arXiv:2611.30854v1 [cs.LG] 23 Nov 2026

深度网络如何编码并选择性地使用信息？视觉与语言模型中的共享编码与条件符号读出

预印本

Kaixiang Shu独立研究者
614729197@qq.com

摘要

训练后的深度网络如何在内部表征之间保留上下文信息，同时在决策中选择性地使用这些信息？我们通过对视觉分类器进行受控的背景纹理编辑，以及对冻结语言模型进行实体绑定编辑，研究这一问题。在两种模态中，因子扰动都会在预先固定且彼此不相交的隐藏单元块之间保留可识别的结构：对于同一张图像，视觉通道块在相同纹理编辑下与其互补块的匹配程度高于与其他图像的匹配程度；而语言块则将在一个查询角色下拟合的关系方向迁移到另一个角色，用于在留出名称对中对关系状态进行排序。随后，我们将局部输出变化分解为有符号的通道贡献和通路贡献。在卷积分类器中，净通道级增强（$R^{(512)} > 1$）与单张图像上正权重组和负权重组之间的抵消（$R^{(2)} < 1$）并存。在固定权重、配对关系编辑以及红色/蓝色答案词的情况下，切换提示呈现会使通义千问2.5中的平均有限选择效应反转。两项留出测试验证了这一机制。减弱开发集选定的拮抗单元会放大未见输入上的编辑效应；在视觉 Transformer 中，在一次前向传播中减弱同一组冻结单元，会增大背景编辑效应在一个类别间隔上的平均幅度，并减小其在另一个类别间隔上的平均幅度；对于第二组单元，则呈现相反模式（四单元配对交互效应+78.30 $\times 10^{-5}$ logit）。最后，将条件通道方差与固定的下游读出权重相结合，可以预测十次切割中的单通道剪枝恢复。这些结果表明，在视觉与语言模型中，共享编码与条件符号读出可以并存。

关键词 机制可解释性 · 共享编码 · 条件符号读出 · 视觉与语言模型 · 因果干预 · 结构化剪枝

1 引言

研究训练后的深度神经网络时，一个核心科学问题是：内部表征如何在广泛保留输入信息的同时，使输出仍能高度选择性地服务于当前任务。传统直觉往往将推理描绘为逐步向上游进行的过滤过程：早期层提取丰富的结构，更深层丢弃与任务无关的上下文，而最终层则读出一个干净、孤立的目标变量。然而，对视觉和语言进行的实证探测以及源空间可视化一再表明，深度隐藏状态保留的上下文和关系结构，远多于任何单一输出决策表面上所使用的部分 [4, 40, 43, 45, 58]。视觉与语言模型如何在内部组织输入信息？它们又如何将广泛保留的表示转化为任务选择性的输出？

两项具体的跨模态观察将这一问题进一步明确为一个可检验的机制问题。在深度图像分类器中，局部伴随校正器（LAC） [43] 揭示，携带正分类头权重的通道和携带负分类头权重的通道都保留了丰富的场景与背景结构，但它们的符号组合却将能量高度集中于前景对象；随后，结合共享轴残差化的闭式源数据锚定反演 [46] 进一步定位了候选区域敏感通道子集。在执行多实体关系绑定的冻结自回归语言模型中 [10, 45], ，在中间隐藏表征内部，一个可迁移的关系分量在不同查询角色之间仍可被识别；然而，在固定模型权重和关系编辑的同时，改变提示呈现或候选答案对比，会调节该编辑的决策增益，甚至反转其符号。

将这两项发现并列来看，它们提出了一个横跨推理两方面的共同架构问题：深度网络如何在内部单元中编码 某个因子，而下游计算如何有选择地利用 它?

深度学习中的两种常见抽象并未充分说明这种共存现象。在编码方面，分布式表征这一经典概念 [41] 允许不同单元承载互补的谜题片段，只有将它们共同组合起来时，这些片段才变得可辨认。我们最初在视觉发现中称之为“全息式”编码的现象[43]——以及我们如何在不同模态中对其进行原生操作化，称之为共享编码——其定义更为具体：对受控输入因子的响应，会在多个预先固定、彼此不相交的坐标块中留下可辨认的结构。在读出方面，即使同一因子在不相交块中仍可辨认，选择性也未必依赖上游擦除。下游通道和非线性路径可以根据当前读出组合有符号贡献，从而使广泛保留的信息在一项任务中受到抑制，在另一项任务中得到放大，或在不同查询条件下发生反转。

为了在同等基础上形式化这两个阶段，我们提出一个用于训练网络推理的统一原始–对偶框架（图1）。在任意内部切口 T，上游映射 f_{up} 会在原始表征空间 $\mathcal{H}_T$中编码受控因子扰动 $\delta H_T^{(f)}$，而下游映射 $f_{\mathrm{down}}(\cdot;\, x, q)$则定义一个任务条件化对偶读出余向量 $\ell_q(x) = \nabla_{H_T} m_q(x) \in \mathcal{H}_T^*$。将 $\ell_q(x)$ 分解到正交单元分区和切空间计算路径上，可得出一条统一的局部有符号读出定律：净方向效应为有符号总和 $\tau_q(x, f) = \left\langle \ell_q(x),\ \delta H_T^{(f)} \right\rangle = \sum_\alpha t_\alpha$，其幅度恰好可分解为 $|\tau_q| = \sqrt{E_{\mathrm{incoh}}}\sqrt{\mathcal{R}_{\mathrm{int}}}$，将非相干响应能量 $E_{\mathrm{incoh}} = \sum_\alpha t_\alpha^2$与尺度不变的干扰比 $\mathcal{R}_{\mathrm{int}} = \tau_q^2 / E_{\mathrm{incoh}}$ 区分开来。根据这一规律，静态分类头抵消 [43] 这一视觉难题，以及提示条件化增益调制和符号反转 [45] 这一语言难题，成为同一规则的两个互补投影：固定任务，可以分解出单个通道和路径如何在一次读出中相互增强或抵消；固定因子并改变读出条件，则可揭示同一编码信息如何在不同任务中呈现不同幅度或相反符号。这以一个可检验的陈述解答了我们开篇提出的问题：深度网络能够在彼此不相交的表征单元间共同保留输入信息，同时以依赖读出的幅度与符号组合这些单元的贡献——使广泛保留的表示能够产生高度选择性的输出。

要建立机制性解释，必须将假设发现与因果证明区分开来。在本研究中，源坐标反演工具（LAC 和闭式源数据锚定反演 [43, 46]）仅用于在可读输入坐标中呈现候选空间结构和通道结构；所有机制性主张都在隐藏切空间中通过原生方式，并借助留出前向干预进行判定，同时明确界定源反演在数学和实证上的适用边界。至关重要的是，真正的工作原理不能只是在事后为描述性分解提供合理解释，还必须在神经网络受到主动扰动时，经受高风险样本外预测的检验。我们在大型独立确认队列中对该框架进行了两项互补的预测测试：

1. 前瞻性拮抗门控与同一前向传播中的读出翻转。如果一组响应强烈的单位确实会在目标读出下抵消其他贡献，那么在开发集划分上锁定该组成员，并在未见输入上将其编辑响应衰减 10%，就应当提高该因子的净决策影响。我们在冻结的 Qwen2.5-1.5B 和 Llama-3.2-1B 的注意力头及末层多层感知机中，以及原始的 ImageNet 预训练 ViT-B/16 中，均验证了这一反直觉预测；在这些模型中，以分散的功能单元取代连续的通道索引块，能够揭示出显著的留出对抗性。最有力的证据来自 ViT-B/16（第10块，264 张留出图像）：从同一个门控输出中读取两个类别间隔可见，衰减一个冻结的 384 单位组，会提高间隔 A 的平均背景编辑效应（$+19.25 \times 10^{-5}$ logit），同时降低间隔 B 的该效应（-10.90×10^{-5} logit）；而另一个冻结组则呈现相反模式（-7.64 与 $+40.50 \times 10^{-5}$ logit），由此得到四格配对组别与读出的交互作用 $+78.30 \times 10^{-5}$ logit。因此，一组单位是促进还是拮抗某个因子，只能结合目标读出来定义。

2. 结构化剪枝中通道恢复的定量预测。将干预从衰减转为恢复后，我们把同一有符号读出定律投影到固定线性后继下被移除通道的条件创新方向上，推导出一个精确的三因子 Frisch–Waugh–Lovell（FWL）恢复恒等式，将条件残差方差与固定下游读出耦合联系起来 [44]。在涵盖 ResNet-50、ViT-S/16 和 Swin-T 的 10 个切分上，对 4,096 张独立 ImageNet 图像进行实验后发现，纳入固定读出耦合可将留出候选恢复排名相关系数提升至 0.865–1.000（六个视觉 Transformer 切分上的数值为 0.975–0.999），并能在 Swin-T 中正确选出实际的顶级通道。

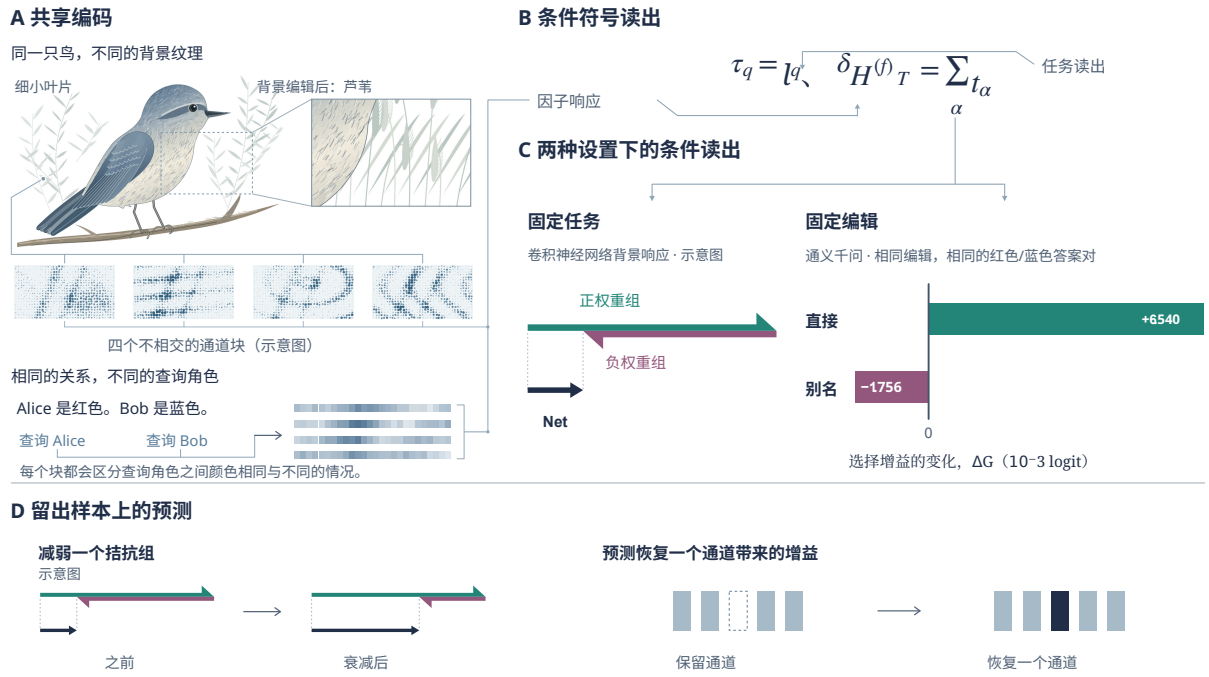


图1：共享编码与条件符号读出。A, 跨模态的共享编码：在视觉中（上），对同一只鸟的背景纹理进行编辑后，在四个预先固定且两两不相交的通道块中，鸟的结构仍清晰可辨（示意图）；在语言中（下），多实体关系状态在不同查询角色下，仍可通过四个互不相交的隐藏坐标块加以区分。**B,** 在内部切口 T处，条件符号读出定律 $\tau_q = \langle \ell_q,\ \delta H_T^{(f)} \rangle = \sum_\alpha t_\alpha$ 将原始因子响应 $\delta H_T^{(f)}$ 与任务条件化对偶读出余向量 ℓ_q配对。**C,** 两种互补的读出设置探测同一定律：固定任务（左侧，卷积神经网络背景示意图）时，正权重（蓝绿色）组与负权重（梅紫色）组部分相互抵消，形成较小的净效应；而在冻结的 Qwen2.5 中固定编辑内容和红色/蓝色答案对（右侧）时，提示切换会使测得的选择增益从 Direct（$+6.540$）反转为 Alias（-1.756×10^{-3} logit）。**D,** 两项样本外预测测试验证了这一机制：削弱开发集选出的拮抗组会增大留出输入上的净编辑效应（左侧，示意图）；结合条件通道方差与固定读出耦合，则可预测在结构化剪枝中恢复一个省略通道所带来的增益（右侧）。

对机制理论而言，同样重要的是明确其适用边界。我们的大样本实验揭示了简化说法何处不成立：仅凭架构标签无法决定抵消现象（Swin-T 中出现了建设性的纹理干扰，$R = 1.113$；而在 ViT-B/16 中，实测纹理方向和随机方向都产生了强烈抵消，纹理方向相较随机方向并未呈现出可确认的额外抵消，$+0.006\,[-0.004, +0.017]$）；按粗粒度结构路径或连续索引对单元分组，会削弱可观察到的拮抗现象；在 Pets ResNet-18 中，输入模糊会增强相对背景抵消（$R^{(2)}$ 从 0.906 降至 0.619），即使绝对相干背景能量上升（$0.00351 \to 0.00559$），分类准确率却急剧下降（87.6% $\to$ 25.9%）；局部单切分 FWL 恢复本身无法解释不同架构之间全网络多层剪枝的差异；在对有符号输入编辑进行排序时，闭式源反演也不能替代原生对偶梯度。

贡献总结。

- **统一原始–对偶理论（第3节，补充材料章节S1–S5）：** 我们在单一尺度不变合成定律下，对预先固定的不相交块上的共享编码和条件符号读出进行了形式化，推导出精确的双尺度通道桥接、适用于各向同性高斯随机基线的路径梯度 Gram 积分、前瞻性门控准则、精确的 FWL 通道恢复恒等式，以及源坐标反演的交换缺陷边界。
- **编码与读出的跨模态原生验证（第4.2节 至 第4.5节）：** 在大规模留出的视觉数据集（1,024张 CUB 图像和444张 Pets 图像）与语言数据集（256个零重叠名称对）队列中，我们分别确立了互不相交分块共享编码，展示了 $R_{(512)} > 1$ 与 $R_{(2)} < 1$在同一图像中的共存现象（444张 Pets 图像中的248张），绘制了六种架构中的多路径格拉姆几何与由粗到细的分组稀释，

并记录了视觉领域中有条件的振幅/组成变化，以及 Qwen2.5 中由提示诱发的符号反转和 Llama-3.2 中依赖词汇的调制。

- **双重样本外因果与定量预测（第 4.6 节和 4.7）：** 我们验证了视觉与语言领域中跨注意力头和功能性多层感知机组的前瞻性拮抗门控——最终在 ViT-B/16 中实现同一单元、同一次前向传播下的交叉读出符号翻转——并在 4,096 张 ImageNet 图像的 10 个剪枝切口上，确认了单通道恢复预测的定量准确性。
- **明确的机制与观察边界（第 5 节，补充章节 S6–S14）：** 我们记录了分组尺度依赖性、图像损坏下相对与绝对指标的脱钩、单次剪枝与全网络剪枝之间的差距，以及跨算子来源反演的局限，并通过完整队列复用审计和数值精确性审计加以佐证。

2 相关工作

我们的框架连接了四个大体上平行发展的研究传统：分布式表征的机制研究、特征反演与归因、上下文路由与背景选择性，以及结构化网络剪枝。

2.1 分布式表征、叠加与机制可解释性

自分布式表征的奠基性框架提出以来， [41], 人们一直认为，神经网络是在单元群体中编码信息，而非严格依赖局部化的“祖母”神经元。近期的机制可解释性研究通过线性表征假设 [20, 37], 过完备特征空间中的多义性与叠加 [9], 概念激活向量和网络解剖 [3, 23], 以及通过因果中介分析、激活修补和路径修补进行的电路级分解 [6, 13, 16, 32, 53, 55, 60]，进一步细化了这一图景。探测研究同样凸显了一个长期存在的差距：某项属性能否从隐藏状态中解码出来，与模型在执行任务时是否实际依赖该属性，并非一回事 [4, 40]。

我们的工作从两个具体方面推进了这一研究脉络。首先，在编码方面，我们区分了一般的分布式编码——不同子集可能分别承载某个因子互不重叠的部分——与跨预先固定的不相交块的共享编码；后者在视觉领域通过原生配对实例匹配验证，在语言领域则通过独立跨角色解码验证。其次，在读出方面，我们不再枚举彼此孤立、针对特定任务的电路，也不假设各个分量具有固定的功能标签，而是提出一条统一的有符号读出综合定律，涵盖通道与非线性路径；并通过同次前向传播的跨读出门控直接证明，一组单元究竟会促进还是拮抗某个因子，是单位组与读出的联合属性。

2.2 特征反演、归因与观察边界

大量研究致力于通过梯度显著性 [2, 42, 47,49], 基于优化的原像反演和特征可视化 [8, 29, 35, 36, 59], 嵌入反演 [34],以及闭式伴随或基于来源的估计器 [43, 46]，使隐藏表示在输入坐标中变得可解释。与此同时，合理性检查 [1], 对将注意力权重视为解释的批评 [22, 56], 以及线性解码中的模式与滤波器的区别 [19, 57] 均表明，可读的输入空间模式不能自动等同于因果敏感性。
我们将这一方法论洞见直接融入理论架构。局部伴随校正器 (LAC)[43] 以及结合共享分量残差化的基于源坐标的反演 [46] 在发现候选通道组抵消和空间区域偏好方面发挥了真正的探索作用。然而，公理 2 和定理 3 正式地将源反演限定为一种假设发现工具：在样本无关线性链中，理想的维纳–豪费交换构成充分条件；而非线性网络会产生经验交换缺陷，使全场反演与原生有符号决策梯度相脱离。因此，本研究中所有因果与机制判断均仅依据原生切空间测量和留出前向干预。

2.3 语言模型中的视觉背景选择性与上下文路由

在计算机视觉领域，大量证据表明，即使深度分类器达到了较高的前景准确率，它们仍会受到场景背景和虚假相关的影响 [14, 58]。在自回归语言模型中，实体绑定和事实回忆分析表明，关系结构会在不同词元位置间传递，而后期注意力头和多层感知机则会相互竞争、补偿，或主动抑制较早的信号——这在Hydra自修复

效应 [31], 复制抑制 [30], 条件共消融 [17], 和事实回忆电路 [10, 15]。关于激活引导和推理时干预的平行研究 [27, 52, 61] 日益发现，固定的引导向量在不同提示词下表现出强烈的上下文依赖性、不可靠性，甚至会发生符号反转 [5, 26, 50, 51]。

我们的原始–对偶框架将这些视觉与语言现象统一起来。在卷积神经网络中静态的正/负头部权重抵消、视觉 Transformer 中相互竞争的残差/注意力/多层感知机路径，以及冻结的大语言模型中的提示条件调制或符号翻转，都可视为同一有符号读出定律在逐步丰富的切割自由度下的特例。此外，我们的由粗到细分组实验阐明了：当单元按照架构路径或连续通道索引进行聚合时，为何拮抗结构经常会被遗漏。

2.4 结构化修剪、功能冗余与局部恢复

结构化和非结构化神经网络压缩传统上会依据激活值/权重幅值及逐层自适应启发式方法 [25, 48], 、二阶损失敏感度 [18, 24],, 或局部逐层重构和最小二乘权重折叠 [11, 44]对单元进行排序。无标签方差或协方差准则考察某个单元的表征中有多少无法从保留单元重构，而任务感知准则则根据下游敏感度对扰动赋予权重。

将我们的有符号读出定律投影到候选单元在固定线性后继下的条件创新方向上，我们得到了一个精确的三因子 Frisch–Waugh–Lovell 恒等式 [12, 28]，从而解答了我们配套剪枝研究 [44]中提出的一个未决问题：若不同时考虑固定的下游线性读出如何汇总其有符号贡献（条件读出），仅知道某个通道相对于保留单元的条件残差方差，还不足以对其恢复价值进行排序。与此同时，我们在 ViT-S/16 和 Swin-T 上进行的直接切片诊断清楚地表明了单切局部线性化的适用范围，也指出全网络多层剪枝仍是一个尚待解决的问题。

3 统一的原始–对偶框架

我们将训练完成、参数冻结的可微神经网络中的推理形式化为两阶段原始–对偶计算。为使正文精炼，并直接联系可检验的预测，我们在此陈述核心定义、主要定律、可检验推论和观测边界；完整证明、精确的解析示例和积分推导则留待补充章节S1–S5。

3.1 两阶段推理切分与原始–对偶分离

固定输入实例 $x \in \mathcal{X}$、内部表征切分 T 以及任务或查询对比 $q \in \mathcal{Q}$。我们在切分 T 处将前向映射分解为上游表征阶段 f_{up} 和下游向量值输出阶段$F_{\text{down}}(\cdot;\ x,\ q) \in \mathbb{R}^{d_{\text{out}}}$ (类别或词元逻辑值)，并由对比向量 $u_{\text{out}}(q) \in \mathbb{R}^{d_{\text{out}}}$ 投影为标量输出裕度 $m_q(x) := u_{\text{out}}(q)^\top F_{\text{down}}(H_T(x);\ x,\ q) \in \mathbb{R}$:

$$x \in \mathcal{X} \xrightarrow{f_{\text{up}}\ \text{(encoding)}} H_T(x) \in \mathcal{H}_T \xrightarrow{F_{\text{down}}(\cdot;\, x, q)} \mathbb{R}^{d_{\text{out}}} \xrightarrow{u_{\text{out}}(q)^\top} m_q(x) \in \mathbb{R}. \tag{1}$$

在视觉分类器中，f_{up} 与头部所选择的类别对比无关 q；在自回归语言模型中，查询词元出现在 x中，但切割处的中间隐藏状态 T仍可保留跨不同查询角色共享的关系结构 [45]。受控因子编辑 f 会引发原始扰动 $\delta H_T^{(f)} \in \mathcal{H}_T$（该扰动可以通过 $\delta H_T^{(f)} = J_{\text{up},\ T}(x)\ \delta x$ 从输入编辑 δx 正向传播而来，δx 也可以直接注入切割处 T）。与此同时，下游神经网络定义了对偶读出余向量

$$\ell_q(x) := \nabla_{H_T} m_q(x) = J_{F_{\text{down}}}(x;\ q)^*\, u_{\text{out}}(q) \in \mathcal{H}_T^*, \tag{2}$$

其中 $J_{F_{\text{down}}}(x;\ q)$ 是关于 H_T 的 F_{down} 的雅可比矩阵，而 * 表示伴随转置。

公理 *1*（严格原空间—对偶空间分离）。在任意切割 T 处，激活 $H_T(x)$ 和因子扰动 $\delta H_T^{(f)}$ 属于原空间 $\mathcal{H}_T$，而下游读出敏感度 $\ell_q(x)$ 属于对偶空间 $\mathcal{H}_T^*$。在对角坐标重缩放 $H_{T,\ k} \mapsto d_k H_{T,\ k}$ $(d_k \neq 0)$ 下，原始扰动按 d_k 缩放，而对偶梯度 $\ell_{q,\ k}$ 和线性状态估计器 $\mathcal{A}_x$ 均按 d_k^{-1} 缩放；因此，对偶协向量不能在未经协变度量转换的情况下直接输入原始状态估计器。此外，只要切割 T 构成一个完全顶点割，截断从 x 到 m_q 的所有计算路径，输入梯度就满足 $\nabla_x m_q(x) = J_{\text{up}}(x)^* \ell_q(x)$，并给出精确的一阶方向导数 $\langle \nabla_x m_q(x)\ \delta x \rangle$（如果 T 位于部分分支上，$\nabla_x m_q$ 会包含一个额外的旁路项；参见 *S2* 节）。

3.2 支柱 1：在不相交块之间实现共享编码

为了将共享编码与通用的分布式表征区分开来——在后者中，互补子集可能各自只包含无法使用的片段——我们使用一个 预先固定的、与任务无关且彼此正交的坐标块投影族 $\{\Pi_m\}_{m=1}^{M}$ 来划分 $\mathcal{H}_T$，该投影族满足 $\Pi_m\Pi_{m'} = \delta_{mm'}\Pi_m$ 和 $\sum_{m=1}^{M}\Pi_m = I_{\mathcal{H}_T}$（在我们的实验中， $M = 4$ 为不重叠的通道块或隐藏维度块）。

定义 1（互不相交分块共享编码）。 若表示切割 T 在受控因子 f 上的扰动响应 $\delta H_T^{(f)}$ 在每个预先固定的不相交块$\Pi_m\mathcal{H}_T$（$m = 1$，……， M）中都保留可识别的结构，则该表示切割表现出共享编码。由于视觉和语言中的受控编辑在结构上有所不同，我们以两种适合各自领域的方式原生地将这一要求操作化：

1. 视觉中的配对实例匹配： 对于给定的纹理对比，在减去开发集平均响应后，每个不相交块 Π_m 仅使用自身的原始通道——而其补集（$I - \Pi_m$）则独立地仅使用自身的通道——来估计预先固定的低维开发子空间中的坐标；对于同一图像，由此得到的坐标估计相比于由相同纹理编辑的不匹配图像，彼此的对齐程度显著更高（检验的是共享的实例级结构，而非单块图像身份解码）。2. 语言中的独立跨角色解码： 利用在开发集名称上基于一个查询角色估计出的关系方向，每个不相交块 Π_m 都能独立地在以不同查询角色评估的留出名称配对上判别关系因素。

由定义 1可直接得出两项方法学保障措施。首先，对每种纹理对比减去开发集平均响应，可以消除共同的纹理对均值偏移，同时保留图像间的实例特异性变化（不过，共享轴上的振幅差异仍可能存在）。其次，划分 $\{\Pi_m\}$必须在读出评估之前固定：事后依据任务表现选择区块，会混淆编码与读出。

3.3 支柱二：条件有符号读出定律

给定切割 T 处的因子扰动 $\delta H_T^{(f)}$，我们通过一个完整的正交单位或群划分$\{\Pi_k\}_{k=1}^{K}$（$\sum_k \Pi_k = I_{\mathcal{H}_T}$）对 $\mathcal{H}_T$ 进行分解，并通过一个完整的切空间包含—排除计算路径集 $\ell_q\,(x) = \sum_{\pi\in\mathcal{P}} \ell_q^{(\pi)}\,(x)$ 分解下游敏感度（例如，在 Transformer 中的旁路、仅注意力、多层感知机专用和混合路径 D、 A、 M、 J，或 ResNet 中的残差／卷积路径 SS、SF、 FS、 FF；具体表达式见 S2 节）。以 $\Omega(T) = \{\,(k,\ \pi)\,\}$ 对活跃分量进行索引，每个分量都对应一个有符号方向贡献

$$t_{k,\pi}(x, q;\, f) \;:=\; \left\langle \ell_q^{(\pi)}(x),\, \Pi_k\, \delta H_T^{(f)} \right\rangle . \tag{3}$$

定理 1 (条件符号读出综合定律**)**。 对于任意可微切分 T、完备正交分划$\{\Pi_k\}$和完备切空间路径分解 $\mathcal{P}$，因子 f对间隔 $m_q(x)$ 的一阶方向效应 $\tau_q(x, f)$ 满足

$$\tau_q(x, f) \;=\; \left\langle \ell_q(x),\, \delta H_T^{(f)} \right\rangle \;=\; \sum_{\alpha\in\Omega(T)} t_\alpha(x, q;\, f). \tag{4}$$

定义读出加权非相干能量 $E_{\mathrm{incoh}}(f;\ x, q)$: $= \sum_{\alpha\in\Omega(T)} t_\alpha(x, q;\ f)^2$。每当 $E_{\mathrm{incoh}} > 0$ 时，无量纲相位相干干扰比

$$\mathcal{R}_{\mathrm{int}}^{\Omega(T)}(f;\, x, q) \;:=\; \frac{\left(\sum_{\alpha\in\Omega(T)} t_\alpha\right)^2}{\sum_{\alpha\in\Omega(T)} t_\alpha^2} \;=\; 1 + \sum_{\alpha<\beta} \frac{2\, t_\alpha t_\beta}{E_{\mathrm{incoh}}} \tag{5}$$

满足精确的振幅–组成因式分解

$$|\tau_q(x, f)| \;\equiv\; \sqrt{E_{\mathrm{incoh}}(f;\, x, q)}\, \sqrt{\mathcal{R}_{\mathrm{int}}^{\Omega(T)}(f;\, x, q)}. \tag{6}$$

此外：*(i)* 每一项 t_α、 E_{incoh} 和 $\mathcal{R}^{\Omega(T)}{}_{\mathrm{int}}$ 在任何与 $\{\Pi_k\}$ 交换的分块可逆基变换 $M = \bigoplus_k M_k$（包括任意非零坐标缩放和符号翻转）下都严格不变，且无需除以基线通道范数；*(ii)* 为各分量赋予相互独立的拉德马赫符号 $s_\alpha \in \{-1, +1\}$，得到的期望基线恰为 $\mathbb{E}_s[\mathcal{R}^{\Omega(T)}{}_{\mathrm{int}}\ ] = 1$，只要 $E_{\mathrm{incoh}} > 0$（而单次符号实现可能偏离1）。

表 1：条件符号读出定律的切割特化 **层级**。视觉与**语言架构中的不同经验行为**

语言架构 ap $\ell_q(x) = J_{F_{\text{down}}}(x;q)^* u_{\text{out}}(q)$。

截断机制	$\ell_q(x)$ **的约束**	**活跃轴**	**代表性模型及其行为**
第三型（完全动态）	无约束：两者均为 $J_{F\,\text{down}}(x;\,q)$ 以及 $u_{\text{out}}(q)$ 会随上下文和查询	$k \times p \times \pi \times q$	冻结的大语言模型（通义千问2.5、Llama-3.2）：竞争路径（D，A，M，J），提示条件化增益调制以及符号翻转（第 4.4 节和4.5）。
类型 II（非线性静态-q）	固定读出对比度 q，实例依赖的非线性 $J_{F_{\text{down}}}(x)$ 和多条路径 π	$k \times p \times \pi$	中间 CNN 和 ViT/Swin 剪枝切分：依赖骨干网络的路径格拉姆矩阵几何、类别对重加权、函数式多层感知机门控（第 4.4 节到 4.6）。
I 型（固定线性头）	恒定下游雅可比矩阵 $J_{F_{\text{down}}} \equiv W$ 和单一路径 $\lvert\mathcal{P}\rvert = 1$	k（单元／组）	最终卷积神经网络头（Pets ResNet-18 双尺度合成，第 4.3 节）和 GELU 后 fc2 剪枝切割（第 4.7 节）。
类型 0（结构性零）	$\forall\alpha \in \Omega(T)\ \ t_\alpha \equiv 0$ 在子空间 $\mathcal{S}$（$E_{\text{incoh}} \equiv 0$，$\mathcal{R}_{\text{int}}$ 未定义）	在 $\mathcal{S}$ 上无	纯 ViT-B/16 最终层的非 [CLS] 仅由 [CLS] 头部处理的图像块 (第 4.4 节).

定理 1 中的两个量回答了不同的问题。带符号的总和 τ_q 给出了该因子对当前输出的一阶净效应，而 $\mathcal{R}^{\Omega(T)}{}_{\text{int}}$ 则衡量指定划分沿该读出的贡献强化（$\mathcal{R}_{\text{int}} > 1$）或抵消（$\mathcal{R}_{\text{int}} < 1$）的程度。我们还区分了结构性类型 *0* 消失——即对子空间 $\mathcal{S}$（$\forall\alpha \in \Omega(T)\ \ t_\alpha \equiv 0$）中的扰动而言，每个单独分量都恒为零，因此 $E_{\text{incoh}} \equiv 0$ 和 $\mathcal{R}_{\text{int}}$ 未定义；纯CLS头视觉变换器中，非CLS末端补丁便是如此——与精确的破坏性抵消（$E_{\text{incoh}} > 0$、$\tau_q = 0$、$\mathcal{R}_{\text{int}} = 0$）相区别。

关键的是，不同架构中看似各异的读出现象，都是单一余向量 $\ell_q(x) = J_{F_{\text{down}}}(x;q)^* u_{\text{out}}(q)$ 在逐步收紧的结构约束下的特化形式（表 1）。卷积神经网络中的静态头部权重抵消（第 4.3 节）、视觉骨干网络中的非线性多路径合成（第 4.4 节），以及冻结的大语言模型中的提示条件调制或符号反转（第 4.5 节）均遵循 公式（4）和（6）。

3.4 四项可检验的机制性推论

四项定量 预测可直接由定理 1 推出，并指导我们在第 4.

1. 精确双尺度桥接与分组尺度稀释。给定任意互不相交的粗粒度分组 $\mathcal{G} = \{g_1, \ldots, g_M\}$ 对细粒度分量 $\Omega(T)$ 进行划分，且粗粒度组和为 $U_m := \sum_{\alpha \in g_m} t_\alpha$，令 $R_{\text{within}} := \sum_m \sum_{\alpha \neq \beta \in g_m} t_\alpha t_\beta / E_{\text{incoh}}$ 并令 $R_{\text{between}} := \sum_{m \neq m'} U_m U_{m'} / E_{\text{incoh}}$。只要 $1 + R_{\text{within}} > 0$，细粒度和粗粒度干扰比就满足以下精确恒等式（证明见第 S4 节）

$$\mathcal{R}_{\text{int}}^{\text{fine}} \;=\; 1 + R_{\text{within}} + R_{\text{between}}, \qquad \mathcal{R}_{\text{int}}^{\text{group}} \;:=\; \frac{\left(\sum_m U_m\right)^2}{\sum_m U_m^2} \;\equiv\; \frac{\mathcal{R}_{\text{int}}^{\text{fine}}}{1 + R_{\text{within}}} \;=\; 1 + \frac{R_{\text{between}}}{1 + R_{\text{within}}}. \tag{7}$$

公式（7）得出两个直接推论。首先，只要组内协同效应大于组间抵消的幅度（$R_{\text{within}} > |R_{\text{between}}|$ 其中 $R_{\text{between}} < 0$），细尺度净增强（$\mathcal{R}_{\text{int}}^{\text{fine}} > 1$）与粗尺度净抵消（$\mathcal{R}^{\text{group}}{}_{\text{内部}} < 1$就会在同一扰动中严格共存（同一扰动（第 4.3 节）。其次，将对立单元合并为一个粗粒度路径区间或连续坐标块，会把它们的负交叉项纳入粗粒度分母，从而削弱可观测抵消或拮抗（第 4.4 节 和 4.6）。

2. 各向同性高斯随机基线的路径梯度 Gram 几何。在具有路径读出梯度的多路径切面上 $a_i := J_i^* u_q (i = 1, \ldots\ldots, P)$，定义 $P \times P$路径梯度 Gram矩阵 $G_{ij} := a_i^\top a_j$ 其特征分解为 $G = U\,\text{diag}(\lambda)U^\top$。对于各向同性高斯分布（等价于均匀球面分布）随机扰动v_{rand}，路径响应向量在尺度变换下服从高斯分布，因此可得到方向期望干涉比的精确一维求积公式（第 S4 节）：

$$\mathbb{E}[\mathcal{R}_{\text{int}}(v_{\text{rand}})] \;=\; \sum_{j=1}^{P} (\mathbf{1}^\top U_{:j})^2 \int_0^\infty \frac{\lambda_j}{1 + 2t\lambda_j} \prod_{k=1}^{P} (1 + 2t\lambda_k)^{-1/2}\, dt. \tag{8}$$

因此，在各向同性高斯（或均匀球面）扰动下，Gram矩阵给出了精确的总体基线；而特定纹理或关系因素的干涉比，还取决于该因素在路径梯度上的投影方式 $\{a_i\}$(第 4.4 节)。

3. 真实前向干预下的前瞻性门控。 为因果性地而非描述性地检验符号抵消，考虑将候选单元组 G 仅针对编辑增量乘以一个较小因子 $\lambda \in (0,1)$ 进行衰减$\delta h_G := h_G(x_{\text{edit}}) - h_G(x_{\text{clean}})$ 同时保持干净运行点不变（物理输出缩放作为配套控制进行评估，见 S11 节）。定义留出配对门控终点

$$H_G(\lambda) := |\Delta m_G(\lambda)| - |\Delta m(0)|, \tag{9}$$

其中， $\Delta m_G(\lambda)$ 是在衰减 λ 下编辑后与干净状态之间的间隔变化。在一个线性化单步模型中（$\tau_G(\lambda) = \tau(0) - \lambda t_G$），只要 $0 < \lambda < |\tau(0)|/|t_G|$，衰减促进组（其符号与 $\tau(0)$ 相同）就会保留原始间隔符号；而只要 $0 < \lambda < 2|\tau(0)|/|t_G|$，就会严格降低 $|\tau_G(\lambda)|$ （S5 节）。在真实的非线性神经网络中，我们使用开发集局部门控斜率 $\kappa_G := \partial|\Delta m_G(\lambda)|/\partial\lambda|_{\lambda=0}$ 来选定并冻结候选促进组和拮抗组，预测在冻结的小剂量（$\lambda_0 = 0.1$）下，衰减拮抗组会使 $H_G(\lambda_0) > 0$ （削弱该组会放大该因子的决策影响），而衰减促进组会使 $H_G(\lambda_0) < 0$；我们还进一步预测，在同一次前向传播中，同一冻结组会使两个读出 q_A, q_B 的平均值 H_G 朝相反方向变化。虽然对于任何完整分解，局部总和方程（4）都按定义成立，但开发阶段选定的组能否预测留出输入上有限非线性门控的符号和幅度，是一个可直接证伪的经验预测（第 4.6 节）。

4. 结构化剪枝中的固定读出通道恢复（FWL恒等式）。最后，考虑一个 I 型切分，其中校准激活为中心化的 $A \in \mathbb{R}^{N\times C}$，固定线性后继为 $Y = AW^\top \in \mathbb{R}^{N\times d_{\text{out}}}$，半正定输出度量为 $\mathbf{H}_{\text{out}} \succeq 0$，并通过最优最小二乘权重折叠 $W'_{\text{keep}} = W_S + W_R B_S^\top (B_S = A_S^+ A_R)$ 映射到保留子集 $S \subset \{1,\dots,C\}$，其重建损失为 $L(S) := \frac{1}{N}\|(I - P_S)Y\|_F^2$，$\mathbf{H}_{\text{out}}$ [44]。恢复单个被省略的单位 $j \notin S$ 会引入条件残差 $r_{j|S} := (I - P_S)a_j$，其条件残差方差为 $v_j := \frac{1}{N}\|r_{j|S}\|_2^2$ （相对于 S），并产生条件创新向量 $d_{j|S} := A^\top r_{j|S}/\|r_{j|S}\|_2^2$ （满足 $(d_{j|S})_j = 1$）。令 $z_k := W\Pi_k d_{j|S}$ 表示分块读出贡献，则有符号读出定律给出一个精确的恢复恒等式（第 S5 节）：

定理 2 (FWL 三因子通道恢复恒等式)。 在固定线性后继 $Y = AW^\top$、半正定度量 $\mathbf{H}_{\text{out}} \succeq 0$以及最优最小二乘权重折叠下，只要 $v_j > 0$ 且 $\sum_k \|z_k\|_{\mathbf{H}_{\text{out}}}^2 > 0$，恢复候选通道 $j \notin S$ 即可降低输出损失，降幅为

$$L(S) - L(S \cup \{j\}) \equiv \underbrace{v_j}_{\substack{\text{conditional residual}\\ \text{variance}}} \cdot \underbrace{\left(\sum_k \|z_k\|_{\mathbf{H}_{\text{out}}}^2\right)}_{\substack{\text{incoherent readout}\\ \text{energy}}} \cdot \underbrace{\mathcal{R}_{\text{int}}^{\mathbf{H}_{\text{out}}}(d_{j|S})}_{\substack{\text{signed readout synthesis}\\ \text{along } d_{j|S}}}. \tag{10}$$

定理 2 可直接从线性后继和最小二乘折叠几何得出（无需满足 定义 1中的跨块可识别性条件）：要预测恢复某个通道能挽回多少局部输出误差，需要同时考虑 它相对于保留单元的条件残差方差（v_j），以及固定的下游线性读出如何组合其带符号贡献（第 4.7 节）。

3.5 基于源坐标反演的观测范围

由于我们的视觉假 设源自局部伴随校正器和闭式源数据锚定反演 [43, 46], 我们明确地 界定源坐标估计器能够与不能够证明的内容。

公理 2 （源几何观测与原生因果性的界限）。 以源为基础的线性估计器$\mathcal{A}_x(U)$反映了（$H_T(x)$ $J_{\text{up}}(x)$ Θ_{cal}）的联合几何，并有助于揭示可读输入坐标中的候选共享结构或通道特定结构（例如，通过 *Shu* 中的共享轴残差化 $r_q = u_q - a_q b$ *[46]*）。然而，一个单行反例表明，雅可比条件回拉本身无法证明功能承载：对于 $h(x) = \|x\|_2 (x \neq 0)$, 它会丢弃所有 $(d-1)$ 角自由度， $J_x = x^\top/\|x\|_2$ 满足$J_x^\top c = c\,x/\|x\|_2$，对于任意正种子 $c > 0$均如此。因此，所有因果性或功能性主张都必须依据原生切空间测量和留出前向干预进行判定，而不能依据源空间视觉相似性。

定理 3 (充分线性对易与非线性对易缺陷)。 作为精确对易的一个充分条件，设 $H_u = J_v H_v$是样本无关线性映射，且满足 $\Sigma_u = J_v \Sigma_v J_v^*$。那么，精确线性均方误差（维纳）状态估计器 $\mathcal{W}_{v\leftarrow u} = \Sigma_v J_v^* (J_v \Sigma_v J_v^*)^+$与双重反向传播恒等对易（补充材料第 *S3* 节）：

$$\mathcal{W}_{v\leftarrow u}(\Sigma_u \ell_u) \equiv \Sigma_v (J_v^* \ell_u) = \Sigma_v \ell_v. \tag{11}$$

在深度非线性网络中，这一恒等式会产生经验性上游交换缺陷

$$\Delta_{\mathrm{comm}}(x;\, q) \;:=\; \mathcal{A}_x\big(\Sigma_T^{\mathrm{global}}\ell_q(x)\big) - \Sigma_X J_{\mathrm{up}}(x)^*\ell_q(x), \tag{12}$$

结合有限样本估计、全局与局部统计失配、岭正则化和近似误差；在实证分析中，我们通过标量归一化交换缺陷$\|\Delta_{\mathrm{comm}}(x;\ q)\|_2/\|\Psi(x)\|_2$来概括每张图像的这一向量差异，其中 $\Psi(x)$: $= \Sigma_X\nabla_x m_q(x)$。因此，带符号输入方向预测由原生的 $\nabla_x m_q(x)$ （及其协方差加权的原始对偶项 $\Psi(x)$）支配，而定理 *3*则校准闭式源反演的非线性操作边界（第 *5.4* 节、*S14* 节）。

4 实验验证

我们使用大型、预先分离的确认队列，在七种视觉与语言架构上评估原始–对偶框架。实验顺序与论文结构相呼应： 第 4.1 节确立原生操作规程、队列来源和数值审计；第 4.2 节在不相交块中原生检验共享编码；第 4.3 节至 4.5 检验条件符号读出在固定任务分解和变任务切换下的表现；而第 4.6 节和 4.7 则通过前瞻性因果门控和定量剪枝恢复预测对该框架进行检验。

4.1 实验规程、队列与精确性审计

所有机制性主张均直接在参数冻结的预训练网络或任务训练网络上进行检验，不依赖特征反演器。表2总结了八项原生实验规程、其中固定与变化的量，以及各自独立的统计单位；涵盖所有终点和范围边界的整合确认台账存档于表 S4(第 S6 节)。

队列分离、复用披露与统计惯例。独立重采样单位始终是完整的源图像（视觉）或完整的实体名称对（语言）；每张图像的多个供体编辑、每个名称对的多个提示词，以及同一次前向传播上的多个读出，均先在各自单位内汇总，再进行配对自助法重采样（10，000 次抽样为默认值；20，000 次抽样用于语言模型注意力头门控和 CUB 区域带符号皮尔逊相关；4，000 次抽样用于 ImageNet 剪枝；以及1，000 次抽样用于 Pets 剂量级导数到有限差分的斯皮尔曼校准）。除非明确标注为经过多重性调整（97.5%、98.75%或99.375%），方括号表示双侧95%配对自助法置信区间。为避免夸大样本数量，我们始终将初始探索性队列与最终确认队列分开，并明确披露所有跨章节复用情况（完整审计见第 S6 节）：

