答案。生成更长解释或显式推理链的生成式系统需要额外的序列解码步骤；当所请求的输出是类型化决策时，这也为实现更大幅度的节省创造了机会。更大的骨干网络可能会增加计算量。要量化这些影响，需要在匹配的硬件、输出要求和任务质量评估条件下，与经过任务训练的生成式脑电图模型及面向推理的系统进行比较。

训练的生成式脑电图模型和面向推理的系统进行比较。

证据范围。 结果来自一次使用单个随机种子的训练运行，区间仅描述受试者抽样的不确定性。没有通过消融实验分别考察语言模型、问题文本或联合训练的贡献。B1 使用相同的冻结特征，在 Stieger2021 上达到 61.95%，且可训练参数为 38 万，因此，当前结果并未表明语言模型能改善解码；它在此处提供的是一个支持三种答案类型的接口。每种应用仅使用一种答案类型，因此答案类型与应用相互混杂。测试问题的措辞和候选项顺序也曾出现在训练中，因此，模型对新措辞的鲁棒性以及向新问题的迁移尚未经过测试。所有窗口均使用整个记录的统计量进行预处理（第 4.1 节）；在在线使用场景中面对新受试者时，如果滤波和标准化仅使用当前窗口，性能是否仍能保持尚未经过测试。五个运动想象数据源按照协议在训练集和测试集中包含相同受试者；九个疼痛数据源均来自同一个研究组，且部分参与者有所重叠。癫痫发作结果基于五名和四名测试患者，并采用 5 秒窗口计算合并后的窗口级指标：事件级敏感度、每 24 小时误报数和检测延迟均未评估。仅有五名患者时，自助法区间较为粗略，应将其视为描述性结果。睡眠分期每次使用一个时段。开发数据从第 3,600 次更新起影响了抽样计划，而训练预算也在此时延长；展示的癫痫发作指标则是在评估后选定的（第 4.3 节）。我们的数据源中至少有七个（均非 NeuralBench 核心来源）被 NeuralBench 列为 REVE[29] 的预训练数据；尚未核实 REVE-L 是否见过这些数据，且任何收益都会同样惠及这三个模型。

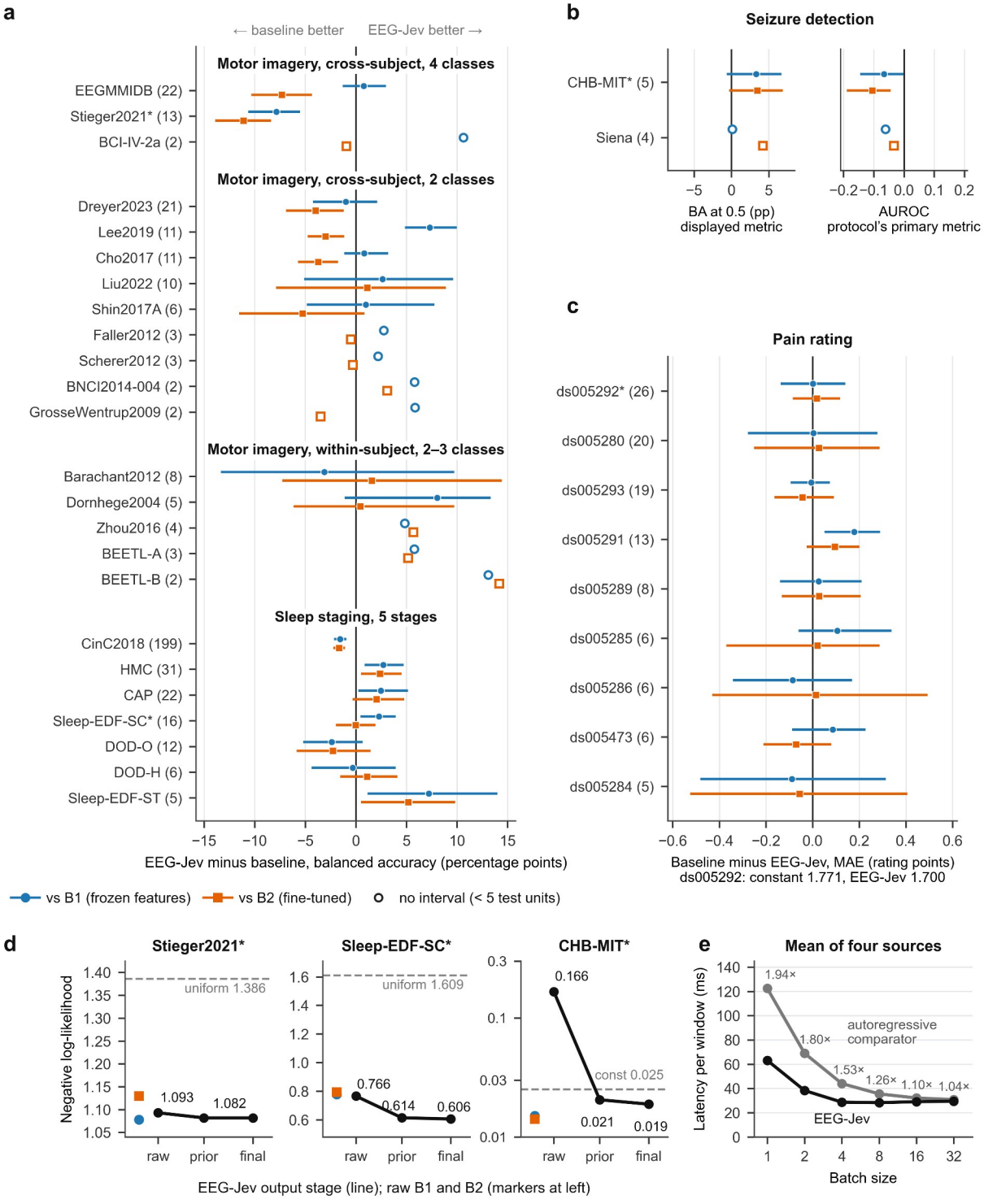

图 4. 各数据源的测试结果、概率分数和延迟。 **（a）** 对于 24 个运动想象和睡眠分期数据源，EEG-Jev 与各逐来源基线在平衡准确率上的差值（原始输出；B1，冻结特征读出；B2，微调编码器；基线值为随机种子 7 和 8 的均值）。柱状图表示基于测试受试者进行 2,000 次重采样得到的配对 95% 区间；空心

