arXiv:2602.01167v1 [cs.AI] 1 Feb 2026

所有单独层都有帮助吗？视觉语言模型中任务干扰层的实证研究

Zhiming Liu[1] Yujie Wei[2] Lei Feng[3] Xiu Su[4] Xiaobo Xia[5] Weili Guan[1] Zeke Xie[6] Shuo Yang[1]✉

[1]哈尔滨工业大学（深圳） [2]哈尔滨工业大学 [3]东南大学[4]中南大学 [5]新加坡国立大学[6]香港科技大学（广州）

2023年312314@stu.hit.edu.cn,
shuoyang@hit.edu.cn,https://mikuz12.github.io/Do_全部_个体_层_有帮助吗/

摘要

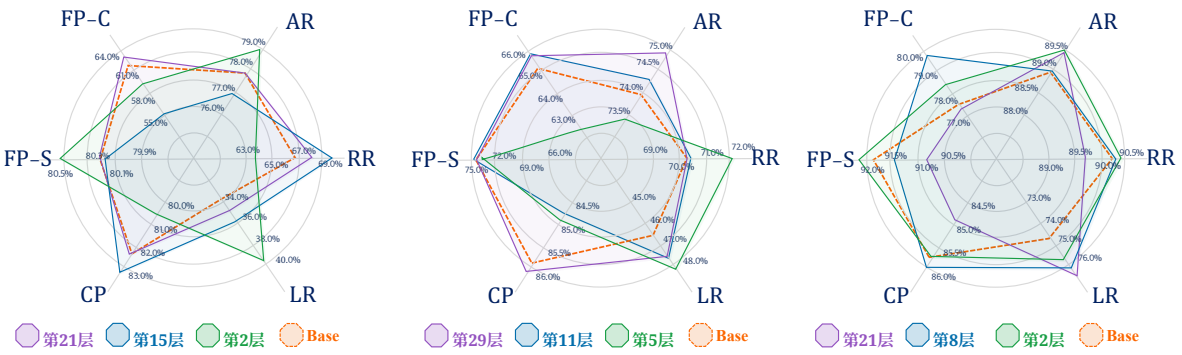
当前视觉语言模型（*VLMs*）已在广泛的多模态任务中展现出卓越的能力。通常，预训练视觉语言模型默认会启用所有层来对下游任务进行预测。令人惊讶的是，我们发现，对单层进行干预（例如将其参数置零）能够提升某些任务的性能，这表明有些层非但无助于下游任务，反而会妨碍其表现。为了理解这种现象何时发生以及为何发生，我们通过层干预（例如参数置零）系统地研究各个层如何影响不同任务。具体而言，我们衡量对每一层进行干预后相较于基础模型的性能变化，并观察到绕过特定层时性能有所提升。这种提升可能在不同模型和数据集间泛化，表明存在会损害下游任务性能的任务干扰层。为进一步分析这一现象，我们提出任务-层交互向量，用于量化给定任务下对视觉语言模型各层进行干预所产生的影响。重要的是，这些任务干扰层呈现出任务特定的敏感性模式：需要相似能力的任务在层干预下表现出一致的响应趋势，这一点可由其任务-层交互向量的高度相似性所证实。受这些发现启发，我们提出 *TaLo*（任务自适应层剔除）（*TaLo*），这是一种无需训练的测试时适应方法，可动态识别并绕过给定任务中干扰最大的层，以验证并落实我们的观察结果。在不更新参数的情况下，*TaLo* 能够持续提升各种模型和数据集上的性能，甚至使 *Qwen-VL* 在 *ScienceQA* 的地图任务上的准确率提高多达 *16.6%*，为这一现象的实际影响提供了概念验证