- **CUB-200-2011 [54]:** 1，024 张独立确认图像用于 ResNet-18、ResNet-50、Swin-T 和 ViT-B/16 的编码、路径及读出分析（1，020 张图像符合S14 节中的空间区域资格筛选条件），另有一组独立的 512 张图像开发集，以及一组独立的264 张图像保留集，用于 ViT-B/16 功能组留出确认（同一组 264 张图像也用于S11 节中的经 CUB 微调的 ViT-B/32 比较）。
- **Oxford-IIIT Pets [38]:** 两组独立的新鲜队列，分别包含 128 张和 444 张源图像（总计 572 张，每张均与四张专属供体图像配对；其中 442 张符合弱暖色/冷色色调探测的条件），并与历史上的 16 + 48 = 64 张图像探索性档案严格区分。

- **冻结的大语言模型（通义千问2.5-1.5B [39] 和 Llama-3.2-1B [33]）：** 256 对零重叠确认名称（共 512 个不同名称）由两个模型共同用于跨角色编码、四路径分解和注意力头门控；其中，52 对名称在两个分词器及两种词序位置上均符合单词元分词条件，构成 2×2因子队列；另有一组独立的 256 对额外零重叠名称（共 512 个不同名称），用于独立确认最后一层 MLP 功能分组和由粗到细的路径拆分。
- **ImageNet-1K [7] 结构化剪枝：** 4，096 张确认图像，涵盖全部1，000 个类别，在 ResNet-50、视觉变换器-S/16 和 Swin-T 的全部 10 种切割中共享（257 张图像与 Shu 历史上的 3，200张全网络评估图像重叠 [44]，另行报告于表 S10）。

数值闭合与物理手术审计。 在评估机制假设之前，我们验证了每个分解量都与最终确认队列上的精确正向执行相匹配（表 S3）。切空间路径总和的闭合误差在 5.26×10^{-7} 以内（CUB ResNet-18 的 `layer3` （1，024 幅图像； 4.20×10^{-8} 以内（`layer4.0`）；对角坐标重缩放遵循公理 1，在 2.32×10^{-6} 以内，覆盖444 幅宠物图像；快速 FWL 评分与显式最小二乘重新拟合的结果一致，误差为 $\leq 9.43 \times 10^{-18}$，涵盖视觉 Transformer 的早期/中间层剪切点（$9.11/8.81 \times 10^{-18}$ 在中间剪切点； $\leq 1.005 \times 10^{-14}$ 覆盖4，096 幅图像的全部六个剪切点）；有限差分与方向导数的符号一致率

表2：操作性方法：八项原生实验。每项测试都冻结模型参数，并隔离一种受控的扰动或读出变化，并以独立源图像或完整实体名称对为单位进行总结。

实验（§）	冻结的模型与切割	固定量与变化量	原生端点与单位
共享编码，视觉（4.2）	CUB ResNet-18（layer3, layer4.0）、ViT-B/16（第6块）；Pets ResNet-18	固定纹理对与开发集16维基；比较同图块到补集与不匹配的同纹理图像	跨4项的均值中心化余弦优势不相交块（1,024 个 CUB）；弱色调探针（442 个 Pets）；源图像
共享编码，语言（4.2）	通义千问2.5-1.5B（第14块），Llama-3.2-1B（第8块）	冻结的开发集关系方向；评估在 4 个互不重叠的隐藏维度块中的查询角色之间区块	组内名称对跨角色排序 AUC & 余弦相似度（256 个名称对）；名称对
通道合成（4.3）	Pets ResNet-18 最终 512 通道	固定源图像、干净样本的 top-1 类别，4 供体背景/前景编辑；头部符号划分与随机划分	有符号贡献，双尺度 $R^{(2)}$ & $R^{(512)}$，有限前向秩相关（444 + 128 图像）；源图像
通路几何（4.4）	ResNet-18/50、Swin-T, 视觉变换器-B/16（CUB）；Qwen2.5, Llama-3.2	固定输入与读出；视觉：子空间投影、单位归一化来自背景编辑的通道向量（在空间/标记位置间广播）与匹配的随机通道向量；大语言模型：投影到冻结的开发子空间（查询标记注意力输出）；粗粒度与功能性 MLP 划分	方向 R、Gram 积分、6 个归一化交叉项、截断前向传播、粗粒度–细粒度 ΔR；图像/名称对
读出切换、视觉（4.5）	CUB ResNet-18（layer3）	固定的干净图像，子空间投影单位归一化背景通道向量在 layer3、剂量和相同的 8 个鸟类类别；重新配对组内与跨组栖息地	单位剂量 $\|\Delta m\|$ 双重差分（完整/半剂量）& $\|\tau\| = \sqrt{E}\sqrt{R}$ 分解（1,024 张图像）；源图像
读出切换，语言（4.5）	通义千问2.5（第14块），Llama-3.2（第8块）	冻结的跨名称关系轴（$0.5ab\,v$, 查询令牌注意力输出）；2×2 呈现（直接/别名）×答案对（RB/OP）	有符号有限选择对比 ΔG（10^{-3} logit, 52 组双分词器配对）；名称对
前瞻性门控（4.6）	通义千问2.5、Llama-3.2（注意力头及最后一个多层感知机）；ViT-B/16（第6块, 10）	开发集选定的冻结单元；10%的编辑增量衰减（$\lambda_0 = 0.1$, 干净点固定）；ViT同次前向传播读出翻转	留出 $H_G = \|\Delta m_G\| - \|\Delta m(0)\|$ 对比连续且强度匹配的对照组；图像/名称对
剪枝恢复（4.7）	ResNet-50（4 个瓶颈块），ViT-S/16 和 Swin-T（早期/中期/后期多层感知机）	冻结的保留集 S、普通最小二乘法折叠、线性后继；逐个恢复每个候选通道逐个	线性后继均方误差降幅、候选通道斯皮尔曼相关系数 ρ，单块边际平均绝对误差（4,096 张图像）；源图像

达到 98。27%（早期）和 99。44%（后期），适用于 CUB ResNet-18（1，024 幅图像）；在 ViT-B/16 中，主剂量/半剂量下达到 99。988%/99.994%（1，024 幅图像）；对 21，504 个语言 MLP 门控条件进行的全网络钩子审计与批次评估完全一致；在全部 36 个物理剪枝手术配置中，$\times 4$，096 幅图像（147，456 个配对）中的每一对均通过类别间隔容差检验（在后期 Swin GELU 后直接切片条件下出现的 732 个 float32 局部均方误差例外，在 float64 中重放后，最大差异降至 6.22×10^{-15}）。

4.2 不相交表征块间的共享编码

我们首先检验支柱 1（ 定义 1）：受控输入因子是否会在预先固定的区块中留下可识别的结构 ，彼此不相交的坐标 块中，且独立于任何特征反演器，并处于原生表征空间。

视觉发现的切入点，以及为何逐纹理的跨图像平均方向并不足够。对 Pets ResNet-18 进行单通道 LAC 反演，揭示了各个深层通道内部包含大量场景和背景结构（图 S3，上）。为在不进行反演的情况下原生测量背景编码，我们在 CUB-200-2011 上构造受控背景编辑：仅在鸟类分割掩码之外以 5% 的混合权重混合自然纹理，所有前景像素均保持不变（图 2A）。一个自然的初步假设是，共享编码仅仅反映了每种纹理对在跨图像中固定不变的平均响应。我们在 ResNet-18 末端全局平均池化（GAP）处，为每种纹理对估计一个固定的跨图像平均方向（四种纹理对共四个方向），并在 1,024 张留出的 CUB 图像上进行评估，结果否定了这一简单图景：尽管匹配纹理对的对齐度高于不匹配纹理对（0.0891 对 0.0020，优势为 $+0.0870\,[0.0840, 0.0901]$），但每种纹理对的跨图像平均方向平均只能解释该纹理对留出图像 GAP 响应能量的 1.52%，且仅在 54.86% 的试验中能预测类别间隔变化的符号。相比之下，逐图像原生方向导数的符号一致率达到 73.47%；当其与不同纹理对不匹配时，一致率降至 50.50%；当其与重新分配的读出配对时，则降至 49.77%（图 S2）。

跨不相交块的实例基础视觉匹配。视觉背景响应并非由末端 GAP 处低能量的每个纹理对跨图像平均方向（1.52%）所捕获，而是在不相交的通道块之间保留了连贯的实例特定结构。在 1,024 张留出的 CUB 图像上，减去

针对每种纹理对比的开发集平均响应后，四个预先固定的不相交通道块中的每一个都仅使用自身的原始通道（通过对块限制的16维开发集基进行伪逆映射，其中将16 选作分析维度，而非声称这是内在维度）来估计16维坐标；与此同时，互补通道仅使用自身的原始通道独立估计坐标。随后，我们检验同图块到补集的余弦对齐是否超过使用完全相同纹理编辑的不匹配图像（图2B）。在所测试的三个视觉切面上，每个块均以较大的正余弦优势通过检验；每个切面的子空间能量覆盖率均单独报告，并与每个纹理对末端 GAP 方向的1.52% 平均能量占比并列：

- **ResNet-18** layer3： 在四个块中，相同与不匹配之间的余弦优势为 $+0.555$ 至 $+0.595$ （99。375% 多重性调整置信区间严格为正；16维子空间覆盖了42。14%的响应能量）；
- **ResNet-18** layer4.0： 余弦优势为 $+0.397$ 至 $+0.422$ （14。09%的能量覆盖率）；
- **ViT-B/16 Block 6**： 余弦优势为 $+0.874$ 至 $+0.884$ （49。26%的能量覆盖率）。

即使两侧各自只能看到自身的不相交通道，并且移除了开发集纹理-对比均值，块及其补集仍能在预先固定的 16维子空间内恢复出高度对齐的实例特异性坐标。为探究因子强度的下限，我们还使用每个块的冻结线性探针对 Pets ResNet-18 中 1,024 个选定背景像素施加的微弱暖色与冷色局部背景色调进行评估：在 442 张新符合条件的留出图像上，四个不相交块分别取得 248、 269、250、 和 253 个正确预测（56。1%–60。9%，联合单侧 $p = 0.0042$；而较小的 64张图像探索性批次有 $p = 0.546$； 图 S2）。因此，微弱的局部背景色调在不相交块之间仍可在较大规模上被检测到，但其效应量远小于结构化自然纹理。

冻结语言模型中的跨角色关系共享编码。我们将定义 1 中的语言操作化方案应用于多实体属性绑定任务中的冻结的 Qwen2.5-1.5B（第14块）和Llama-3.2-1B（第8块）[45]。我们在开发集名称上针对一种查询角色估计关系绑定方向，并在另一种查询角色下，跨四个预先固定且互不重叠的隐藏维度块进行评估，使用256 组零重叠留出名称对（图2C）。具体而言，对于每组留出名称对和留出查询角色，我们计算名称对内、跨角色排序 *AUC*，比较两个同属性状态（$a = b$）与两个不同属性状态（$a \neq b$），允许针对每组名称对使用特定的分数基线，并对所有角色和名称对取平均。两个模型的八个块均实现了名称对内跨角色排序$\mathrm{AUC} = 1.000$，其中连续跨角色余弦对齐在 Qwen2.5 中介于0.792 至0.816 之间，在 Llama-3.2 中介于0.827 至0.844 之间。因此，在视觉与语言架构中，受控输入因素在彼此不相交的隐藏坐标块中仍可被识别。在阐明因素如何编码之后，我们转而探讨下游读出如何有选择地使用这些因素。

4.3 固定任务通道合成与theTwo-ScaleBridge

如果多个通道块保留了背景结构，固定线性分类器如何在保留前景的同时抑制背景？我们直接在定理1的 I 型特例上进行测试，使用 Oxford-IIIT Pets ResNet-18（512 个最终通道），固定每张源图像干净样本的 top-1 类别，并严格按照分类头权重的符号将通道划分为正组和负组。每张源图像在前景掩码或背景掩码内接受四种互斥的供体扰动（先对每张图像的供体能量求和）。

对组间背景抵消进行大样本独立确认。在两个独立的确认队列中，双组干扰比 $R^{(2)}$ 能清晰地区分背景抵消与前景增强（图 3A；队列来源（包括较早的16、48和64 个探索性批次，其中 $R^{(2)}_{\text{背景}} =$0。817、0。839、0。834，详见图S4）：

- **首个新鲜队列（128 张图像）：** 背景 $R^{(2)}_{\text{背景}} = 0.858\,[0.787, 0.929]$ $(81/128 < 1)$，前景$R^{(2)}_{\text{前景}} = 1.122\,[1.058, 1.189]$;
- **扩展的新鲜队列（444 图像）：** 背景 $R^{(2)}_{\text{背景}} = 0.906\,[0.868, 0.944]$ $(266/444 < 1)$，前景$R^{(2)}_{\text{前景}} = 1.089\,[1.054, 1.125]$。

将固定读出从干净样本Top-1 logit切换为固定的Top-1减Top-2间隔，仍能保持背景抵消（$R^{(2)}_{\text{背景}} = 0.879$ 在这 444 张图像上），且原生方向导数能紧密跟踪实际的有限剂量logit变化（斯皮尔曼相关系数0。871，完整RGB剂量为 0。125；半剂量为 0。936，剂量为 0。0625；图S4）。

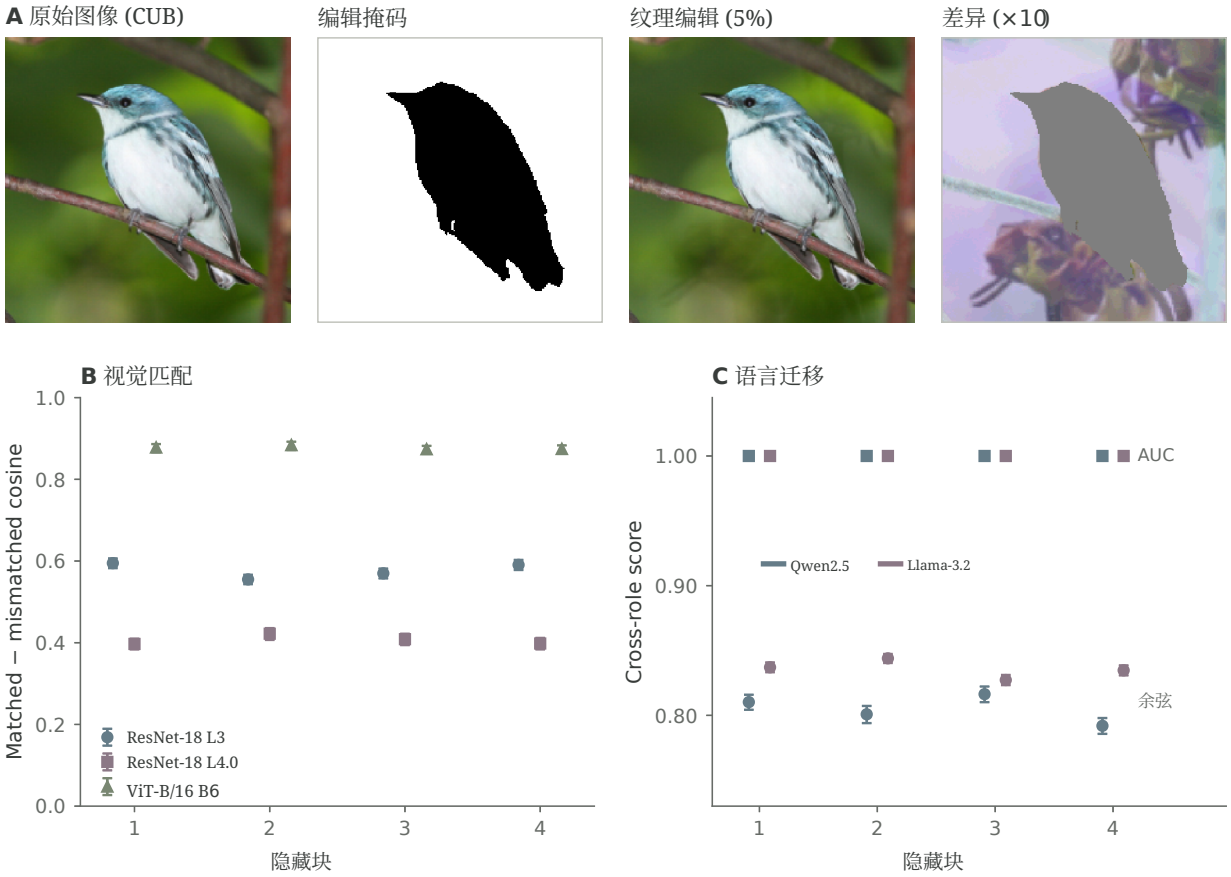


图2：视觉与语言模型中不相交表征块的共享编码。A、在 CUB-200-2011 上进行的受控视觉背景纹理编辑：原始鸟类图像、可编辑背景掩码（白色：可编辑背景；黑色：未修改前景）、5% 的自然纹理背景编辑，以及以中灰色为中心的 $10\times$ 放大带符号像素差（整个鸟类轮廓上的变化均为零；Pets LAC 发现字段见图 S3）。B、在 1,024 张留出的 CUB 图像上进行的原生视觉配对匹配：对每种纹理对比减去开发集平均响应后，每个不相交通道块及其补集都仅使用各自的原始通道，独立估计 16维坐标（预先固定的 16 维子空间在 ResNet-18 layer3、layer4.0 和 ViT-B/16 Block 6 的响应能量覆盖率分别为 42.14%、14.09% 和 49.26%）；同图像对齐显著高于具有相同纹理的不匹配图像（柱状图：99.375% 源图像自助法置信区间）。C、每个模型在 256 个留出名称对上进行的原生语言跨角色迁移：冻结的Qwen2.5-1.5B（第14块）和 Llama-3.2-1B（第8块）中的全部四个不相交隐藏块，均实现了名称对内、跨角色排序 AUC $= 1.000$（方块：在每个名称对内对两个 $a = b$ 与两个 $a \neq b$ 状态进行排名）以及较高的连续余弦对齐（圆圈；柱状图：95% 名称对置信区间）。

归因于权重符号分组和同图像双尺度共存。 为验证背景抵消确实反映了实际的头部权重符号划分，而非任意的双组平均，我们在每张图像上固定完全相同的 512个带符号通道贡献，并将真实的头部权重划分与 1,024 个规模匹配的随机双组划分进行比较（图3B）。在这 444 张图像上，与随机划分相比，真实的符号划分降低了背景 $R^{(2)}$，降幅为 $-0.792\,[-0.821, -0.761]$（-0.816 在 128张图像组成的队列中），从而将这种抑制明确归因于学习到的权重符号归属。

与此同时，在完全相同的512 个单独通道上评估完全相同的444 个背景扰动，得到的细尺度干扰比远高于 1：$R^{(512)} = 1 + 9.200 - 0.937 = 9.263$ (图 3C)，其中正向的组内符号协同（+9.200）超过了负向的组间符号抵消（-0.937）。根据公式（7）中的精确双尺度桥接，这两项测量在逐张图像层面代数一致。具体而言，在444张确认图像中， $266/444$ 张图像同时呈现正向组内交叉项和负向组间交叉项（等价地， $R^{(2)} < 1$）；并且 $248/444$ **张单张图像同时满足** $R^{(2)} < 1$**（组间抵消）和** $R^{(512)} > 1$ **（在**512**通道分解中的净增强，** $426/444$）。因此，组内协同与组间抵消原生地共存于同一输入中。接下来我们将探讨

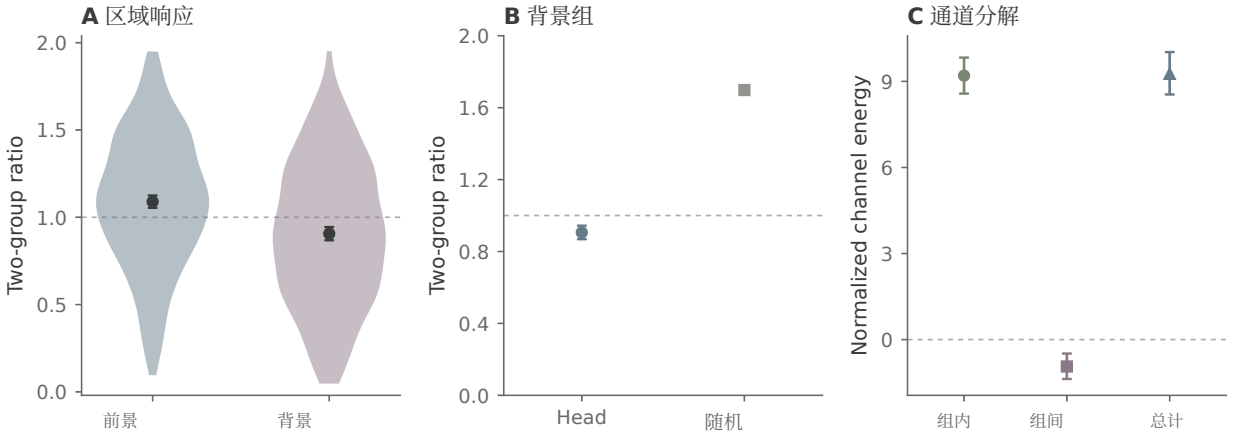


图 3：Pets ResNet-18 中两个分组尺度下的固定任务通道合成（444 张留出图像）。A, 固定干净样本的 top-1 类别头符号划分下，双组干扰比 $R^{(2)}$ 的分布与图像均值（95% 自助法置信区间）：前景编辑带来增强（$1.089\,[1.054, 1.125]$），而背景编辑则产生抵消（$0.906\,[0.868, 0.944]$）。**B,** 在完全固定相同的 512 个带符号背景贡献时，头部权重符号划分的抵消程度明显强于大小匹配的随机划分（配对差异 $-0.792\,[-0.821, -0.761]$）。**C,** 在相同的 444 张图像上进行精确的 512 通道分解（$R^{(512)} = 1 + 9.200 - 0.937 = 9.263$）： $266/444$ 张图像的组内符号交叉项为正、组间符号交叉项为负；$248/444$ 张单张图像同时满足 $R^{(2)} < 1$ 和 $R^{(512)} > 1$（在 512 通道分解中净增强）。

当下游读出经过多个非线性模块和架构路径时，这种有符号合成如何扩展。

4.4 固定任务多路径几何与由粗到细的分组尺度

从单个线性头（I 型）转向中间非线性切面（类型 II/III）后，我们将下游对偶协向量分解为四条完整的切空间计算路径：ResNet 中的残差/卷积路径（SS、SF、FS、FF），以及 Transformer 中的旁路、仅注意力、仅多层感知机和混合路径（D、A、M、J）（S2 节）。在 CUB 视觉路径实验中（图 4A），我们首先施加5% 仅背景像素纹理编辑，以估计切面处的通道均值响应差 T，将该差异投影到该切面冻结的16维发育子空间上，并将所得通道向量归一化为单位 L_2 范数，然后将其注入干净图像的隐藏切面（在 ResNet 和 Swin-T 中沿空间位置均匀广播，并作为相同的768维向量添加到 ViT-B/16 的全部197 个词元中）；匹配随机方向则在完全相同的通道空间中抽取。因此，这些实验测量了下游路径如何读取根据背景纹理编辑估计出的隐藏层通道方向。

四种视觉骨干网络中的路径梯度 Gram 基线与架构多样性。在相同的1,024张留出的 CUB 图像（图 4A）上，路径梯度 Gram 积分（公式 (8)）能够预测每个受测视觉切面的各向同性高斯随机方向干扰基线，同时揭示出显著的架构差异：

- **ResNet-18（layer3）：** 纹理 $R = 1.158$ 与冻结随机方向库 $R = 1.133$（解析 Gram 积分为 **1。1488**；配对逐图积分减随机方向库差值 $+0.0156\,[0.0119, 0.0193]$，反映了有限随机方向库的影响，但两者仍具有很强的建设性）；梯度能量高度集中于两条对齐的卷积路径（FF 90.6%、SF 9.3%）。
- **ResNet-50：** 纹理 $R = 1.754$ 与随机基线 $R = 1.377$（纹理减随机基线 $+0.376\,[0.348, 0.405]$）。
- **Swin-T：** 纹理 $R = 1.113$ 与随机基线 $R = 0.992\,[0.986, 0.998]$（纹理减随机基线 $+0.121\,[0.104, 0.138]$）。这确立了一个重要边界：*Transformer* 并不会普遍抵消纹理扰动。
- **ViT-B/16（第6块）：** 纹理 $R = 0.284\,[0.272, 0.295]$ 与随机基线 $R = 0.277\,[0.274, 0.280]$（Gram 积分 0.2779）；四条路径的能量均衡，且混合路径 J 与其余路径之间存在强烈对立（通过原生前向钳制得到验证：在全剂量、半剂量和一阶条件下，归一化的混合路径与其他路径的交叉项均四舍五入为 -0.396——95% 置信区间 $[-0.408, -0.384]$）。

在同一个冻结的 ViT-B/16 中，另外两项控制实验验证了表 1 中的分类法。在终端层编辑非 [CLS] 补丁，并采用仅含 [CLS] 的头部时，所有 1,024 张图像的间隔梯度和有限 logit 变化均精确为零（结构性类型 0）。在第6块中，将注意力/多层感知机模块的完全相同总导数重新归组为前三个与后三个深度区块后，纹理 R 从 0.284 变为 0.409（随机基线 $0.277 \to 0.331$，格拉姆 0.332；图 S8），这证实即使总敏感度固定，$\mathcal{R}_{\text{int}}$ 也取决于所选的分组边界。此外，尽管测得的纹理和匹配随机方向在 ViT-B/16 中都表现出强烈的净抵消，但纹理相较随机方向是否存在额外抵消仍无法确定：在这 1,024 张图像上，纹理减去随机基线 R 为 $+0.006\,[-0.004, +0.017]$（随机方向中有 4,096 个图像—读出组合中的 4,094 个同时涵盖 $R < 1$ 和 $R > 1$；图S9）。

视觉与语言 Transformer 中反复出现的负向通路交叉项。 检查 ViT-B/16（$\sum_{i<j} 2t_i t_j / E_{\text{incoh}} = R - 1$）中的全部六个归一化成对交叉项（1,024 个 CUB 图像）、通义千问2.5-1.5B（256个确认名称对）和 Llama-3.2-1B（256 个确认名称对）揭示了两个一致的跨模态负向交互效应（图 4B）：

- **仅注意力 × 混合写入**（$A \times J$）在三个模型中均呈强负向：$-0.157\,[-0.170, -0.145]$ 在 ViT-B/16 中，$-0.262\,[-0.284, -0.240]$ 在通义千问2.5 中，以及 $-0.332\,[-0.350, -0.315]$ 在 Llama-3.2 中。
- **旁路 × 仅MLP写入**（$D \times M$）在三个模型中同样均为负值：$-0.061\,[-0.067, -0.054]$在 ViT-B/16 中，$-0.213\,[-0.233, -0.194]$ 在 Qwen2.5 中，以及 $-0.108\,[-0.126, -0.090]$ 在 Llama-3.2 中。

相比之下，$A \times M$ 在 ViT-B/16 中略为负（-0.020），在 Qwen2.5 中也略为负（-0.023），而在 Llama-3.2 中穿过零点（$-0.006\,[-0.018, +0.005]$；完整的六对表格见表 S5）。

粗粒度通路聚合会稀释可观测的关系抵消。在语言通路测试中，每个名称对自身的关系交互作用都会被投影到冻结开发子空间，并注入查询词元处的注意力模块输出中（表2）。在粗粒度四路径分解（D，A，M，J）下，两个语言模型整体上都表现出净抵消（$R < 1$），然而，相对于匹配范数的随机方向，关系方向的额外抵消（$\Delta R := R_{\text{rel}} - R_{\text{rand}}$）幅度较小或尚无定论：在第一个256对队列中，$\Delta R = -0.0192\,[-0.0545, +0.0162]$ 在 Qwen2.5 中，以及 $-0.0700\,[-0.1041, -0.0360]$ 在 Llama-3.2 中；在额外的256对队列中，$\Delta R = -0.0315\,[-0.0592, -0.0025]$ 在 Qwen2.5 中，以及 $-0.0479\,[-0.0822, -0.0123]$ 在 Llama-3.2 中。

受分组尺度原则（公式（7））的启发，我们测试了将相互对立的多层感知机神经元合并为单一粗粒度 M 和 J 通路是否会稀释 ΔR（这一设计是在观察到粗粒度四路径结果后提出的，并在另外 256 组零重叠名称对上独立验证）。每次试验均保持完全相同的输入和编辑，并且总方向导数不变；我们将最后一层多层感知机（通义千问第 28 层，1，120/8，个单元；Llama 第 16 层，1，024/8，个单元）划分为开发阶段选定的功能单元、其余同层单元以及绕过层的通路（图 4C）。将功能单元分离后，两个模型中的关系减随机抵消均显著加深：

- **Qwen2.5-1.5B（新增 256 组配对）：** ΔR从 -0.0315 加深至 $-0.1068\,[-0.1339, -0.0794]$，得到配对的粗粒度减细粒度对比 $B = +0.0753\,[0.0564, 0.0948]$（与重放队列上的 $+0.0751$ 非常接近）。对 B 贡献最大的单一交叉项，是选定单元与其余同层单元之间的 MLP 内部交互（$+0.0836\,[0.0668, 0.1018]$）。相较于严格的 ±1% 强度匹配随机对照，功能组整体的粗粒度减细粒度对比 B 仍保有 $+0.0699\,[0.0563, 0.0851]$的配对优势。
- **Llama-3.2-1B（新 256 对）：** ΔR从 -0.0479 加深至 $-0.1706\,[-0.1943, -0.1471]$，其中 $B = +0.1227\,[0.0876, 0.1595]$（在重放队列上为$+0.1124$）。同样，主导分量是组内选定与剩余部分之间的交叉项（$+0.1039\,[0.0786, 0.1304]$，相应的 J-路径交叉项贡献了 $+0.0376\,[0.0256, 0.0499]$），且功能组的总体 B超过严格的强度匹配对照组 $+0.1372\,[0.1061, 0.1702]$。

因此，将作用相反的单元合并为单条粗粒度通路，会掩盖 MLP 层内部显著的关系特定抵消。在剖析了跨通道、通路和尺度的固定任务读出后，我们现在反转实验约束：固定因子，改变读出条件。

4.5 在固定因子下改变读出条件

如果选择性源于条件符号读出，而不是硬编码的上游擦除，那么固定完全相同的隐藏状态扰动、仅改变读出条件 q，就应当重塑该因子的增益，甚至反转其符号。

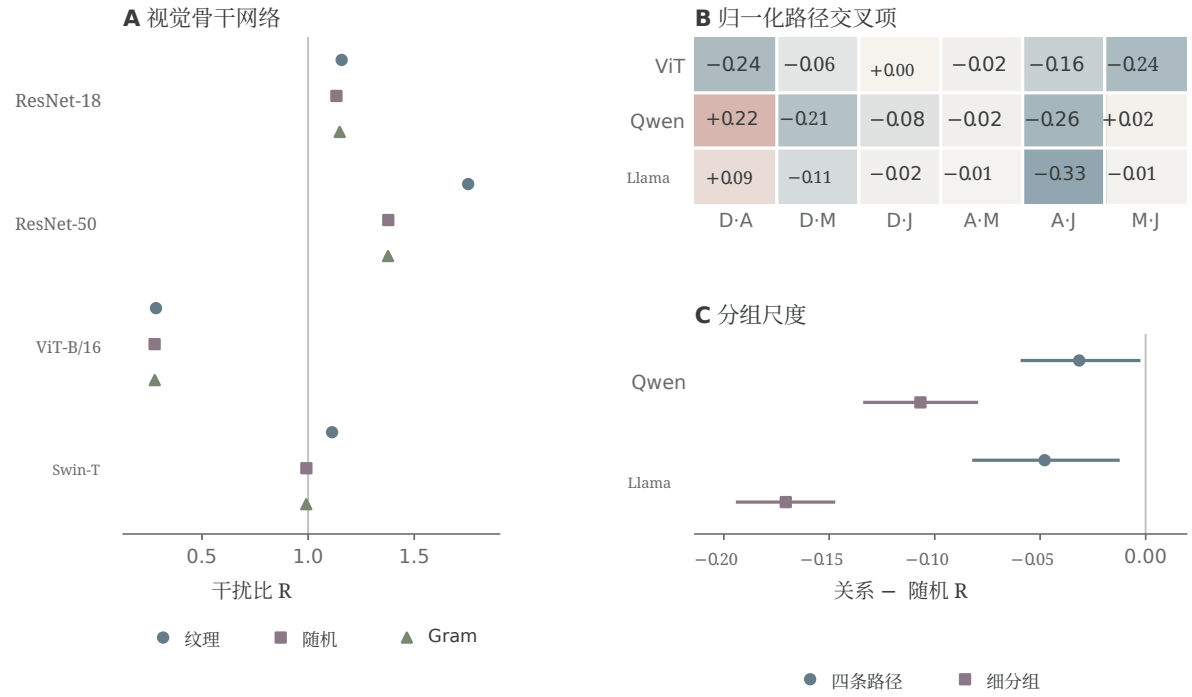


图 4：多路径干扰几何与由粗到细的分组尺度。A，在相同的 1,024 张 CUB 图像上，对测得的纹理编辑、匹配范数的随机方向以及精确的路径梯度 Gram 积分（公式 (8)），计算四种冻结视觉骨干网络的平均方向干扰比 R；对于测得的纹理方向，ResNet-18/50 和 Swin-T（1.113，而 Swin-T 的随机基线略低于 1，为 0.992）高于 1，而 ViT-B/16 则表现出强烈的抵消（0.284）。B，在 ViT-B/16（1,024 张图像）、通义千问2.5-1.5B 和 Llama-3.2-1B（各有 256 对确认名称对）中，平均归一化成对交叉项：在这三种 Transformer 中，$A \cdot J$ 和 $D \cdot M$ 均为负值。C，在额外的 256 对零重叠名称对上，以相同的总方向导数进行粗粒度四路径与精细功能性多层感知机分解：隔离开发阶段选定的多层感知机单元，会使通义千问2.5中的关系减随机 ΔR 加深 $+0.0753$，并使 Llama-3.2 中的关系减随机 ΔR 加深 $+0.1227$（95% 配对名称级置信区间）。

视觉读出重新配对与幅度—组成分解。 在 1,024 张留出的 CUB 图像上，使用与 第 4.4 节完全相同的隐藏层注入协议，我们固定干净图像、ResNet-18 `layer3`切割、子空间投影的单位归一化背景通道向量（在空间位置上进行广播）、编辑剂量以及 完全相同的八种鸟类物种，仅改变这八个类别配对为类别间隔对比的方式：栖息地内配对与跨栖息地配对（每个类别在两组中均恰好出现一次）。减去匹配范数的随机方向所对应的跨栖息地减栖息地内差异，即可分离出读出条件化背景偏移（图5A）。在真实的有限前向传播中，重新配对这八个类别会使绝对背景间隔增益发生双重差分（DiD）为 $+0.05127\,[0.04250, 0.06016]$ 每单位通道L_2 剂量对应的 logit（半剂量：$+0.05097\,[0.04215, 0.05989]$）。将精确的对称乘积分解应用于一阶总 DiD（$|\tau| = \sqrt{E_{\text{incoh}}}\,\sqrt{\mathcal{R}_{\text{int}}}$（公式 (6)，第 S4 节）可在较大样本量下清晰解析出+0.05091两个分量：路径幅度项贡献 $+0.03608$，有符号组成项贡献 $+0.01484$，且幅度显著大于组成（配对差异 $+0.02124\,[0.01238, 0.03001]$）。因此，切换视觉类别对比不仅会调节相同背景扰动的幅度，还会调节其有符号相位合成。

提示条件下冻结的大语言模型中的符号反转与词汇交互。在语言模型中，我们冻结预训练权重和单个跨名称关系轴 v，并将 $0.5ab\,v$ 注入查询词元注意力模块输出，其中 a、$b \in \{-1, +1\}$ 编码两个实体的属性，而 $r \in \{-1, +1\}$ 指定查询角色。对于任意匹配的（a、b、r）元组，直接提示和别名提示接收完全相同的注入。将 m_{edit} 和 m_{clean} 分别记作有符号答案对数几率差（蓝色减红色，或紫色减橙色），每个名称对的有限选择对比定义为 ΔG：$=$ 均值$_{a、b、r}\big[\,(a-b)\,r\,(m_{\text{edit}} - m_{\text{clean}})\,\big]$。我们在52个零重叠名称对上执行析因设计，交叉考察提示呈现（直接提示与别名提示）和候选答案对（红–蓝与橙–紫）；这些名称对在两种分词器中均保持为单个词元（图5B–C；名称对之间的平均 ΔG 位于 10^{-3} logit中）：

- **Qwen2.5-1.5B：** 在固定的红–蓝 答案对下，将提示呈现方式从直接切换为别名，会使平均有限选择对比的符号从 $+6.540\,[5.066, 8.031]$ 反转为 $-1.756\,[-2.586, -0.878]$

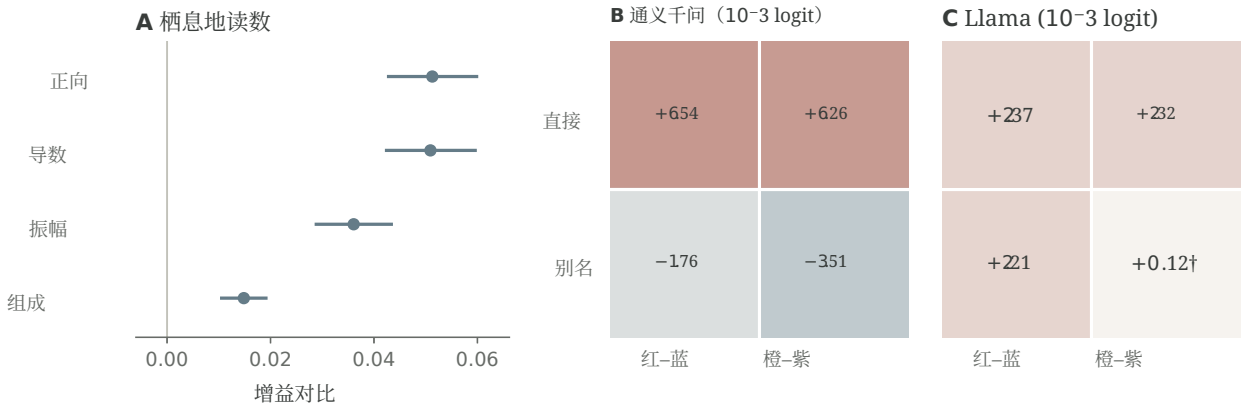


图5：固定因子编辑下的条件读出变化。A、在 1,024 张 CUB 图像（ResNet-18 layer3）上的视觉读出切换：将完全相同的八个鸟类类别从栖息地内对比重新配对为跨栖息地对比，产生正的背景减随机增益对比（$+0.05127$ 有限前向传播； $+0.05091$ 一阶导数），并对称地分解为振幅（$+0.03608$）和符号组成（$+0.01484$）贡献（95% 图像级置信区间）。**B–C、**语言 2×2 析因设计（52 个共享双分词器名称对；对于每个 (a, b, r) 元组，配对注入 $0.5ab\,v$ 在直接和别名之间完全相同；平均选择对比 ΔG 以 10^{-3} logit 表示，共享以零为中心的尺度）：在固定的红–蓝答案词下，将直接切换为别名会翻转冻结的 Qwen2.5 中的平均有限对比（$+6.54 \to -1.76$），而冻结的 Llama-3.2 则表现出强烈的呈现方式与 $\times$ 词汇交互作用（$+2.050$；匕首符号 † 标记唯一一个其 95% 置信区间包含零的单元格： $+0.12\,[-0.39, +0.64]$）。

（在橙–紫条件下，也同样从 $+6.263$ 变为 -3.513 ）。主效应和交互作用都清晰可辨：呈现方式 $-9.035\,[-10.873, -7.250]$、词汇 $+1.017\,[0.242, 1.770]$，以及交互作用$+1.480\,[0.106, 2.872]$。

- **Llama-3.2-1B：**在红–蓝条件下，两种呈现的结果均为正（$+2.368\,[1.311, 3.432]$ 在直接条件下，$+2.214\,[1.476, 2.915]$ 在别名条件下）；而在橙–紫条件下，平均对比值从 $+2.321$降至 $+0.117\,[-0.394, +0.641]$（跨过零点）。在 2×2 设计中，呈现主效应（$-1.179\,[-2.235, -0.151]$）、词汇主效应（$+1.072\,[0.646, 1.495]$）以及呈现与词汇的交互作用（$+2.050\,[1.192, 2.914]$）均达到统计显著。

关键的是，配对答案词之间的二选一候选准确率证实，通义千问 Qwen2.5 中的符号反转并不意味着别名提示下发生了任务崩溃：通义千问 Qwen2.5 的直接条件二选一准确率为 77.2%–78.8%，Llama-3.2 为 90.4%–91.1%；而在核心红–蓝别名条件下，二选一准确率仍然较高，通义千问 Qwen2.5 为 81.7%，Llama-3.2 为 79.8%（较低的 67.1% 和 63.2% 准确率仅出现在橙–紫别名条件下；见图 S14）。由于提示切换在有限的$0.5ab\,v$ 注入下同时改变了措辞、序列长度和查询位置，我们直接报告测得的有限反转，而不将其完全归因于单点局部雅可比矩阵。总之，两种模型均证实，在冻结权重下存在条件读出——其表现形式分别为通义千问 Qwen2.5 中明确的有限符号翻转，以及 Llama-3.2 中依赖词汇的增益调制。

4.6 条件读出的前瞻性门控与留出测试

第 4.3 至 4.5 节中的分析刻画了固定和切换任务中的有符号读出。现在，我们检验在真实前向传播过程中主动干预冻结单元组时，这一机制能否产生准确且出人意料的样本外 预测。在所有主要门控测试中，候选单元组仅依据开发集划分进行选择，在开放留出评估集划分前冻结，随后通过真实前向传播进行评估，仅对其编辑减清洁激活增量施加 $\lambda_0 = 0.1(10\%)$ 衰减（保持干净工作点不变；配套的物理输出缩放、六剂量曲线和漂移检查见 S11 节）。具体而言，ViT-B/16 和语言注意力头组针对给定编辑直接依据开发集局部门控斜率进行选择 κ_G ；而通义千问 Qwen2.5 和 Llama-3.2 中的末层 MLP 功能组，则依据在干净工作点计算的一阶局部分数进行选择。两种情况下，所有报告的评估指标均来自实际的有限 10% 衰减前向传播，并将同队列重放（固定权重、层和先前的测试输入，仅改变单元成员资格）与在新样本上进行的独立留出确认区分开来。正的 H_G (公式 (9)) 表示削弱该组会放大编辑对决策的影响（真正拮抗）； 负的 H_G 则表示促进。由于实验涵盖原始前瞻性测试、同权重后续成员替换和敏感性控制，我们明确标注了每项比较的来源和匹配协议。

冻结语言模型中的双向注意力头门控。 使用 16 组成开发名称对，以锁定由两个头部构成的促进组、拮抗组和强度匹配随机组，我们在 256 个零重叠确认名称对上评估真实前向门控（图 6A–B；单位为 10^{-3} logit）。两个冻结模型均满足全部五项联合标准——对于两个选定组，其表现均优于不匹配的 128 组随机强度库；在促进衰减下，$|\Delta m|$ 下降；在拮抗衰减下，$|\Delta m|$上升；并且优于强度匹配的随机对照：

- **Qwen2.5-1.5B：** 促进 $H = -0.132\,[-0.139, -0.125]$、拮抗 $H = +0.036\,[+0.029, +0.043]$，以及拮抗减去强度匹配的随机基线 $+0.093\,[+0.084, +0.103]$。
- **Llama-3.2-1B：** 促进 $H = -0.179\,[-0.187, -0.171]$、拮抗 $H = +0.153\,[+0.145, +0.161]$，以及拮抗减去强度匹配的随机基线 $+0.261\,[+0.247, +0.275]$。

分离直接提示词与别名提示词后，每个单元格中都保留了两种符号（图 6D）。此外，用真实的词汇关系词语替换取代合成隐藏向量编辑后，在更大的尺度上仍保留了双向符号结构（Qwen2.5：$-43.49/+17.61 \times 10^{-3}$ logit；Llama-3.2：$-58.66/+70.42 \times 10^{-3}$ logit；图 S15）。