由于数据源的独立测试单位少于五个，因此标记器没有区间。括号内的数字表示测试单位数，星号表示核心数据源。**(b)** 在图示指标（阈值为0.5时的平衡准确率，BA）和协议的主要指标（曲线下面积，AUROC）下，癫痫发作检测的相同差异。**(c)** 九个疼痛数据源的平均绝对误差（MAE）中，基线减去 EEG-Jev；数值为正表示 EEG-Jev 表现更好。**(d)** EEG-Jev 在三个核心测试集上的负对数似然（NLL），分别对应原始、先验校正和最终输出阶段；Stieger2021 没有温度步骤，因此其最终阶段与先验校正阶段相同。左侧标记给出了 B1 和 B2 的原始 NLL。虚线表示类别上的均匀预测（Stieger2021、Sleep-EDF-SC），或测试集患病率下的恒定预测（CHB-MIT，对数坐标轴）。**(e)** 在采用优化执行的四数据源开发阶段计时面板上，展示每个窗口的摊销延迟随批量大小的变化；均值基于四个数据源，数字表示自回归对照模型与 EEG-Jev 的比值。数据来源：表 3 、 A3 和 A4，以及 results/metrics-long.csv。

未来方向。第一个方向是规模：扩展更多数据集和应用，同时改进架构以及跨数据来源的训练分配。第二个方向是更加灵活的语言条件决策，包括训练中未出现的问题、候选项集合和量表。第三个方向是训练时的概率质量，尤其是校准决策强化学习，即 TypeSafe 为 Jev[8,22] 命名的训练方法。第四个方向是部署：将类型化决策提炼到能够在设备上运行的小型模型中。

7 结论

EEG-Jev 将脑电图选择、二元判断和有序评分统一到一个共享决策模型中。其类型化读出层可在单次前向传播中生成含义明确的概率，并支持标准的事后校准。在35个数据来源中，根据所展示的应用指标，其点估计在26个来源上优于B1，在19个来源上优于B2；在具有配对区间的26个来源中，与B1相比，有6个来源的区间在有利于该模型的一侧排除零值，另有2个来源的区间在不利于该模型的一侧排除零值；与B2相比，相应数量分别为2个和6个。剩余差距最明显的领域是跨受试者运动想象和癫痫发作排序。先验校正和温度缩放有助于改善睡眠与癫痫发作概率。在包含4个数据源的计时测试中，直接读出的平均单请求延迟为63.11毫秒，批次大小为4时每个窗口的摊销处理时间为28.68毫秒，其延迟比为

相对于匹配骨干生成式比较器，延迟比分别为1.94和1.53。这些发现为实现高效、可校准的脑电图决策提供了切实依据，并推动进一步探索训练分配、泛化和模型架构。

声明

数据可用性。全部 35 个数据来源均为公共数据集（附录 A）。其中没有任何数据集要求提交应用申请或签署数据使用协议；其中两个数据集（Shin2017A、Dornhege2004）要求用户在下载页面接受相关条款。数据集访问信息和许可协议详情通过项目网站提供。托管于 PhysioNet 的数据来源均按照其页面要求进行引用[35]。项目网站不再分发任何原始脑电图记录、衍生特征、逐窗口测试预测或模型权重；汇总结果以机器可读文件形式发布。随附的演示仅对允许在注明出处的情况下再分发其脑电图数据许可所涵盖数据来源的开发集划分记录，回放其中的短时显示片段及这些片段对应的已保存模型输出。各个数据来源的许可歧义和第三方条款均记录在注册表中。

伦理与使用。本研究对公开的去标识化脑电图记录进行了二次分析，未开展新的受试者招募或数据收集。该模型及其结果仅供研究使用，不用于临床诊断、监测或报警。

致谢。 我们感谢为本研究所用脑电图数据集提供公开访问的研究人员和数据共享倡议，也感谢 EEG-Jev 所依托的开源软件和模型的开发者与维护者，包括 NeuralBench、REVE 和通义千问。

CRediT 作者 贡献 声明

Zhifei Zhu：构思、方法、软件、验证、正式分析、调查、可视化、撰写初稿；Hongtao Zhang：构思、方法、软件、验证、审阅与编辑；Yu Ren：数据整理、验证、审阅与编辑；Wei Zhou：资源、验证、审阅与编辑；Zeping Xie：资源、验证、审阅与编辑；Yi He：构思、监督、经费获取、资源、调查、审阅与编辑。

利益冲突。 作者声明不存在任何利益冲突。

基金资助。 本研究获得了中国科学院战略性先导科技专项（XDB0930000）、深圳市科技计划（JSGGKQTD20210831174329010）、深圳市杰出人才培养基金、国家重点研发计划（2023YFF0714200、2024YFC3406602）、中国科学院基础研究青年科学家项目（YSBR-114）、广东省磁共振与多模态成像重点实验室（2023B1212060052）以及国家自然科学基金委员会（编号：82505533）的资助。

参考文献

1.El Ouahidi, Y. 等 REVE：脑电图基础模型——通过对 25,000 名受试者进行大规模预训练，适配任意设置。载于 神经信息处理系统进展 **38**, 25583–25619 (2025)。2.Jiang, W.-B., Zhao, L.-M. & Lu, B.-L. 用于学习通用表征的大型脑模型：基于海量脑机接口脑电图数据。载于 国际学习表征会议论文集（*ICLR*） (2024)。3.Wang, J. 等 CBraMod：用于脑电解码的交叉脑基础模型。载于 国际学习表征会议论文集（*ICLR*） (2025)。4.Jiang, W.-B., Wang, Y., Lu, B.-L. & Li, D. NeuroLM：弥合语言与脑电信号之间差距的通用多任务基础模型。载于 国际学习表征会议论文集（*ICLR*）(2025)。5.Lu, W. 等 UniMind：释放大语言模型在统一多任务脑解码中的潜力。预印本见 https://arxiv.org/abs/2506.18962 (2025)。6.Holtzman, A., West, P., Shwartz, V., Choi, Y. & Zettlemoyer, L. 表面形式竞争：为什么概率最高的答案并不总是正确答案。载于 *2021* 年自然语言处理经验方法会议论文集（*EMNLP*）, 7038–7051 (2021)。7.Zhao, Z., Wallace, E., Feng, S., Klein, D. & Singh, S. 使用前先校准：提升语言模型的少样本性能。载于 第 *38* 届国际机器学习会议论文集（*ICML*）（机器学习研究论文集 **139**）, 12697–12706 (2021)。