。我们的工作揭示了预训练视觉语言模型中一种出乎意料的模块化形式，并提供了一种即插即用、无需训练的机制，可在推理时释放隐藏能力。源代码将公开发布。

1. 引言

视觉语言模型（VLMs）凭借强大的跨模态理解与生成能力，在医学 [25, 26, 50]，自动驾驶 [16, 41]，和创意产业 [19, 45]，等多个领域取得了显著成功。在实际部署中，人们通常认为视觉语言模型中的每一层都会被积极利用，因此有理由使用完整模型，并在推理过程中执行完整的计算流程 [53]。然而，我们的实证研究揭示了一种违反直觉的现象：有选择地跳过预训练模型中的单层，可能会显著提升其在某些任务上的性能。这一观察自然引出了一个根本问题：预训练视觉语言模型中的每个单独层，在特定任务中都能发挥积极作用吗？

为了解答这一问题，我们首先引入层干预，以量化层对特定任务的层贡献：如果对某一层进行干预后，任务性能有所提升，我们便推断该层此前妨碍了该任务。具体而言，我们将每一层的自注意力模块置零，同时保留残差连接，从而绕过注意力机制，使该层所学知识失效。如图1所示，在不同模型中，将特定层置零可显著提升某些任务的性能；图2a 则提供了更系统、全面的分析，展示了



(a) Qwen2-VL-2B (b) LLaVa-NEXT-8B (c) InternVL-40B

图1。**任务干扰层现象概述**。每个坐标轴对应一个任务类别：属性推理（AR）、关系推理（RR）、逻辑推理（LR）、粗粒度感知（CP）、细粒度感知 [单实例]（FP-S）以及细粒度感知 [跨实例]（FP-C）。每张图展示了将单层置零后的模型性能（实线），橙色虚线表示基线性能（未进行干预）。在若干任务中，移除层后性能有所提升，这直接证明了任务干扰层的存在。

不同层受到干预时，每项任务的性能变化都呈现出特征性模式，反映了任务对层功能的任务特定敏感性。随后，我们引入**任务-层交互向量**，以进一步形式化这一分析。该向量通过衡量层干预下的性能变化，量化任务与各模型层之间的交互。将每项任务编码为这样的向量后，其对层操控的独特敏感性便成为可计算、可比较的表征。在此基础上，对这些交互向量之间的相关性进行聚类分析（见图2b）发现，需要相似能力的任务（如数学推理）具有高度相似的任务-层交互向量。这表明，模型内部的功能组织与其所执行任务的认知属性之间存在密切关系，也为揭示模型功能专门化背后的机制提供了线索。

尽管这一现象凸显了视觉语言模型内部复杂的功能组织，但目前仍缺乏系统性的探索。近期研究指出，对模型中的层进行干预（例如参数置零或均匀缩放）会改变其通用能力 [6, 56]。然而，这些研究主要关注模型总体性能的下降，却忽视了某些下游任务能力同时增强的现象。更重要的是，现有研究不仅未能透彻理解这一现象，也忽略了其潜在价值。我们的工作旨在弥补这一关键空白。我们的关注点不仅在于识别我们所称的**任务干扰层**（即其存在会主动限制模型在特定任务上的潜力的层），还在于揭示这一现象背后的机制，并探索其实际应用。

基于这些动机，我们进一步提出**TaLo**（**Ta**sk-Adaptive**L**ayer **Ko**ckout，任务自适应层剔除），这是一个极简的免训练框架，旨在作为验证我们观察结果的概念验证。TaLo 会针对给定任务，在推理过程中动态选择要剔除的层，从而有效增强模型在该任务上的特定能力。我们在多个视觉语言模型和基准测试中验证了该方法的有效性。对于 LLaVA [24]，，应用 TaLo 方法使其在 MMMU [54] 的科技与工程任务上性能最高提升 4.7%。值得注意的是，在 ScienceQA [28]，的自然地理任务上，该方法在完全不更新参数或进行额外训练的情况下，实现了令人瞩目的 10.4% 提升，清楚地验证了任务干扰层现象不仅在统计上可观察到，而且具有重要的实际意义。