功能性MLP分组解决了视觉和语言中的连续索引稀释问题。 在我们最初的前瞻性 ViT-B/16 测试中，候选组被限定为 48 个连续通道索引子块（384/3，072 个多层感知机单元，分布于第 6–11 块）。虽然最初基于 32张图像的试点选择得到了一个连续的第6块候选组，但在 512 张 CUB 图像上重新进行连续开发集选择后，发现了 0/24 个符合强度条件的正向连续候选组（图 S20），而原始的第6块连续对抗组在 1，024 张确认图像上的评估结果仍未得到解决（$H = +0.46\,[-0.34, +1.24] \times 10^{-5}$ logit，尽管促进组通过了 $-6.80\,[-8.13, -5.55]$，对抗组减随机组通过了 $+6.13\,[+4.48, +7.81]$，且配套物理缩放通过了全部三个标准，达到 $-9.88/+1.70/+4.67$）。

为确定连续索引本身是否削弱了拮抗效应，我们在原始 ImageNet ViT-B/16、通义千问 Qwen2.5-1.5B 和 Llama-3.2-1B 上进行了受控成员替换测试，采用统一的 1/8<style id='4'>-层多层感知机预算和 10% 增量衰减（图 6C；单位 10^{-5} logit）：

- 原始 ImageNet ViT-B/16（在 1,024 张图像上进行相同权重重放 **+**；在 264 张保留图像上进行留出确认）：固定确切的 ImageNet-1K 检查点、层、384 单元预算、四个鸟类类别边际，以及来自尚未解决的第6块连续测试的 1,024 张确认图像（功能成员资格在独立的 512 张图像开发集划分上拟合），仅将连续成员资格替换为 384 个分散的功能单元（均匀分布在全部八个索引分段中：49、44、50、45、55、48、49、44），即可将留出拮抗效应从 $+0.46$ 提高到 $+30.86\,[28.09, 33.68]$（相较于最佳连续分组的配对优势：$+30.41\,[27.77, 33.06]$）。在独立于 512 张图像开发集划分和 1,024 张图像重放划分的 264 张图像保留集上（也用于 S11 节中的经 CUB 微调的 ViT-B/32 比较），主要的第10块功能组（各分段数量为 41、56、58、46、41、53、46、43）达到 $H = +22.23\,[17.84, 26.57]$，显著优于最佳连续分组（$-2.39\,[-4.63, -0.30]$；配对 $+24.62\,[20.21, 29.04]$）和 ViT 强度匹配的随机对照（-11.13；配对 $+33.36\,[28.75, 38.01]$）；而第6块达到 $+29.68\,[24.25, 35.44]$（比最佳连续分组高 $+28.57\,[23.58, 33.70]$；比强度匹配对照高 $+32.27\,[27.34, 37.46]$）。
- **冻结的 Qwen2.5（第 28 层，1，120/8，960）和 Llama-3.2（第 16 层，1，024/8，192）在 256 个额外共享名称对上：** 功能性多层感知机组分别达到 $H = +13.74\,[12.54, 15.02]$（Qwen2.5）和 $+6.46\,[5.83, 7.10]$（Llama-3.2），分别比最佳连续区块高出 $+8.88\,[8.00, 9.80]$ 和 $+5.71\,[5.12, 6.32]$。由于粗略的 64 档强度分层在 Qwen2.5 中仍留下 16.5% 的平均开发集强度差距（控制后，配对优势从 $+8.47$ 缩小），我们评估了 32. ±1% 强度匹配的随机对照（实际开发集强度为功能组的 99.07%–100.98%；在 Qwen2.5 中作为事后敏感性分析进行，并在 Llama-3.2 中于留出集评估前锁定）。在两个模型中，功能性选择相较于严格强度匹配对照组仍保持明显的正向优势：Qwen2.5 中为 $+2.22\,[1.93, 2.54]$（重放时为 $+2.72$），Llama-3.2 中为 $+4.49\,[3.94, 5.05]$（重放时为 $+4.92$）；同时，组方向与编辑方向的交互作用均为正（Qwen2.5 中为 $+0.66\,[0.38, 0.97]$；Llama-3.2 中为 $+2.77\,[2.15, 3.43]$）。

同一单位、同次前向传播中的交叉读出符号翻转。最后，我们在原始 ImageNet ViT-B/16（第10块，264 张保留图像；图 6E，个单元 10^{-5} logit）中检验有符号读出的条件性所导致的最显著因果结果。我们在开发集划分上锁定两个单任务功能组，每组包含384 个多层感知机单元：A组根据类别间隔A（美洲白骨顶减去燕雀）进行选定，B组根据类别间隔B（黑天鹅减去环颈松鸡）进行选定。对于每张留出图像、每种纹理编辑和每个门控组，两个类别间隔都从完全相同的输出对数几率向量中读出（先对每张图像的四种纹理编辑取平均，再对264 张留出图像取平均）：

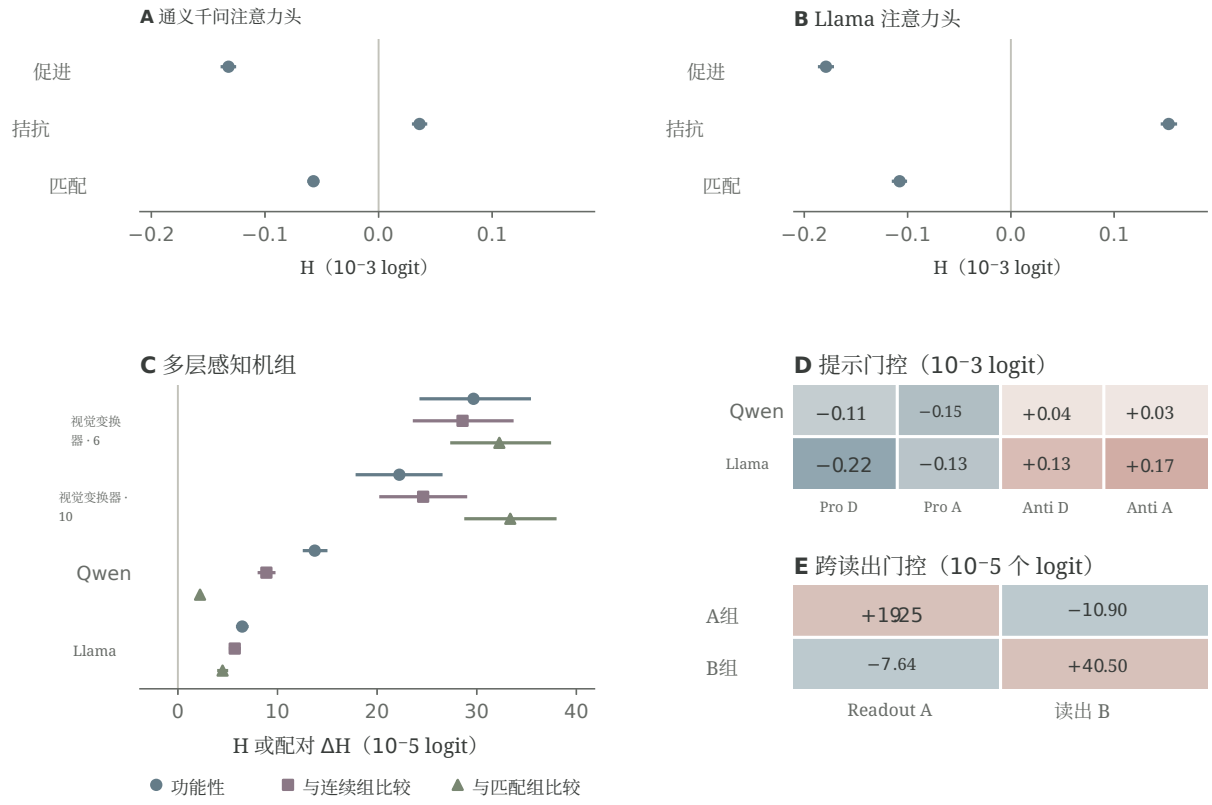


图 6：前瞻性门控与同一前向传播跨读出符号反转。A–B，冻结的 Qwen2.5-1.5B 和 Llama-3.2-1B 中的双向注意力头门控（256 个留出名称对，10% 编辑增量衰减，10^{-3} logit）：削弱促进头会降低 $|\Delta m|$ ($H < 0$)，而削弱拮抗头会放大 $|\Delta m|$ ($H > 0$)，并超过强度匹配的随机头。C，原始 ImageNet ViT-B/16 中的功能性多层感知机组（1/8层预算，10^{-5} logit）及其相对于最佳连续区块和强度匹配对照组的配对优势（264 张保留图像；ViT 匹配协议；第6块的 1,024 张图像重放在 S11 节中得到 $+30.41$，相对于连续区块），以及 Qwen2.5/Llama-3.2 中的结果（256 个额外共享名称对，严格的 ±1% 匹配：Qwen 中为事后敏感性分析，Llama 中为前置冻结）。D，按直接（D）和别名（A）提示词分层的双向注意力头门控（10^{-3} logit）。E，ImageNet ViT-B/16 第10块中留出均值的同一前向传播跨读出矩阵（264 张保留图像，10^{-5} logit）：对于每个组，两个间隔均从同一门控输出中读出；衰减 A 组会增加其对读出 A 的平均效应（$+19.25$），并降低其对读出 B 的平均效应（-10.90），而 B 组则呈现相反模式（-7.64 相对于 $+40.50$），从而产生 $+78.30\,[64.73, 91.92]$ 的四单元配对组别与读出交互作用（填充色调编码符号，打印条目给出幅度）。

- 减弱冻结的**A组**会增加读出A上的平均绝对背景编辑效应（$H = +19.25\,[8.66, 29.72]$），同时降低读出B上的该效应（$H = -10.90\,[-15.86, -6.24]$；就单张图像而言，$97/264$张图像呈现读出A正向偏移与读出B负向偏移同时发生的情况）。

- 减弱冻结的**B组** 呈现相反的模式：降低对 Readout A 的平均效应（$H = -7.64\,[-11.99, -3.34]$），同时提高对读出 B 的平均效应（$H = +40.50\,[28.96, 51.37]$；$115/264$ 图像分别表现出 Readout A 同时负向偏移、Readout B 正向偏移）。

综合 A组和 B组的全部四个单元格，得到配对组别与读出的交互作用，数值为 $+78.30\,[64.73, 91.92] \times 10^{-5}$ logit（在更大尺度上也得到重复验证，数值为 $+352.59\,[319.71, 386.80]$，见经 CUB 微调的 ViT-B/；图 S）。由于两个类别间隔均是针对同一张固定图像、背景编辑和冻结的单元组，从完全相同的门控前向传播中提取的，因此该实验直接证实了单元作用对读出的依赖性（补充了第 4.2 节中的编码证据）：一组单元是促进还是拮抗某个因子，并非单元固有的标签，而只由该组单元与目标读出共同决定。

4.7 结构化剪枝中通道恢复的定量预测

现在，我们将干预从衰减拮抗单元反转为在固定线性后继下恢复单个被移除通道，以检验定理 2 （FWL 三因子恒等式）在跨越 ResNet-50、ViT-S/16 和 Swin-T 的10 个切割位置上对4，096 张留出的 ImageNet 图像是否成立。在每个切割位置，我们冻结保留的通道子集 S 以及基于开发集激活进行的最优最小二乘权重折叠，在留出的图像上逐一恢复每个被省略的候选通道，并将测得的线性后继均方误差降幅与两个冻结的开发集预测量进行比较：仅使用条件残差方差 v_j （表示不可替代性），或使用纳入固定下游读出耦合的完整 FWL 分数（公式 (10)）。

跨架构排序确认：覆盖 10 个切分点。 纳入固定的下游读出耦合后，在 10 个切分点中的每一个切分点上，图 7A、C；完整表格见 表 S8）均改善了留出候选恢复排序，斯皮尔曼秩相关达到 0。865–1。000，覆盖全部 10 个切分点；在视觉 Transformer 的六个切分点上达到 0。975–0。999：

- ResNet-50（四个瓶颈切分点 layer1.0–第4层4.0卷积2）：仅方差相关性 0.867、0.420、0.569、0.564 提升至 FWL 相关性 1.000、0.997、0.994、0.865。关键的是，仅消融第 2 阶段中的带符号交叉项，就会使相关性从 0.997 骤降至 0.330（图 7B，上），直接证实了有符号读出合成，而非单独的无符号能量，决定了候选恢复价值。
- **视觉变换器-S/16（早期、中间层、后期 GELU 后 `fc2` 输入；每个有614 个候选）：** 相关性从 $0.919 \rightarrow 0.999$ ($+0.080$）、 $0.935 \rightarrow 0.996$ ($+0.061$）和 $0.948 \rightarrow 0.981$ ($+0.032$）上升。
- **Swin-T（早期、中间层、后期 GELU 后`fc2` 输入；** $307/614/614$ **个候选）：** 相关性从$0.728 \rightarrow 0.995$ ($+0.267$）、 $0.842 \rightarrow 0.992$ ($+0.150$）和 $0.867 \rightarrow 0.975$ ($+0.108$）上升。在中间剪切点，FWL 将顶部候选选择从通道$\#653$ （仅依据方差选择，恢复了最优增益的 85.3% ）反转为通道 $\#826$，后者达到了留出最优值的 100.0%（图 7B，下方）。

从 64 到 4,096 张图像的嵌套样本量变化轨迹表明，在每个深度和模型中，FWL 排名增益都保持稳定（图 7C）。

局部均方误差恢复与类别间隔误差解耦。 最后，我们考察在 40% 通道移除（图 7D）条件下，局部剪枝恢复如何传递至中间 MLP 块处最终干净的第一名减第二类类别间隔。在直接通道切片（无 OLS 折叠）下，将通道选择统计量从 GELU 前移至 GELU 后，会降低两种架构中的归一化局部 `fc2`输出均方误差，两种架构均如此（ViT-S/16：$0.1724 \rightarrow 0.1071$；Swin-T：$0.4376 \rightarrow 0.3784$）。然而，ViT-S/16 中最终类别边际平均绝对误差大幅下降（后置减前置：$-0.1756\,[-0.1929, -0.1583]$ logit），而在 Swin-T 中则在统计上保持平稳（$+0.0009\,[-0.0120, +0.0130]$ logit）。逐图像对称一阶分解 $|J_{\text{down}}\delta h|$ 在我们的读出框架内直接解释了这种解耦（图 7D，右）。在 ViT-S/16 中，将选择移至 GELU 后会同时降低扰动幅度（-0.1106logit）和梯度对齐程度（-0.0351 logit）。在 Swin-T 中，较小的局部扰动会使预测的间隔误差降低-0.0266 logit，但残余扰动与下游间隔梯度的对齐程度更高，又会使误差增加 $+0.0310$ logit，几乎抵消了局部均方误差的收益。当两个选择器都与普通最小二乘法权重折叠配合使用时，从 GELU 前选择改为 GELU 后选择会显著降低两个模型中的真实类别间隔平均绝对误差：Swin-T 降低 $-0.0306\,[-0.0370, -0.0241]$ logit，ViT-S/16 降低 $-0.1094\,[-0.1181, -0.1004]$ logit。因此，第 4.7 节的核心启示与全文主旨如出一辙：即使知道某个通道的条件残差方差，若要预测恢复该通道的局部函数能够恢复多少，仍需了解固定的下游读出如何组合其有符号贡献。

5 讨论、范围与机制边界

机制解释的实用性取决于其边界是否清晰。我们的大样本实验不仅证实了视觉与语言架构中的共享编码与条件有符号读出，还精准指出了朴素概括失效的四种情形。

5.1 分组尺度与架构非普适性

在我们的实验中，出现了两种截然不同的聚合效应。在冻结语言模型中（第 4.4 节），将粗粒度结构路径（M 和 J）中的相互对立的最后一层 MLP 单元合并，会掩盖关系特定抵消；在总导数相同的情况下，将功能单元分开，会使通义千问2.5 中的 ΔR 加深 $+0.0753$，并使 Llama-3.2 中的 $+0.1227$ 加深。

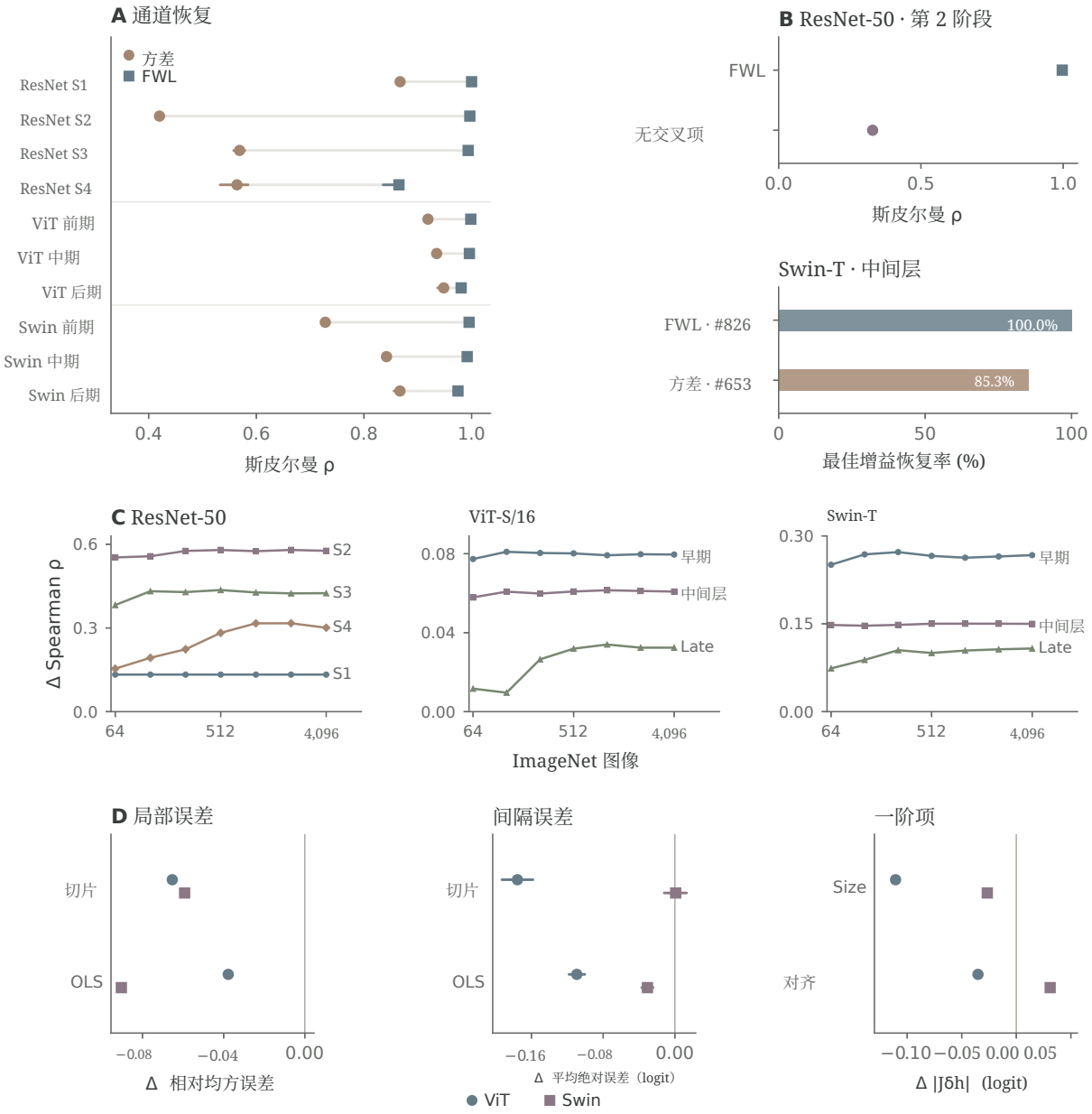


图 7：结构化剪枝中单通道恢复的定量预测（4，096 张 ImageNet 图像）。A, 在 ResNet-50、ViT-S/16 和 Swin-T 的全部 10 个切割中进行的留出候选恢复排名相关系数：加入固定下游读出耦合（FWL，方块）始终优于仅使用条件残差方差（圆点），整体达到 0。865–1。000，在六个 ViT/Swin 剪枝切分中达到 0。975–0。999（97。5%/98.75% 家族校正置信区间）。**B,** ResNet-50 第 2 阶段中的带符号交叉项消融（$0.997 \to 0.330$，顶部），以及 Swin-T 中期多层感知机中最佳候选从按方差选定的 #653（85。3%）切换为按 FWL 选定的 #826（100。0%，底部）。**C,** 在嵌套 ImageNet 子集（64 至 4，096 张图像）中，FWL 减方差的斯皮尔曼增益的稳定性。**D,** 在直接切片下，中间层从 GELU 前到 GELU 后的相对局部均方误差、类别边际平均绝对误差，以及对称一阶扰动幅度相对于梯度对齐分解的变化（这解释了为什么 Swin-T 的 -0.0266 幅度下降与 $+0.0310$ 对齐增加在直接切片下相互抵消；而在 OLS 折叠下，两个模型中 GELU 后选择相较于 GELU 前选择都会降低边际平均绝对误差）。

在原始 ImageNet ViT-B/16 中（第 4.6 节），将多层感知机神经元分组成连续的坐标索引块，会削弱留出对抗性对 $+0.46 \times 10^{-5}$ logit 的影响；在权重、层、预算和图像完全相同的情况下，以分散的功能单元取代连续索引，则会释放 $+30.86 \times 10^{-5}$ logit。综合来看，这些发现表明

有符号拮抗沿着分布于坐标索引上的功能组合进行组织，而非存在于连续区块或粗略的层级分组之内。

同样重要的是，干扰的极性并非架构固有的口号。尽管 ViT-B/16 表现出强烈的路径抵消（$R = 0.284$），Swin-T 则表现出净建设性纹理干扰（$R = 1.113$，随机基线 0.992）；在 ViT-B/16 中，测得的纹理方向和随机方向都强烈抵消，而纹理相较于随机方向是否存在额外抵消仍未确定（$+0.006\,[-0.004, +0.017]$ 在 1,024 张图像上）。切割是增强还是抵消某个因子，取决于训练权重、切割位置、任务读出和分组尺度的联合几何，而非卷积网络与 Transformer 之间的二元区别。

5.2 相对抵消 与分布下的绝对敏感性相比 布偏移

在分布偏移下，无量纲干扰比 $\mathcal{R}_{\mathrm{int}}$ 与绝对输出敏感度会显著分离。我们考察了 Oxford-IIIT Pets ResNet-18 在七个工作点上的表现——干净基线，以及强度为 1、3 和 5 的高斯噪声与高斯模糊——并在 444 张确认图像中固定干净样本的 top-1 类别、头部权重符号划分和供体编辑（完整剂量曲线及在 128 张图像队列上的重复实验见第 S13 节）。随着腐蚀严重程度上升，分类准确率从 87.6%（干净）骤降至 34.7%（最高噪声）和 25.9%（最高模糊），而背景和前景的双组比例均急剧降至 1 以下：等权重的最高严重程度背景偏移为 $-0.300\,[-0.347, -0.254]$（噪声 -0.313，模糊 -0.287；在 128 张图像队列上的重复结果为 -0.296 ），前景双组比例 $R^{(2)}$ 则从 1.089 降至 0.777 （噪声）和 0.897 （模糊）。

最清楚地说明为什么 $\mathcal{R}_{\mathrm{int}}$ 必须结合绝对响应能量来解读的，是最严重的模糊条件：背景 $R^{(2)}$ 从 0.906 降至 0.619（组间相对抵消更强），然而净相干背景响应能量却从 0.00351 升至 0.00559（图 S37）。将精确的双组恒等式 $R^{(2)} = 1 + bc$ （第 S4 节）应用于等权重的最严重背景偏移，可将 -0.300 的下降分解为：约 83% 来自正权重组与负权重组响应向量之间更强的角度对立（-0.2495；平均余弦值从 -0.140 降至 -0.452），17% 来自更接近的范数平衡（-0.0503；平衡度从 0.676 升至 0.829）；在 128 张图像的批次上也得到了高度一致的结果（-0.2527 和 -0.0432；图 S36）。因此，相对抵消增大并不必然意味着残余敏感性降低，也不必然意味着在腐蚀条件下鲁棒性更高。

5.3 单切局部线性化与全网络剪枝

对于具有固定线性后继、半正定输出度量和最优最小二乘权重折叠的单个 I 型切分，定理 2 是精确的；在所有 10 个测试切分中，它都提升了留出候选恢复排序（4,096 张 ImageNet 图像；第 4.7 节）。超出这一单切机制后，仍有两个明确的边界。第一，在检验 Swin-T 的九个移位空间窗口是否能改善中间块普通最小二乘补偿时，实际窗口分层优于等容量随机分层（均方误差分别为 0.017068 和 0.017425；差异为 $+0.000357\,[0.000347, 0.000367]$），但全局位置共享普通最小二乘取得了最低的留出误差（0.016910，与实际窗口相比差异为 -0.000158 ；第 S12 节），这反映了冻结校准样本下的偏差-方差权衡，以及窗口与空间坐标之间的协变关系。第二，单切直接切片本身无法解释六个架构家族中历史上的全网络多层剪枝差异（表 S10；3,200 张历史 ImageNet 图像，其中 257 张与我们的本地队列重合）：相较于 L_2，无标签剪枝带来的 40% 增益从 ViT-S/16 中的 $+24.56$ 个百分点变为 Swin-T 中的 -10.69 个百分点；任务条件化排序则取决于比较基线（比 Fisher 高 $+4.66$ 个百分点，比 Wanda 低 -7.56 个百分点）。如何将固定读出 FWL 理论扩展至耦合多层非线性压缩，仍是一个开放问题。

5.4 源坐标反演作为发现工具

我们在 S14 节进行的大样本边界检验，证实了公理 2 和定理 3 的两个方面。在发现方面，采用共享轴残差化的基于源坐标的反演（Shu [46] 的公式 15）为定位候选区域敏感通道子集提供了真实且可复现的信号：在 1,020 张符合条件的 CUB 确认图像中，去共享化残差组在全部四个级联端点上均优于强度匹配的随机对照——7×7 空间均方根（$+0.02812$）、GAP响应（$+0.01096$）、无符号通道权重和（$+0.00362$），以及匹配组减去不匹配组的绝对间隔恢复幅度（先计算每组内的有符号总和，再取恢复后的间隔变化的绝对值，并减去强度匹配的随机对照： $+0.00199\,[0.00026, 0.00373]$；去共享化结果与完整映射之间的差异 $+0.00122\,[-0.00102, +0.00345]$ 仍未有定论；图 S39）。对在 32 张图像的开发集划分上校准的有符号分组预测（冻结投影斜率为 0.03187）进行评估后发现，在 1,020 张留出图像中，每张图像的组级有符号皮尔逊相关系数均值较小但为正（ $+0.0485\,[0.0114, 0.0859]$ ）。同样，在 444 张 Pets 图像中，将 LAC 加权从 $\nabla_z \odot z$ 改为 ∇_z，可稳定地使前景能量占比提高 $+9.43$ 个百分点（56.62% $\rightarrow$66.06%；图 S39）。

然而，在边界方面，基于源坐标的反演无法取代原生双重敏感性。在 444 张 Pets 图像中，协方差加权的原生梯度 $\Psi = \Sigma_X \nabla_x m_q$ 对有符号候选像素编辑的排序，其斯皮尔曼相关系数为 0.9936；相比之下，全场维纳反演仅达到 $0.0139\,[-0.0272, +0.0556]$（平均归一化交换缺陷为 $\|\Delta_{\mathrm{comm}}\|_2/\|\Psi\|_2 = 1.0001\,[0.9992, 1.0011]$；图 S39）。LAC 与原生双组 $R^{(2)}$ 之间的图像级秩相关较弱（前景为 0.114，背景为 0.088，且跨越零点）。此外，跨算子首个共享轴投影证实，不同线性反演器呈现出不相容的有符号干涉结构：LAC $\to$LAC 得到 $R_b = 0.500\,[0.435, 0.578]$（通过了符号翻转、置换和错误类别零假设检验），而维纳 $\to$ 维纳则呈现强烈的建设性效应（Pets 中为 33.93，CUB 中为 25.58）；将 LAC 权重投影到同一图像的维纳首轴上，得到 $R_b = 1.282\,[1.207, 1.361]$，尽管无符号轴余弦相似度很高（匹配组为 0.585，不匹配组为 0.102；图 S38）。因此，基于源坐标的反演最适合作为生成假设的显微镜，其揭示的结构线索随后应通过原生方法进行检验。

5.5 局限性与开放方向

四项局限性限定了我们的结论，并指向具体的后续工作。第一，全部实验都研究参数冻结的训练后推理；共享编码与条件有符号抵消如何在优化动态、预训练数据规模扩展或微调过程中形成，仍有待研究。第二，我们的受控因子侧重于视觉中的空间背景编辑，以及语言中的结构化实体绑定关系；将原生不相交块和功能门控测试扩展到更丰富的语义组合与生成式多词元展开，是自然的后续方向。第三，功能性MLP分组是在 ViT-B/16、Qwen2.5 和 Llama-3.2 中执行的；将同样经过开发选择的功能门控应用于中间非线性卷积层，仍有待直接测试。第四，正如第 5.3 节所述，将单切分 FWL 恢复推广到非线性交互下的多层联合压缩，仍是一项开放的理论与工程挑战。

6 结论

我们探讨了一个关于训练后深度神经网络中推理的基础性问题：内部表征如何能够在不相交的单元之间广泛保留输入信息，同时生成针对当前任务高度选择性的输出？在视觉分类器和冻结语言模型中，我们的原始–对偶框架与大样本原生实验通过两种相互耦合的机制解决了这一问题。深度网络无需在上游消除上下文或关系因素，便能表现出选择性：多个不相交的表征单元可以共同保留同一因子的可识别结构（**共享编码**），而下游通道与非线路径则以依赖读出的幅度与符号组合这些单元的贡献（**条件符号读出**）。

将固定任务分解与变化任务切换视为同一局部综合定律的两个投影（$\tau_q = \sum_\alpha t_\alpha$、$|\tau_q| = \sqrt{E_{\mathrm{incoh}}}\sqrt{\mathcal{R}_{\mathrm{int}}}$），从而统一子卷积神经网络中的静态类别头背景消除、Transformer 中的多路径格拉姆几何与分组尺度稀释，以及冻结语言模型中的提示条件调制或符号反转。最重要的是，该框架经受住了真实干预下具有风险的样本外测试：削弱开发阶段选定的拮抗单元，会放大某一因子对未见输入的决策影响；在单次 ViT-B/16 前向传播期间削弱同一冻结单元，会增大背景编辑效应在一个类别间隔上的平均幅度，同时减小其在另一个类别间隔上的平均幅度（对于第二个冻结组则呈现相反模式；四单元配对交互效应$+78.30 \times 10^{-5}$ logit）；将相同的读出耦合投影到固定线性后继上，则可预测 10 次切割和 4、096 张 ImageNet 图像中的单通道剪枝恢复。结合针对分组尺度、分布偏移、多层剪枝和源坐标反演所明确界定的边界，这些结果为深度网络如何编码并选择性使用信息提供了一个有依据且具预测性的解释。

代码、数据和补充材料。所有主要表格及全部86 张非重复补充图版（涵盖所有确认队列、早期/中间层/后期剪枝操作表以及历史探索性对照）均直接嵌入在第S1–S14 节中，作为 图 S1–图 S40 （仅省略了 图 7C 的一张独立全宽重复图）。冻结的确认摘要、分析代码和绘图脚本可在以下网址公开获取：
`https://github.com/k-Jayus/How-Do-Deep-Networks-Encode-and-Selectively-Use-Information-`。

参考文献

[1] Julius Adebayo、Justin Gilmer、Michael Muelly、Ian Goodfellow、Moritz Hardt 和 Been Kim。显著性图的合理性检查。载于神经信息处理系统进展，第31卷，2018年。[2] Marco Ancona、Enea Ceolini、Cengiz Öztireli 和 Markus Gross。探讨如何更好地理解深度神经网络的基于梯度的归因方法。载于国际学习表征会议，2018年。

[3] David Bau、Bolei Zhou、Aditya Khosla、Aude Oliva 和 Antonio Torralba。网络解剖：量化深层视觉表征的可解释性。载于*IEEE*计算机视觉与模式识别会议论文集，第3319–3327页，2017年。[4] Yonatan Belinkov。探测分类器：前景、局限与进展。计算语言学，48(1)：207–219，2022年。[5] Joschka Braun、Carsten Eickhoff、David Krueger、Seyed Ali Bahrainian 和 Dmitrii Krasheninnikov。理解语言模型中引导向量的（不）可靠性。*arXiv*预印本 *arXiv:2505.22637*，2025年。[6] Arthur Conmy、Augustine Mavor-Parker、Aengus Lynch、Stefan Heimersheim 和 Adrià Garriga-Alonso。迈向机制可解释性的自动化电路发现。载于神经信息处理系统进展，第36卷，第16318–16352页，2023年。[7] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大规模分层图像数据库。载于*IEEE*计算机视觉与模式识别会议论文集，第248–255页，2009年。[8] Alexey Dosovitskiy 和 Thomas Brox。利用卷积网络逆向视觉表征。载于*IEEE*计算机视觉与模式识别会议论文集，第4829–4837页，2016年。[9] Nelson Elhage、Tristan Hume、Catherine Olsson、Nicholas Schiefer、Tom Henighan、Shauna Kravec、Zac Hatfield-Dodds、Robert Lasenby、Dawn Drain、Carol Chen、Roger Grosse、Sam McCandlish、Jared Kaplan、Dario Amodei、Martin Wattenberg 和 Christopher Olah。叠加的玩具模型。*Transformer Circuits Thread*，2022年。[10] Jiahai Feng 和 Jacob Steinhardt。语言模型如何在上下文中绑定实体？载于国际学习表征会议，2024年。[11] Elias Frantar 和 Dan Alistarh。SparseGPT：大规模语言模型可以通过一次性剪枝实现精确剪枝。载于第*40*届国际机器学习会议论文集，第202卷，第10323–10337页，2023年。[12] Ragnar Frisch 和 Frederick V. Waugh。与个体趋势相比的部分时间回归。《计量经济学》，1(4)：387–401，1933年。[13] Atticus Geiger、Hanson Lu、Thomas Icard 和 Christopher Potts。神经网络的因果抽象。载于神经信息处理系统进展，第34卷，第9574–9586页，2021年。[14] Robert Geirhos、Jörn-Henrik J acobsen、Claudio Michaelis、Richard Zemel、Wieland Brendel、Matthias Bethge 和 Felix A. Wichmann。深度神经网络中的捷径学习。《自然·机器智能》，2(11)：665–673，2020年。[15] Mor Geva、Jasmijn Bastings、Katja Filippova 和 Amir Globerson。剖析自回归语言模型对事实关联的回忆。载于*2023*年自然语言处理实证方法会议论文集，第12216–12235页，2023年。[16] Nicholas Goldowsky-Dill、Chris MacLeod、Lucas Sato 和 Aryaman Arora。利用路径修补定位模型行为。*arXiv*预印本 *arXiv:2304.05969*，2023年。[17] Zhiren Gong、He Lu、Tiantong Wang、Yichi Zhang、Yixin Wang、Zihao Zeng、Ming Xiao、Chau Yuen 和 Wei Yang Bryan Lim。条件共消融：恢复变换器电路中的自修复备份。*arXiv*预印本 *arXiv:2607.01940*，2026年。[18] Babak Hassibi 和 David G. Stork。网络剪枝的二阶导数：最优脑外科医生。载于神经信息处理系统进展，第5卷，第164–171页，1993年。[19] Stefan Haufe、Frank Meinecke、Kai Görgen、Sven Dähne、John-Dylan Haynes、Benjamin Blankertz 和 Felix Bießmann。关于多变量神经影像学中线性模型权重向量的解释。《神经影像》，87：96–110，2014年。[20] Evan Hernandez、Arnab Sen Sharma、Tal Haklay、Kevin Meng、Martin Wattenberg、Jacob Andreas、Yonatan Belinkov 和 David Bau。变换器语言模型中关系解码的线性性。载于国际学习表征会议，2024年。[21] Michael F. Hutchinson。拉普拉斯平滑样条影响矩阵迹的随机估计量。《统计通讯——模拟与计算》，18(3)：1059–1076，1989年。[22] Sarthak Jain 和 Byron C. Wallace。注意力不是解释。载于*2019*年计算语言学协会北美分会人类语言技术会议论文集，第3543–3556页，2019年。

[23] Been Kim、Martin Wattenberg、Justin Gilmer、Carrie Cai、James Wexler、Fernanda Viégas 和 Rory Sayres。超越特征归因的可解释性：使用概念激活向量进行定量测试（TCAV）。载于《第35届国际机器学习大会论文集》，第2668–2677页，2018年。[24] Yann LeCun、John S. Denker 和 Sara A. Solla。最优脑损伤。载于《神经信息处理系统进展》第2卷，第598–605页，1989年。[25] Jaeho Lee、Sejun Park、Sangwoo Mo、Sungsoo Ahn 和 Jinwoo Shin。面向基于幅值剪枝的层自适应稀疏性。载于《国际学习表征会议》，2021年。[26] Jiaqian Li、Yanshu Li 和 Kuan-Hao Huang。用于大型语言模型上下文感知推理时控制的引导向量场。arXiv预印本 arXiv:2602.01654，2026年。[27] Kenneth Li、Oam Patel、Fernanda Viégas、Hanspeter Pfister 和 Martin Wattenberg。推理时干预：从语言模型中引出真实回答。载于《神经信息处理系统进展》第36卷，第41451–41530页，2023年。[28] Michael C. Lovell。经济时间序列的季节调整与多元回归分析。《美国统计学会会刊》，58(304):993–1010，1963年。[29] Aravindh Mahendran 和 Andrea Vedaldi。通过反演深度图像表征来理解它们。载于《IEEE计算机视觉与模式识别会议论文集》，第5188–5196页，2015年。[30] Callum McDougall、Arthur Conmy、Cody Rushing、Thomas McGrath 和 Neel Nanda。复制抑制：全面理解语言模型注意力头中的一种模式。载于《第7届BlackboxNLP研讨会：分析与解释用于自然语言处理的神经网络论文集》，第337–357页，2024年。[31] Thomas McGrath、Matthew Rahtz、Janos Kramar、Vladimir Mikulik 和 Shane Legg。九头蛇效应：语言模型计算中的涌现式自我修复。arXiv预印本 arXiv:2307.15771，2023年。[32] Kevin Meng、David Bau、Alex Andonian 和 Yonatan Belinkov。在GPT中定位并编辑事实关联。载于《神经信息处理系统进展》第35卷，第17359–17372页，2022年。[33] Meta。Llama 3.2模型卡，2024年。网址 https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md。[34] John Morris、Volodymyr Kuleshov、Vitaly Shmatikov 和 Alexander Rush。文本嵌入揭示的信息几乎不亚于文本本身。载于《2023年自然语言处理实证方法会议论文集》，第12448–12460页，2023年。[35] Charlie Nash、Nate Kushman 和 Christopher K. I. Williams。使用自回归神经密度模型反演监督式表征。载于《第22届国际人工智能与统计学会议论文集》，第1620–1629页，2019年。[36] Chris Olah、Alexander Mordvintsev 和 Ludwig Schubert。特征可视化。《Distill》，2(11):e7，2017年。[37] Kiho Park、Yo Joong Choe 和 Victor Veitch。线性表征假设与大型语言模型的几何结构。载于《第41届国际机器学习大会论文集》，《机器学习研究论文集》第235卷，第39643–39666页，2024年。[38] Omkar M. Parkhi、Andrea Vedaldi、Andrew Zisserman 和 C. V. Jawahar。猫与狗。载于《IEEE计算机视觉与模式识别会议论文集》，第3498–3505页，2012年。[39] Qwen团队。Qwen2.5技术报告。arXiv预印本 arXiv:2412.15115，2024年。[40] Abhilasha Ravichander、Yonatan Belinkov 和 Eduard Hovy。探究探测范式：探测准确率是否意味着任务相关性？载于《第16届计算语言学协会欧洲分会会议论文集》，第3363–3377页，2021年。[41] David E. Rumelhart、Geoffrey E. Hinton 和 Ronald J. Williams。通过误差反向传播学习表征。《自然》，323(6088):533–536，1986年。[42] Ramprasaath R. Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedantam、Devi Parikh 和 Dhruv Batra。梯度加权类激活映射：通过基于梯度的定位解释深度网络。《IEEE国际计算机视觉会议论文集》，第618–626页，2017年。[43] Kaixiang Shu。伴随反演揭示CNN分类器中的全息叠加与相消干涉。arXiv预印本 arXiv:2604.27529，2026年。[44] Kaixiang Shu。用于结构化剪枝的统一响应几何。arXiv预印本 arXiv:2609.18239v2，2026年。[45] Kaixiang Shu。共享关系，特定选择：语言模型中的条件敏感性与竞争通路。预印本手稿，2026年。

[46] Kaixiang Shu。从原像搜索到基于源的特征反演。*arXiv*预印本 *arXiv:2607.12526*，2026年。[47] Karen Simonyan、Andrea Vedaldi 和 Andrew Zisserman。深入卷积网络：图像分类模型与显著图的可视化。*arXiv*预印本 *arXiv:1312.6034*，2013年。[48] Mingjie Sun、Zhuang Liu、Anna Bair 和 J. Zico Kolter。一种简单有效的大型语言模型剪枝方法。载于国际学习表征会议，2024年。[49] Mukund Sundararajan、Ankur Taly 和 Qiqi Yan。深度网络的公理化归因。载于《第*34*届国际机器学习大会论文集》，第70卷，第3319–3328页，2017年。[50] Daniel Tan、David Chanin、Aengus Lynch、Brooks Paige、Dimitrios Kanoulas、Adrià Garriga-Alonso 和 Robert Kirk。分析引导向量的泛化能力与可靠性。载于《神经信息处理系统进展》第37卷，第139179–139212页，2024年。[51] Max Torop、Aria Masoomi 和 Jennifer Dy。引导向量中的反向检测与控制。*arXiv*预印本 *arXiv:2608.02957*，2026年。[52] Alexander Matt Turner、Lisa Thiergart、Gavin Leech、David Udell、Juan J. Vazquez、Ulisse Mini 和 Monte MacDiarmid。使用激活工程引导语言模型。*arXiv*预印本 *arXiv:2308.10248*，2023年。[53] Jesse Vig、Sebastian Gehrmann、Yonatan Belinkov、Sharon Qian、Daniel Nevo、Yaron Singer 和 Stuart Shieber。使用因果中介分析研究语言模型中的性别偏见。载于《神经信息处理系统进展》第33卷，第12388–12401页，2020年。[54] Catherine Wah、Steve Branson、Peter Welinder、Pietro Perona 和 Serge Belongie。加州理工学院-加州大学圣迭戈分校鸟类200-2011数据集。技术报告 CNS-TR-2011-001，加州理工学院，2011年。[55] Kevin Ro Wang、Alexandre Variengien、Arthur Conmy、Buck Shlegeris 和 Jacob Steinhardt。真实环境中的可解释性：GPT-2 small中用于间接宾语识别的电路。载于国际学习表征会议，2023年。[56] Sarah Wiegreffe 和 Yuval Pinter。注意力并非解释。《2019年自然语言处理实证方法会议论文集》，第11–20页，2019年。[57] Norbert Wiener。平稳时间序列的外推、插值与平滑：工程应用。麻省理工学院出版社，1949年。[58] Kai Xiao、Logan Engstrom、Andrew Ilyas 和 Aleksander Madry。噪声还是信号：图像背景在物体识别中的作用。载于国际学习表征会议，2021年。[59] Matthew D. Zeiler 和 Rob Fergus。可视化并理解卷积网络。载于欧洲计算机视觉会议，第818–833页，2014年。[60] Fred Zhang 和 Neel Nanda。迈向语言模型激活修补的最佳实践：指标与方法。载于国际学习表征会议，2024年。[61] Andy Zou、Long Phan、Sarah Chen、James Campbell、Phillip Guo、Richard Ren、Alexander Pan、Xuwang Yin、Mantas Mazeika、Ann-Kathrin Dombrowski、Shashwat Goel、Nathaniel Li、Michael J. Byun、Zifan Wang、Alex Mallen、Steven Basart、Sanmi Koyejo、Dawn Song、Matt Fredrikson、J. Zico Kolter 和 Dan Hendrycks。表征工程：一种自上而下的人工智能透明度方法。*arXiv*预印本 *arXiv:2310.01405*，2023年。

补充材料

概述。 第 S1–S5 节提供统一无冲突符号表，以及支持 第 3节的完整数学证明、推导和解析模型。 第 S6–S14 节记录检查点来源、队列重复使用披露、数值精度审计，以及支持 第 4节和 第 5节的完整扩展表格和图表。全部分析代码、冻结的确认摘要和图表生成脚本均公开发布于 https://github.com/k-Jayus/How-Do-Deep-Networks-Encode-and-Selectively-Use-Information-。

S1 统一无冲突符号

表 S1 汇总了正文（第 3 节至5）和补充材料中使用的所有数学符号，并根据 公理 1，明确区分原表示对象与对偶敏感度对象。

表 S1：无冲突符号表。 原对象随坐标重缩放协变变换$H_{T,\ k} \mapsto d_k H_{T,\ k}$，而对偶对象则逆变换（$d_k^{-1}$）。

符号	空间 / 类型	定义
$x \in \mathcal{X}$	原输入	源输入（$L_0 \times C_0$：视觉中的像素、语言中的词元嵌入）；经验二阶矩 / 协方差矩阵 Σ_X。
$T,\ H_T(x) \in \mathcal{H}_T$	原始切割状态	内部切口 T和隐藏激活向量/张量 $H_T(x) = f_{\text{up}}(x)$。
$\delta H_T^{(f)} \in \mathcal{H}_T$	原始扰动	一阶推前 J_{u} $_{\text{p},T}(x)\delta x$ 或由受控因子 f引起的直接隐藏注入。
$\{\Pi_k\}_{k=1}^K,\ \mathcal{G}$	正交投影	完整的互相正交坐标/块划分（$\Pi_j\Pi_k = \delta_{jk}\Pi_k, \sum_k \Pi_k = I_{\mathcal{H}_T}$）；粗粒度分组 $\mathcal{G} = \{g_1, \ldots, g_M\}$。
$J_{\text{up}}(x)$ $J_{F\ \text{down}}(x;\ q)$	线性算子	在工作点 x处评估的上游和下游输出映射雅可比矩阵；* 表示伴随转置。
$m_q(x) \in \mathbb{R},\ u_{\text{out}}(q)$	标量裕度	由对比 $m_q(x) = u_{\text{out}}(q)^\top F_{\text{down}}(H_T(x); x,\ q)$选定的任务或查询条件标量裕度向量 $u_{\text{out}}(q)$。
$\ell_q(x) \in \mathcal{H}^*{}_T$	对偶读出余向量	下游敏感度 $\ell_q(x) := \nabla_H$ $_T\, m_q(x) = J_{F\ \text{down}}(x; q)^* u_{\text{out}}(q)$，按路径分解 $\ell_q = \sum_{\pi\in\mathcal{P}} \ell_q^{(\pi)}$.
$t_{k,\pi}(x, q; f) \in \mathbb{R}$	标量配对	分量 $\alpha = (k, \pi) \in \Omega(T)$ 的带符号贡献 $\left\langle \ell_q^{(\pi)}(x)\ \ \Pi_k \delta H_T^{(f)} \right\rangle$。
$\tau_q(x, f)$ E_{incoh} $\kappa^{\Omega(T)}$ 整数 $(f; x, q)$	净效应 / 能量 无量纲比率	净一阶敏感度 τ $_q = \sum_\alpha t_\alpha$ 和读出加权非相干能量 $E_{\text{incoh}} = \sum$ $_\alpha t_\alpha^2$. 相位相干干涉比 $\tau_q^2/E_{\text{incoh}} = 1 + \sum_{\alpha<\beta} 2t_\alpha t_\beta/E_{\text{incoh}}$ $(E_{\text{incoh}} > 0)$。
$G_{ij} = a_i^\top a_j$ $H_G(\lambda)$	$P \times P$半正定矩阵 门控端点	由路径读出梯度 $a_i = J_i^* u_q$ 构建的路径梯度格拉姆矩阵。 留出配对变化 $\lvert\Delta m_G(\lambda)\rvert - \lvert\Delta m(0)\rvert$ 在编辑增量衰减 λ下；局部偏导斜率 $\kappa_G = \partial\lvert\Delta m_G\rvert/\partial\lambda\rvert_{\lambda=0}$。
$A,\ W,\ \mathbf{H}_{\text{out}}$	剪枝切割对象	中心化校准激活 $A \in \mathbb{R}^{N\times C}$，固定线性后继 $Y = AW^\top$，半正定输出度量 $\mathbf{H}_{\text{out}} \succ 0$。
$r_{j\|S},\ v_j,\ d_{j\|S}$	FWL 创新项	条件残差 $r_{j\|S} = (I - P_S)a_j$，条件残差方差 $v_j =$ $-$ $\frac{1}{N}\lVert r_{j\|S}\rVert_2^2$，条件创新向量 $d_{j\|S} = A^\top r_{j\|S}/\lVert r_{j\|S}\rVert_2^2$ 与 $(d_{j\|S})_j = 1$。
$\nabla_x m_q,\ \ \Psi(x)$	对偶／原始输入	原生输入梯度 $\nabla_x m$ $_q \in \mathcal{X}^*$ 以及协方差加权的原始对偶项 $\Psi(x) := \Sigma_X \nabla_x m_q(x) \in \mathcal{X}$。
$\mathcal{A}_x,\ \ \mathcal{W}_{v\leftarrow u},\ \ \Delta_{\text{comm}}$	源反演对象	闭式源基础估计器 $\mathcal{A}_x$、精确线性最小均方误差估计器 $\mathcal{W}_{v\leftarrow u}$、交换缺陷向量 $\Delta_{\text{comm}}(x; q)$，以及标量归一化缺陷 $\lVert\Delta_{\text{comm}}\rVert_2/\lVert\Psi\rVert_2$。
$u_q = a_q b + r_q$	单位源映射	沿第一右奇异轴 b分解的单位归一化单通道反演映射，以及正交残差 r $_q \perp b$.