8.Almeida, D. System One 模型与 Jev 介绍。TypeSafe AI 博客，https://typesafe.ai/blog/introducing-system-one-models-and-jev (2026)。9.TypeSafe AI。TypeSafe 文档：原语（选择、评分、Noul）、System One 模型及 System One API 参考。TypeSafe AI 文档，https://docs.typesafe.ai/introduction (2026)。10.brain-bzh。REVE-L：模型卡片与配置（brain-bzh/reve-large）。Hugging Face，https://huggingface.co/brain-bzh/reve-large (2025)。11.Hu, E. J. 等。LoRA：大语言模型的低秩适配。载于国际学习表征会议论文集（ICLR）（2022）。12.Qwen 团队。Qwen3.5：迈向原生多模态智能体。Qwen 博客，https://qwen.ai/blog? id=qwen3.5 (2026)。13.Wang, G. 等。EEGPT：用于脑电图通用且可靠表征的预训练 Transformer。载于神经信息处理系统进展 37, 39249–39280 (2024)。14.Yang, C., Westover, M. B. & Sun, J. BIOT：用于真实场景跨数据学习的生物信号 Transformer。载于神经信息处理系统进展 36, 78240–78260 (2023)。15.Banville, H. 等。NeuralBench：神经人工智能模型基准测试的统一框架。预印本见 https://arxiv.org/abs/2605.08495 (2026)。16.Meta Platforms, Inc. 及其关联公司。NeuralBench-EEG v1.0：核心任务结果表。官方项目页面，https://facebookresearch.github.io/neuroai/neuralbench/index.html (2026)。17.Ma, F. 等。E2-LLM：弥合神经信号与可解释情感分析之间的鸿沟。预印本见 https://arxiv.org/abs/2601.07877 (2026)。18.Pradeepkumar, J., Chen, Z. & Sun, J. 神经信号在真实场景中生成临床笔记。预印本见 https://arxiv.org/abs/2601.22197 (2026)。19.Cho, M.-J., Shin, H.-B., Lee, S.-H. & Lee, S.-W. 通过语义对齐进行连续潜变量预测建模，应用于脑电-语言基础模型。预印本见 https://arxiv.org/abs/2608.11656 (2026)。20.Jiang, M. 等。LEAF：面向脑机接口的语言-脑电对齐基础模型。预印本见 https://arxiv.org/abs/2509.24302 (2025)。21.Zhang, S. 等。EEG-PRIME：通过多级条件控制进行原型对齐表征学习，以实现脑电解码。预印本见 https://arxiv.org/abs/2608.13072 (2026)。22.Gao, Z. & Wang, P. OpenJev-RLCD：可运行的 RLCD 实现。预印本见 https://arxiv.org/abs/2609.38850 (2026)。23.Brier, G. W. 以概率形式表达的预测验证。月刊天气评论 78, 1–3 (1950)。24.Gneiting, T. & Raftery, A. E. 严格适当评分规则、预测与估计。美国统计学会杂志 102, 359–378 (2007)。25.Matheson, J. E. & Winkler, R. L. 连续概率分布的评分规则。管理科学 22, 1087–1096 (1976)。26.Saerens, M., Latinne, P. & Decaestecker, C. 将分类器输出调整为新的先验概率：一种简单方法。神经计算 14, 21–41 (2002)。27.Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. 现代神经网络的校准。载于第 34 届国际机器学习会议论文集（ICML）（机器学习研究论文集 70），1321–1330 (2017)。28.Qwen 团队与 Hugging Face Inc. 团队。Qwen3.5 实现（modeling_qwen3_5.py）。Hugging Face Transformers，https://github.com/huggingface/transformers/blob/v5.13.1/src/transformers/models/qwen3_5/modeling_qwen3_5.py (2026)。29.Meta Platforms, Inc. 及其关联公司。NeuralBench 源代码，版本 0.4.0（任务配置）。官方源代码（MIT 许可证），https://github.com/facebookresearch/neuroai/tree/7aaccf34b11540e00cca2425c6b9094cae396bf6/neuralbench-repo (2026)。