S2 定理1(i) 的证明、路径容斥原理与不变性

S2.1 完整切空间路径分解

在切割 T处，将下游模块线性化到干净工作点 x。第 4 节使用了三种典型分解：

1. **多块 Transformer 四路径展开（$D,\ A,\ M,\ J$）：** 对于单个线性化残差块（$I + M$）（$I + A$）$v = v + Av + Mv + MAv$，这四项分别对应旁路、仅注意力、仅多层感知机和混合路径。对于多个下游块，令 $\ell_q^{(0)}$表示完整的下游敏感度，$\ell_q^{\mathcal{A}}$表示在切空间中将所有指定的注意力写入钳制为零后的敏感度，$\ell_q^{\mathcal{M}}$表示将所有指定的多层感知机

写入钳制为零后的敏感度，而 $\ell_q^{\mathcal{AM}}$表示将两者都钳制为零后的敏感度。容斥原理得出精确的四路划分：

$$\ell_q^{(D)} = \ell_q^{\mathcal{AM}}, \qquad \ell_q^{(A)} = \ell_q^{\mathcal{M}} - \ell_q^{\mathcal{AM}},$$
$$\ell_q^{(M)} = \ell_q^{\mathcal{A}} - \ell_q^{\mathcal{AM}}, \qquad \ell_q^{(J)} = \ell_q^{(0)} - \ell_q^{\mathcal{A}} - \ell_q^{\mathcal{M}} + \ell_q^{\mathcal{AM}}. \tag{S1}$$

将这四个表达式逐项相加，即可还原 $\ell_q^{(D)} + \ell_q^{(A)} + \ell_q^{(M)} + \ell_q^{(J)} \equiv \ell_q^{(0)}$。请注意，对于多个区块，$\ell_q^{(J)}$汇集了全部至少经过一个指定注意力写入和一个指定多层感知机写入的路径，无论二者的时间顺序如何（先注意力后多层感知机，或先前置多层感知机后后期注意力）。将单层多层感知机（如第 4.4 节中所示）拆分为选定的功能单元 S、剩余的同层单元 R和旁路后，$\ell_q^{(M)}$与 $\ell_q^{(J)}$会进一步划分，同时保持 $\ell_q^{(0)}$与 τ_q 完全不变。

2. **ResNet 双块四路径展开（SS，SF，FS，FF）：**对于两个连续的残差块（$S_1 + F_1$）（$S_0 + F_0$）（其中 S_0 包含 layer4.0 中的 1×1 下采样投影），展开切线组成可得 $\ell_q = \ell_q^{(SS)} + \ell_q^{(SF)} + \ell^{(FS)}$ $_q$ $+ \ell_q^{(FF)}$；对于单个残差块（layer4.0 $\to$ layer4.1），$\ell_q = \ell_q^{(\text{skip})} + \ell_q^{(F_1)}$。

3. **注意力内部路由与值分解：**线性化 $P(x)V(x)$ 将 $\delta(PV) = (\delta P)V + P(\delta V)$，得到$\ell_q = \ell_q^{\text{route}} + \ell_q^{\text{val}}$。

S2.2 完全顶点割集、输入编辑特化与八条件语言协议

完全顶点割与部分分支。当 $\delta H_T^{(f)} = J_{\text{up},T}(x)\delta x$ 时，每个分量的贡献可改写为 $t_{k,\pi} = \langle \gamma_{k,\pi}(x;\ q)\ \delta x \rangle$，其输入空间余向量为 $\gamma_{k,\pi}(x;\ q) := J_{\text{up},T}(x)^* \Pi_k^* \ell_q^{(\pi)}(x) \in \mathcal{X}^*$。若 T 构成一个完全顶点割，截断了从 x 到 m_q 的每条计算路径，则链式法则保证 $\sum_{k,\pi} \gamma_{k,\pi}(x;q) \equiv \nabla_x m_q(x)$。

如果 T 只截断部分路径（例如单个残差子分支），那么 $\nabla_x m_q = \nabla$ $_x$ m_q，而直接切入注入恒等式 $\sum_{k,\ \pi} t_{k,\ \pi} = \langle \ell_q,\ \delta H_T \rangle$ 则无条件成立。 $^{(T)}_{x} m_q + \nabla^{(\text{b通过}\ T)}$

八条件语言协议。在 Shu 的八条件实体绑定协议中，[45], 条件由 $q = (a,\ b,\ r) \in \{-1, +1\}^3$ 索引，其中关系注入为 $\delta H_T(q) = ab\,v$，选择汇总权重为 $(ar - br)$。由于 $a^2 = b^2 = 1$，组合权重满足 $\chi_G(q) = (ar - br)$ $ab \equiv br - ar$，因此，平均有符号选择敏感性在各分量间恒等分解为 $\beta_G = \sum_{k,\ \pi} \mathbb{E}_q[(br - ar) t_{k,\ \pi}(x_q,\ q;\ v)]$。

S2.3 精确不变性性质的证明

定理 *1(i)* 的证明。设 $M = \bigoplus_{k=1}^{K} M_k$ 是任意一个与正交˜划分 $\{\Pi_k\}$($M\Pi_k = \Pi_k M$对所有 k) 可交换的可逆线性变换。在坐标变换 $H_T = MH_T$下，原始扰动按协变方式变换为 $\delta\tilde{H}_T = M\delta H_T$，而每个路径协向量均按逆变方式变换为 $\tilde{\ell}_q^{(\pi)} = (M^{-1})^* \ell_q^{(\pi)}$(等价地，在欧几里得坐标中为 $M^{-\top}\ell_q^{(\pi)}$)。因此，对每个 $\alpha = (k,\ \pi) \in \Omega(T)$，

$$\tilde{t}_{k,\pi} = \left\langle (M^{-1})^* \ell_q^{(\pi)},\ \Pi_k M \delta H_T \right\rangle = \left\langle \ell_q^{(\pi)},\ M^{-1} M \Pi_k \delta H_T \right\rangle = \left\langle \ell_q^{(\pi)},\ \Pi_k \delta H_T \right\rangle = t_{k,\pi}. \tag{S2}$$

由于每个单独项 t_α都具有不变性，$\tau_q = \sum_\alpha t_\alpha$、$E_{\text{incoh}} = \sum_\alpha t_\alpha^2$ 和 $\mathcal{R}^{\Omega(T)}{}_{\text{int}} = \tau_q^2 / E_{\text{incoh}}$在任意块可逆缩放和符号翻转下都严格不变。此外，由于 t_α 的定义将切向扰动直接与对偶协向量配对，而不除以基线激活范数$\|H_{T,\ k}\|_2$，因此它避免了零范数除法奇异性，同时忠实反映任何真实的下游敏感度（包括经由 LayerNorm 或 RMSNorm 的雅可比衰减）。 □

示例1（尺度不等时的 2×2 精确抵消）。设 $x = (b,\ f)^\top \in \mathbb{R}^2$ 分别表示背景 b和前景f，并将它们编码到两个尺度不等的隐藏单元 $h_1 = 10b + f$和 $h_2 = -b + f$中。在此线性读出 $m_q(x) = f = \frac{1}{11}h_1 + \frac{10}{11}h_2$下，分量输入协向量分别为 $\overline{\gamma_1} = \frac{1}{11}(\overline{10},\ 1)^\top$和 $\gamma_2 = \frac{1}{11}(-10,\ 10)^\top$。它们的总和 $\overline{\gamma_1} + \gamma_2 = (0,\ 1)^\top$ 恰好抵消背景（$\tau_b = 0$, $E_{\text{incoh}}(b) = \frac{200}{121} > 0$, $\mathcal{R}_{\text{int}}(b) = 0$），同时以建设性方式保留前景（$\tau_f = \overline{1}$, $E_{\text{incoh}}(f) = \frac{101}{121}$, $\mathcal{R}_{\text{int}}(f) = \frac{121}{101} > 1$）。

S3 定理 3 的证明与理想线性交换图

定理 3 的证明。设 $H_u = J_v H_v$ 是两个表示 H_v 和H_u之间的样本无关线性映射，它们的二阶矩矩阵 Σ_v和 $\Sigma_u = J_v \Sigma_v J_v^*$ 均为半正定。定义 $B := \Sigma_v^{1/2} J_v^*$。则 $\Sigma_v J_v^* = \Sigma_v^{1/2} B$ 和 $\Sigma_u = B^* B$。由于 $\ker(B^* B) = \ker(B)$，摩尔－彭若斯伪逆满足如下普适恒等式 $B(B^* B)^+ (B^* B) \equiv B$，而无需对 J_v 或 Σ_v作任何满秩假设。左乘 $\Sigma_v^{1/2}$并将等式两边作用于任意对偶协向量 ℓ_u，可得

$$\mathcal{W}_{v \leftarrow u}(\Sigma_u \ell_u) = \Sigma_v J_v^* (J_v \Sigma_v J_v^*)^+ \Sigma_u \ell_u = \Sigma_v^{1/2} B (B^* B)^+ (B^* B) \ell_u \equiv \Sigma_v^{1/2} B \ell_u = \Sigma_v (J_v^* \ell_u) = \Sigma_v \ell_v. \tag{S3}$$

□

理想线性交换图。 沿着 方程（11） 在样本无关线性链上迭代 $x = H_0 \to H_1 \to \cdots \to H_T$ 可得到一个两行交换图，通过层协方差度量连接双重反向传播（上行）与原始最小均方误差维纳回拉（下行） Σ_k:

$$\begin{array}{lccccc} \text{Dual space:} & \ell_T & \xrightarrow{J_T^*} & \ell_{T-1} & \xrightarrow{J_{T-1}^*} \cdots \xrightarrow{J_1^*} & \ell_0 = \nabla_x m_q \\ & \downarrow \Sigma_T & & \downarrow \Sigma_{T-1} & & \downarrow \Sigma_0 = \Sigma_X \\ \text{Primal space:} & U_T & \xrightarrow{\mathcal{W}} & U_{T-1} & \xrightarrow{\mathcal{W}} \cdots \xrightarrow{\mathcal{W}} & U_0 \equiv \Sigma_X \nabla_x m_q = \Psi(x) \end{array} \tag{S4}$$

三个推论。

1. 传递的复合回拉的基不变性： 对于任意可逆坐标变换 $H_T \mapsto M H_T, \Sigma_T \mapsto M \Sigma_T M^*$和 $\ell_q \mapsto (M^{-1})^* \ell_q$，因此 $\Sigma_T \ell_q \mapsto M \Sigma_T \ell_q$作为原始向量协变变换；当相同的线性估计器 $\mathcal{A}_x$ 协变地传递到新坐标中时（$\tilde{\mathcal{A}}_x = \mathcal{A}_x M^{-1}$），$\tilde{}$ $\tilde{}$ $\tilde{}$复合回拉 $\mathcal{A}_x(\Sigma_T \ell_q) = \mathcal{A}_x(\Sigma_T \ell_q)$ 严格不变。（相比之下，在缩放后的坐标中使用不变的各向同性标量惩罚项重新拟合岭估计器，会定义一个不同的算子。）

2. 一阶全局协方差加权： 原始向量 $\Psi(x) = \Sigma_X \nabla_x m_q(x)$ 使用全局经验输入协方差 Σ_X 对原生输入梯度进行加权（而不是在 x处执行精确的非线性流形投影）。

3. 上游交换缺陷与下游非线性的不等价性： 经验缺陷$\Delta_{\text{comm}}(x;\ q) = \mathcal{A}_x(\Sigma_T^{\text{global}} \ell_q(x)) - \Sigma_X J_{\text{up}}(x)^* \ell_q(x)$ 衡量了在 0 与 T之间偏离样本无关线性和估计精确性的上游程度；它在概念上不同于 T之外的下游 II/III 型非线性，且在没有专门消融实验（第 S14 节）的情况下，不能将其归因于单一误差来源。

定理 1（ii）的证明、双尺度桥接、格拉姆积分与精确分解

S4.1 拉德马赫期望基线与符号采样规则

定理 *1*（*ii*）的证明。设 $s_\alpha \in \{-1, +1\}$ 为相互独立的拉德马赫随机变量（$\mathbb{E}[s_\alpha] = 0$，$\mathbb{E}[s_\alpha s_\beta] = \delta_{\alpha\beta}$）。每当 $E_{\text{incoh}} = \sum_\alpha t_\alpha^2 > 0$时，请注意，随机化符号会使分母 $\sum_\alpha (s_\alpha t_\alpha)^2 = \sum_\alpha t_\alpha^2 = E_{\text{incoh}}$ 保持确定性。对分子取期望可得 Hutchinson 恒等式 [21]:

$$\mathbb{E}_s\left[\mathcal{R}_{\text{int}}^{\Omega(T)}(s \odot t)\right] = \frac{\mathbb{E}_s\left[\left(\sum_\alpha s_\alpha t_\alpha\right)^2\right]}{E_{\text{incoh}}} = \frac{\sum_\alpha t_\alpha^2 + \sum_{\alpha \neq \beta} \mathbb{E}[s_\alpha s_\beta]\, t_\alpha t_\beta}{E_{\text{incoh}}} = 1. \tag{S5}$$

对于符号的任意一次抽样 s，实际得到的比率可能高于或低于 1；精确的基线 1 是对独立符号翻转取期望的结果。

□

为测试纯单元干扰，仅随机化 s_k；为测试纯路径干扰，仅随机化 s_π；为测试单元–路径联合干扰，独立抽取$s_{k,\ \pi}$。输出对比符号反转（$A - B \leftrightarrow B - A$）会对每个 t_α 进行相同的取反，因此使 $\mathcal{R}_{\text{int}}$ 严格保持不变（作为区别于真正任务切换 $q \to q'$的对称性检查）。

S4.2 目标能量比（TER）分解与五种模式表

对于在度量 Σ 下具有相干和非相干区域能量的互不相交前景（M_{fg}）与背景（M_{bg}）支持集，相干前景能量占比 $\mathrm{TER}_{\mathrm{coh}} := E^{\Sigma}_{\mathrm{coh}}(M_{\mathrm{fg}})/(E^{\Sigma}_{\mathrm{coh}}(M_{\mathrm{fg}}) + E^{\Sigma}_{\mathrm{coh}}(M_{\mathrm{bg}}))$ 满足

$$\mathrm{TER}_{\mathrm{coh}} = \left[1 + \frac{\mathcal{R}^{\Sigma}_{\mathrm{int}}(M_{\mathrm{bg}})}{\mathcal{R}^{\Sigma}_{\mathrm{int}}(M_{\mathrm{fg}})} \cdot \frac{E^{\Sigma}_{\mathrm{incoh}}(M_{\mathrm{bg}})}{E^{\Sigma}_{\mathrm{incoh}}(M_{\mathrm{fg}})}\right]^{-1}. \tag{S6}$$

因此，$\mathrm{TER}_{\mathrm{coh}}$ 的增加本身无法证明背景抵消：对于两个向量 $v_1 = (1,1,0)$和 $v_2 = (1,0,1)$（第一个坐标为前景，后两个为背景），$\mathcal{R}_{\mathrm{int}}(M_{\mathrm{fg}}) = 2.0 > 1$而 $\mathcal{R}_{\mathrm{int}}(M_{\mathrm{bg}}) = 1.0$，仅通过增强前景来提高 $\mathrm{TER}_{\mathrm{coh}}$。表 S2 列举了全部五种诊断模式。

表 S2：区域前景能量集中度（$\mathrm{TER}_{\mathrm{coh}}$）的五种干涉模式。分别测量前景和背景上的 $\mathcal{R}_{\mathrm{int}}$，可以将抵消与增强区分开来。

Mode	$\mathcal{R}_{\mathrm{int}}(M_{\mathrm{bg}})$	$\mathcal{R}_{\mathrm{int}}(M_{\mathrm{fg}})$	$\mathrm{TER}_{\mathrm{coh}}$	物理机制
模式 1	< 1	≈ 1	增加	纯背景消除
模式 2	≈ 1	> 1	增加	纯前景增强（例如 $v_1 = (1, 1, 0)$、$v_2 = (1, 0, 1)$）
模式 3a	< 1	> 1	显著增加	联合背景抵消 + 前景增强（Pets 原生 第 4.3 节）
模式 3b	$< \mathcal{R}_{\mathrm{int}}(M_{\mathrm{fg}}) < 1$	< 1	增加	两个区域均被衰减，背景衰减更强（LAC 映射，S14 节）
模式 4	≈ 1	≈ 1	未改变	非相干／随机相位叠加

S4.3 双尺度桥接恒等式的证明

证明方程(7)。令 $\mathcal{G} = \{g_1, \ldots, g_M\}$划分$\Omega(T)$为若干互不相交的组，其总和为 $U_m = \sum_{\alpha \in g_m} t_\alpha$。展开细尺度分子和各组的平方，得到

$$\sum_{m=1}^{M} U_m^2 = \sum_{m=1}^{M} \left(\sum_{\alpha \in g_m} t_\alpha^2 + \sum_{\alpha \neq \beta \in g_m} t_\alpha t_\beta \right) = E_{\mathrm{incoh}} \left(1 + R_{\mathrm{within}}\right), \tag{S7}$$

$$\left(\sum_{m=1}^{M} U_m \right)^2 = \sum_{m=1}^{M} U_m^2 + \sum_{m \neq m'} U_m U_{m'} = E_{\mathrm{incoh}} \left(1 + R_{\mathrm{within}} + R_{\mathrm{between}}\right). \tag{S8}$$

将第二个恒等式除以 E_{incoh}，得到 $\mathcal{R}^{\mathrm{fine}}_{\mathrm{int}} = 1 + R_{\mathrm{within}} + R_{\mathrm{between}}$。将第二个恒等式除以第一个恒等式（只要 $1 + R_{\mathrm{within}} > 0$），得到

$$\mathcal{R}^{\mathrm{group}}_{\mathrm{int}} = \frac{\left(\sum_m U_m\right)^2}{\sum_m U_m^2} = \frac{1 + R_{\mathrm{within}} + R_{\mathrm{between}}}{1 + R_{\mathrm{within}}} \equiv \frac{\mathcal{R}^{\mathrm{fine}}_{\mathrm{int}}}{1 + R_{\mathrm{within}}} = 1 + \frac{R_{\mathrm{between}}}{1 + R_{\mathrm{within}}}. \tag{S9}$$

$\hat{}\,\hat{}$ 按符号划分权重时，精确为零的权重通道会保留在一个单独的零组 Γ_0 中，因此 $\Gamma = \hat{}\,\hat{}\,\hat{}\,\Gamma_+ + \Gamma_- + \Gamma_0$ 仍然精确。 □

S4.4 路径梯度格拉姆矩阵的精确高斯求积公式

令 $a_1, \ldots, a_P \in \mathbb{R}^D$表示切割 T处的 P路径读出梯度，其 Gram矩阵为 $G_{ij} = a_i^\top a_j$，特征分解为 $G = U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_P)U^\top$。对于各向同性高斯分布（等价地，均匀球面分布）的随机方向 v_{rand}（注意，仅指定协方差与 I_D 成正比，并不能唯一确定比值的期望），$\mathcal{R}_{\mathrm{int}}(v_{\mathrm{rand}})$关于 v_{rand}是零次齐次的，因此我们可以在标准高斯向量 $z \sim \mathcal{N}(0, I_D)$下计算期望。于是，路径贡献向量 $g := (a_1^\top z, \ldots, a_P^\top z)^\top$服从 $\mathcal{N}(0, G)$，并且 $y := U^\top g \sim \mathcal{N}(0, \mathrm{diag}(\lambda))$的各坐标相互独立，且具有 $\sum_i g_i^2 = \sum_j y_j^2$和 $\sum_i g_i = \sum_j c_j y_j$，其中$c_j := \mathbf{1}^\top U_{:j}$。利用拉普拉斯恒等式 $1/w = \int_0^\infty e^{-tw}\,dt$ $(w > 0)$和富比尼定理：

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_{\mathrm{int}}(v_{\mathrm{rand}})] &= \mathbb{E}\left[\frac{\left(\sum_j c_j y_j\right)^2}{\sum_k y_k^2}\right] = \int_0^\infty \mathbb{E}\left[\left(\sum_j c_j y_j\right)^2 \exp\Big(-t\sum_k y_k^2\Big)\right] dt \\
&= \sum_{j=1}^{P} c_j^2 \int_0^\infty \mathbb{E}\left[y_j^2 e^{-t y_j^2}\right] \prod_{k \neq j} \mathbb{E}\left[e^{-t y_k^2}\right] dt.
\end{aligned} \tag{S10}$$

计算初等高斯积分 $\mathbb{E}[e^{-ty_k^2}] = (1+2t\lambda_k)^{-1/2}$ 以及 $\mathbb{E}[y_j^2 e^{-ty_j^2}] = \lambda_j(1+2t\lambda_j)^{-3/2}$ 可得 公式（8）：

$$\mathbb{E}[\mathcal{R}_{\text{int}}(v_{\text{rand}})] = \sum_{j=1}^{P} (\mathbf{1}^\top U_{:j})^2 \int_0^\infty \frac{\lambda_j}{1+2t\lambda_j} \prod_{k=1}^{P} (1+2t\lambda_k)^{-1/2}\, dt. \tag{S11}$$

通过 256 点或 384 点高斯–勒让德求积进行评估，并在代入 $t = u/(1-u)$后得到。该总体积分与四种视觉骨干网络上的有限蒙特卡罗随机方向库高度吻合（在 ResNet-18 上，两者均预测相长干涉：解析积分为 1。1488，冻结的随机方向库为 1。1332，图像配对差异为 $+0.0156\,[0.0119, 0.0193]$；第 4.4 节，第 S9 节）。

S4.5 精确对称乘积分解

对于任意从状态 0 变为状态 1 的乘积 $Z = XY$，精确的对称沙普利/中点分解将$\Delta Z := Z_1 - Z_0$ 拆分为两个相加项，且余项为零:

$$\Delta Z \;\equiv\; \underbrace{\frac{Y_0+Y_1}{2}\,(X_1 - X_0)}_{\text{change due to } X} \;+\; \underbrace{\frac{X_0+X_1}{2}\,(Y_1 - Y_0)}_{\text{change due to } Y}. \tag{S12}$$

我们在本研究的三个部分应用 公式 (S12)：

1. **振幅与组成（第 4.5 节）：** 令 $|\tau| = \sqrt{E_{\text{incoh}}}\sqrt{\mathcal{R}_{\text{int}}}$，设置 $\overline{X} = \sqrt{E_{\text{incoh}}}$ 和$Y = \sqrt{\mathcal{R}_{\text{int}}}$，即可将一阶 CUB 读出切换双重差分（$+0.05091$）精确分解为振幅（$+0.03608$）和组成（$+0.01484$）。

2. **双组腐化几何（第 5.2 节，第 S13 节）：** 设 $P,\ N \in \mathbb{R}^4$ 分别表示源图像上四种供体编辑对应的正权重组和负权重组响应向量。双组干扰比满足以下精确恒等式

$$R^{(2)} = \frac{\|P+N\|_2^2}{\|P\|_2^2 + \|N\|_2^2} \equiv 1 + b\,c, \qquad b := \frac{2\|P\|_2\|N\|_2}{\|P\|_2^2+\|N\|_2^2} \in [0,1], \qquad c := \frac{P^\top N}{\|P\|_2\|N\|_2} \in [-1,1], \tag{S13}$$

其中 b 衡量范数平衡，c 衡量组余弦。将公式 (S12) 逐图应用于 bc，可将腐化偏移 $\Delta R^{(2)} = -0.2998$（444 幅图像）精确分解为角度变化（-0.2495）和范数平衡变化（-0.0503）。

3. **一阶剪枝裕度分解（第 4.7 节，第 S12 节）：** 记 $|J_{\text{down}}\delta h| = \|\delta h\|_2 \, \|J_{\text{down}}\|_2 \cdot |\text{余弦}(\delta h,\ J_{\text{down}}^*)|$，并将公式 (S12) 应用于扰动范数与梯度对齐，以揭示为何 GELU 前后切片在 Swin-T 中相互抵消（-0.0266 大小与 $+0.0310$ 对齐），而在 ViT-S/16 中相互增强（-0.1106和 -0.0351）。

S5 证明定理2的证明，统一二维模型、门控过冲与剪枝审计 t

S5.1 FWL 三因子通道恢复恒等式的证明

定理2的证明。设 $A \in \mathbb{R}^{N\times C}$ 为固定线性后继$Y = AW^\top \in \mathbb{R}^{N\times d_{\text{out}}}$ 和输出度量$\mathbf{H}_{\text{out}} \succeq 0$下 I 型切分处的中心化校准激活。对于任意保留子集 $S \subset \{1,\dots,C\}$ 及其补集$R = S^c$，最优最小二乘折叠使用 $B_S = A_S^+ A_R$ 和 $W'_{\text{keep}} = W_S + W_R B_S^\top$。由于 $A_S B_S = P_S A_R$其中 $P_S := A_S A_S^+$ 是到$\text{col}(A_S)$上的正交投影算子，折叠预测等于

$$\widehat{Y}_S = A_S(W_S + W_R B_S^\top)^\top = A_S W_S^\top + P_S A_R W_R^\top = P_S(A_S W_S^\top + A_R W_R^\top) \equiv P_S Y, \tag{S14}$$

得到重构损失 $L(S) = \frac{1}{N}\|(I-P_S)Y\|_{F,\,\mathbf{H}_{\text{out}}}^2$。现在恢复一个省略通道 $j \notin S$及其条件残差 $r_{j|S} = (I-P_S)a_j$。若 $v_j := \frac{1}{N}\|r_{j|S}\|_2^2 = 0$，则 $a_j \in \text{col}(A_S)$、$P_{S\cup\{j\}} = P_S$ 和 $L(S) - L(S\cup\{j\}) = 0$。当 $v_j > 0$时，格拉姆–施密特正交化给出秩一投影算子增量

$$P_{S\cup\{j\}} - P_S \;=\; \frac{r_{j|S} r_{j|S}^\top}{\|r_{j|S}\|_2^2}. \tag{S15}$$

根据 $P_{S\cup\{j\}} - P_S$ 与 $I - P_{S\cup\{j\}}$的正交性，Frisch–Waugh–Lovell（FWL）定理 [12, 28] 得出

$$L(S) - L(S\cup\{j\}) = \frac{1}{N}\left\|(P_{S\cup\{j\}} - P_S)Y\right\|_{F,\mathbf{H}_{\text{out}}}^2 = \frac{1}{N}\left\|r_{j|S} b_{j|S}^\top\right\|_{F,\mathbf{H}_{\text{out}}}^2 = v_j \left\|b_{j|S}\right\|_{\mathbf{H}_{\text{out}}}^2, \tag{S16}$$

其中 $b_{j|S} := Y^\top r_{j|S}/\|r_{j|S}\|_2^2 = WA^\top r_{j|S}/\|r_{j|S}\|_2^2 = Wd_{j|S} = \sum_k z_k$ 且 $z_k := W\Pi_k d_{j|S}$。每当 $\sum_k \|z_k\|_{\mathbf{H}_{\text{out}}}^2 > 0$，提取非相干输出能量 $\sum_k \|z_k\|_{\mathbf{H}_{\text{out}}}^2$ 即可得到精确的三因子恒等式 公式（10）：

$$L(S) - L(S \cup \{j\}) \equiv v_j \cdot \left(\sum_k \|z_k\|_{\mathbf{H}_{\text{out}}}^2\right) \cdot \mathcal{R}_{\text{int}}^{\mathbf{H}_{\text{out}}}(d_{j|S}), \tag{S17}$$

其中 $\mathcal{R}^{\mathbf{H}_{\text{out}}}{}_{\text{积分}}(d_{j|S}) := \|\sum_k z_k\|_{\mathbf{H}_{\text{out}}}^2 / \sum_k \|z_k\|_{\mathbf{H}_{\text{out}}}^2$。中心化会将常数截距吸收到偏置或批次归一化移位中；严格的零截距切分会以非中心化二阶矩替代中心化协方差。在数值上，快速 FWL 分数与显式最小二乘重新拟合结果一致，达到 $\leq 9.43 \times 10^{-18}$ 在早期/中间层视觉 Transformer 切分上，并且$\leq 1.005 \times 10^{-14}$ 在 4,096 张 ImageNet 图像的全部六个切分上均一致（第 4.1 节，表 S3）。 □

S5.2 统一二维解析玩具模型

令 s、t 为相互独立且均值为零的标量因子，其被编码到两个共享单元 $h_+ = s + t$ 和$h_- = s - t$中。考虑欧几里得度量$\mathbf{H}_{\text{out}} = 1$下的任务条件化线性读出 $Y_c = \eta_c s + t = \frac{\eta_c+1}{2} h_+ + \frac{\eta_c-1}{2} h_-$：

1. **共享编码主导性：** 第一个主轴 s 占解释方差比$_1 = \sigma_s^2/(\sigma_s^2 + \sigma_t^2) \to 1$，而$h_+$ 和 h_- 各自都携带 s。

2. **通过部分抵消实现正向因子效用：** 在 $\eta_c = 0$，s 完全抵消（$\mathcal{R}_{\text{int}}(s) = 0$），而 t则起到增强作用（$\mathcal{R}_{\text{int}}(t) = 2$）。在 $\eta_c = 0.5$（$Y = 0.75h_+ - 0.25h_-$），$+0.5$ $\mathcal{R}_{\text{int}}\ s = 0.5^2/\ 0.75^2 + 0.25^2 = 0.4 < 1$ 产生了明显的正净效应（ ），尽管其干扰比严格小于一（ ( ) ( ) ），这表明模型可以通过部分抵消利用正向因子。

3. **读出条件化符号反转与剪枝顺序翻转：** 反转符号(η_c)会同时翻转决策敏感度 $\partial Y_c/\partial s = \eta_c$ 和第一步 FWL 剪枝恢复顺序：

$$L(\emptyset) - L(\{h_\pm\}) = \frac{(\eta_c\sigma_s^2 \pm \sigma_t^2)^2}{\sigma_s^2 + \sigma_t^2}. \tag{S18}$$

因此，当 $\eta_c > 0$ 时，应保留 h_+ 这一较优单位；而当 $\eta_c < 0$ 时，h_- 变得更优，尽管上游激活协方差相同。

S5.3 线性化促进组符号保持与幅度缩减界限

在线性化模型中，总编辑效应为 $\tau(0) = \sum_\alpha t_\alpha \neq 0$，而削弱总贡献为 $t_G = \sum_{\alpha \in G} t_\alpha$ 的单位组 G 会使净效应变为 $\tau_G(\lambda) = \tau(0) - \lambda t_G$：

- 如果 G 是一个促进 组（$\text{sign}(t_G) = \text{sign}(\tau(0))$），那么 $\tau_G(\lambda)$ 保持其原始符号且不会穿过零点，当且仅当

$$0 < \lambda < \frac{|\tau(0)|}{|t_G|}, \tag{S19}$$

而较弱的幅度缩减条件 $|\tau_G(\lambda)| < |\tau(0)|$（该条件允许符号反转穿过零点，但不会越过 $-\tau(0)$，例如 $\tau(0) = 1$，$t_G = 3$，$\lambda = 0.5 \Rightarrow \tau_G = -0.5$）成立，当且仅当

$$0 < \lambda < \frac{2\,|\tau(0)|}{|t_G|}. \tag{S20}$$

对于超过 $2|\tau(0)|/|t_G|$（当 $|t_G| > 2|\tau(0)|$时可能出现），减弱一个促进组可能会翻转间隔符号并增加 $|\tau_G(\lambda)|$，因此我们在门控方案（第 4.6 节）中采用较小的主要剂量（$\lambda_0 = 0.1$）。

- 如果 G 是一个拮抗 组（$\text{sign}(t_G) = -\text{sign}(\tau(0))$），则线性化衰减会对所有 $\lambda > 0$严格增大 $|\tau_G(\lambda)| = |\tau(0)| + \lambda|t_G|$；然而，在真实的非线性网络中，有限剂量交互作用和样本间异质性可能破坏线性，因此，开发阶段选定的组能否预测 $\lambda_0 = 0.1$ 在留出输入上的有限门控符号及其幅度，是一项真正可证伪的实证检验。

S5.4 ViT-S/16 和 Swin-T 的剪枝接口审计

对 Shu 发布的结构化剪枝代码库进行审查后 [44] 确认，ViT-S/16 和 Swin-T 都会剪除 `mlp.fc1` 的输出行，并同时剪除 `mlp.fc2` 的输入列$W = W_{\mathrm{fc2}}$（因此，这两种切割都属于线性后继的 I 型切割）。然而，发布的无标签选择器在*GELU* 之前收集激活统计数据（`fc1` 输出），且未应用岭回归/普通最小二乘权重折叠，而定理 2 要求观测 *GELU* 之后的激活（即 `fc2`的直接输入），并与 OLS 折叠配对。我们在第 4.7 节和第 S12 节采用的六组受控方案，在早期、中期和后期模块中清晰地区分了 GELU 前与 GELU 后的选择，以及直接切片与 OLS 折叠。

S6 队列、来源、复用矩阵和精度审计

本节记录规范以及数 检查点来源、严格的队列分离和复用披露、统计重采样支持第 4.1 节的数值闭合检查。

检查点来源与队列复用。视觉实验使用经过任务训练的 Oxford-IIIT Pets ResNet-18 (resnet18_pets.pth)、CUB-200-2011 ResNet-18 (resnet18_cub200-1.pth) 和 ViT-B/32 (vit_b_32_cub200-1.pth)，以及根据实验指定的 ImageNet-1K 预训练检查点：用于 CUB 通路几何分析（图 4）的 ViT-B/16 (Torch Hub SHA-256 c867db91d3e12c6cbadabb610d73c24a546bf82d8c03a9fea34f43a712ddb0e9) 和 torchvision ResNet-50 (resnet50_imagenet-1.pth)，以及用于 ImageNet 剪枝分析（图 7）的 timm ResNet-50 (resnet50_a1_0)、ViT-S/16 和 Swin-T（图 4 中也使用同一 Swin-T 检查点）。语言实验使用冻结的预训练通义千问2.5-1.5B 和 Llama-3.2-1B 基础模型，搭配原始分词器和确定性 CUDA 配置；配对自助法重采样默认进行 10,000 次抽样，明确的例外包括：语言模型注意力头门控和 CUB 区域带符号皮尔逊相关进行 20,000 次抽样，ImageNet 剪枝进行 4,000 次抽样，以及 Pets 剂量级有限差分校准斯皮尔曼相关系数进行 1,000 次抽样。所有独立统计单位均为完整源图像或完整实体名称对；共享评估队列（视觉骨干网络共用的 1,024 张 CUB 图像；通义千问2.5 和 Llama-3.2 共用的 256 个确认名称对及另外 256 个 MLP 分组对；ViT-B/16 和 ViT-B/32 功能比较共用的 264 张 CUB 保留图像；以及全部 10 个剪枝切口共用的 4,096 张 ImageNet 图像，其中 257 张与包含 3,200 张图像的历史整体网络基准重叠）绝不会被累加为独立重复样本。图 S1 展示了 Pets、语言模型和历史先导研究中嵌套子集的样本量稳定性；表 S3 总结了核心数值闭合审计，表 S4 则汇总了涵盖所有研究终点的确认台账。

表 S3：最终确认队列中数值闭合、梯度与有限差分一致性及物理手术精度审计的代表性结果。样本量（n）在每一行中均有明确列出。

审计检查	模型／切割（队列规模）	测量值	单位／范围
切空间路径和闭合	CUB ResNet-18 layer3 / layer4.0 ($n = 1,\ 024$)	5.26×10^{-7} / 4.20×10^{-8}	最大绝对误差
切空间路径和闭合	Qwen2.5 / Llama-3.2 4 条路径（$n = 256$ 配对）	$\leq 2.78 \times 10^{-17}$	最大绝对误差
坐标缩放协变性检查	Pets ResNet-18 维纳输运（$n = 444$）	2.32×10^{-6}	最大相对 L_2
双尺度精确桥接误差	Pets ResNet-18（$n = 128,\ 896$ 个点）	4.44×10^{-16}	最大绝对误差
格拉姆求积（256 与 384 个节点）	R50 / Swin-T / ViT-B/16（$n = 1,\ 024$）	$\leq 1.16 \times 10^{-14}$	最大绝对值 $\mathcal{R}_{\mathrm{int}}$
有限差分与导数符号匹配	CUB ResNet-18 早期 / 后期（$n = 1,\ 024$）	98.27%/99.44%	分数（主要）
有限差分与导数符号匹配	ViT-B/16 Block 6（$n = 1,\ 024$，主要 / 一半）	99。988%/99.994%	分数
快速 FWL 与显式 OLS 重新拟合	ViT-S/16 和 Swin-T 中间 MLP（$n = 4,\ 096$）	$9.11/8.81 \times 10^{-18}$	最大绝对均方误差（$\leq 1.01 \times 10^{-14}$ 全部 6)
物理手术裕度检查	36 个设置 $\times 4,\ 096$ 张图像	147，456/147，456	通过裕度容差
后期 Swin 直接切片双精度浮点数重放	732 个 float32 均方误差异常（$n = 4,\ 096$）	6.22×10^{-15}	最大绝对均方误差（双精度浮点数）
完整神经网络与批处理多层感知器门控	通义千问 / Llama（每个 21，504 个单元）	精确匹配	条件级别 H

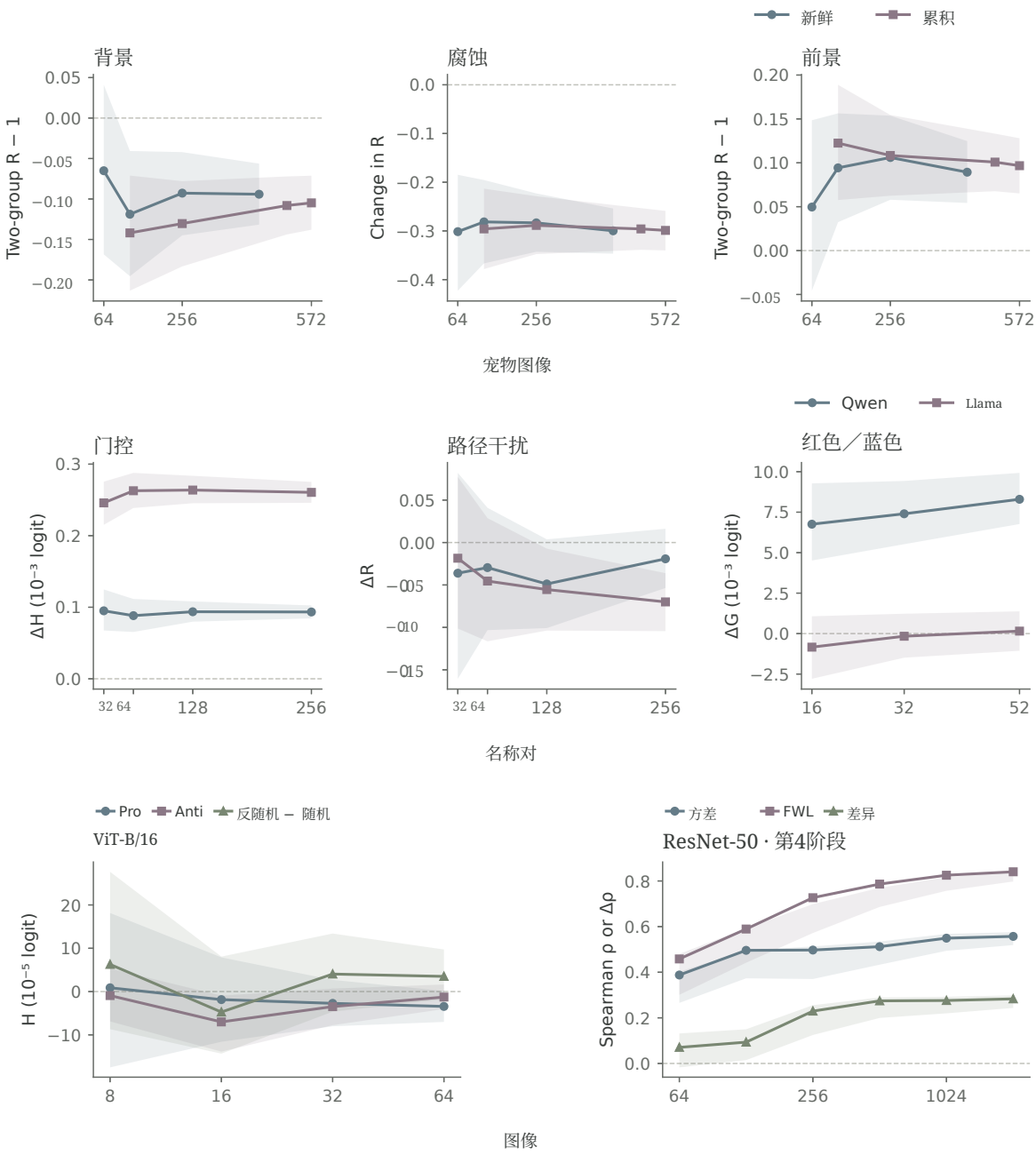


图 S1：Pets（上）、语言确认队列（中）和历史试点（下）的样本量稳定性。上：在 444 张新鲜 Pets 图像和累计 572 张图像的嵌套子集中，背景/前景双组干扰比的均值及最高腐蚀偏移（95% 逐点置信区间）。中：在通义千问2.5和Llama-3.2的嵌套名称对前缀中，拮抗组减匹配组头部门控、四路径关系减随机基线 ΔR，以及红色/蓝色直接减别名对比。下：历史试点样本数敏感性，分别来自原始的 64 张图像 CUB 视觉变换器-B/16 连续门控试点（左）和 2，048 张图像 ResNet-50 第4阶段剪枝试点（右）。

表 S4：汇总 **确认台账和范围边界**。 全部主要估计值均使用最终的大样本确认队列；探索性试点分别归档于补充章节 S6–S14。

问题（§）	主要确认结果（除非另有说明，均为 95% 置信区间）	范围与明确边界
跨……的共享编码不相交块（4.2）	CUB (1,024)：中心化区块匹配 $+0.555$– $+0.595$ (R18 L3，42.1% 能量)， $+0.397$– $+0.422$(L4.0，14.1%)，$+0.874$– $+0.884$ （ViT B6，49.3%；99.375% 置信区间）。Pets 定位色调（442）：联合 $p = 0.0042$（56.1%–60.9%）。大语言模型（256）：全部 4 个区块的配对内跨角色 AUC $= 1.000$ （余弦相似度 0.792–0.844）。	每个纹理对的跨图像 GAP 平均方向平均覆盖仅 1.52% 的能量；视觉证明了配对互补匹配位于预先固定的 16 维子空间内（而非内在维度或单区块身份解码）。
固定任务通道突触-论文（4.3）	Pets（444）：背景 $R^{(2)} = 0.906$ [0.868, 0.944] 对比前景 1.089 [1.054, 1.125]（128 批次：0.858 对比 1.122）；符号对比随机划分 -0.792 $[-0.821, -0.761]$；512 通道背景 $1 + 9.200 - 0.937 = 9.263$；248/444 张图像同时满足 $R^{(2)} < 1$和 $R^{(512)} > 1$（266/444 个具有正的组内和负的组间项）。	在固定干净样本的 top-1 类别、4 个供体编辑和头部符号的条件下分组；替代的 top1−top2 间隔给出背景 $R^{(2)} = 0.879$。
多路径几何 & 尺度 (4.4)	CUB (1,024) tex/随机基线 R: R18 $1.158/1.133$ (格拉姆 1.1488, 差异 $+0.0156$), R50 $1.754/1.377$, Swin-T $1.113/0.992$, 视觉变换器 $0.284/0.277$ (格拉姆 0.2779；截断后 J 轮次为 -0.396)。在 ViT/通义千问/Llama 中，两者均为 $A \times J$ （$-0.157/-0.262/-0.332$）和 $D \times M$ （$-0.061/-0.213/-0.108$）均为负值。拆分多层感知机使 ΔR 加深：通义千问增加 $+0.0753$，Llama 增加 $+0.1227$。	Swin-T > 1 反驳了 Transformer 普遍抵消的说法；ViT 纹理减去随机基线后为 $+0.006$ $[-0.004, +0.017]$；ViT 末端非-[CLS]精确 0（类型 0）。
固定因子，变化读出（4.5）	CUB（1,024）：8 个物种重新配对的双重差分 $+0.05127$ $[0.04250, 0.06016]$；一阶拆分为振幅 $+0.03608$ + 组成 $+0.01484$(差异 $+0.02124$)。大语言模型 (52 对配对，10^{-3} logit)：通义千问 RB 平均 ΔG翻转 $+6.540 \to -1.756$；Llama RB $+2.368 \to +2.214$ 与交互作用 $+2.050$ $[1.192, 2.914]$。	大语言模型中的有限 $0.5ab\,v$ 注入；提示切换与措辞相关，长度和查询位置；别名 RB 的二选一准确率仍然很高（81.7% 通义千问，79.8% Llama）。
前瞻性门控 & 读出翻转 (4.6)	大语言模型注意力头 (256, 10^{-3})：通义千问 正向 -0.132/ 反向 $+0.036$ ($+0.093$与随机基线对比)；Llama -0.179/ $+0.153$($+0.261$)。多层感知机功能组 (10^{-5})：ViT-B/16 B6 重放 $+30.86$($+30.41$ 与连续组比较）；B10 留出集 $+22.23$ （$+24.62$ 与连续组比较，$+33.36$ 与匹配组比较）；通义千问/Llama 以紧密匹配的 ± 1% 击败 $+2.22$ / $+4.49$。ViT B10 同次前向传播均值：A组 $+19.25/-10.90$ （97/264 张图像），B组 $-7.64/+40.50$ （115/264），四单元交互作用 $+78.30$。	通义千问 ± 1% 的匹配是事后敏感性分析；Llama ± 1% 已锁定前置留出集；ViT 264 保留集与 B/32 比较共享。
剪枝通道恢复（4.7）	ImageNet（4,096，10 次切割）：FWL 斯皮尔曼相关系数整体为 0.865–1.000（6 个 ViT/Swin 剪枝切分中为 0.975–0.999）；R50 第 2 阶段跃升 $0.420 \to 0.997$（无交叉项 0.330）；Swin 中层切换至最高位单位 #653 (85.3%) $\to$ #826 (100%)。	对于单切分固定线性后继 + 普通最小二乘法成立；在直接切片下 Swin 后置–前置 MAE 保持平坦（$+0.0009$），而在普通最小二乘法下，两者后置–前置下降（Swin -0.0306，视觉变换器 -0.1094）；全网络表 S10打开。
损坏与反演界限（5、S13–S14）	Pets 模糊5（444）：背景 $R^{(2)}$ 下降 $0.906 \to 0.619$，而相干背景能量上升 $0.00351 \to 0.00559$ （准确率 87.6% $\to$ 25.9）。Pets 维纳编辑排序 0.0139，相较于原生 Ψ 0.9936（平均归一化缺陷 $\|\Delta_{\text{comm}}\|_2/\|\Psi\|_2 = 1.0001$ $[0.9992, 1.0011]$）；CUB 去共享绝对间隔优势 $+0.00199$ $[0.00026, 0.00373]$，组平均符号 $r = +0.0485$。	模糊条件下，相对抵消与绝对敏感性相互独立；源反演能够发现候选区域，但无法取代原生读出。

S7 扩展共享编码分析

图 S2 和图 S3 从四项互补诊断扩展了第 4.2 节：（i）未中心化与均值中心化视觉匹配，以及截断特定的 16 维子空间能量覆盖率（ResNet-18 layer3 为 42.14%，layer4.0 为 14.09%，ViT-B/16 Block 6 为 49.26%）；（ii）比较末端全局平均池化处逐纹理对的跨图像平均方向（平均仅占留出响应能量的 1.52%，输出符号一致率为 54.86%；而逐图像原生方向导数的输出符号一致率为 73.47%，在不匹配纹理对下降至 50.50%，在重新分配的读出下则降至 49.77%）；（iii）弱 Pets 局部暖色/冷色调探针的统计功效递进：从原始的 64 张图像（37、40、39、32；联合 $p = 0.546$）扩展至 442 张新符合条件的图像（248、269、250、253；56.1%–60.9%，联合 $p = 0.0042$）；以及（iv）在独立的 63 张图像探索性 CUB 队列和 1,024 张图像确认性 CUB 队列中进行重复验证，并附上源图像视觉示例（图 S3）。

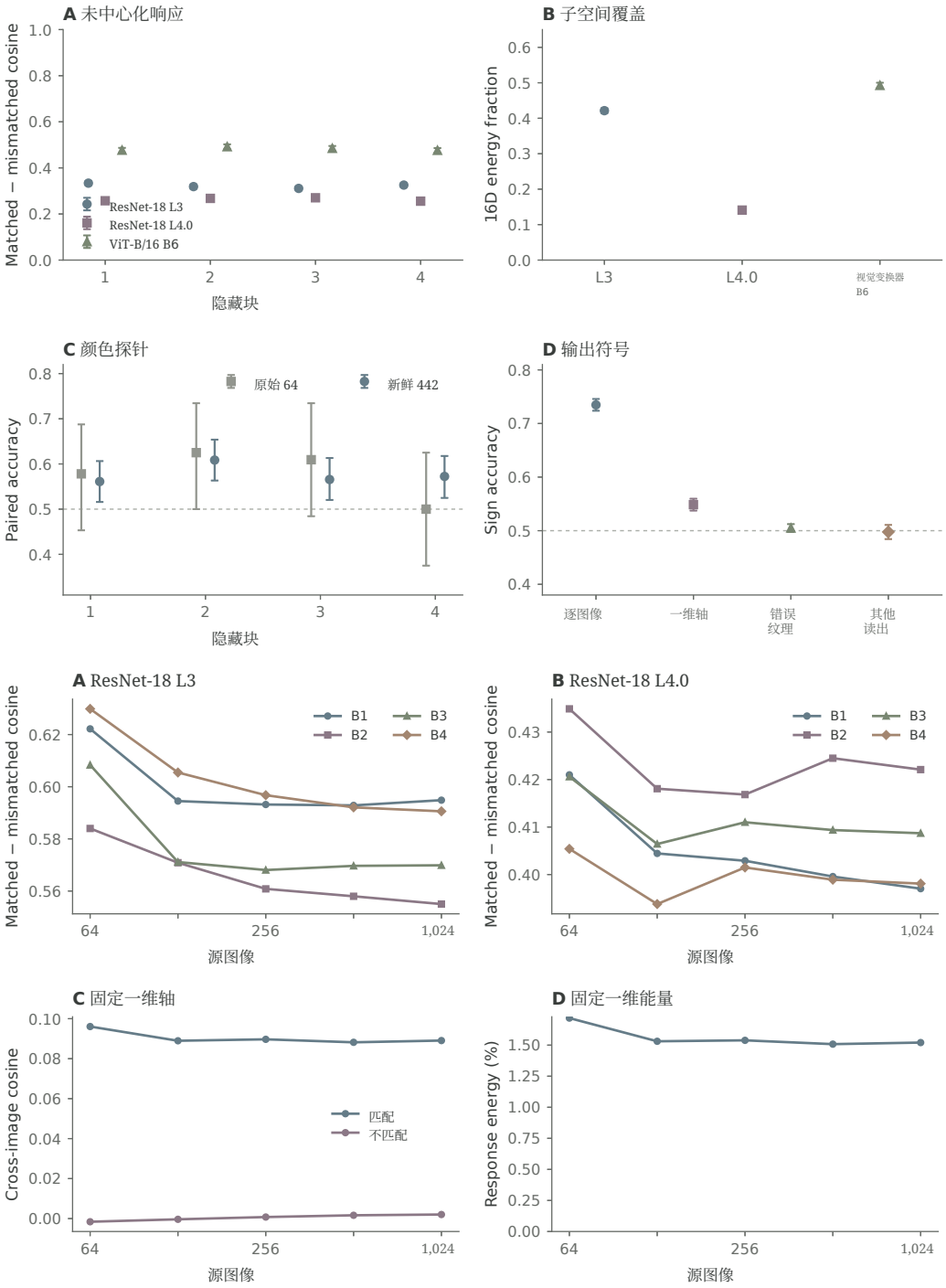


图 S2：视觉共享编码对照（上）与样本量稳定性（下）。上：未中心化块匹配（99。375% 置信区间）、切分特异性16维子空间能量分数、弱局部暖/冷 Pets 色调探针（$n = 64$ 与 $n = 442$），以及 CUB 输出符号预测对比（$n = 1,\ 024$）。下：用于在 ResNet-18 layer3/layer4.0 和末端全局平均池化中进行中心化块匹配、一维轴对齐及能量占比分析的嵌套 CUB 前缀。

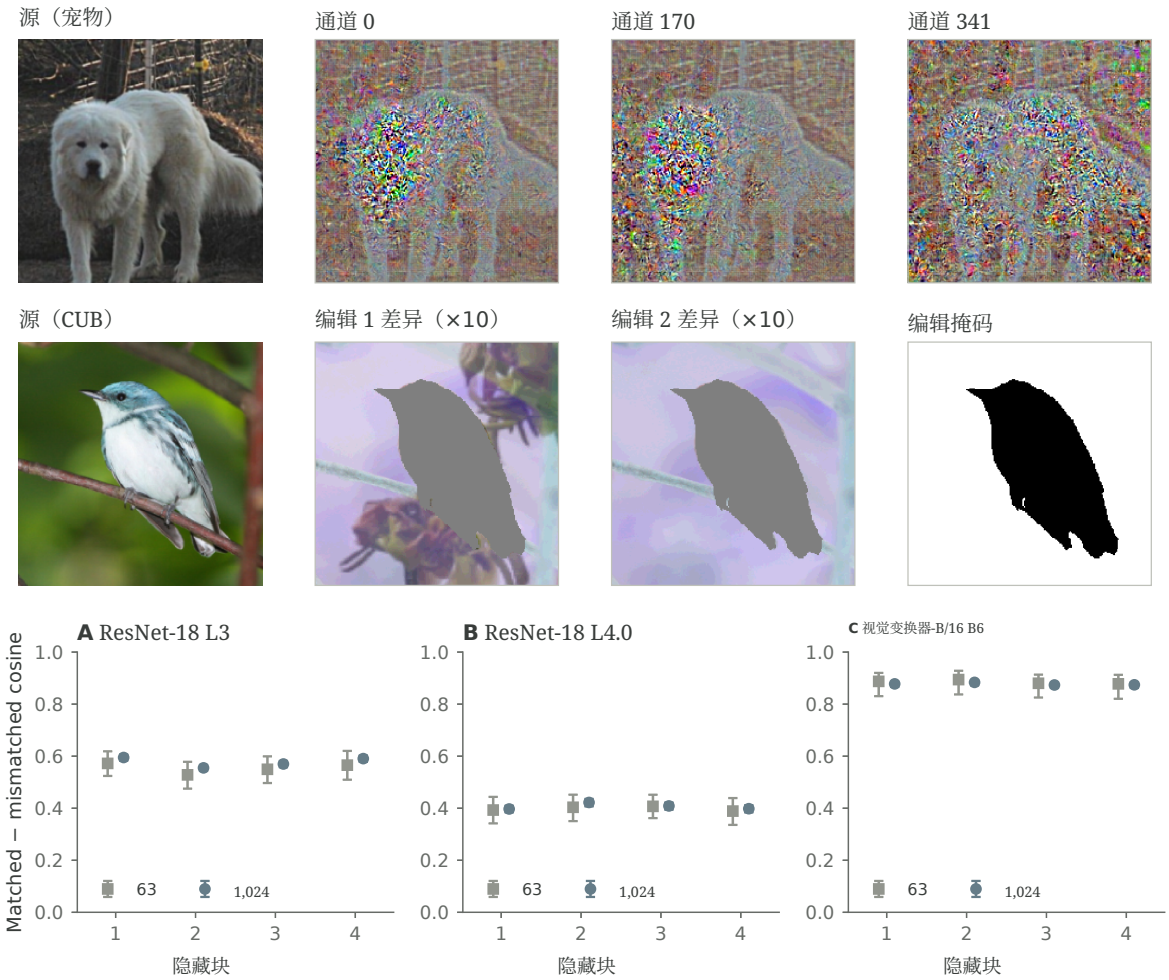
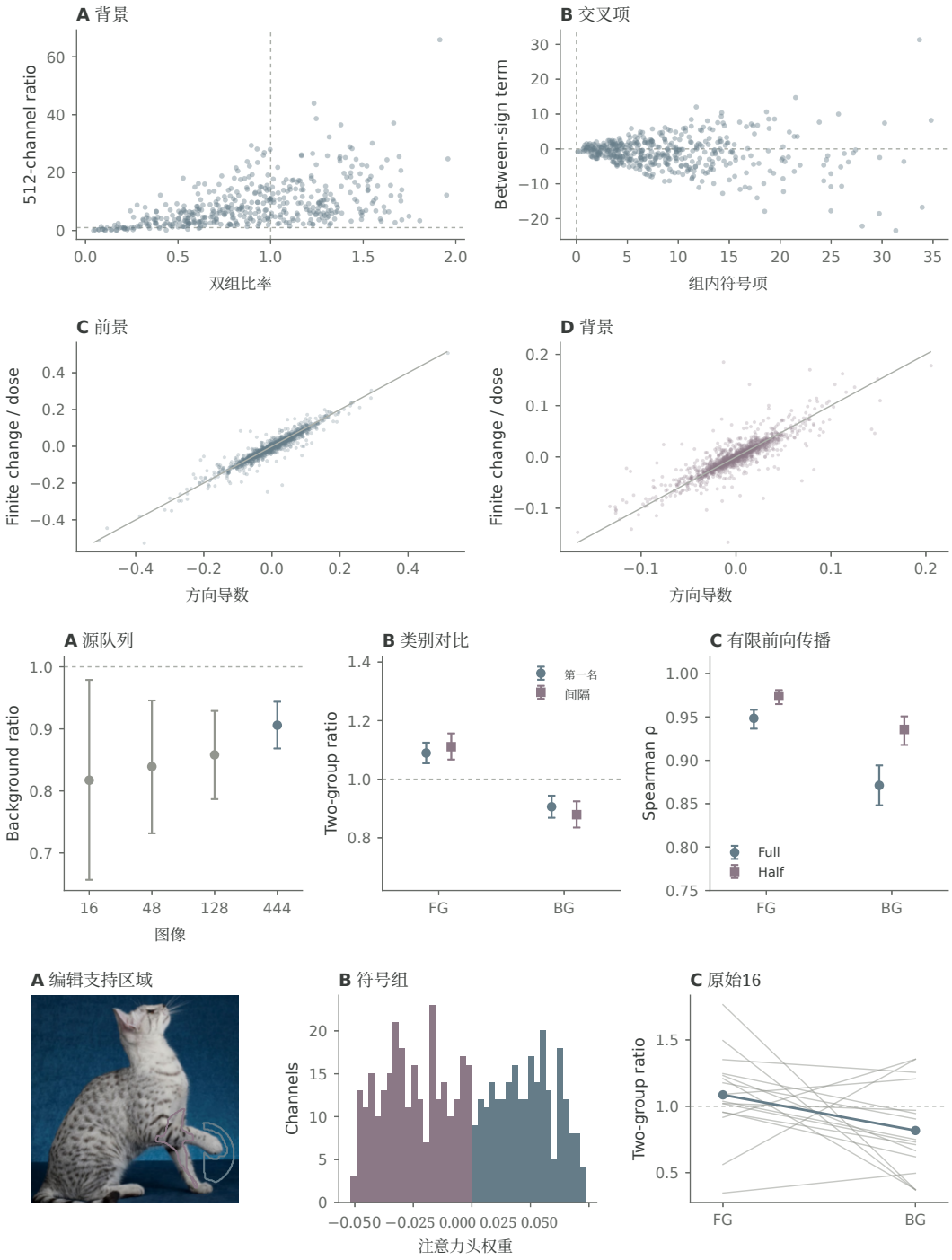


图 S3：视觉源字段（上）和独立队列复现（下）。上：冻结的 Pets 大白熊犬_图像_9的通道 0、170和 341 的 LAC 字段（以中灰为中心，并按显示用的第 99 百分位幅值缩放），随后是 CUB-09631, $10\times$ 针对两种不同的 5% 背景纹理编辑得到的放大有符号像素差异（以中灰为中心），证实前景变化严格为零；以及可编辑背景掩码。下：在独立的 63 张图像探索性 CUB 队列和 1、024 张图像确认性 CUB 队列之间比较中心化不相交区块匹配（99。375% 置信区间）。

S8 Pets 上的扩展双尺度通道合成

图S4 展示了支持第 4.3 节的完整队列沿革、替代读出对照、编辑支持区域和图像级验证。在各队列中，历史背景 $R^{(2)}_{\text{背景}}$ 值分别为 0。817（原始16张图像）、0。839（原始 64 张图像批次中剩余的 48 个源）和 0。834（合并后的原始 64 个源；不将其重复计为三次独立复现），随后是独立确认批次 $0.858\,[0.787, 0.929]$（128 个新源）和 $0.906\,[0.868, 0.944]$（444 个新源）。将读出改为固定的 Top-1减 Top-2间隔后，抵消仍然存在（$R^{(2)}_{\text{背景}} = 0.879$）；导数与有限差分的 Spearman 秩相关分别达到 0。871（RGB 剂量 0。125）和 0。936（半剂量 0。0625）。



图S4：图像级通道分解与有限剂量验证（顶部）、队列/读出对照（中部），以及带有原始16张图像发现集的编辑支持区域（底部）。顶部：逐图像双组与512通道干扰比及带符号交叉项（$n = 444$； $248/444$ 同时满足 $R^{(2)} < 1$ 和 $R^{(512)} > 1$），以及供体级方向导数与全剂量（0.125）有限逻辑值变化的比较。中部：队列来源（16、48、128、444）、干净的第一名与第一名和第二名之差的间隔读出，以及全剂量/半剂量导数相关性。底部：首个确认来源（埃及_Mau_9），包括前景/背景供体支持集、注意力头权重直方图（243个正值、269个负值），以及原始16个源之间的配对比率。

S9 扩展多路径几何、控制与证伪档案

本节归档了全部三种 Transformer 的完整六对归一化交叉项表（表 S5）以及全部十四张多路径诊断图版（图 S5 至 S10）：四种骨干网络的完整路径梯度格拉姆几何、信号与随机基线的六对交叉项（以及初始与最终复现）、ViT-B/16 的模块与深度重新分组及钳制前向验证、两项透明归档但缺乏支持/尚未解决的探索性测试——ViT-B/16 中的极性对齐盲测（内部归档为 H-path；纹理减去随机基线 $+0.006\,[-0.004, +0.017]$ 在1,024 张 CUB 图像上，包含 4,094/4,096 种组合，涵盖 $R = 1$两侧）以及 ResNet-18 中同图像与错误图像的纹理注入（其中 R 差异在早期和 -0.0026 早期 $+0.0007$ 后期切割处均跨越零值，而后期绝对增益显示同图像优势为 , 97.5% 置信区间 $[0.0090, 0.0242]$）——原始的 63 张图像 ResNet-18 注入试点、完整的 Qwen2.5 与 Llama-3.2 粗粒度与细粒度多层感知机路径分解，以及初始队列格拉姆汇总。

表 S5：视觉与语言 Transformer 中六组完整归一化路径交叉项。 每个条目均根据全精度汇总独立四舍五入至小数点后三位（未舍入的精确交叉项之和为 $+1$ 至 R）。ViT-B/16 使用 1,024 张 CUB 图像；Qwen2.5-1.5B 和 Llama-3.2-1B 使用首个包含 256 对样本的确认队列。

模型与方向	$D \times A$	$D \times M$	$D \times J$	$A \times M$	$A \times J$	$M \times J$	R
ViT-B/16 纹理（1,024）	-0.240	-0.061	$+0.003$	-0.020	-0.157	-0.242	0.284
ViT-B/16 随机基线	-0.304	-0.082	$+0.029$	$+0.058$	-0.121	-0.303	0.277
Qwen2.5 关系（256）	$+0.221$	-0.213	-0.077	-0.023	-0.262	$+0.016$	0.662
Qwen2.5 随机基线	$+0.231$	-0.170	-0.084	-0.022	-0.295	$+0.020$	0.681
Llama-3.2 关系（256）	$+0.090$	-0.108	-0.021	-0.006	-0.332	-0.010	0.614
Llama-3.2 随机基线	$+0.044$	-0.064	-0.004	$+0.051$	-0.303	-0.040	0.684

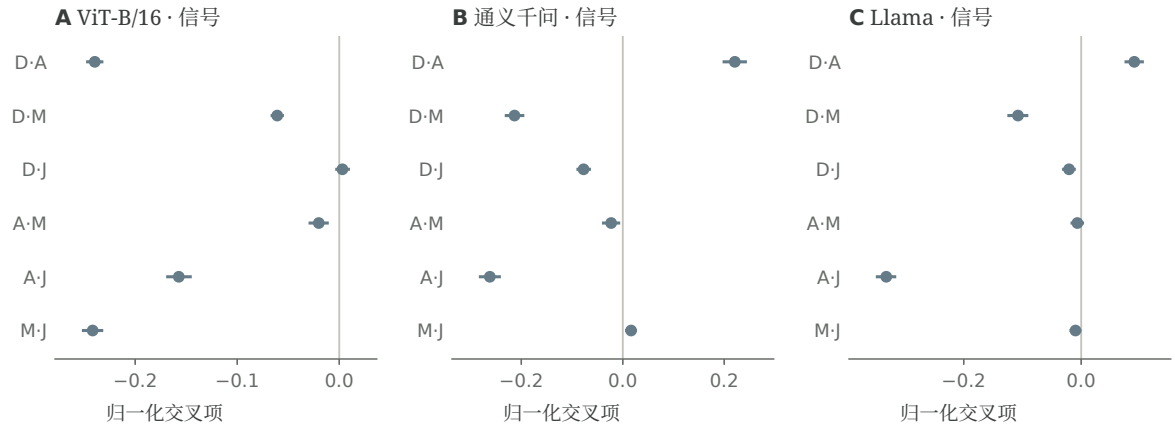


图 S5：视觉与语言 Transformer 中六组配对的信号通路交叉项。测得的纹理（ViT-B/16，1，024 张图像）和关系（通义千问Qwen2.5-1.5B、Llama-3.2-1B、256 组配对）的归一化交叉项及其95% 置信区间，与表 S5相对应。

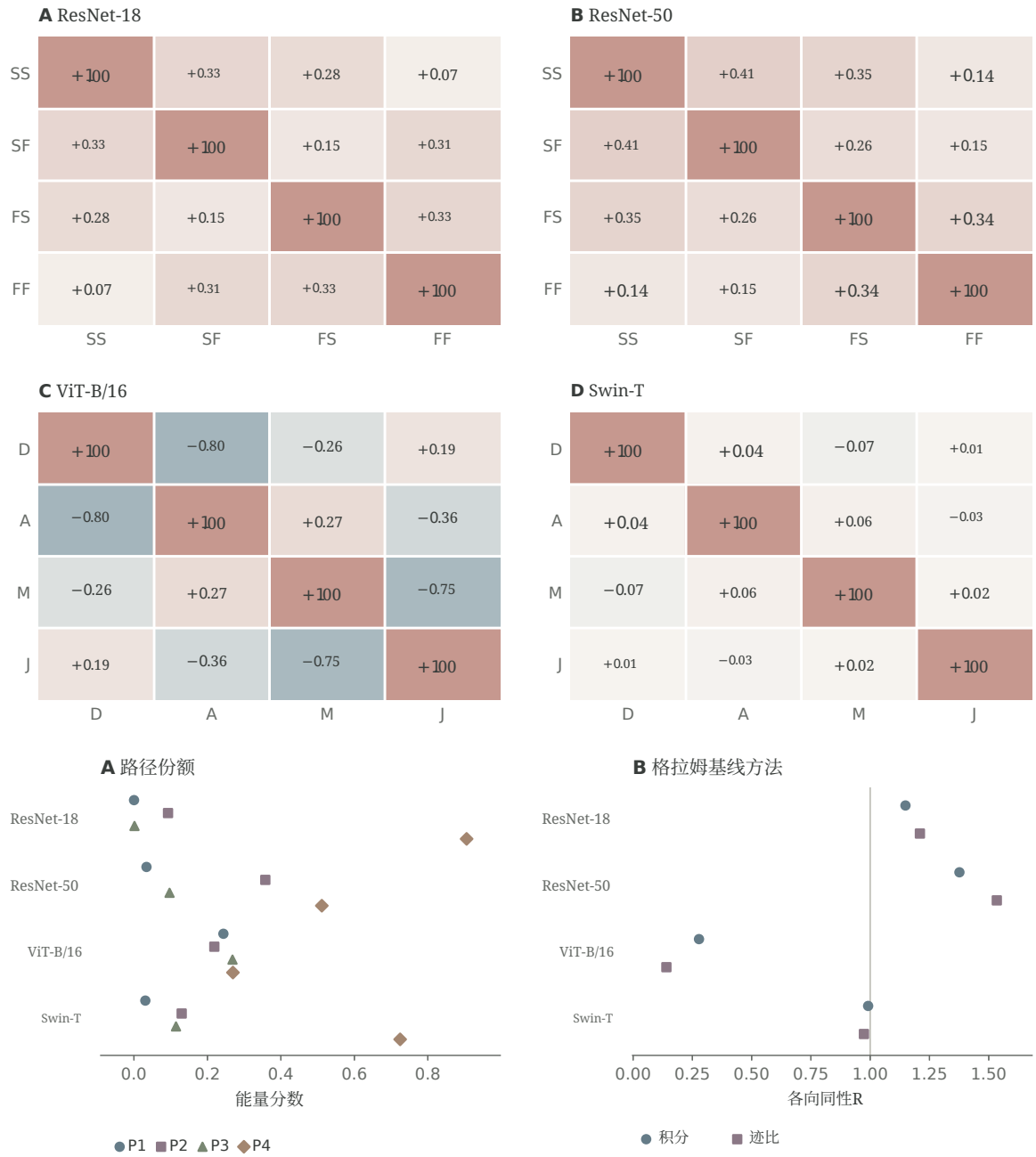


图 S6：四种视觉骨干网络中的完整路径梯度格拉姆几何（1，024 CUB 图像）。ResNet-18、ResNet-50（路径 SS、SF、FS、FF）、Swin-T 和 ViT-B/16（路径 D、A、M、J）中的成对路径梯度余弦相似度、梯度能量占比、精确高斯积分期望以及迹比近似。

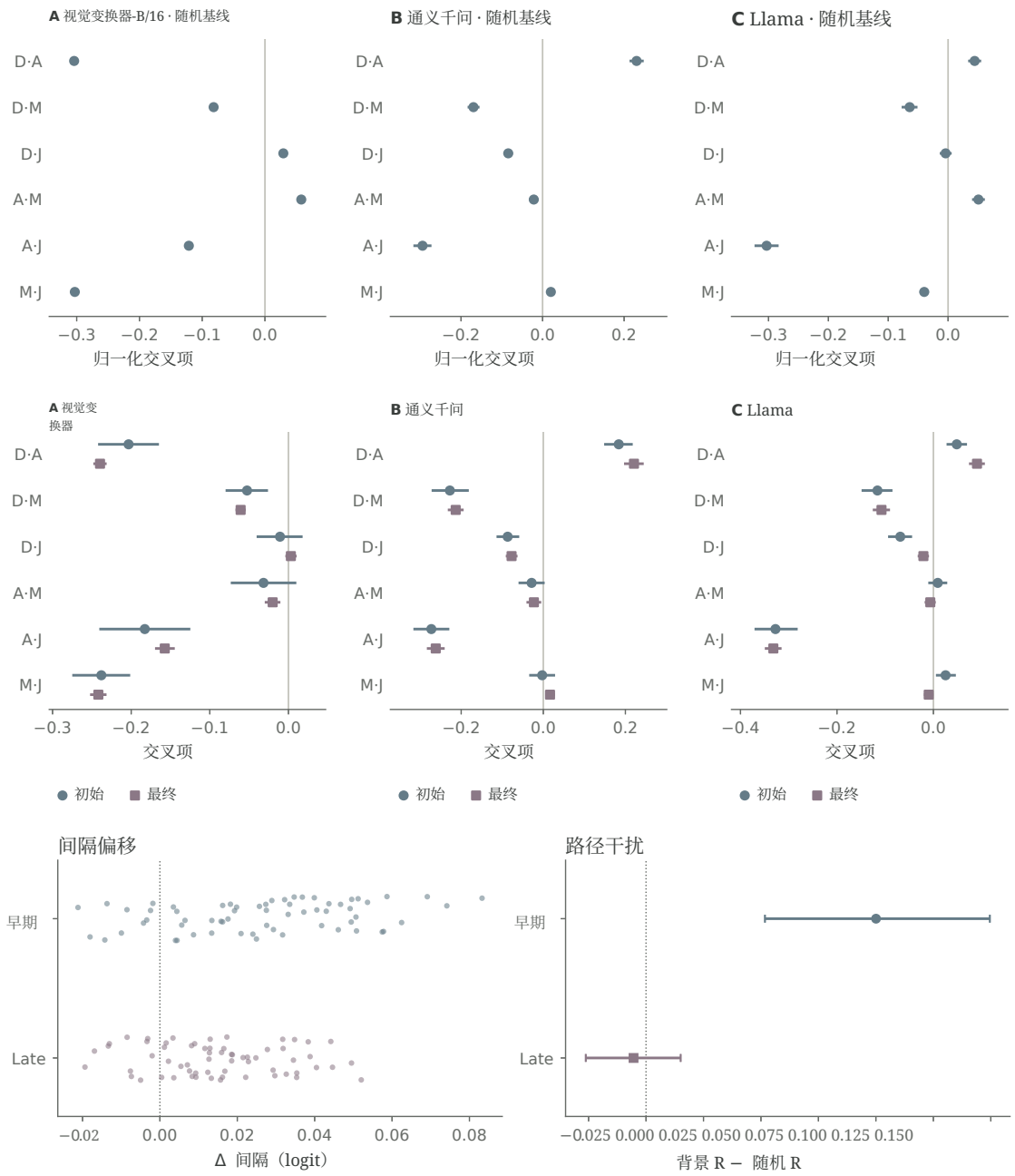


图 S7：随机方向六对交叉项（顶部）、初始与最终队列复现（中部）以及探索性 ResNet-18 试点（底部）。顶部：匹配的随机方向归一化交叉项（95% 置信区间）。中部：带符号信号交叉项的初始与最终队列复现。底部：原始的 63幅图像的 ResNet-18 注入试点（P1、早期四路径和晚期双路径切分）。

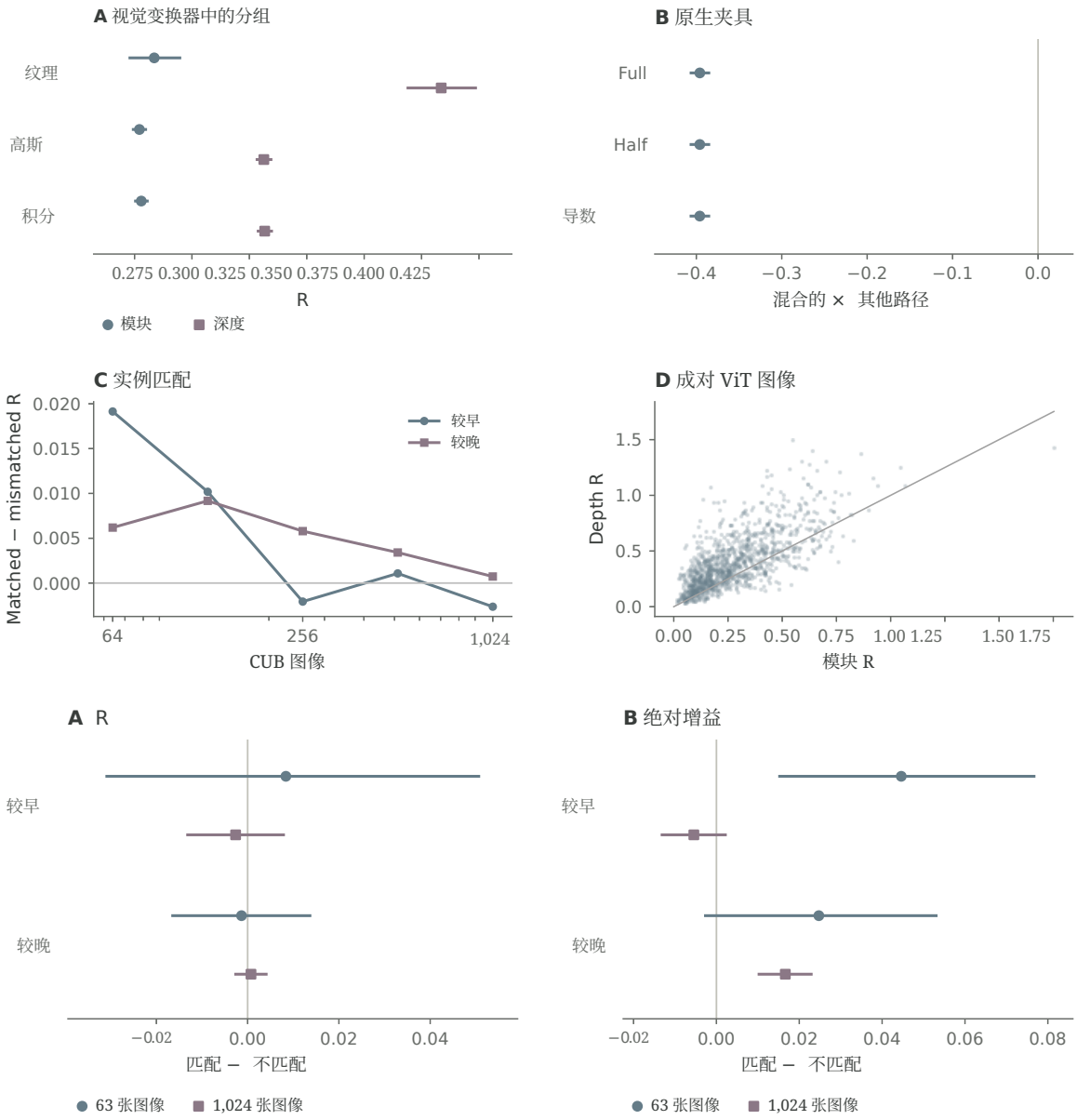
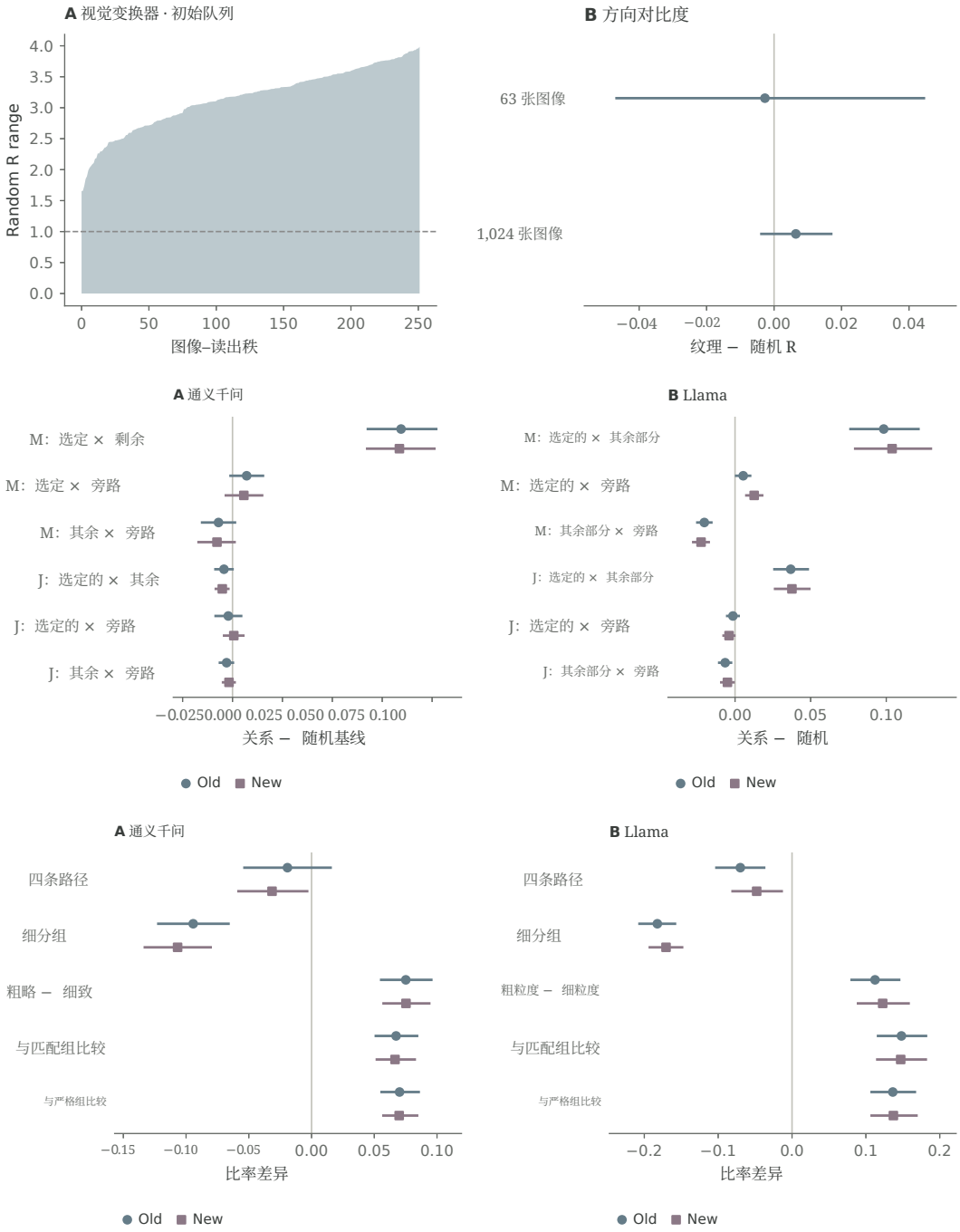


图 S8：ViT 中的重新分组、前向钳制以及同图像与错误图像对照。上：ViT-B/16 第6块中的模块与深度重新分组、原生钳制混合路径交叉项（四舍五入至 −0.396）、ResNet-18 实例匹配，以及配对 ViT 来源比较。下：两个独立 CUB 队列中同图像效应减去错误图像效应的比较。



图S9：视觉变换器随机方向范围（顶部），以及通义千问2.5和Llama-3.2中的由粗到细多层感知机路径拆分（中部和底部）。顶部：每种图像–读出组合的256个随机方向的范围（最终队列中，4，094/4，096个方向覆盖 $R = 1$ 的两侧），以及初始/最终纹理减随机检验。中部：对 M路径和 J路径交叉项进行拆分，以区分选定的功能性多层感知机单元、其余同层单元和旁路路径。底部：完整的旧版和新版256对由粗到细对比，以及相对于分层和严格的 $\pm 1\%$强度匹配对照组的配对优势。