30.Stieger, J. R., Engel, S. A. & He, B. 基于连续感觉运动节律的大规模群体脑机接口学习。《Scientific Data》8, 98 (2021)。 31.Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A. C. & Oberyé, J. J. L. 睡眠依赖性神经反馈环路分析：脑电图的慢波微连续性。《IEEE生物医学工程汇刊》47, 1185–1194 (2000)。 32.Dan, J. & Shoeb, A. BIDS CHB-MIT头皮脑电数据库。Zenodo, https://zenodo.org/records/10259996 (2023)。 33.Shoeb, A. 机器学习在癫痫发作起始检测与治疗中的应用。博士论文，麻省理工学院 (2009)。 34.Zhao, X. 等. 用于疼痛研究的激光诱发电位综合脑电数据集。《Scientific Data》12, 1536 (2025)。 35.Pollard, T. 等. PhysioNet：生物医学研究的全球平台。《Nature Health》1, 792–795 (2026)。 36.Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N. & Wolpaw, J. R. BCI2000：通用脑机接口（BCI）系统。《IEEE生物医学工程汇刊》51, 1034–1043 (2004)。 37.Tangermann, M. 等. 第四届脑机接口竞赛综述。《神经科学前沿》6, 55 (2012)。 38.Dreyer, P., Roc, A., Pillette, L., Rimbert, S. & Lotte, F. 一个包含用户档案信息的大型脑电数据库，用于运动想象脑机接口研究。《Scientific Data》10, 580 (2023)。 39.Lee, M.-H. 等. 用于三种脑机接口范式的脑电数据集和OpenBMI工具箱：对脑机接口文盲现象的调查。《GigaScience》8, giz002 (2019)。 40.Cho, H., Ahn, M., Ahn, S., Kwon, M. & Jun, S. C. 运动想象脑机接口脑电数据集。《GigaScience》6, gix034 (2017)。 41.Liu, H. 等. 用于急性卒中患者脑机接口的脑电运动想象数据集。《Scientific Data》11, 131 (2024)。 42.Shin, J. 等. 用于脑电+近红外光谱单试次分类的开放获取数据集。《IEEE神经系统与康复工程汇刊》25, 1735–1745 (2017)。 43.Faller, J., Vidaurre, C., Solis-Escalante, T., Neuper, C. & Scherer, R. 自动校准与循环适应：迈向即插即用的在线事件相关去同步脑机接口。《IEEE神经系统与康复工程汇刊》20, 313 319 2012– ( )。 44.Scherer, R. 等. 脑机接口技术：整体不止是各部分之和。《软计算》17, 317–331 (2012)。 45.Steyrl, D., Scherer, R., Faller, J. & Müller-Putz, G. R. 随机森林在非侵入式感觉运动节律脑机接口中的应用：一种实用便捷的非线性分类器。《生物医学工程／生物医学技术》61, 77–86 (2016)。 46.Leeb, R. 等. 脑机通信：探索虚拟公寓的动机、目标与影响。《IEEE神经系统与康复工程汇刊》15, 473–482 (2007)。 47.Grosse-Wentrup, M., Liefhold, C., Gramann, K. & Buss, M. 非侵入式脑机接口中的波束形成。《IEEE生物医学工程汇刊》56, 1209–1219 (2009)。 48.Barachant, A. 博士论文《基于异步脑电脑机接口的执行器鲁棒控制》中的脑电运动想象数据集。Zenodo, https://doi.org/10.5281/zenodo.806023 (2012)。 49.Blankertz, B. 等. 第三届脑机接口竞赛：验证针对实际脑机接口问题的替代方法。《IEEE神经系统与康复工程汇刊》14, 153–159 (2006)。 50.Dornhege, G., Blankertz, B., Curio, G. & Müller, K.-R. 通过特征组合和多类别范式提高非侵入式脑电单试次分类的比特率。《IEEE生物医学工程汇刊》51, 993–1002 (2004)。 51.Zhou, B., Wu, X., Lv, Z., Zhang, L. & Guo, X. 用于优化基于运动想象的脑机接口的全自动试次选择方法。PLOS ONE 11, e0162657 (2016)。

52.Wei, X. 等 2021年BEETL竞赛：推动受试者独立性与异质性脑电数据集的迁移学习。载于*NeurIPS 2021*竞赛演示赛道论文集（机器学习研究论文集 **176**），205–219 (2022)。 53.Alvarez-Estevez, D. & Rijsman, R. M. 深度学习方法用于自动睡眠评分的数据库间验证。*PLOS ONE* **16**，e0256111 (2021)。 54.Ghassemi, M. M. 等 小睡即获胜：PhysioNet/Computing in Cardiology 2018挑战赛。载于*2018*年计算心脏病学会议（*CinC*） **45**，1–4 (2018)。 55.Terzano, M. G. 等 人类睡眠中周期性交替模式（CAP）评分的图谱、规则和记录技术。睡眠医学 **2**，537–553 (2001)。 56.Guillot, A., Sauvet, F., During, E. H. & Thorey, V. Dreem开放数据集：用于比较人工与自动睡眠分期的多评分睡眠数据集。*IEEE*神经系统与康复工程汇刊 **28**，1955–1965 (2020)。 57.Detti, P., Vatti, G. & Zabalo Manrique de Lara, G. 用于癫痫发作预测的脑电同步分析：基于非侵入性记录数据的研究。*Processes* **8**，846 (2020)。

A 数据来源与分区

表A1列出了每个数据源及其协议、读出、分区和测试范围。疼痛核心源 ds005292 在获得任何疼痛测试结果之前，依据元数据规则选定：在首轮更新时已纳入训练集、采用该系列通用实验设计且没有记录显示接触过其测试数据的疼痛数据源中，它拥有最多的训练与开发集受试者。九个疼痛数据源的逐试次评分不属于 OpenNeuro记录；这些评分取自数据集作者的补充存储库，而该存储库未附带许可声明（见注册表）。

表A1。35个数据来源。协议：NeuralBench 代码0.4.0版的核心任务或完整基准测试（NB 核心、NB 完整版），或补充来源（补充材料）。读出层及选项数量（选择）或等级（评分）。划分：受试者在各分区间互不重叠（交叉），或相互共享（内部）。受试者：癫痫发作数据中的患者；测试计数和测试窗口采用留出测试结果中的数据。*：核心来源。[a]所有比较模型均存在的输入通道缺陷；[b]ds005284 和 ds005292 共享参与者；[c]在构建分区前读取评分；[d]在第3,601次更新时加入训练。参考文献：主要数据集论文（全部九个激光疼痛数据集共享综合数据描述符[34]）；确切的存储库DOI、版本和访问入口均记录在注册表中。许可协议：数据许可协议的简称（[e] 进一步的声明或条件记录在注册表中；对于九个疼痛数据源，脑电图记录采用CC0许可，而每次试验的疼痛评分来自数据集作者的补充存储库，该存储库未附带许可声明）；项目网站上的数据注册表（results/data-source-registry.csv）列出了每个数据源的版本、访问入口、许可文本和条件。