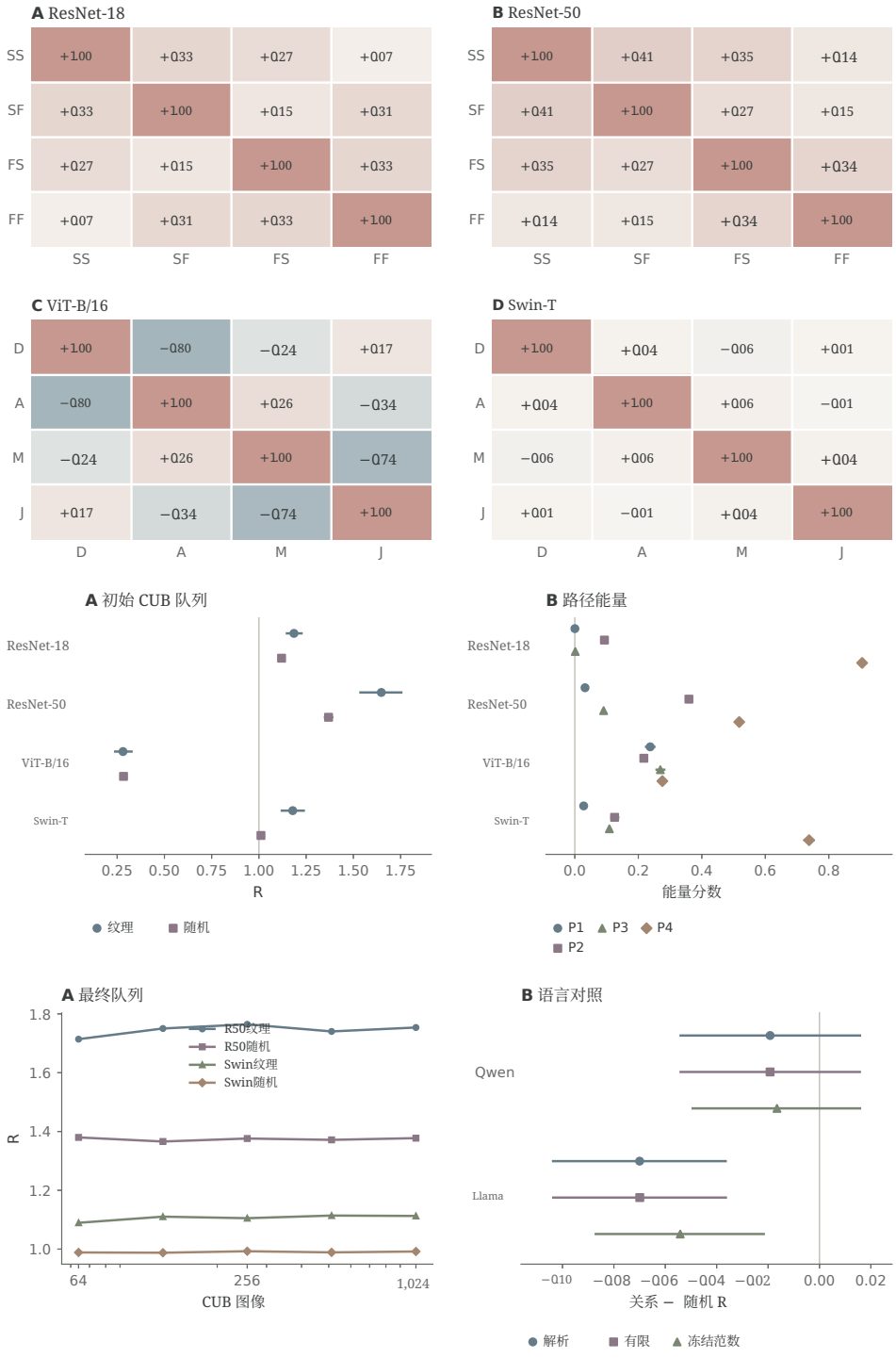


图 S10：初始队列路径几何与嵌套视觉/语言确认稳定性。 上方和中间：视觉骨干网络中的初始队列格拉姆余弦、测得的方向干扰比和梯度能量占比。下方：嵌套视觉确认曲线，以及解析、有限和冻结范数语言 ΔR 比较。

S10 扩展条件读出诊断

图 S11 至 S14 汇总了支持第 4.5 节的全部八个条件读出诊断图板。在 CUB（1,024 幅图像）中，仅考察纹理时，R 跨栖息地与栖息地内配对之间的差异跨越零点，而减去匹配范数的随机方向后得到 R 双重差分 $+0.03445\,[0.01939, 0.04940]$（两种比较所探讨的问题不同）。在语言数据中（52 个双分词器确认配对以及32对探索性队列），图板详细展示了声明式读出与同一前向传播双读出的有限差分和导数估计、每个单元的二选一候选准确率（包括冲突试验）、从32 对到52 对的队列复现、呈现和词汇主效应、交互效应、配对内呈现对比以及初始队列导数检验。

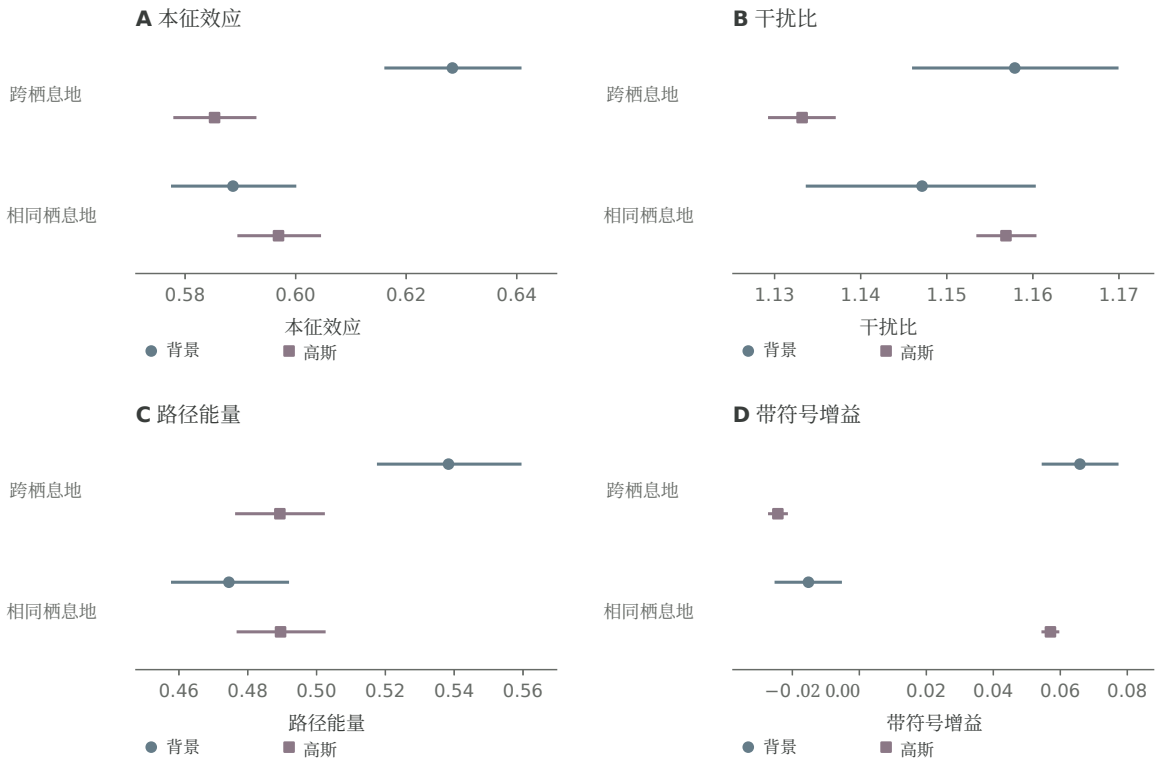


图 S11：CUB 八物种读出重新配对单元诊断（1,024 幅图像）。分别展示跨栖息地和栖息地内单元在背景及匹配高斯编辑下的绝对增益、干扰比 R、非相干路径能量 E和带符号增益（图像级 95% 置信区间）。

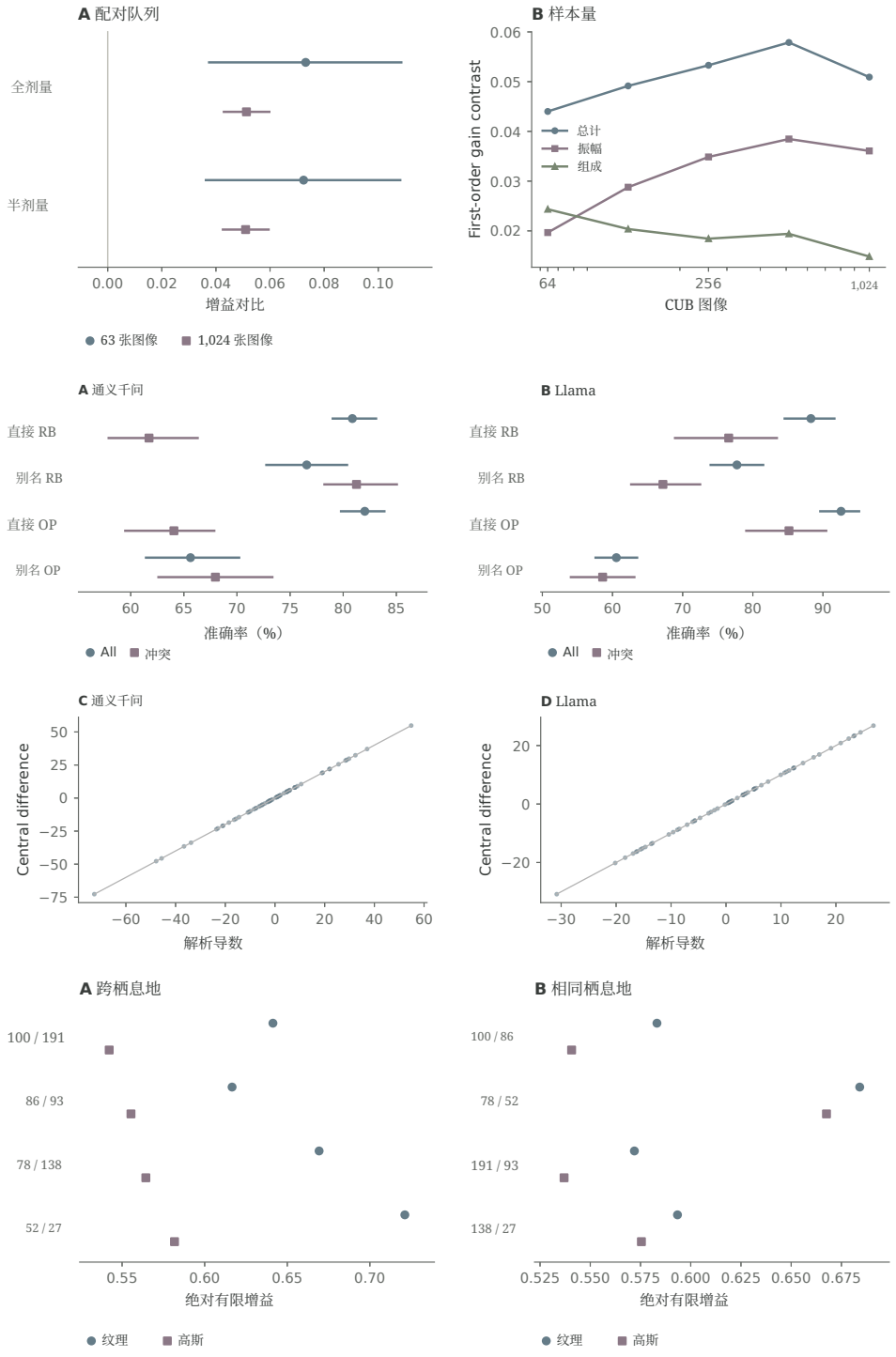


图 S12：CUB 剂量/样本量分解（上）以及语言（中）和视觉（下）的初始队列对照。上：在包含 63 张和 1，024 张图像的 CUB 队列中，全剂量和半剂量下的有限效应，以及对称振幅–组成分解的嵌套轨迹。中：初始 32 对语言队列中的逐单元准确率，以及解析导数与中心差分导数验证（每单位剂量的 10^{-3} logit）。下：初始 63 张图像队列中全部八个固定 CUB 类别对比的绝对有限增益。

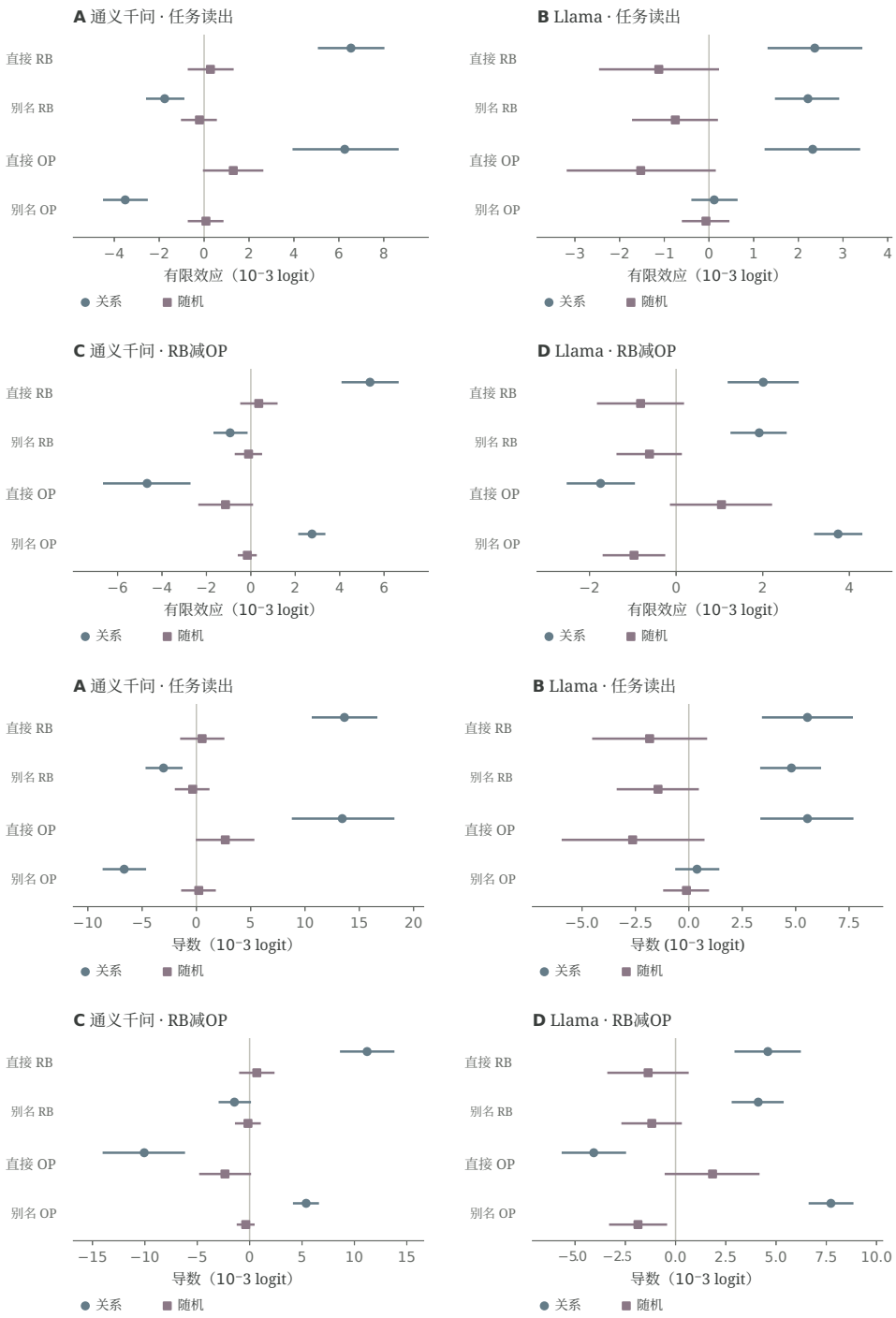


图 S13：语言有限差分与一阶导数读出比较（52 共享配对）。对于通义千问Qwen2.5-1.5B（上图）和 Llama-3.2-1B（下图），展示了关系方向和匹配随机方向的声明读出（上排），以及同一前向传播的两种读出之间的差异（下排）。

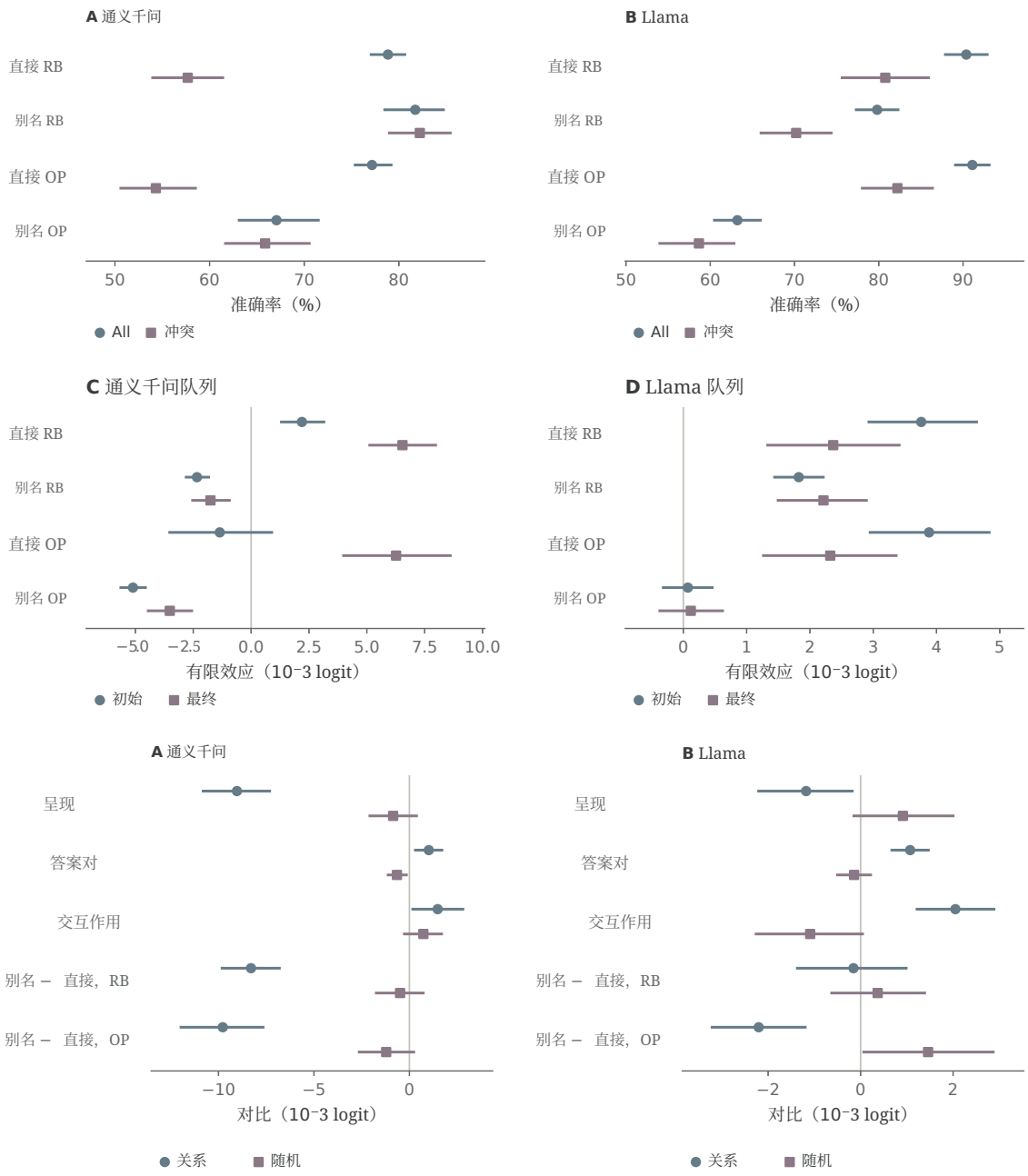


图 S14：语言 2×2 二选一准确率控制与析因对比。 上：通义千问2.5和Llama-3.2的全部四个单元格中的总体及冲突试次二选一候选准确率（52组共享配对；核心别名红色/蓝色准确率分别为81.7%和79.8%），并列展示32组与52组配对的有限效应复现。下：呈现主效应、答案词汇主效应、交互作用，以及词汇内呈现对比（95%配对名称层级置信区间）。

S11 扩展门控曲线、功能分组与严格强度匹配

本节收录了所有完整门控表（表 S6 和 S7）以及全部二十一幅门控和功能分组图版（图 S15 至 S23）：针对隐藏状态编辑和真实词汇关系替换的增量衰减与物理输出缩放六剂量曲线（$\lambda \in \{0, 0.1, 0.25, 0.5, 0.75, 1.0\}$）、翻转率、基线漂移和干净前缀困惑度、逐名称对散点云及代数随机符号检验、基于 3232、64、512、和1、024 幅图像的 ViT-B/16 连续索引开发集与留出集诊断、包括经 CUB 微调的 ViT-B/32 在内的功能性多层感知机分组控制，以及初始的 3232对配对 / 64幅图像对照表。

表 S6：双大语言模型在 256 个重放名称对和 256 个新零重叠名称对上的末层多层感知机功能分组。 门控 H 以 10^{-5} logit 表示（在 $+0.1$ 编辑剂量下，编辑增量衰减 10%）；ΔR 和粗粒度减细粒度对比 B 均为无量纲。通义千问2.5 中进行了严格的强度匹配（99.07%–100.98%），作为事后敏感性检验；在 Llama-3.2 中，则在留出集检验前锁定了该匹配。

度量	通义千问 旧（256）	通义千问 新（256）	Llama 旧（256）	Llama 新（256）
功能组 H（10^{-5} logit）	$+14.81\,[13.52, 16.23]$	$+13.74\,[12.54, 15.02]$	$+7.07\,[6.36, 7.80]$	$+6.46\,[5.83, 7.10]$
减去最佳连续区块	$+9.78\,[8.87, 10.79]$	$+8.88\,[8.00, 9.80]$	$+6.18\,[5.51, 6.86]$	$+5.71\,[5.12, 6.32]$
减去 64 桶分层匹配	$+9.22\,[8.43, 10.08]$	$+8.47\,[7.74, 9.23]$	$+6.19\,[5.50, 6.90]$	$+5.63\,[5.02, 6.24]$
减去严格匹配的 ±1%	$+2.72\,[2.35, 3.08]$	$+2.22\,[1.93, 2.54]$	$+4.92\,[4.30, 5.55]$	$+4.49\,[3.94, 5.05]$
选择 × 编辑交互作用	$+0.96\,[0.61, 1.32]$	$+0.66\,[0.38, 0.97]$	$+3.44\,[2.70, 4.21]$	$+2.77\,[2.15, 3.43]$
粗略四路径 ΔR	$-0.0192\,[-0.0545, +0.0162]$	$-0.0315\,[-0.0592, -0.0025]$	$-0.0700\,[-0.1041, -0.0360]$	$-0.0479\,[-0.0822, -0.0123]$
精细功能性拆分 ΔR	$-0.0943\,[-0.1231, -0.0652]$	$-0.1068\,[-0.1339, -0.0794]$	$-0.1824\,[-0.2082, -0.1568]$	$-0.1706\,[-0.1943, -0.1471]$
粗略–细致对比 B	$+0.0751\,[0.0545, 0.0965]$	$+0.0753\,[0.0564, 0.0948]$	$+0.1124\,[0.0790, 0.1467]$	$+0.1227\,[0.0876, 0.1595]$
B 减去严格 ±1% 匹配	$+0.0702\,[0.0548, 0.0864]$	$+0.0699\,[0.0563, 0.0851]$	$+0.1364\,[0.1061, 0.1680]$	$+0.1372\,[0.1061, 0.1702]$
内部-M 选定 × 剩余	$+0.0844\,[0.0672, 0.1027]$	$+0.0836\,[0.0668, 0.1018]$	$+0.0983\,[0.0740, 0.1238]$	$+0.1039\,[0.0786, 0.1304]$

表S7：原始 ImageNet ViT-B/16 的功能性多层感知机分组（384/3，072 个单元，10% 衰减，10^{-5} logit），在264张保留图像和1，024张图像的原始权重重放中。 在相同的 264 张图像上经 CUB 微调的 ViT-B/32 得到功能性 $+45.67$ 与最佳连续分组相比 -0.51（配对 $+46.19\,[42.61, 49.68]$）以及跨读出交互作用 $+352.59\,[319.71, 386.80]$。

组间或配对对比	第10块（264 张留出集图像）	第6块（264留出集）	第6块（1，024 重放）	第10块（1，024重放）
功能组	$+22.23\,[17.84, 26.57]$	$+29.68\,[24.25, 35.44]$	$+30.86\,[28.09, 33.68]$	$+22.53\,[20.15, 24.84]$
最佳连续区块	$-2.39\,[-4.63, -0.30]$	$+1.10\,[-0.77, +2.93]$	$+0.46\,[-0.34, +1.24]$	$+0.05\,[-0.97, +1.05]$
强度匹配随机均值	$-11.13\,[-12.79, -9.56]$	$-2.59\,[-3.87, -1.28]$	—	—
随机划分中的最佳值	$-3.22\,[-4.27, -2.20]$	$+0.77\,[-0.18, +1.74]$	—	—
功能性 − 最佳连续分组	$\mathbf{+24.62\,[20.21, 29.04]}$	$\mathbf{+28.57\,[23.58, 33.70]}$	$\mathbf{+30.41\,[27.77, 33.06]}$	$\mathbf{+22.47\,[20.20, 24.71]}$
功能性 − 强度匹配	$\mathbf{+33.36\,[28.75, 38.01]}$	$\mathbf{+32.27\,[27.34, 37.46]}$	—	—

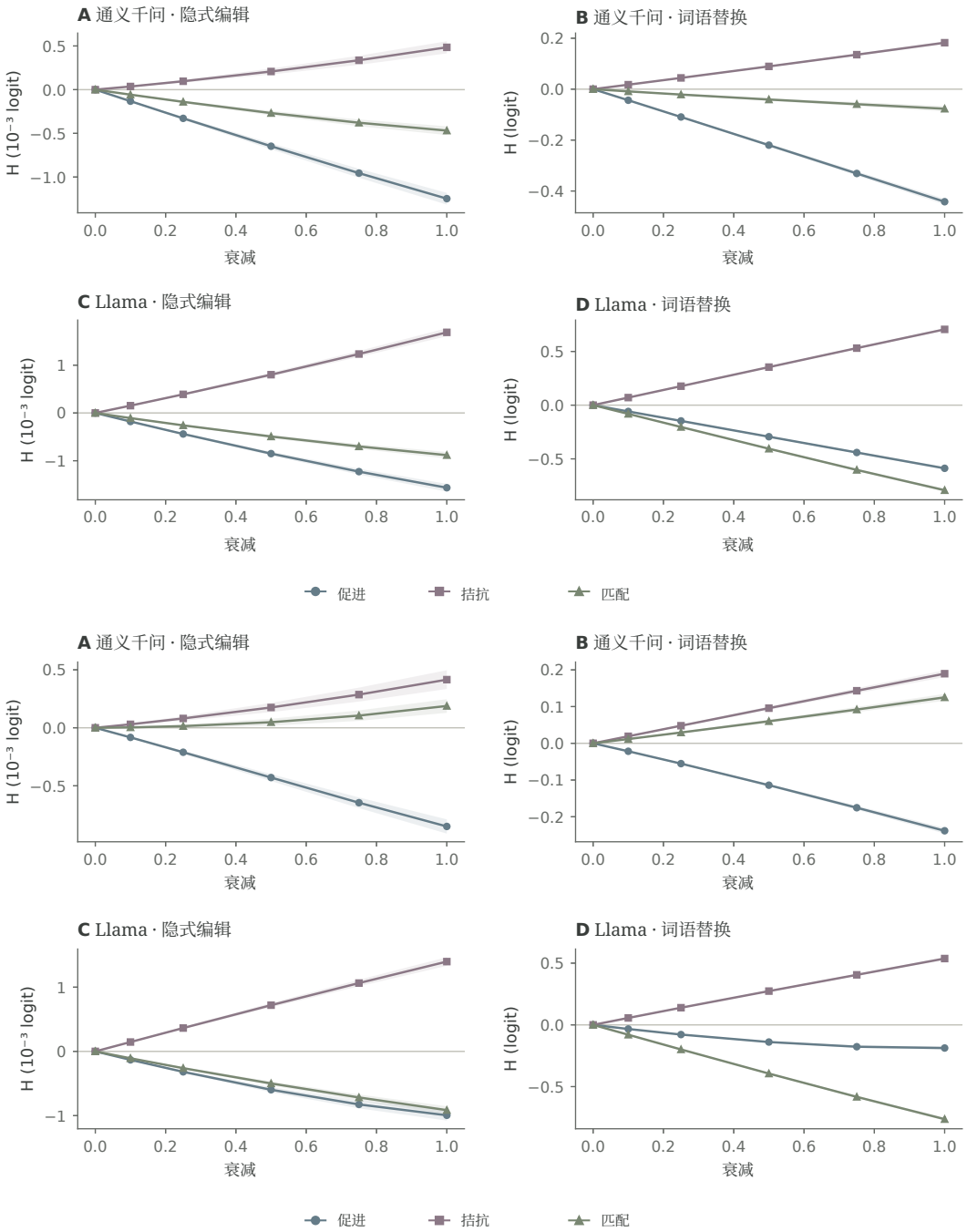


图 S15：增量衰减（上）和物理输出缩放（下）条件下的语言注意力头六剂量曲线。在通义千问2.5-1.5B和Llama-3.2-1B中，针对六种剂量（$\lambda \in \{0, 0.1, 0.25, 0.5, 0.75, 1.0\}$）下的256对留出名称对进行评估，涵盖隐藏编辑（$10^{-3}$logit）和真实的词汇关系词替换（原始logit）（95%名称级置信区间）。

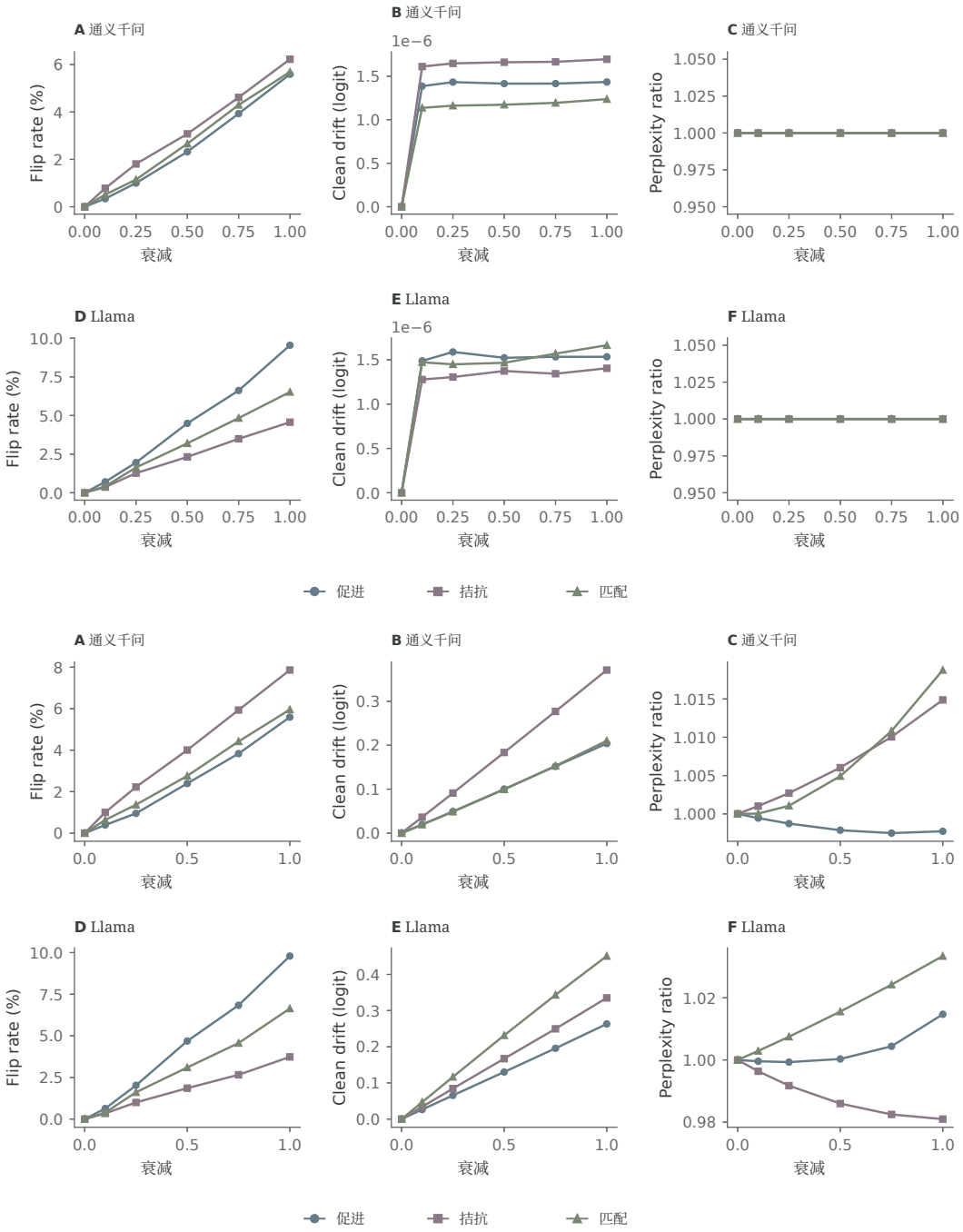


图 S16：冻结语言模型中的清洁工作点稳定性与翻转率诊断（256 组配对）。在增量衰减（上图；漂移 $\leq 1.7 \times 10^{-6}$ logit，困惑度比为 1.000）与物理输出缩放（下图；明显的清洁漂移促使我们将增量衰减作为主要干预）条件下，六种剂量的间隔翻转率、绝对清洁基线漂移和清洁前缀困惑度比。

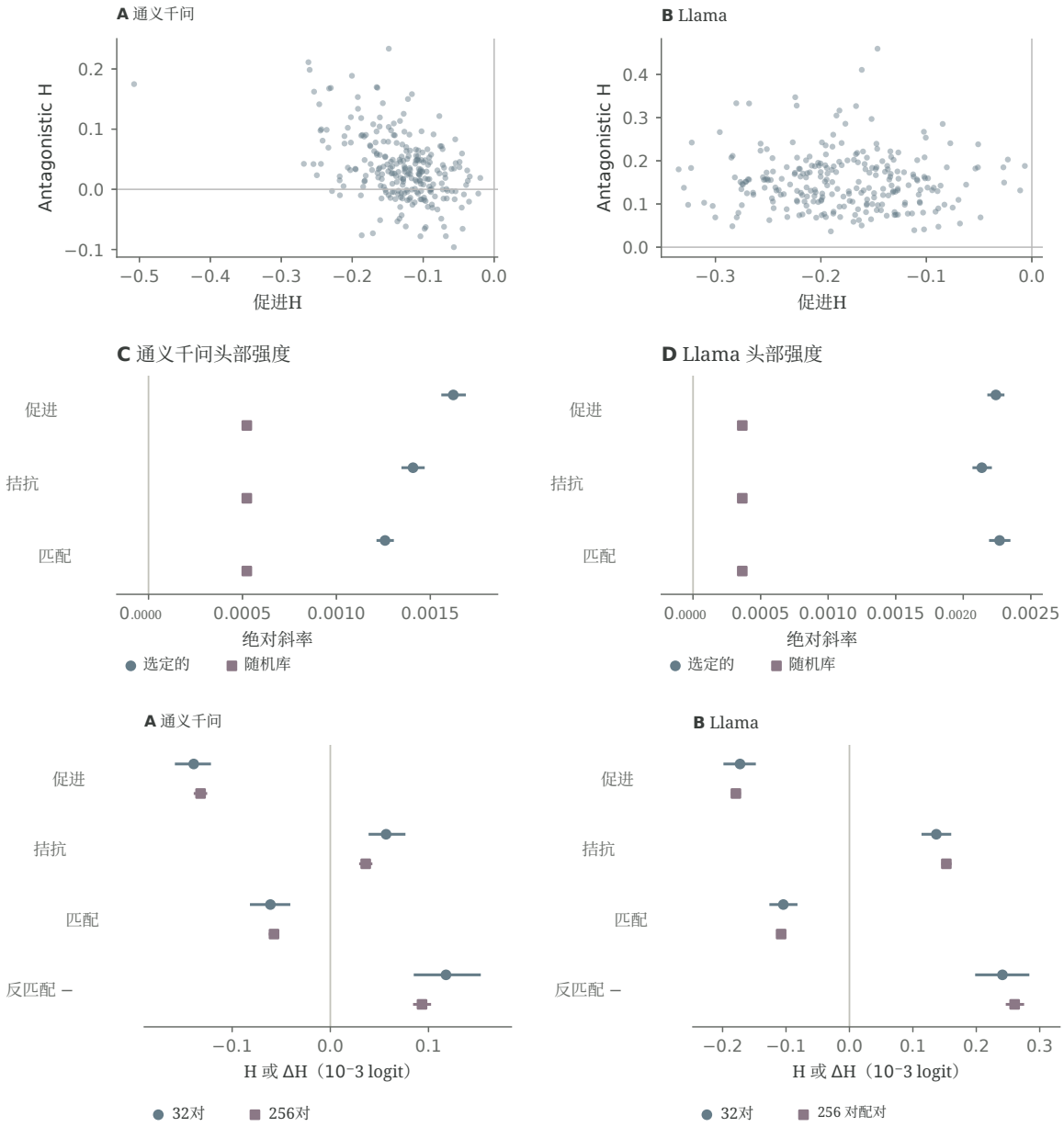


图 S17：语言模型中按名称对划分的门控分布、强度库及队列复现。上：主要剂量 $\lambda_0 = 0.1$（10^{-3} logit）下各名称对的配对 H 值，以及与 128 组未匹配随机库相比的选定头部开发强度。下：在包含 32 对的初始队列和包含 256 对的确认队列中，主要门控估计及拮抗减匹配对比的复现结果。

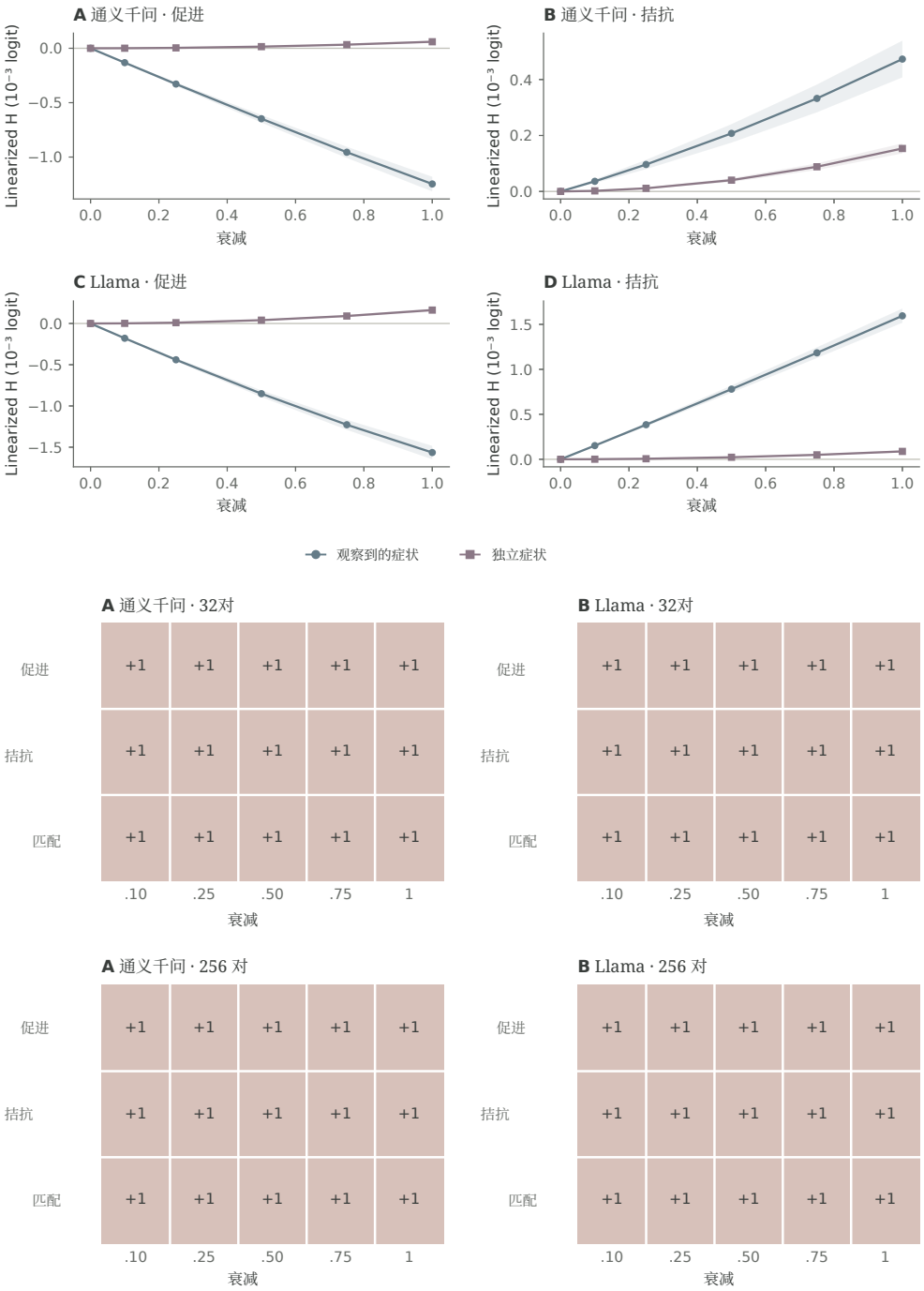


图 S18：代数随机符号对照与隐藏符号和词汇符号一致性。 顶部：比较观测到的原子符号与独立随机符号的代数剂量反应（256 对）。中部和底部：在 32 对初始队列（中部）和 256 对确认队列（底部）中，隐藏编辑与词汇替换均值之间的符号一致性 H（$+1$ 一致，-1 相反，0 表示均值为零）。

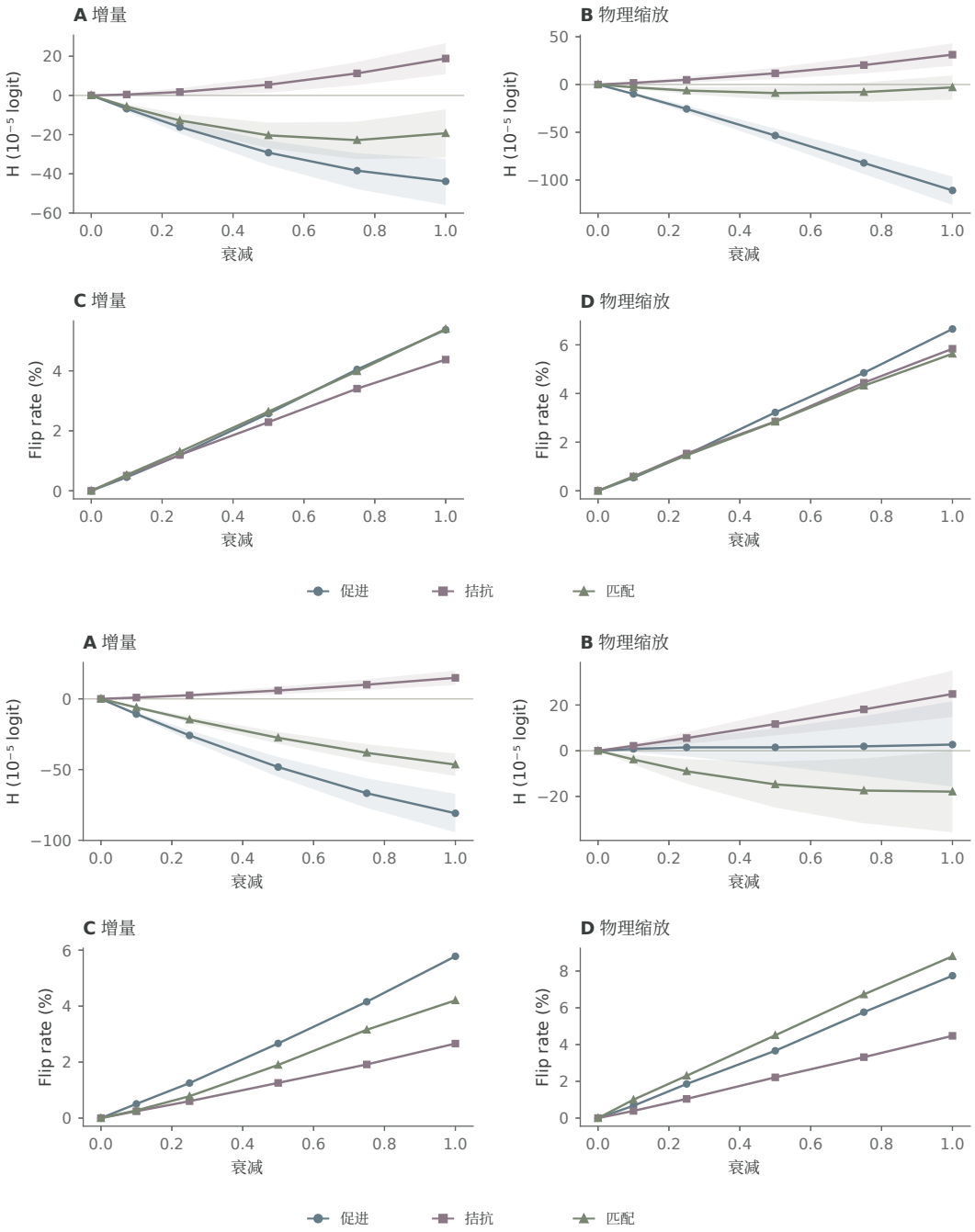


图 S19：OriginalViT-B/16 在以下数据上的连续索引门控：1，024 **张 CUB 确认图像。上：在 1**，024 张确认图像上，评估原始选定的 32 张图像连续第6块组在不同剂量下的表现（增量衰减和物理缩放）。下：在相同的 1，024 张图像上使用 512 图像回退连续选择器（其中 0/24 符合强度条件的连续候选项具有正向平均开发斜率）。