数据源	参考文献	许可协议	协议读出		划分	受试者 训练集 / 开发集 / 测试集	Test 窗口
运动想象							
Stieger2021*	30	CC BY 4.0[e]	NB 核心	选择，4	交叉	36 / 13 / 13	25,340
EEGMMIDB	36	ODC-By 1.0	NB 完整版	选择，4	交叉	63 / 21 / 22	1,980
BCI-IV-2a	37	知识共享署名-禁止演绎 4.0	NB 完整版	选择，4	交叉	5 / 2 / 2	1,152
Dreyer2023	38	知识共享署名 4.0 国际许可协议	NB 完整版	选择，2	交叉	52 / 14 / 21	5,040
Lee2019	39	CC0 1.0	NB 完整版	选择，2	交叉	32 / 11 / 11	2,200
Cho2017	40	CC0 1.0	NB 完整版	选择，2	交叉	30 / 11 / 11	2,200
Liu2022	41	知识共享署名 4.0 国际许可协议	NB 完整版	选择，2	交叉	30 / 10 / 10	400
Shin2017A	42	GPL v3[e]	NB 完整版	选择，2	交叉	17 / 6 / 6	360
Faller2012	43	CC BY-NC-ND 4.0	NB 完整版	选择，2	交叉	6 / 3 / 3	1,200
Scherer2012	44, 45	知识共享署名-禁止演绎 4.0	NB 完整版	选择，2	交叉	8 / 3 / 3	480
BNCI2014-004	37, 46	知识共享署名-禁止演绎 4.0	NB 完整版	选择，2	交叉	5 / 2 / 2	1,440

数据源	参考文献	许可协议	协议读出		划分	受试者训练集 / 开发集 / 测试集	Test 窗口
GrosseWentrup2009[a]	47	知识共享署名 4.0 国际许可协议	NB 完整版	选择，2	交叉	6 / 2 / 2	600
Barachant2012	48	知识共享署名-相同方式共享 4.0 国际许可协议	NB 完整版	选择 ，3	在.....之内	8 / 8 / 8	96
Dornhege2004[a]	49, 50	参见注册表	NB 完整版	选择，2	其中	5 / 5 / 5	112
Zhou2016	51	知识共享署名 4.0 国际许可协议	NB 完整版	选择 ，3	其中	4 / 4 / 4	600
BEETL-A	52	知识共享署名 4.0 国际许可协议	NB 完整版	选择 ，3	在.....之内	3 / 3 / 3	600
BEETL-B	52	知识共享署名 4.0 国际许可协议	NB 完整版	选择 ，3	在.....之内	2 / 2 / 2	400
睡眠分期							
Sleep-EDF-SC*	31	ODC-By 1.0	NB 核心	选择，5	交叉	46 / 16 / 16	41,656
HMC	53	知识共享署名 4.0 国际许可协议	NB 完整版	选择 ，5	交叉	90 / 30 / 31	28,766
CinC2018	54	ODC-By 1.0	NB 完整版	选择，5	交叉	596 / 199 / 199	174,033
CAP[a]	55	ODC-By 1.0	补充材料	选择，5	交叉	63 / 10 / 22	21,678
DOD-H	56	MIT	补充材料	选择，5	交叉	15 / 2 / 6	5,903
DOD-O	56	MIT	补充材料	选择，5	交叉	33 / 5 / 12	11,349
Sleep-EDF-ST	31	ODC-By 1.0	补充材料	选择，5	交叉	13 / 4 / 5	10,296
癫痫发作检测							
CHB-MIT，α*	32, 33	ODC-By 1.0[e]	NB 核心	Noul	交叉	13 / 5 / 5	121,650
锡耶纳[a]	57	知识共享署名 4.0 国际许可协议	补充材料	Noul	交叉	8 / 1 / 4	34,103
疼痛评分							
ds005292*,b	34	CC0 1.0[e]	补充材料	评分，11	交叉	83 / 25 / 26	779
ds005280[a]	34	CC0 1.0[e]	补充材料	评分，11	交叉	60 / 20 / 20	600
ds005284[b]	34	CC0 1.0[e]	补充材料	评分，11	交叉	16 / 5 / 5	80
ds005285	34	CC0 1.0[e]	补充材料	评分，11	交叉	17 / 6 / 6	960
ds005286	34	CC0 1.0[e]	补充材料	评分，11	交叉	18 / 6 / 6	180
ds005289[c]	34	CC0 1.0[e]	补充材料	评分，11	交叉	23 / 8 / 8	78
ds005291	34	CC0 1.0[e]	补充材料	评分，11	交叉	39 / 13 / 13	390
ds005293	34	CC0 1.0[e]	补充材料	评分，11	交叉	57 / 19 / 19	1,243
ds005473[d]	34	CC0 1.0[e]	补充材料	评分，11	交叉	17 / 6 / 6	770

B 训练与配置

模型。 发布的 REVE-L[10] 包含22个宽度为1216的模块，并将每个通道切分为大小为200个采样点、步幅为180的分块；对于3秒、4秒、5秒和30秒的窗口，分别得到3、4、5和33个分块。Qwen3.5-2B包含24层，分为六组，每组三个门控 DeltaNet 层和一个注意力层。LoRA（秩为16，$\alpha = 32$）应用于这些层的全部186个线性映射。这些可训练参数中，连接器占6,690,817个，三个标记器占6,144个，LoRA占16,819,200个，读出层占4,218,892个

（其中 4,194,304 个位于 CHOICE 投影 W中），总计 27,735,053 个；一个继承而来、从未调用的线性层含 22,539 个参数，不计入其中。

训练计划。表 A2 列出了四个阶段。训练按重新生成的计划所设定的固定预算结束，未设置停止准则；在最终阶段，训练损失仍在缓慢下降。优化器为 AdamW（$\beta=(0.9,0.999)$, $\epsilon=10^{-8}$，权重衰减为 0.01），并采用范数为 1.0 的梯度裁剪和 32 位算术。每次更新对 32 个问题的损失取平均值，训练问题总数为 428,800 个。在一台 NVIDIA RTX PRO 6000 Blackwell GPU 上运行耗时 10.4 小时。

表 A2。单次运行的训练阶段（随机种子 8）。连接器、标记器和读出层以及 LoRA 的峰值学习率。每个阶段中，学习率先在预热阶段（更新步数；最终阶段无预热）线性上升，随后按余弦衰减降至该阶段峰值的十分之一。在第 3,600 次更新时重新生成计划之前，最后两个阶段原计划分别进行 2,600 和 1,800 次更新，总计 8,000 次。各项占比表示从运动想象、睡眠分期、癫痫发作检测和疼痛评分中抽取的问题所占比例；第三阶段的占比由重新生成计划中各数据源的问题数量决定。

阶段	更新步	训练模块	峰值学习率	份额 （运动想象/睡眠/癫痫发作/疼痛）	Ramp
预热阶段	1–600	连接器、标记器、读出层	10^{-3} / LoRA 冻结	0.25 / 0.61 / 0.04 / 0.10	10
混合训练	601–3,600	预热阶段的，且 LoRA	3×10^{-4} / 2.5×10^{-5}	0.30 / 0.56 / 0.04 / 0.10	100
困难任务强化	3,601–11,000	与混合训练相同	3×10^{-4} / 2.5×10^{-5}	0.46 / 0.39 / 0.04 / 0.11	100
最终阶段	11,001–13,400	与混合训练时相同	3×10^{-5} / 2.5×10^{-6}	0.25 / 0.62 / 0.04 / 0.09	0

采样。在一个阶段内，确定性调度器会为低于其份额最多的数据源提供服务。在前两个阶段中，运动想象和睡眠数据源的权重按其训练受试者数量的平方根计算，各类别均等抽取，癫痫发作抽样在正负窗口之间保持平衡，疼痛抽样则轮流遍历所有已评分事件。在更新 3,600 时重新生成的训练计划，根据各数据源的训练数据量，为整个运行设置每个数据源的问题数量（适用于

Lee2019 和 Cho2017 根据它们在先前一次运行中的曝光量）并扣除已抽取的部分。对于睡眠，固定的开发子集包含 16,717 个窗口，使用第 3,600 次更新的检查点进行评分，根据各数据源的 BA 设定 0.81 至 1.19 的源乘数，并根据逐类 F1 设定 0.12 至 0.35 的类别占比，其中 N1 期最高。对于癫痫发作，30% 的阳性抽样来自该检查点下 p(是) 最低的五分之一阳性训练窗口；50% 的阴性抽样来自已评分的阴性训练窗口中数值最高的十分之一。第九个疼痛源 ds005473 在第3,601次更新时加入，此时 SCORE 目标也发生变化（第 3.3 节）。各来源问题数量将随代码库一同发布。

基线。 B1 对每个 REVE-L 词元应用多层感知机 ($1216 \to 256 \to 256$)，随后进行均值池化、层归一化和线性头处理。B2 在预处理窗口上运行 REVE-L，并使用相同的头对其最后四个模块进行微调。对于每个数据源，留出五分之一的训练受试者（至少两名）用于模型选择；CHOICE 损失按类别频率倒数加权。这些基线仅供参考：其学习率从两个候选值中选定，除此之外未进行任何调优。它们并非完全统一。七个 B1 数据源（Lee2019、Cho2017 和五个疼痛数据源）复用了此前使用相同架构训练的读出层；采用相同划分、损失和选择规则的随机种子为 7 的 B2 运行来自三条代码路径；另有十次随机种子为 8 的 B2 运行在达到十个训练轮次的上限时结束。

C 问题示例

每个数据源都属于七个问题类别之一，每个类别有五种训练措辞。评估使用其中的第一种:

- 睡眠分期（选择）：“根据数据集标注协议，该 30 秒脑电图时段被标注为哪个睡眠阶段？”

• 运动想象（选择）：“参与者在这次有提示的运动想象试验中想象的是哪只手的运动？”，候选项包括：“候选项 1：左手运动。候选项 2：右手运动。”

• 癫痫发作检测（NOUL）：“是否有任何标注的癫痫发作区间与这 5 秒脑电图窗口重叠，且重叠时长为正？”其中包含：“命题：在此窗口内的某个时间点发生了标注的癫痫发作。”

• 疼痛评分（SCORE）：“参与者在此次激光刺激试验中，根据 0 至 10 量表报告的疼痛强度是多少？”每个等级对应一行：“等级 k：报告的疼痛强度：10 分中的 k 分。”

效率实验的 AR比较器接收相同的问题与答案规范。其系统提示词被替换为“仅用一个简短答案回答有关脑电图的问题。不要解释。”并在问题末尾附加一句说明答案格式的话，例如“仅回答是或否。”所有措辞都将与代码库一同发布。

D 完整效率测量

表 A3 报告了优化配置下每个数据源和批量大小的测量结果。随附的基准测试文件还包含参考执行测量结果，便于在不同内核配置和批量大小之间进行比较。

表 A3。EEG-Jev 与针对短答案调整的匹配骨干网络的自回归对照方法（自回归）之间的推理基准测试；使用相同的图形处理器、精度和内核；每个数据源包含 32 个开发窗口。上方：批量大小为 1 时的请求延迟（每个数据源和路径计时 96 个请求）、生成的词元数量（包括序列结束标记）以及 32 个计时窗口上的任务质量（BA 范围为 0–1；ds005292 使用平均绝对误差），这些并非测试集结果。下方：批量大小为 1–32 时的每个窗口的摊销延迟（总批次时间除以窗口数量），格式为 EEG-Jev / 自回归；最后一列为四个数据源的等权平均值。计时从主机内存中的预处理后的脑电图开始，到中央处理器上得到决策（EEG-Jev）或解析后的答案（自回归）结束。所有数值见：results/inference-benchmark.csv。

	批量大小为 1 时的请求延迟（毫秒）							面板质量	
	EEG-Jev			AR			AR 词元		
数据源	mean	p50	p95	mean	p50	p95	mean	EEG-Jev	AR
Lee2019	61.16	56.30	77.74	143.30	144.22	172.03	3.00	0.781	0.719
Sleep-EDF-SC	59.16	55.93	78.09	128.78	127.75	158.89	2.72	0.629	0.248
CHB-MIT	59.92	58.61	78.00	101.21	95.87	130.41	2.00	0.750	0.875
ds005292	72.20	70.87	80.07	116.45	114.82	132.79	2.03	2.400	2.656
四个来源的均值	63.11			122.43			2.44		