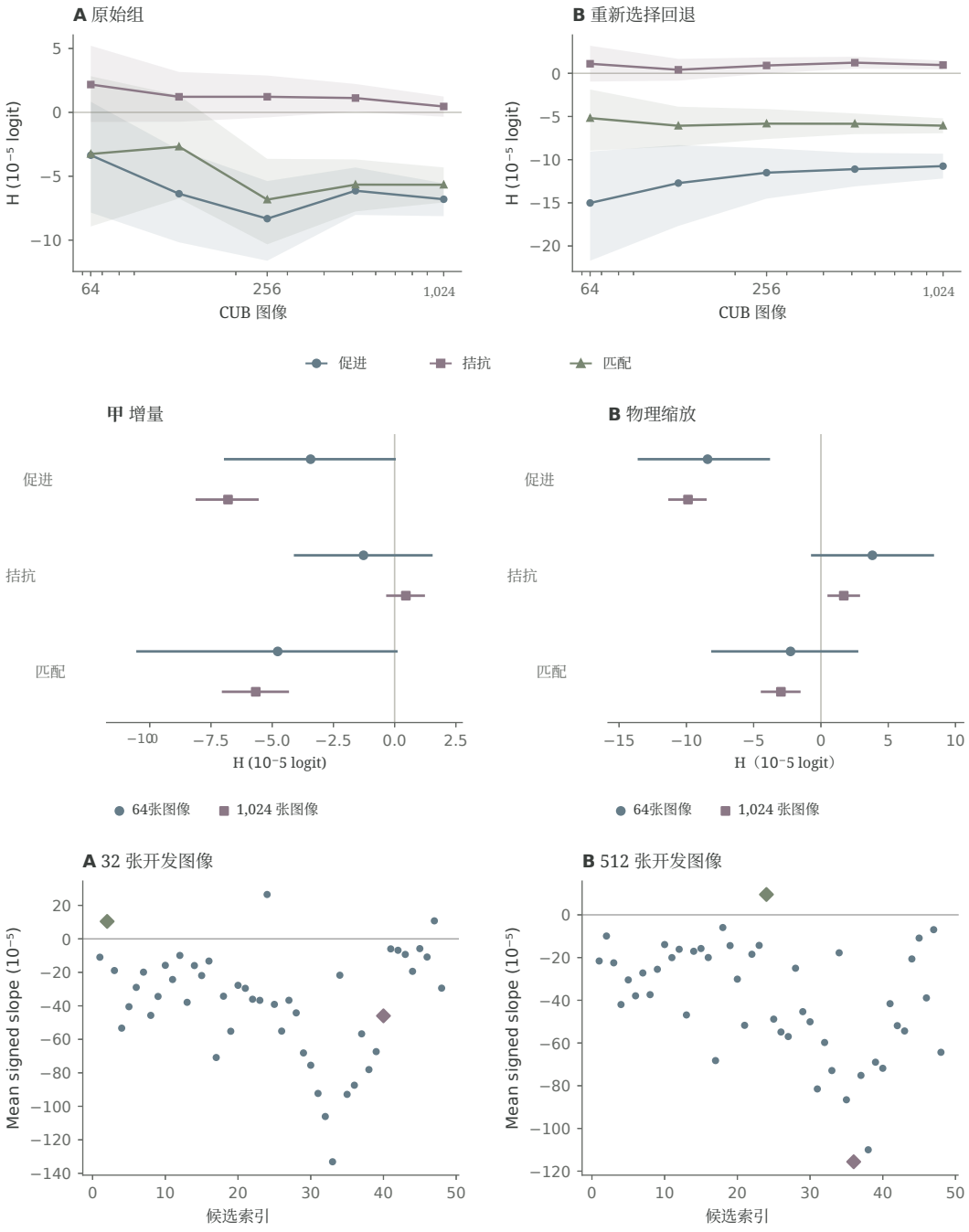


图 S20：ViT-B/16 连续组稳定性，64-对-1,024 比较及开发斜率全景图。顶部：原始连续选择器和备用连续选择器的嵌套样本量曲线。中间：原始连续选择器在 64 张图像上的初始估计值与在 1,024 张图像上的确认估计值。底部：在 32 张与 512 张开发图像上，全部 48 个连续 ViT-B/16 多层感知机候选项的带符号开发斜率（菱形标记选定的促进型/拮抗型候选项；在 512 张图像上，$0/24$ 强度合格的连续候选项均为正值）。

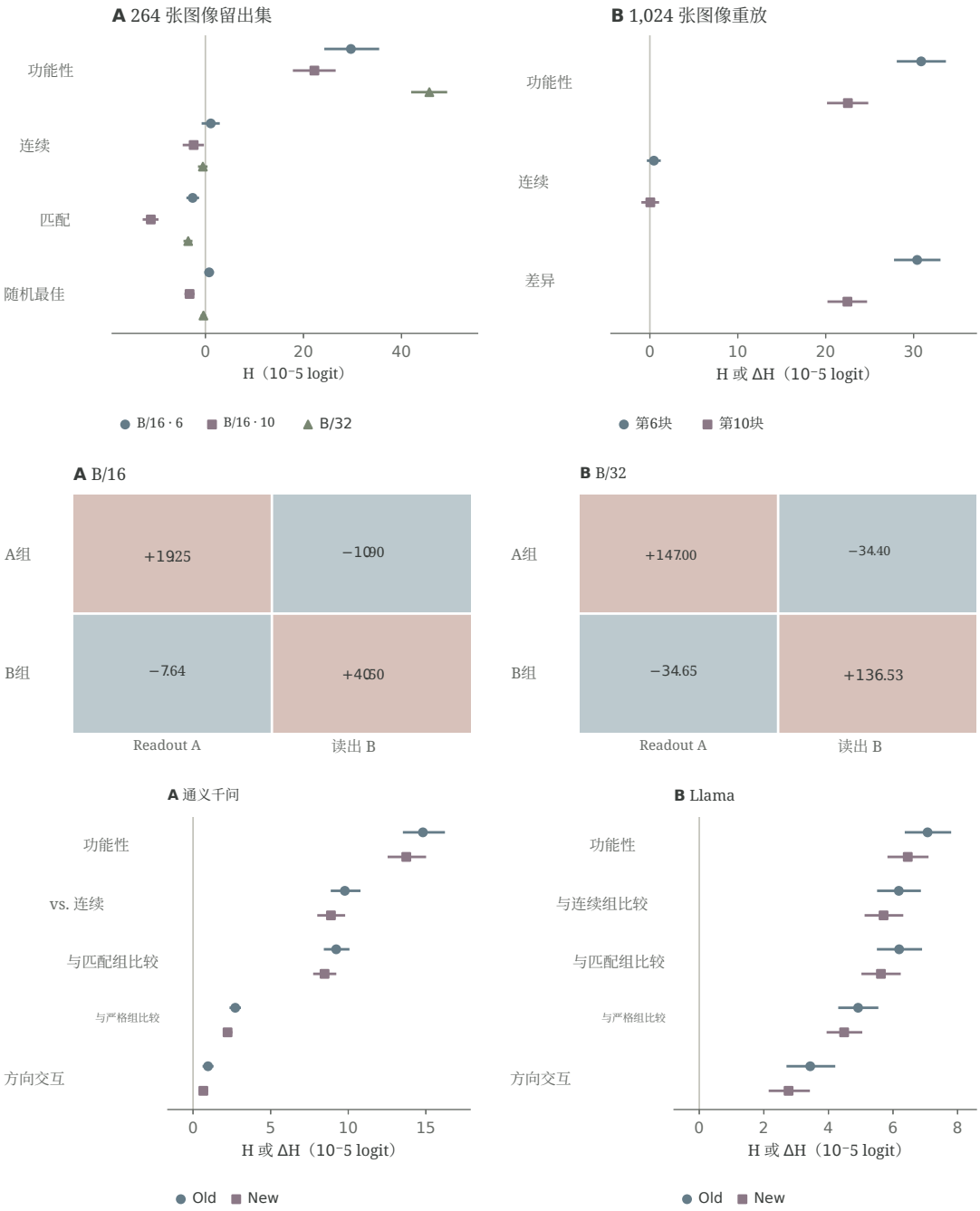


图 S21：视觉与语言中的功能性多层感知机分组控制。 上：在 264 张图像保留集（ViT-B/16 第 6 和第 10 模块，以及经 CUB 微调的 ViT-B/32）和 1,024 张图像的原始权重重放中进行的四种分组控制。中：原始 ImageNet ViT-B/16（交互作用 $+78.30$）和经 CUB 微调的 ViT-B/32（$+352.59 \times 10^{-5}$ logit）中的同一前向传播交叉读出矩阵。下：通义千问2.5和 Llama-3.2 中完整的旧新 256 对多层感知机功能门控对比。

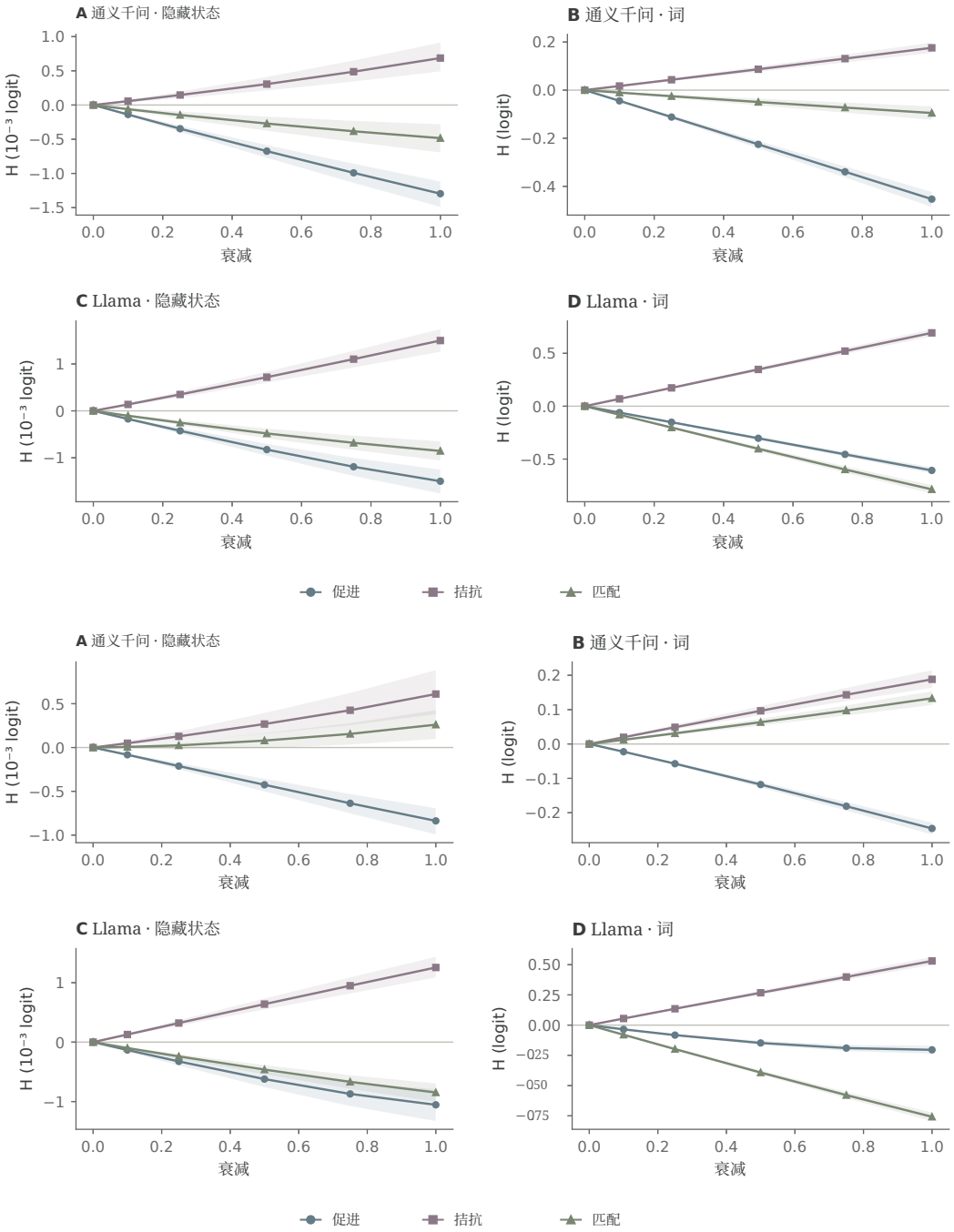


图 S22：初始 32 对语言队列六剂量门控档案。 初始 32 对探索性语言队列中，隐式编辑和真实词汇替换条件下的增量衰减（上）与物理输出缩放（下）。

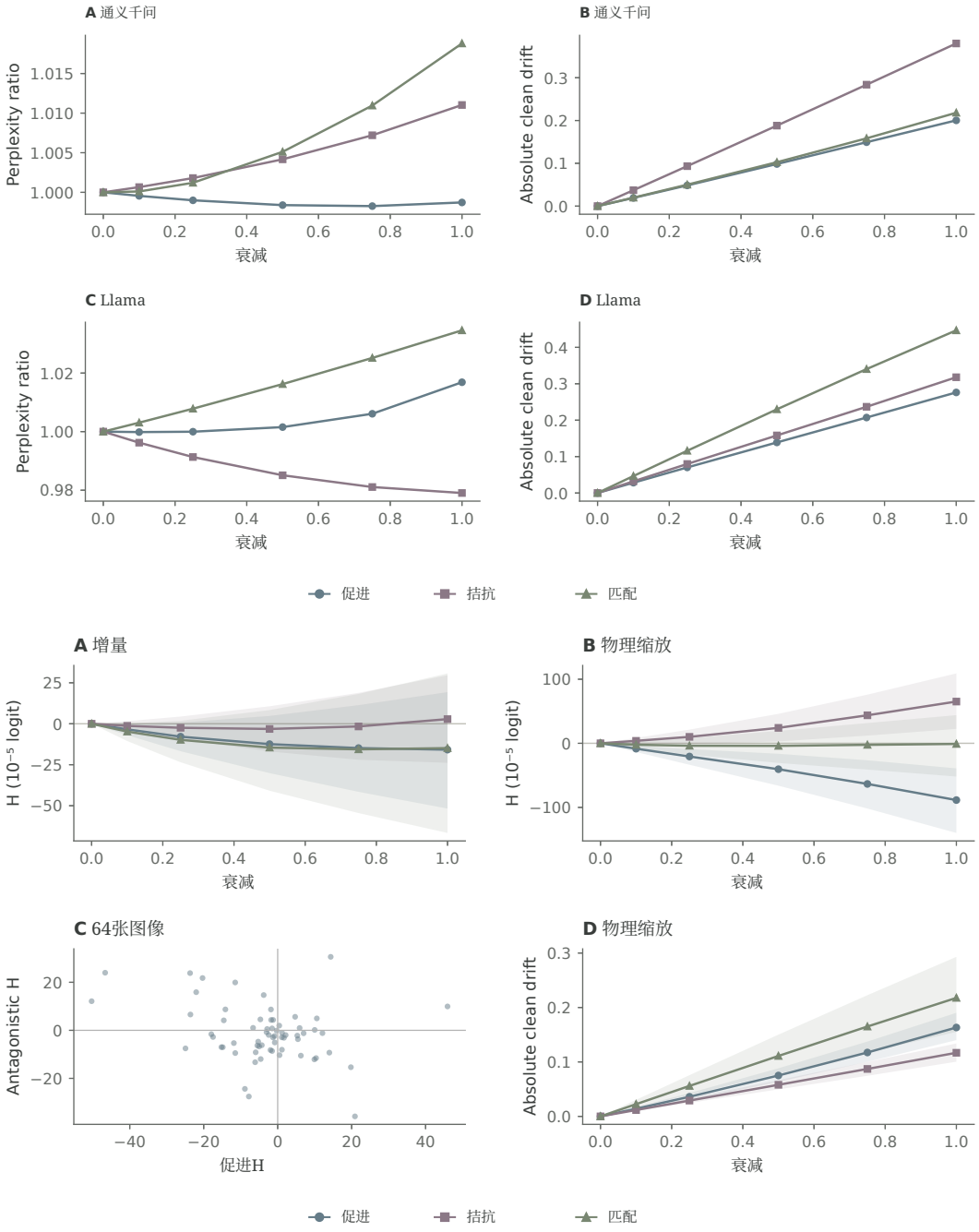


图 S23：语言（32 对）和 ViT-B/16（64 张图像）中的初始队列漂移检查。顶部：初始 32 对语言队列中的物理缩放下的干净困惑度和基线裕度漂移。底部：初始 64 张图像 CUB ViT-B/16 队列中的六剂量曲线、逐图像主要效应（10^{-5} logit）和物理基线漂移。

S12 扩展剪枝恢复、手术、窗口控制与全网络档案

表 S8 记录了全部 10 个剪枝切割（4,096 张 ImageNet 图像）中准确的留出斯皮尔曼相关系数；表 S9 列出了全部六个分支和各深度中观测间隔与线性化间隔之间的皮尔逊相关系数；表 S10 重现了经验证的历史六族全网络准确率档案（来自 Shu [44]，的 3,200 张图像，其中有 257 张与本地 4,096 张图像队列重叠）。图 S24 至 S35 提供了完整的 24 个非重复剪枝图板：初始排名与最终排名、Swin-T 移位窗口分层（全局普通最小二乘法 0.016910 <真实移位窗口0.017068 < 随机分层 0.017425）、Swin 候选增益云图、全网络准确率增益、跨早期、中间层和后期切割的六分支均方误差/平均绝对误差曲线，以及 ViT-S/16 和 Swin-T 在早期、后期、中间层（ 4,096 张图像）和历史中间层（ 1,000 张图像）切割中的每张图像物理手术散点云。

表 S8：在 4，096 张留出的 ImageNet 图像上的全部十个局部剪枝切口。两个主要中间层 Transformer 切口采用 97.5% 的经家族校正的配对自助法置信区间，其余切口采用 98.75% 的置信区间。

模型与切口	图像	方差 ρ	FWL ρ	增益（$\Delta\rho$）	$\Delta\rho$ 的族群置信区间
ResNet-50 第1层1.0卷积2	4,096	0.8671	1.0000	+0.1329	[0.1329, 0.1329]
ResNet-50 第2层2.0卷积2	4,096	0.4201	0.9967	+0.5766	[0.5688, 0.5824]
ResNet-50 第3层3.0卷积2	4,096	0.5689	0.9936	+0.4248	[0.4129, 0.4362]
ResNet-50 第4层4.0卷积2	4,096	0.5641	0.8650	+0.3009	[0.2678, 0.3178]
ViT-S/16 早期多层感知机	4,096	0.9189	0.9985	+0.0796	[0.0787, 0.0807]
ViT-S/16 中期多层感知机	4,096	0.9351	0.9959	+0.0608	[0.0594, 0.0623]
ViT-S/16 后期多层感知机	4,096	0.9482	0.9806	+0.0324	[0.0274, 0.0373]
Swin-T 早期多层感知机	4,096	0.7282	0.9954	+0.2672	[0.2629, 0.2709]
Swin-T 中期多层感知机	4,096	0.8419	0.9918	+0.1499	[0.1475, 0.1526]
Swin-T 后期多层感知机	4,096	0.8669	0.9747	+0.1078	[0.1006, 0.1136]

表 S9：全部物理手术散点云中观测到的类别间隔变化与线性化类别间隔变化之间的皮尔逊相关（图表 S28 至 S35）。

图像	Cut	模型	术前切片	LAMP 切片	术后切片	前置普通最小二乘法	LAMP OLS	后置普通最小二乘法
4,096	早期	ViT-S/16	0.781	0.733	0.800	0.848	0.833	0.918
4,096	早期	Swin-T	0.836	0.762	0.846	0.897	0.894	0.887
4,096	中间层	ViT-S/16	0.897	0.881	0.959	0.953	0.935	0.979
4,096	中间层	Swin-T	0.898	0.881	0.858	0.916	0.901	0.890
4,096	Late	ViT-S/16	0.890	0.827	0.753	0.914	0.873	0.859
4,096	Late	Swin-T	0.961	0.942	0.948	0.973	0.969	0.975
1,000	初始化中期	ViT-S/16	0.904	0.875	0.951	0.951	0.933	0.967
1,000	初始化中期	Swin-T	0.899	0.886	0.844	0.878	0.885	0.841

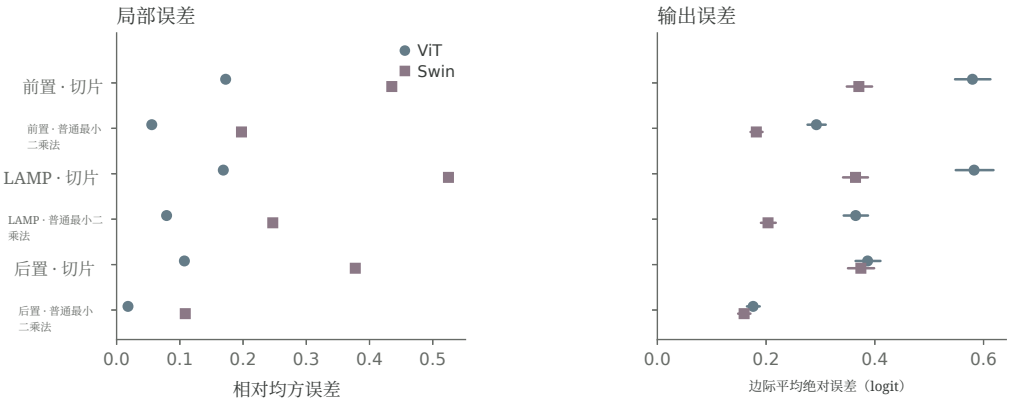


图 S24：初始1，000张图像研究中间层 Transformer 多层感知机切分点的六臂汇总。展示初始的1，000 张图像研究中，六种方案（直接切片与普通最小二乘法下的 GELU 前置、LAMP、GELU 后置）的相对局部均方误差和类别边际平均绝对误差。

表 S10：六种架构在三个移除预算下的历史全网络结构化剪枝准确率档案。 已根据 540 篇已发表记录重新核验，见 Shu [44]（3，200 张 ImageNet 验证图像；257 张与我们的 4，096 张图像本地队列重叠）。无标签（M_u）和任务条件化（M_s）准确率与 L_2、Fisher 及表现最佳的竞争基线方法进行比较。除单切局部线性化之外，全网络多层 Swin 的准确率表现仍有待探究。

架构	预算	M_u (%)	对比 L_2	其他最佳项	对比最佳项	M_s (%)	与 Fisher 相比	其他最佳方法	与最佳方法相比
ConvNeXt-B	30%	71.69	+21.41	APoZ	−0.81	65.69	−1.72	TaylorGate	−6.84
ConvNeXt-B	40%	47.09	+33.06	APoZ	−0.07	49.69	−2.00	FisherGate	−6.28
ConvNeXt-B	50%	3.03	+2.59	APoZ	−0.75	15.22	−4.25	OBD	−8.81
EfficientNet-B3	30%	39.66	−0.31	FPGM	−1.34	61.03	+8.65	OBD	+0.25
EfficientNet-B3	40%	19.38	+4.32	APoZ	+2.63	39.97	+18.78	OBD	+3.31
EfficientNet-B3	50%	5.19	+0.97	APoZ	+0.25	19.88	+12.57	OBD	+6.07
MobileNetV2	30%	15.84	−6.88	L_1	−7.63	42.63	+19.63	OBD	+5.66
MobileNetV2	40%	4.72	−2.72	L_2/LAMP	−2.72	16.25	+9.75	OBD	+4.72
MobileNetV2	50%	1.19	−0.12	L_2/LAMP	−0.12	4.69	+3.28	GradNorm	+2.25
ResNet-50	30%	40.94	+29.03	APoZ	+18.00	50.84	+6.90	GradNorm	+5.93
ResNet-50	40%	18.75	+15.16	APoZ	+13.25	27.87	+8.28	GradNorm	+3.68
ResNet-50	50%	5.41	+3.44	L_2/LAMP	+3.44	9.88	+4.85	GradNorm	+3.04
Swin-T	30%	59.94	−5.31	L_2/LAMP	−5.31	60.03	+4.47	Wanda	−4.94
Swin-T	40%	44.53	−10.69	L_2/LAMP	−10.69	43.03	+4.66	Wanda	−7.56
Swin-T	50%	21.22	−12.50	L_2/LAMP	−12.50	22.03	+2.06	Wanda	−3.34
ViT-S/16	30%	52.53	+32.47	APoZ	+6.97	54.03	+9.94	FisherGate	+0.72
ViT-S/16	40%	31.28	+24.56	APoZ	+9.19	36.34	+14.00	FisherGate	−6.16
ViT-S/16	50%	9.75	+7.22	随机	+6.22	15.75	+8.69	FisherGate	−5.56

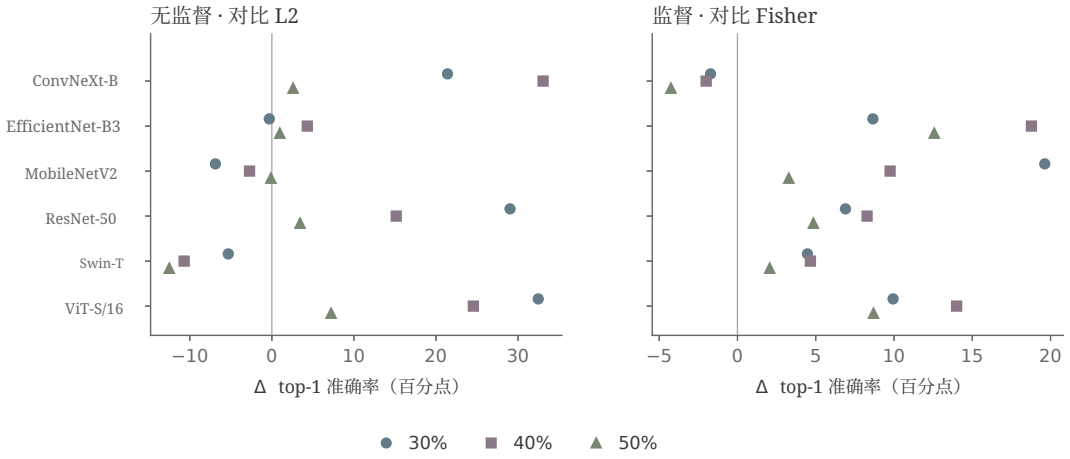


图 S25：三种移除预算下六种架构的历史全网络准确率增益。与表 S10相对应的可视化图示，展示了无标签（M_u 与 L_2）及任务条件化（M_s 与 Fisher）在六种架构家族（3，200张 ImageNet 图像）上的准确率变化。

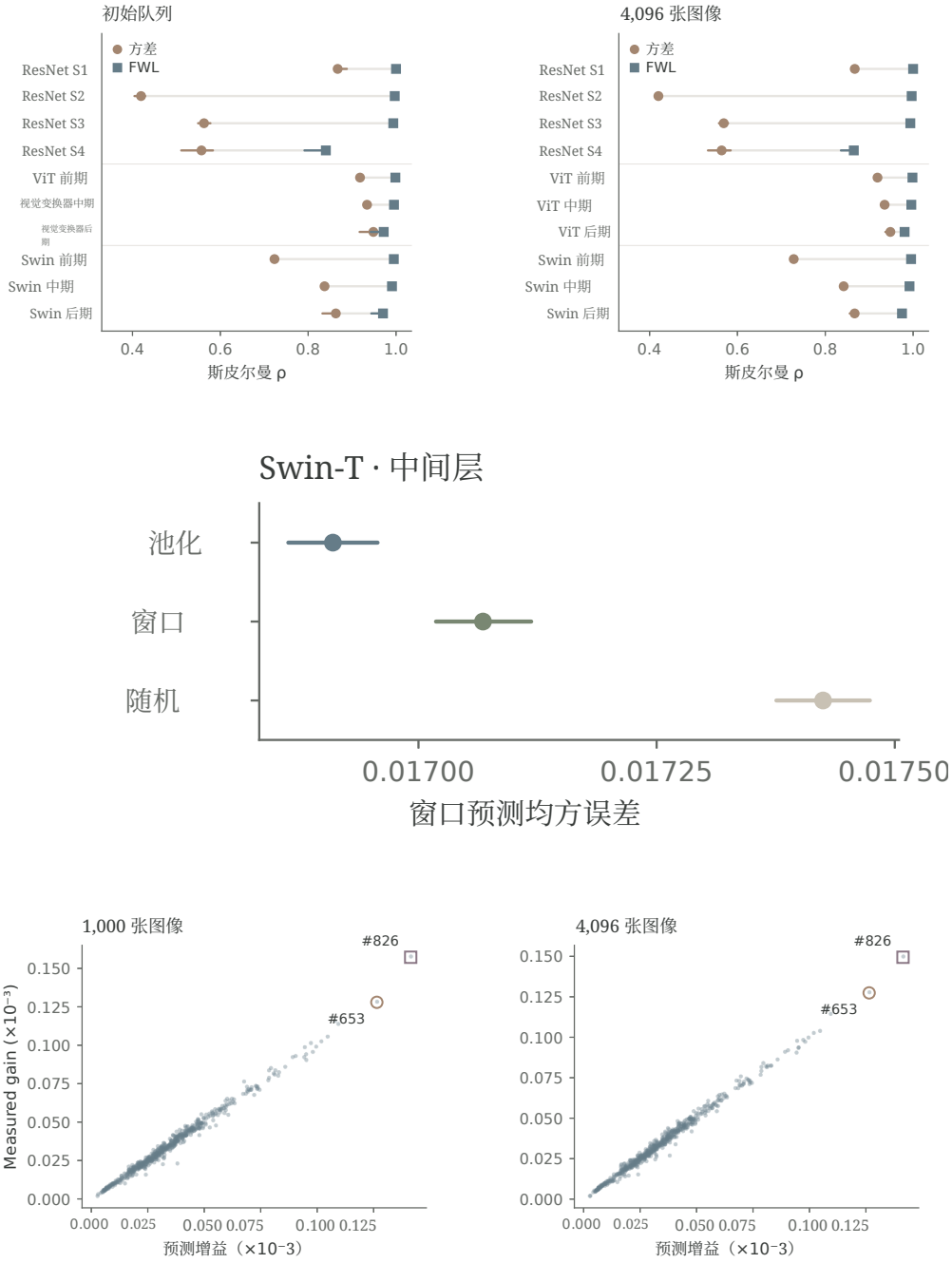


图 S26：初始与最终剪枝排名（顶部）、Swin 窗口分层（中部）以及 Swin 候选增益云图（底部）。 顶部：初始（2，048/1，000 张图像）与最终（4，096 张图像）的候选排名。中部：Swin-T 中间块 OLS 均方误差，对比全局共享 OLS（0。016910）、九个真实移位窗口（0。017068）和等容量随机分层（0。017425）。底部：初始（1，000 张图像）与最终（4，096 张图像）Swin-T 中间剪切点中的逐通道均方误差降幅（每个输出坐标 10^{-3}），突出显示方差选定（#653）和 FWL 选定（#826）的通道。

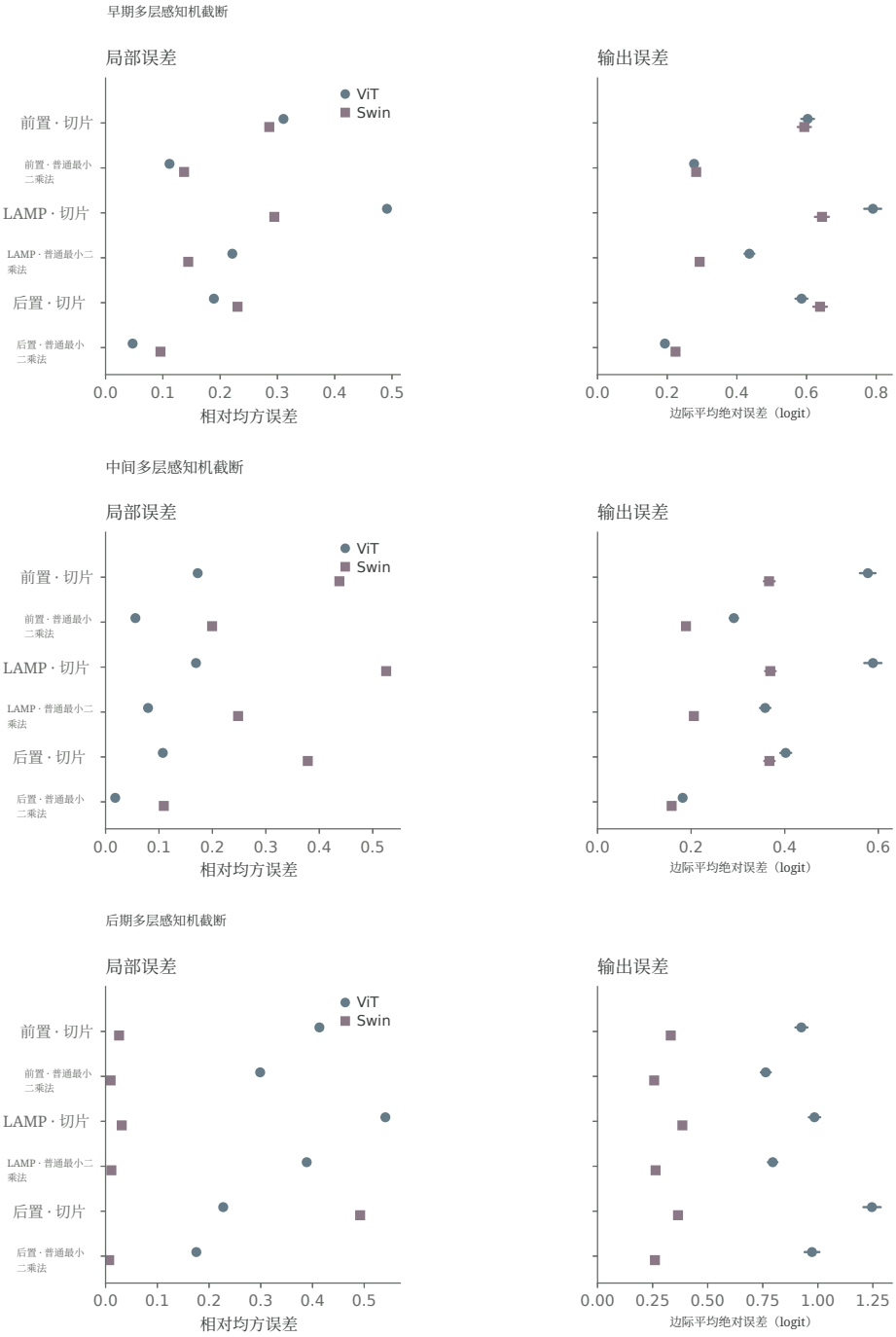


图 S27：在早期（上）、中间层（中）和后期（下）的 Transformer 多层感知机切分点上，六臂相对局部均方误差和类别边际平均绝对误差（4，096 张ImageNet 图像）。比较 Pre-GELU、LAMP 和 Post-GELU 选择器在直接切片与 OLS 权重折叠下的表现（95%配对图像自助法置信区间）。

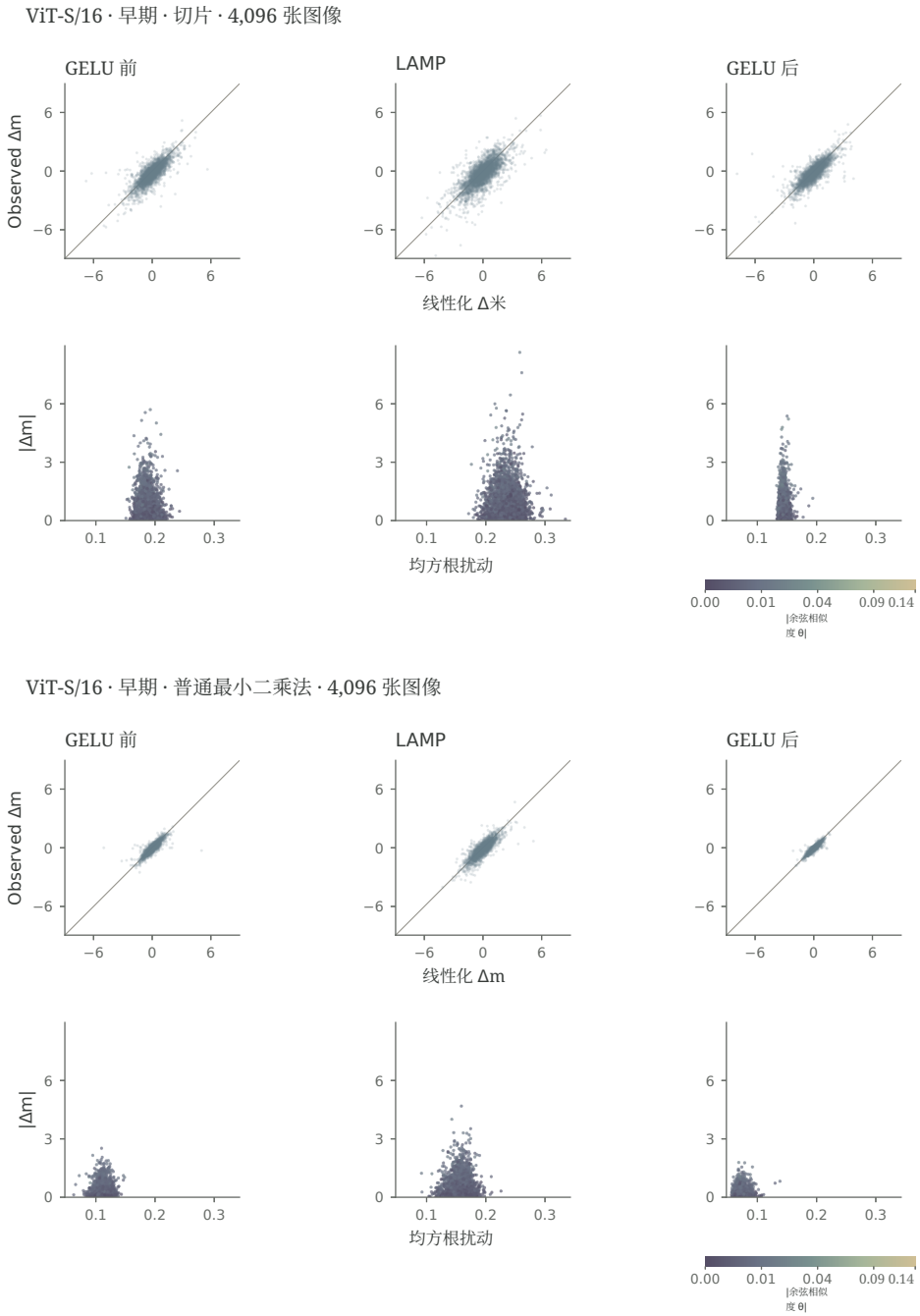


图 S28：在 ViT-S/16 早期多层感知机切口处进行逐图像物理干预（4，096 张 ImageNet 图像）。在 GELU 前、LAMP 和 GELU 后选择器上比较直接切片（上）与 OLS 折叠（下）。颜色表示 $|$局部扰动与干净边际梯度之间的余弦相似度 $\theta|$，在所有干预图表共享的统一平方根尺度上编码 $[0, 0.14]$。

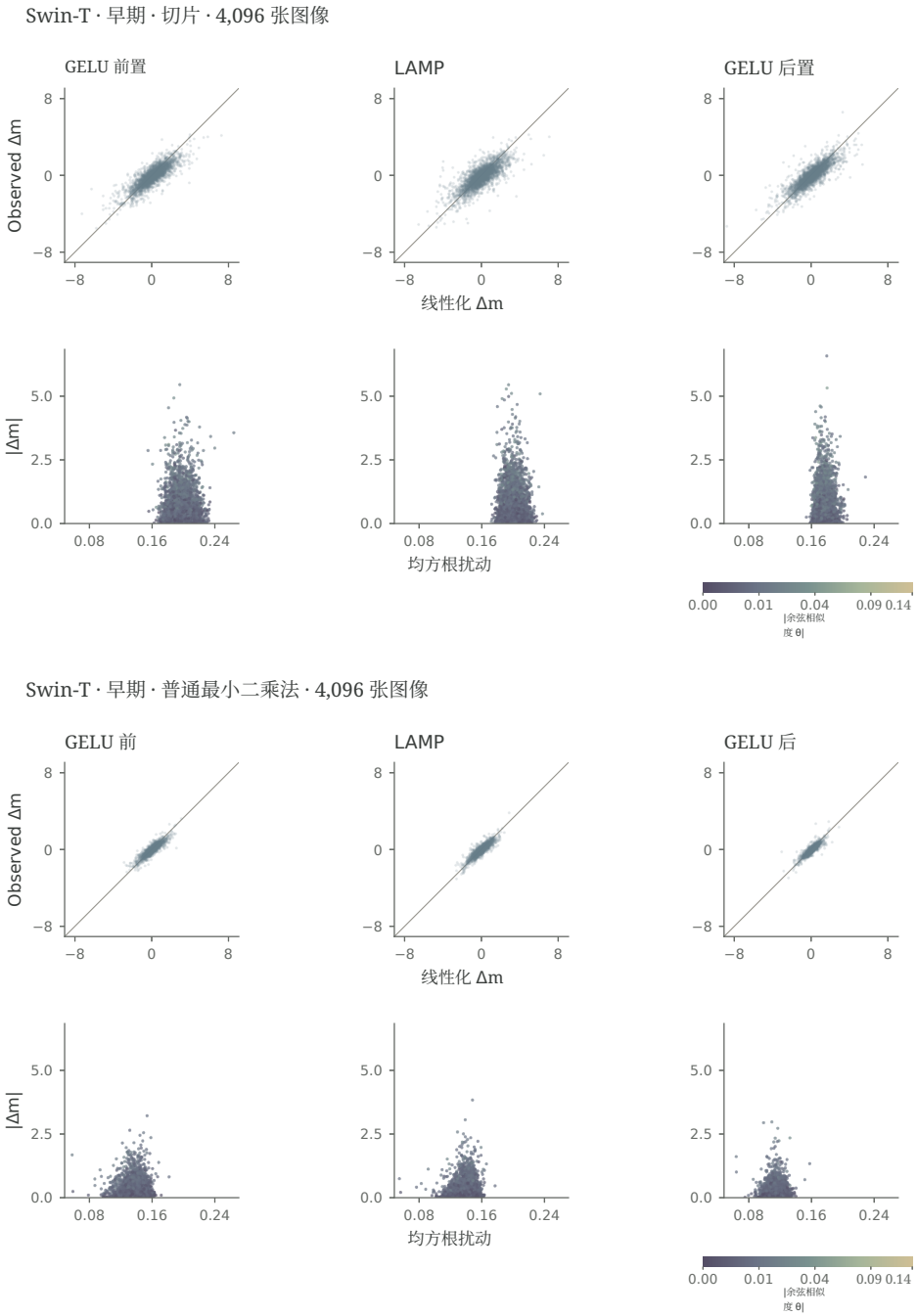


图 S29：在 Swin-T 早期多层感知机切口处进行逐图像物理干预（4,096 ImageNet 图像）。在相同的平方根尺度上比较直接切片（上）与 OLS 折叠（下） |余弦相似度 θ| 尺度。