	每个窗口的摊销延迟（毫秒），EEG-Jev / 自回归				
批次 size	Lee2019	Sleep-EDF-SC	CHB-MIT	ds005292	四者均值 数据来源
1	61.16 / 143.30	59.16 / 128.78	59.92 / 101.21	72.20 / 116.45	63.11 / 122.43
2	40.30 / 80.86	31.81 / 68.94	30.95 / 53.13	50.00 / 72.84	38.27 / 68.94
4	32.09 / 53.61	21.39 / 39.49	21.44 / 31.49	39.81 / 51.47	28.68 / 44.01
8	32.37 / 42.40	19.13 / 26.93	20.42 / 25.05	41.17 / 47.71	28.27 / 35.52
16	33.77 / 38.63	19.24 / 21.10	20.71 / 22.35	42.66 / 46.38	29.09 / 32.12
32	34.46 / 36.14	19.41 / 19.10	20.72 / 21.48	43.41 / 46.52	29.50 / 30.81

计时流程。两种路径均在一台 NVIDIA RTX PRO 6000 Blackwell GPU 上运行，使用 PyTorch 2.8.0、32 位浮点权重和算术运算，并使用一个 CPU 线程；优化配置为语言模型的 Gated DeltaNet 层添加融合内核，并将其设置为 IEEE 算术。预热调用后，两种路径在相同窗口上交替计时，批大小为 1 至 32，且每个批次均由同一数据源中的样本组成；每次计时调用开始前和结束时都会同步 GPU，每次

计时调用。批大小不超过 8 时，每个窗口计时三次；批大小为 16 和 32 时，计时十三次。只有当完整文本是所列候选标识符或名称、“yes”或“no”，或 0 到 10 之间的数字时，自回归答案才有效。

输入与答案。 这些窗口分别具有 62、2、24 和 63 个通道（Lee2019、Sleep-EDF-SC、CHB-MIT、ds005292）。EEG-Jev 的输入行分别有 345、196、213 和 464 个位置，而AR比较器的预填充分别为 333、184、197 和 454 个位置；AR行较短，主要是因为其指令比 EEG-Jev 的系统提示词短。AR答案的平均长度分别为 3.00、2.72、2.00 和 2.03 个词元，其中包括序列结束标记。在一个批次中，已完成的行会一直保留到最后一行结束，因此最长答案决定了该批次所需的时间。

比较器的适配。 在未进行适配的情况下，EEG-Jev 检查点无法生成可用文本：在由 32 个开发窗口组成且与计时面板不重叠的诊断面板上，只有 2 个答案有效，另有 24 个达到 32 个词元的上限。经过 50 个 LoRA更新（800 个训练集问题，每个应用 200 个；AdamW，学习率2×10^{-5}）后，全部 32 个答案均有效，并以序列结束标记结束。更新步的数量是在这个小型面板上选定的。

E 所有来源的结果

表 A4 给出了 EEG-Jev、B1 和 B2 在 35 个数据来源中各自应用主要指标上的原始输出测试结果。配对区间通过在每个数据来源内对测试受试者或患者进行 2,000 次重采样得到，并列出了至少包含五个此类独立单位的 26 个数据来源的区间；其余九个数据来源则省略区间，这一规则在测试评估前便已确定。这些区间描述的是测试群体的抽样情况，不包含训练随机种子之间的变化：EEG-Jev 仅进行了一次训练运行，而每个基线值是两个随机种子的均值，其各自的结果见结果文件。我们不报告应用层面平均值，因为

同一应用的不同数据来源在规模、协议和类别数量上各不相同，而不同应用的指标也无法相互比较。

疼痛参考。 在 ds005292 上，预测训练集均值所得的平均绝对误差为 1.771，而 EEG-Jev、B1 和 B2 分别为 1.700、1.701 和 1.717。EEG-Jev 与所学习基线之间的配对区间包含零。

NeuralBench 最佳结果。 表2中的 NeuralBench 数值，是官方 NeuralBench-EEG v1.0 结果表针对每项核心任务所列条目中的最高测试集平衡准确率均值[15,16] （结果文件来自 NeuralBench代码库的 7aaccf3 提交，对应软件包版本0.4.0；截至 2026年11月12日未作更改），以三个随机种子下的均值和标准差表示：Stieger2021 上的 REVE-Base、Sleep-EDF-SC 上的 CBraMod，以及 CHB-MIT 上的 BDTCN。它们与我们的评估共用数据划分文件、窗口、标签和指标。不同之处在于模型及其训练方式（每项任务使用一个模型，在开发集平衡准确率上进行检查点选择，并以端到端方式训练或微调）、睡眠和癫痫发作条目的信号处理方案，以及随机种子的数量。

完整指标结果。项目网站上的 results/归档文件包含所有已计算的适用指标，涵盖每个数据来源、模型、训练随机种子和输出阶段（原始、先验校正、温度缩放）：metrics-long.csv（每个指标值一条记录，包括阈值、方向、样本量、独立单位数量和区间状态）、metrics-wide-by-source.csv、metrics-by-seed.csv、paired-differences.csv（差异、区间以及每个缺失区间的原因）和 data-dictionary.md（指标定义）。本文各表格及推理基准测试背后的数值均分发在同一归档文件中。不发布逐时间窗测试数据。