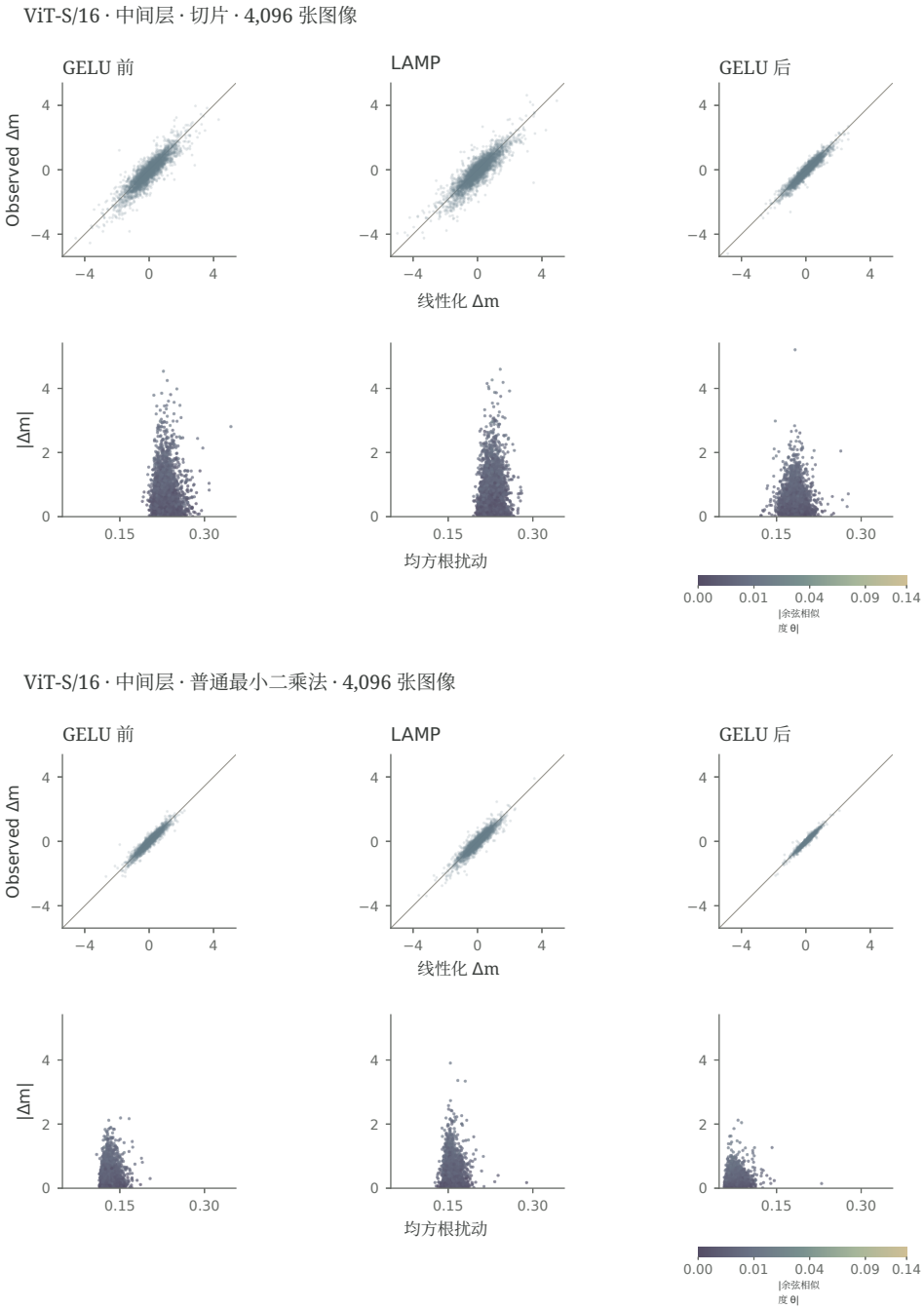


图 S30：在 ViT-S/16 中间多层感知机截断处进行逐图像物理干预（4，096 张 ImageNet 图像）。在相同的平方根 |余弦相似度 θ| 尺度下，对比 Pre-GELU、LAMP 和 GELU 后选择器的直接切片（上）与 OLS 折叠（下）。

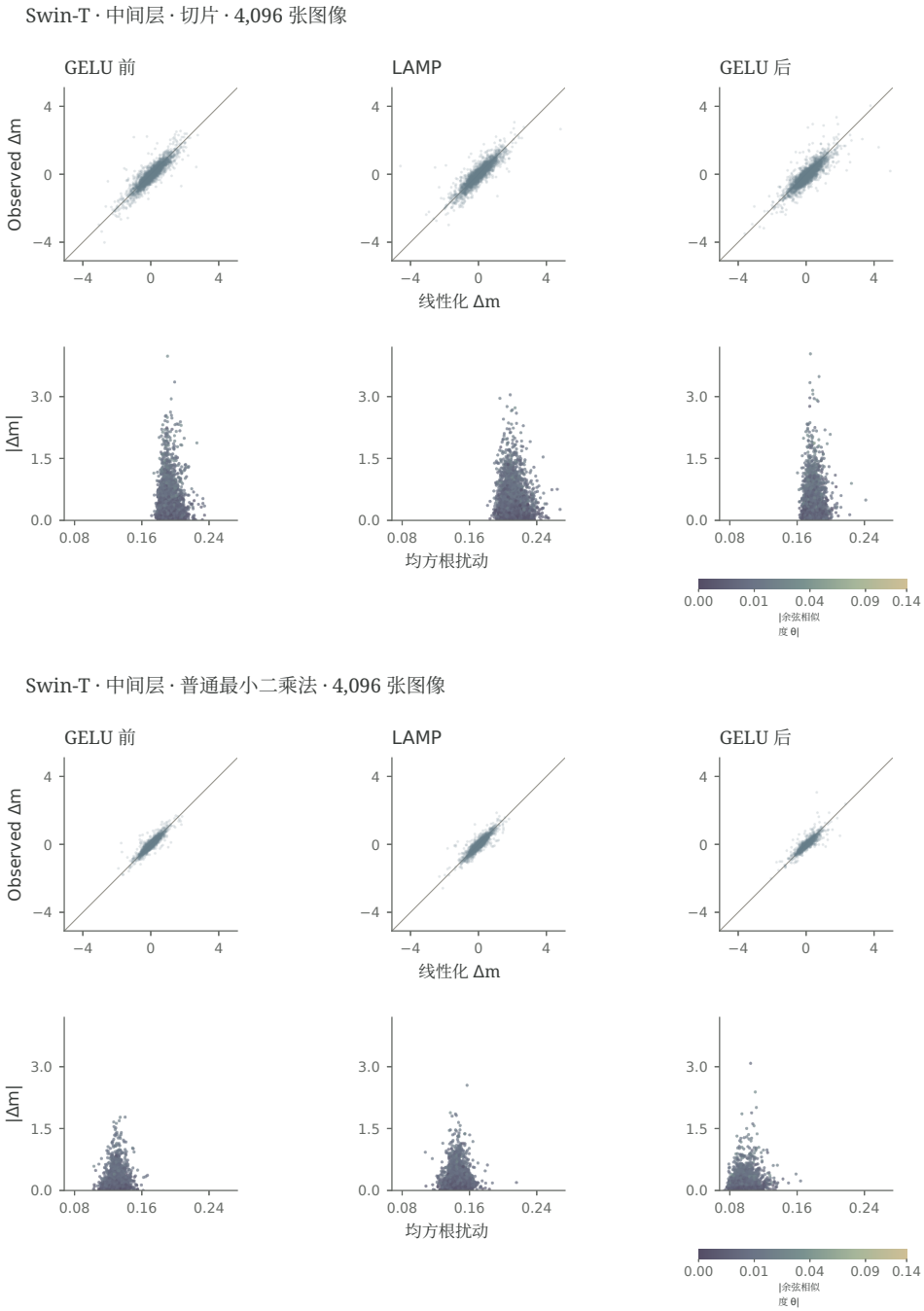


图 S31：在 Swin-T 中间多层感知机截断处进行逐图像物理干预（4，096 张 ImageNet 图像）。在相同的平方根 |余弦相似度 θ| 尺度下，比较直接切片（上）与 OLS 折叠（下），直观展示了 GELU 后的直接切片如何压缩均方根扰动，同时提高 |余弦相似度 θ|，而 $|\Delta m|$ 基本保持不变，直至与 OLS 折叠配合使用。

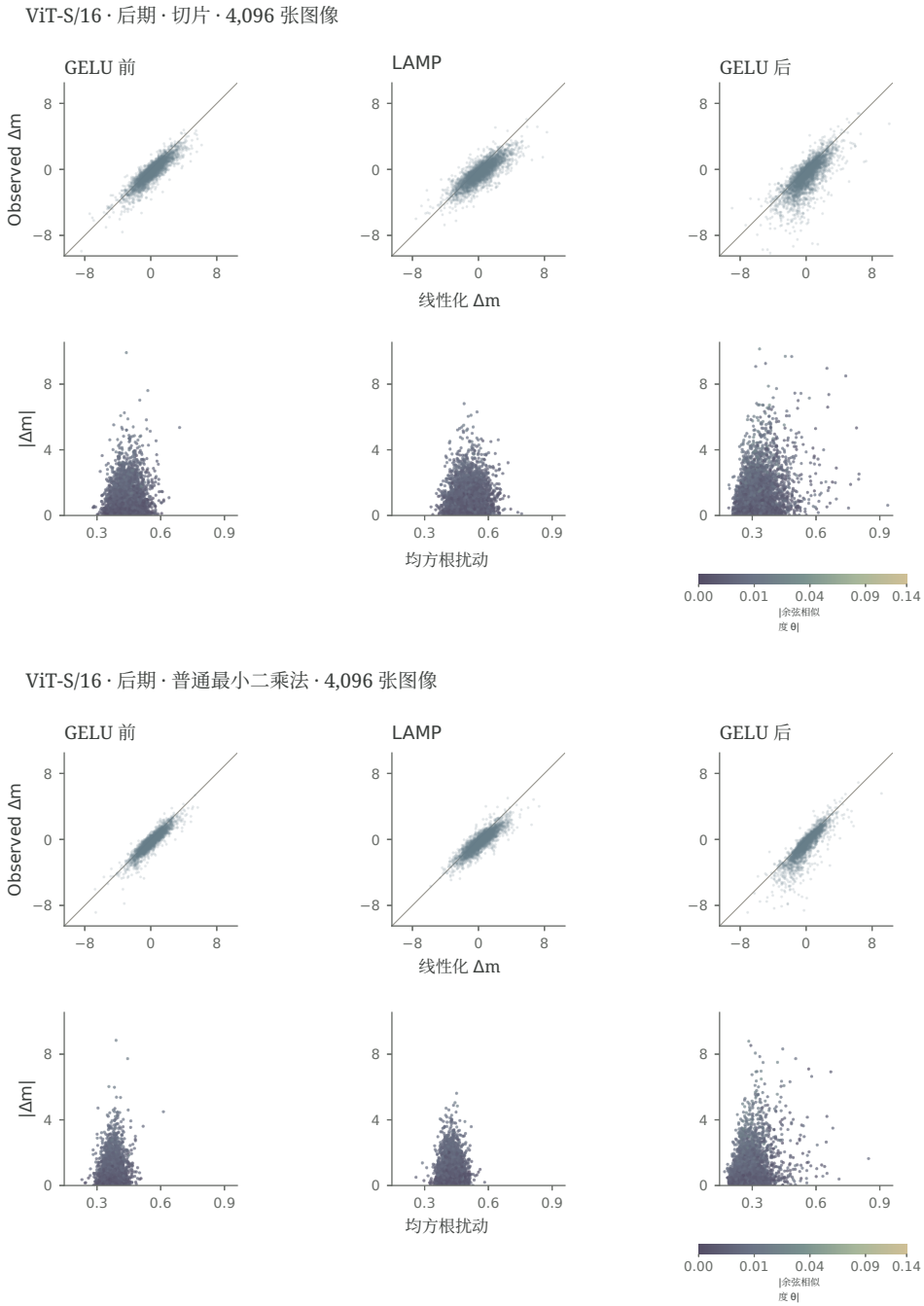


图 S32：在 ViT-S/16 后期多层感知机切口进行逐图像物理干预（4，096 张 ImageNet 图像）。在相同的平方根 |余弦相似度 θ| 尺度上比较直接切片（上）与 OLS 折叠（下）。

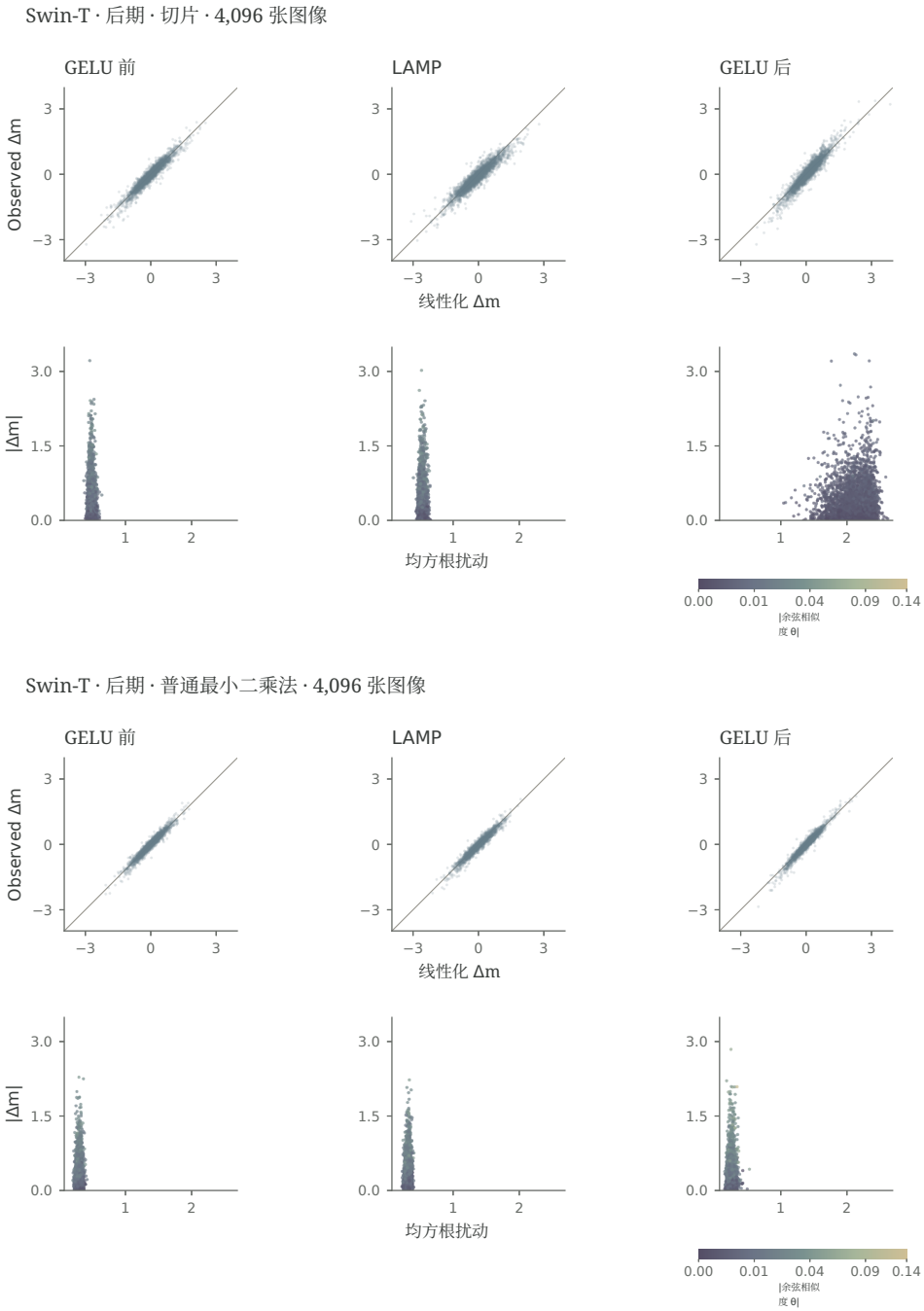


图 S33：在 Swin-T 后期多层感知机切口进行逐图像物理干预（4,096 张 ImageNet 图像）。在相同的平方根 |余弦相似度 θ| 尺度下，比较直接切片（上）与 OLS 折叠（下）。

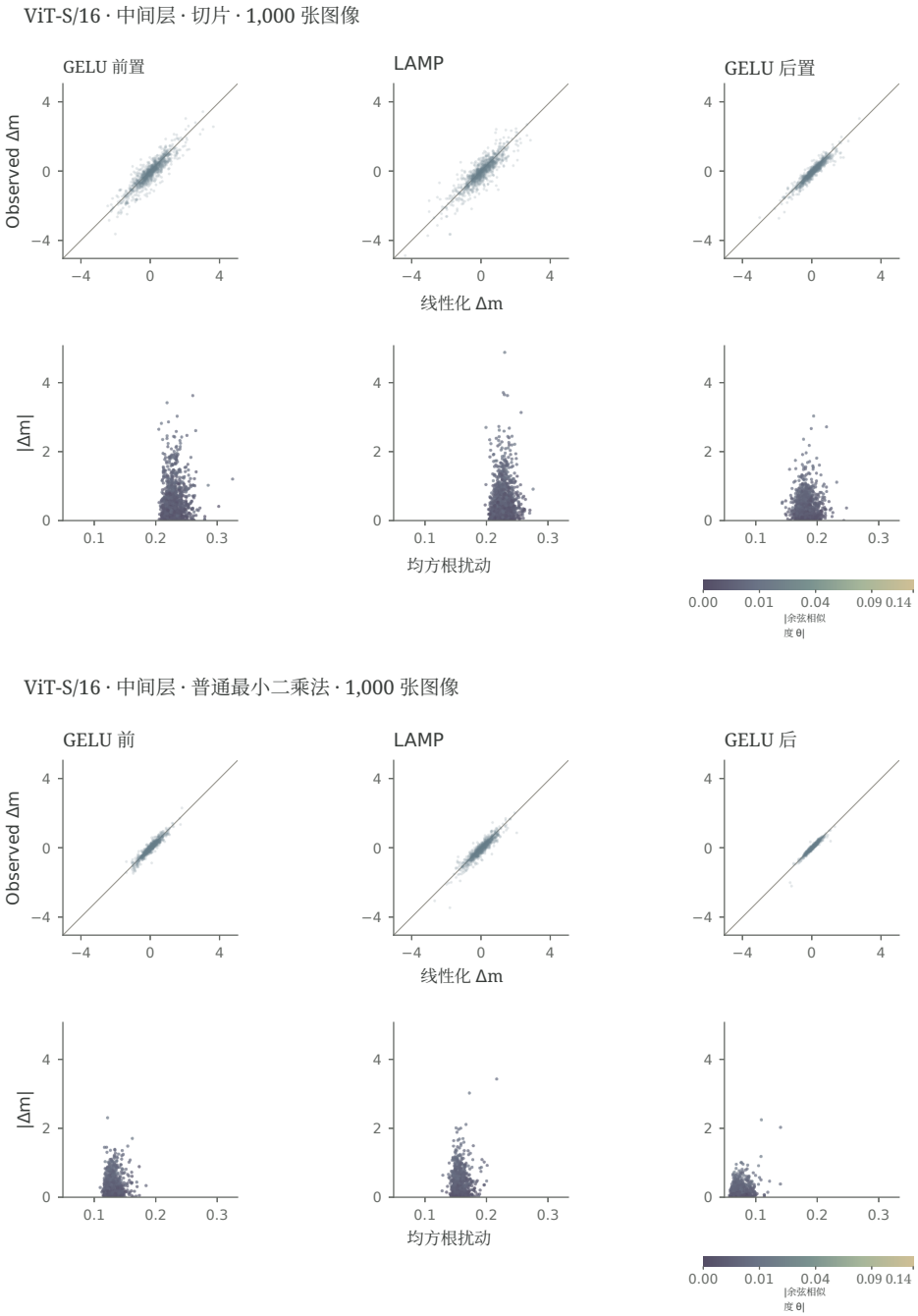


图 S34：历史1,000张图像——在 ViT-S/16 中间多层感知机截断处进行物理手术。在相同的平方根 | 余弦相似度 θ| 尺度上，对比直接切片（上）与 OLS 折叠（下），与4,096张图像的确认云相匹配，见图 S30。

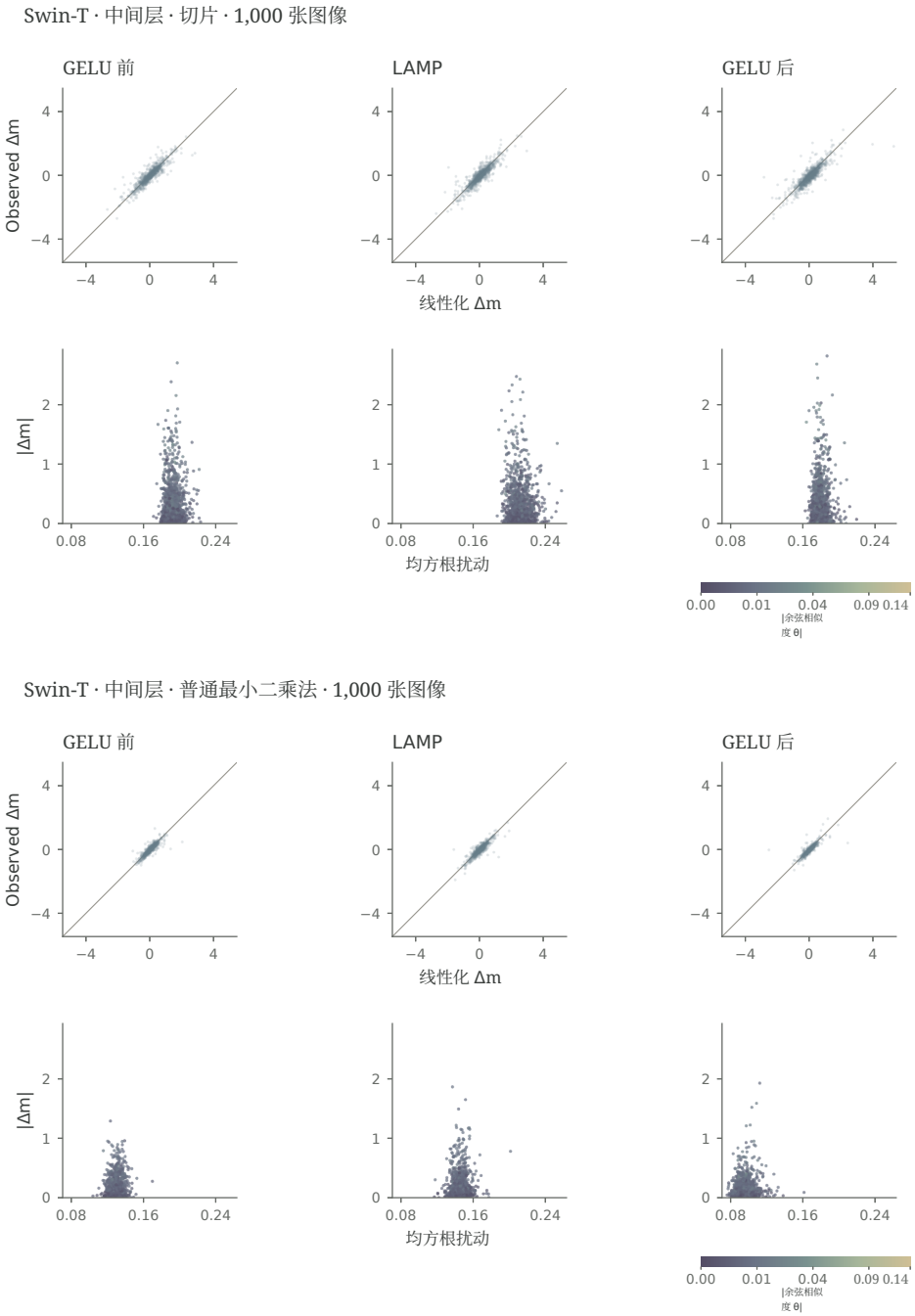


图 S35：历史上的 1,000 张图像规模的 Swin-T 中间多层感知机切口物理手术。 在相同的平方根尺度上进行直接切片（上）与 OLS 折叠（下），$|$余弦相似度 $\theta|$ 尺度，与 4,096 张图像的确认云相匹配，见图 S31。

S13 OOD 腐蚀现象与精确 $R = 1 + bc$ 分解

我们评估了 Oxford-IIIT Pets ResNet-18 在七个工作点上的表现——一个共享的干净基线，以及严重程度为 1 、 3 和 5 的高斯噪声和高斯模糊 —— 并固定干净样本的 top-1 类别、头部权重符号划分，以及每张源图像的四次供体编辑（表 S11、图表 S36 和S37）。为透明地记录研究时间线，最初的启发式预期认为，腐蚀会削弱组间背景抵消（使 $R^{(2)}_{\text{背景}}$升高）；然而，两个独立的确认队列（128 张和 444 张图像）在噪声和模糊条件下都显示 $R^{(2)}_{\text{背景}}$高度一致地下降，因此我们事后提出了对称乘积分解 $R^{(2)} = 1 + bc$(公式 (S13))。在最高严重程度的噪声条件下（noise5），相对 $R^{(2)}_{\text{背景}}$和绝对相干背景能量均下降（$0.00351 \to 0.00150$）；在最高严重程度的模糊条件下（blur5），相对 $R^{(2)}_{\text{背景}}$下降（$0.906 \to 0.619$），尽管绝对相干背景能量上升（$0.00351 \to 0.00559$），且 top-1 准确率骤降（87.6% $\to$ 25.9%）。几何分解将这一变化定位在固定读出处（83% 为角度对立，17% 为范数均衡）；较早卷积块内部的上游因果机制仍有待研究。

表 S11：Pets ResNet-18 腐蚀汇总（444 确认图像，七个工作点）及事后 $R = 1 + bc$ 分解。

条件	Top-1 准确率	类别翻转	前景 $R^{(2)}$	背景 $R^{(2)}$	背景 $\Delta R^{(2)}$ 对比干净	背景相干能量
干净	87。6%	0.0%	1.089 [1.054, 1.125]	0.906 [0.868, 0.944]	0.000	0.00351
噪声等级 5 ($\sigma = 0.15$)	34。7%	63。7%	0.777 [0.732, 0.822]	0.593 [0.554, 0.633]	-0.313 [$-0.365, -0.263$]	0.00150
模糊等级 5 ($\sigma = 3.2$px)	25.9%	74。8%	0.897 [0.858, 0.935]	0.619 [0.578, 0.661]	-0.287 [$-0.343, -0.232$]	0.00559

等权最大腐蚀背景 $R = 1 + bc$ 分解： 在 444 张图像队列中，总偏移 -0.2998 [$-0.347, -0.254$] 分解为角度 -0.2495 [$-0.2975, -0.2037$]（83.2%，c：$-0.140 \to -0.452$）以及范数平衡 -0.0503 [$-0.0632, -0.0379$]（16.8%，b：$0.676 \to 0.829$）。在 128 张图像队列中，总偏移 -0.2959 [$-0.378, -0.213$] 分解为角度 -0.2527 加平衡 -0.0432。

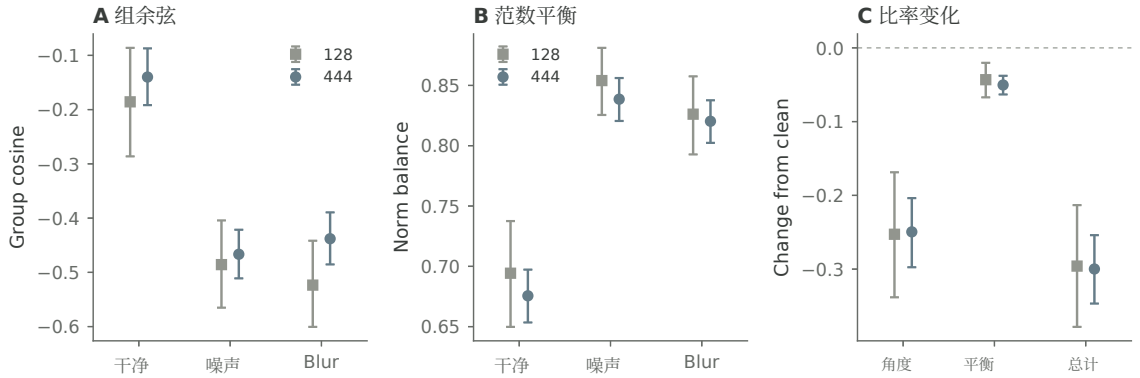


图 S36：双组腐化几何及 $R = 1 + bc$ 在 128 张和 444 张图像 Pets 队列中的分解。 组余弦、范数平衡，以及等权最大腐蚀背景偏移的角度与平衡的对称贡献（-0.2998 分解为 -0.2495 角度和 -0.0503 平衡）。

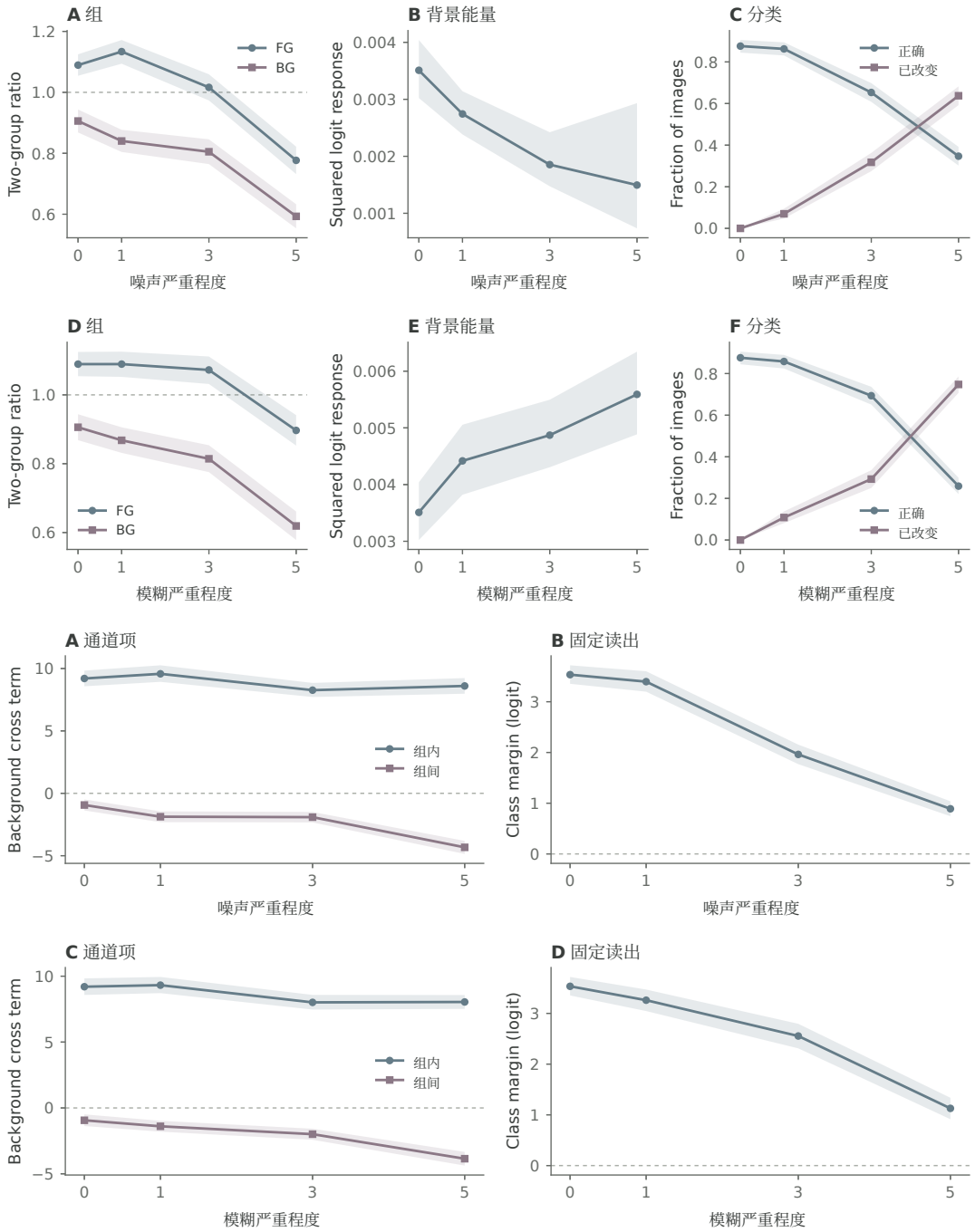


图 S37：分布外腐蚀严重程度曲线（顶部）以及具有固定干净边际的细尺度交叉项（底部）。顶部：444 张 Pets 图像在严重程度（0、1、3、5）下的双组比例曲线、相干背景能量（在噪声下下降，在模糊下上升）以及准确率/翻转率。底部：噪声和模糊下 512 个通道的组内/组间项及固定干净边际。

S14 源坐标反演观测边界

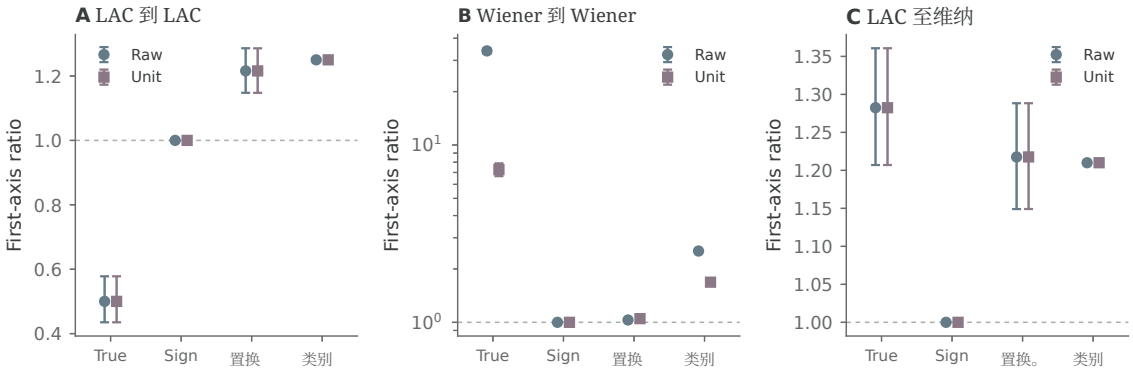
图表 S38 至 S40汇集了支持 公理 2、定理 3和 第 5.4 节的所有实证边界测试：

1. **LAC 系数敏感性与较弱的原生对应关系（444 张宠物图像）：** 将 LAC 系数从 $\nabla_z \odot z$ 调整为 ∇_z，全图前景能量占比从 56.62% 增至 66.06%（配对$+9.43\,[8.52, 10.33]$ 个百分点），但 LAC 与原生 $R^{(2)}$ 之间的逐图秩相关仍然较弱（前景为 0.114；背景为 0.088，置信区间跨越零）。

2. **共享轴去共享残差四阶段级联（Shu 的公式 15 [46]；1,020 张 符合条件的 CUB 图像）：** 残差选择通道在全部四个原生端点上均优于强度匹配的随机对照： 7×7 空间 RMS $+0.02812$、GAP响应 $+0.01096$无符号通道权重和 $+0.00362$，以及匹配组与不匹配组的绝对间隔恢复幅度之差（先计算各组内的有符号总和，再取恢复后的间隔变化的绝对值，并减去强度匹配的随机对照： $+0.00199\,[0.00026, 0.00373]$，而去共享与全图之间的差异 $+0.00122\,[-0.00102, +0.00345]$ 仍未能确定）。在32张图像的开发集划分上校准有符号分组预测（冻结投影斜率为 0.03187），在 $+0.0485\,[0.0114, 0.0859]$ 张1,020 张留出图像上，逐图像组级有符号皮尔逊相关的均值为。

3.**定理3 交换缺陷（444 宠物图像）：** 原生协方差加权梯度 $\Psi = \Sigma_X \nabla_x m_q$ 有符号像素编辑的秩为八，斯皮尔曼相关系数为0。9936，而全场维纳方法达到$0.0139\,[-0.0272, +0.0556]$（维纳方法减去打乱后的 $+0.0155\,[-0.0287, +0.0600]$，平均归一化交换缺陷 $\|\Delta_{\mathrm{comm}}\|_2/\|\Psi\|_2 = 1.0001\,[0.9992, 1.0011]$）。

4. **跨算子首个共享轴边界（存档H-12检查）：** 在444张Pets图像上，LAC $\to$ LAC得到 $R_b = 0.500\,[0.435, 0.578]$（通过了符号翻转、置换和错误类别零假设检验），而维纳方法 $\to$Wiener具有构造性（Pets上为33。93；CUB上为25。58）；将LAC权重跨投影到同图像的维纳首轴上，得到 $R_b = 1.282\,[1.207, 1.361]$，尽管无符号轴余弦相似度很高（匹配时为0。585，不匹配时为0。102）。



图S38：跨算子首个共享轴对照（444 宠物图像）。LAC $\to$ LAC、维纳 $\to$ 维纳和交叉投影 LAC $\to$ 维纳相对于符号翻转、权重置换和错误类别零模型的原始振幅与单位振幅首个共享轴干扰比。

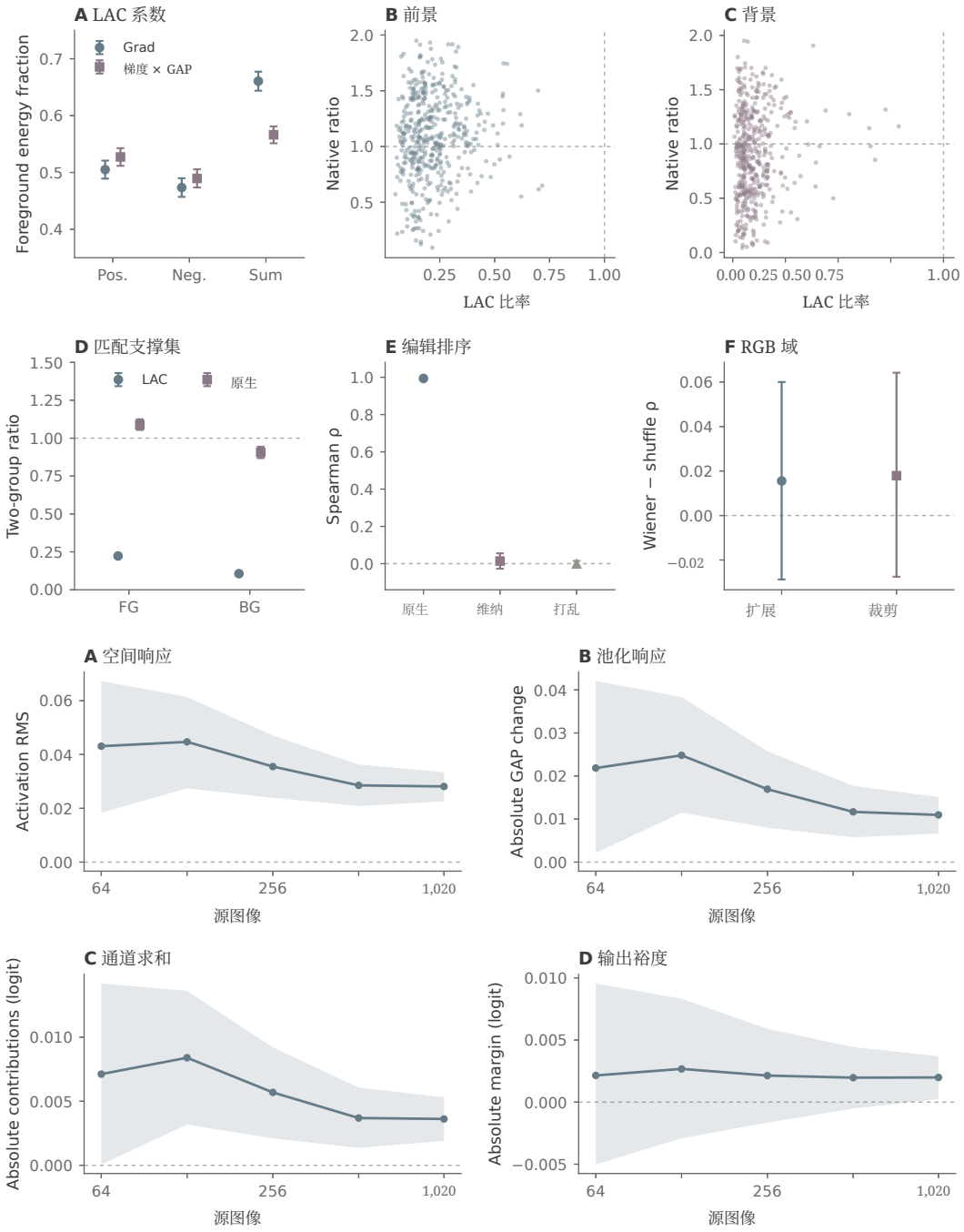


图 S39：源坐标反演边界（上）和 CUB 四阶段原生级联（下）。上：Pets 444张图像的 LAC 系数切换、LAC 与原生方法的散点图、支撑集比较以及有符号像素编辑排序（原生方法 Ψ 与全场维纳方法的对比）。下：去共享通道的四阶段原生响应级联（Shu 的公式 15 [46]），适用于嵌套 CUB 前缀，最多包含1,020 张符合条件的图像。

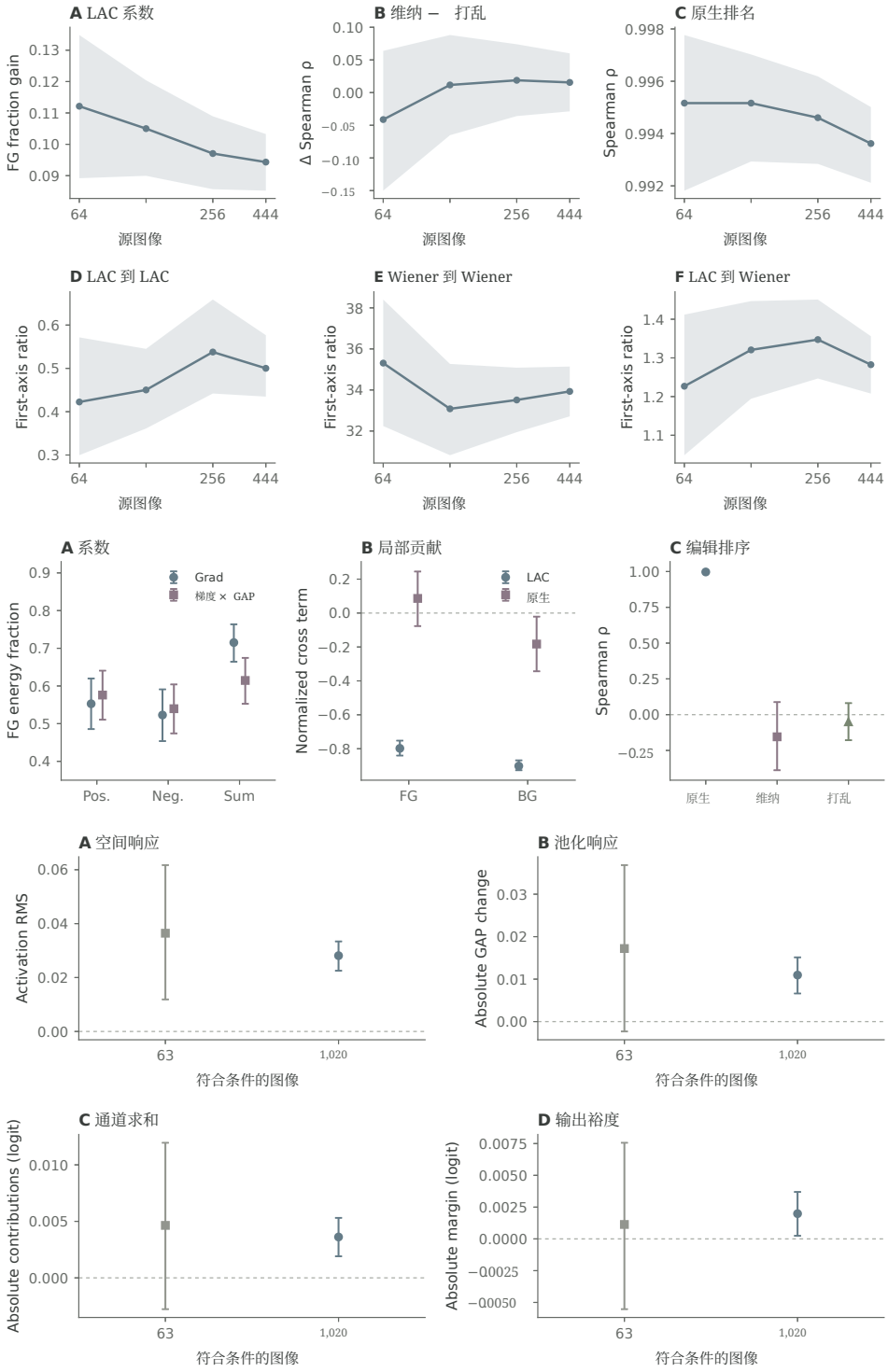


图 S40：反演样本量稳定性与队列比较。顶部：用于 LAC 偏移、维纳减乱序排序、原生 Ψ 排序和原始第一轴比率的嵌套 Pets 前缀（64 至 444 张图像）。中间：原始16张图像来源 Pets 的探索性反演测量。底部：在全部四个原生端点上，对比原始63来源与最终1，020来源 CUB 的区域敏感性。