表 A4。各数据源在展示指标上的测试结果，原始输出：百分比形式的 BA（运动想象、睡眠分期）；原始 p（是）≥0.5 的百分比形式，以及癫痫发作检测中的 AUROC；疼痛评分采用 0–10 量表上的 MAE。单位：独立测试单位（受试者；癫痫发作数据中的患者）。B1、B2：随机种子 7 和 8 的均值。Δ：EEG-Jev 减去基线（MAE：基线减去 EEG-Jev）；当 EEG-Jev 表现更好时为正，单位与该行相同。方括号：在数据源内对测试受试者或患者进行 2,000 次重采样得到的配对 95% 区间。该区间仅描述测试总体抽样造成的变化；不包含训练随机种子之间的差异（ EEG-Jev 只有一个随机种子；各随机种子的基线值见结果文件）。当数据源少于五个独立测试单位时（9 个数据源），不提供区间；该规则在测试评估前已固定。粗体表示每行 EEG-Jev、B1 和 B2 中展示值的最佳值；展示值并列时一并加粗。*：核心来源。其他已计算指标（准确率、宏平均 F1、AUROC、平均精确率、敏感度和特异度、负对数似然、布里尔评分、期望校准误差、CRPS、覆盖率；每个随机种子和输出阶段）见 results/metrics-long.csv；results/README.md 对结果文件进行了索引。

数据源	单位	EEG-Jev	B1	B2	Δ 与 B1 [95% 区间]	Δ 与 B2 [95% 区间]
运动想象，跨受试者，K=4；平衡准确率（%）						
EEGMMIDB	22	59.09	58.32	**66.41**	+0.77 [−1.31, +2.98]	−7.32 [−10.33, −4.35]
Stieger2021*	13	54.08	61.95	**65.19**	−7.87 [−10.66, −5.52]	−11.11 [−13.91, −8.39]
BCI-IV-2a	2	44.18	33.55	**45.14**	+10.63	−0.95
运动想象，跨受试者，K=2；平衡准确率（%）						
Dreyer2023	21	76.35	77.35	**80.35**	−1.00 [−4.26, +2.10]	−4.00 [−6.88, −1.21]
Lee2019	11	80.86	73.57	**83.86**	+7.30 [+4.84, +9.98]	−3.00 [−4.77, −1.16]
Cho2017	11	72.23	71.41	**75.95**	+0.82 [−1.16, +3.20]	−3.73 [−5.70, −1.75]
Liu2022	10	**51.75**	49.13	50.63	+2.63 [−5.13, +9.63]	+1.13 [−7.88, +8.88]
Shin2017A	6	60.28	59.31	**65.56**	+0.97 [−4.86, +7.78]	−5.28 [−11.53, +0.83]
Faller2012	3	65.67	62.92	**66.17**	+2.75	−0.50
Scherer2012	3	58.33	56.15	**58.65**	+2.19	−0.31
BNCI2014-004	2	**77.50**	71.70	74.41	+5.80	+3.09
GrosseWentrup2009	2	55.00	49.17	**58.50**	+5.83	−3.50
运动想象，被试内，K=2–3；平衡准确率 *(%)*						
Barachant2012	8	34.38	**37.50**	32.81	−3.13 [−13.33, +9.72]	+1.56 [−7.28, +14.41]
Dornhege2004	5	**66.07**	58.04	65.63	+8.04 [−1.11, +13.31]	+0.45 [−6.17, +9.74]
Zhou2016	4	**70.50**	65.67	64.83	+4.83	+5.67
BEETL-A	3	**41.44**	35.67	36.28	+5.78	+5.17
BEETL-B	2	**52.00**	38.92	37.83	+13.08	+14.17

数据源	单位	EEG-Jev	B1	B2	Δ 对比 B1 [95% 区间]	Δ 对比 B2 [95% 区间]
睡眠分期，K=5；平衡准确率（%）						
CinC2018	199	74.17	75.74	**75.83**	−1.56 [−2.15, −0.96]	−1.66 [−2.23, −1.08]
HMC	31	**74.87**	72.16	72.51	+2.71 [+0.87, +4.74]	+2.36 [+0.52, +4.52]
CAP	22	**70.32**	67.84	68.29	+2.48 [+0.23, +5.16]	+2.03 [−0.35, +4.77]
Sleep-EDF-SC*	16	70.45	68.17	**70.48**	+2.29 [+0.46, +3.94]	−0.03 [−1.98, +1.93]
DOD-O	12	65.14	**67.54**	67.41	−2.40 [−5.22, +0.68]	−2.27 [−5.85, +1.46]
DOD-H	6	71.32	**71.65**	70.21	−0.33 [−4.38, +3.93]	+1.11 [−1.53, +4.12]
Sleep-EDF-ST	5	**76.73**	69.53	71.55	+7.20 [+1.14, +13.99]	+5.18 [+0.49, +9.79]
癫痫发作检测；阈值为*0.5*时的平衡准确率 *(%)*						
CHB-MIT*	5	**72.11**	68.81	68.65	+3.30 [−0.46, +6.53]	+3.46 [−0.15, +6.73]
锡耶纳	4	**74.53**	74.41	70.34	+0.13	+4.20
癫痫发作检测；曲线下面积						
CHB-MIT*	5	0.825	0.891	**0.929**	−0.066 [−0.141, −0.005]	−0.105 [−0.185, −0.047]
锡耶纳	4	0.804	**0.865**	0.838	−0.061	−0.034
疼痛评分；平均绝对误差						
ds005292*	26	**1.700**	1.701	1.717	+0.001 [−0.131, +0.136]	+0.017 [−0.080, +0.113]
ds005280	20	**1.609**	1.613	1.636	+0.004 [−0.271, +0.273]	+0.027 [−0.247, +0.281]
ds005293	19	1.958	1.953	**1.914**	−0.006 [−0.089, +0.069]	−0.044 [−0.159, +0.086]
ds005291	13	**1.402**	1.581	1.498	+0.178 [+0.057, +0.283]	+0.096 [−0.021, +0.196]
ds005289	8	**1.303**	1.329	1.330	+0.026 [−0.135, +0.205]	+0.028 [−0.126, +0.201]
ds005285	6	**2.225**	2.331	2.245	+0.106 [−0.055, +0.333]	+0.020 [−0.364, +0.282]
ds005286	6	2.037	**1.951**	2.051	−0.086 [−0.336, +0.164]	+0.014 [−0.425, +0.488]
ds005473	6	1.259	1.346	**1.188**	+0.086 [−0.083, +0.221]	−0.071 [−0.205, +0.075]
ds005284	5	1.690	**1.602**	1.634	−0.088 [−0.476, +0.308]	−0.056 [−0.519, +0.401]