我们的贡献总结如下：

1. 通过系统的逐层干预，我们观察到，绕过某些层可以提升任务表现。我们将这些层称为**任务干扰层**，即其预训练知识与特定下游任务的目标不一致的预训练组件。2. 我们通过引入**任务-层交互向量**，建立了一个用于分析任务与模型层之间关系的定量框架，从而进一步考察相似任务对层干预是否表现出一致的响应。

3. 我们开发了一种实用、即插即用的算法 **TaLo**，利用这些洞见在测试时提升模型性能，无需更新任何参数。使用此方法，LLaVA 和 Qwen-VL [47] 在涵盖 5 个基准测试的 10 项任务中，性能最高分别提升了 10.4% 和 16.6%。

2. 相关工作

我们的研究处于模型编辑、剪枝和测试时适应（TTA）的交叉领域。我们借鉴模型编辑和剪枝的理念，通过参数干预来调节模型行为，同时提出了一种独特的方法，专注于动态抑制以提升特定任务的表现。我们的方法 TaLo 还通过在推理时跳过模型层，为 TTA 提供了一种新范式。

2.1. 模型编辑与剪枝

模型编辑旨在修改预训练模型的行为或更新知识，而无需全量重训练。方法主要分为两类。第一类涉及直接参数更新，例如通过约束微调来缓解遗忘 [59] 或利用超网络进行动态参数调整 [10]。然而，由于参数规模较大，这些方法难以扩展到大语言模型。MEND [34] 通过低秩梯度分解实现高效更新，从而解决了这一问题。第二类是定位并编辑，它识别关键参数（例如“知识神经元”）并施加针对性修改 [9, 32, 33]。尽管该方法提升了可解释性，但往往劳动密集且可扩展性有限。另一些方法则保持原始参数不变，并使用辅助模块进行编辑 [35]。在多模态场景中，编辑视觉语言模型（视觉语言模型）需要独特的策略。直接移植大语言模型方法效果不佳；相反，近期工作 [5] 提出通过识别并编辑与目标提示最相关的区域来操控中间视觉表征，从而在保持效率的同时尽量减少对无关特征的干扰。

模型剪枝 不同于模型编辑。它是一种用于模型压缩和加速的技术，通过移除冗余或不重要的组件（如权重、神经元或整个层），在保持性能的同时减小模型大小并提升推理速度 [8, 13, 17,30, 36, 40, 52]。其主要目标是优化效率，而非增强或纠正模型的知识。

相比之下，我们的方法 TaLo 旨在通过在推理期间暂时抑制有害的推理路径来提升任务表现，而不会永久改变模型。这与模型编辑和剪枝都有根本区别。值得注意的是，TaLo 不采用基于梯度的微调或细粒度参数操控[1,58], ——这两者在推理时计算成本高且效率低下——而是通过可逆的测试时干预，在块级别进行操作。我们的结果表明，性能提升不仅可以通过添加或移除模型组件来实现，也可以通过在推理期间有策略地抑制现有组件来实现。

2.2. 测试时适应

测试时适应（TTA）旨在使模型在推理期间动态适应不断变化的数据分布，这是在真实场景中实现稳健部署的关键步骤。现有方法通常使用测试批次 [21, 38, 44, 46, 51], 更新权重或归一化统计量等模型组件；或者，尤其对于视觉语言模型，微调可学习提示 [14, 39, 55]。其他方法 [22] 则利用测试时特征缓存执行零样本分类。我们的工作提出了一种基于层干预的独特 TTA 范式。在一小组无标注测试样本的引导下，TaLo 识别出一个会损害目标任务表现的固定层，并将其置零，以应用于该任务后续的所有实例。重要的是，这种自适应是在任务级别进行且保持静态：一旦确定干预方案，在推理期间无需针对每个实例做出决策，也无需进行参数更新。这种无需训练的策略避免了以往方法中固有的大规模参数更新和复杂的提示修改。因此，它能够以极低的计算开销实现高效的即插即用适配，同时保持原始模型参数完整，并可在不同任务之间重复使用。

3. 发现并刻画任务干扰层

我们首先通过参数干预，揭示普遍存在的任务干扰层现象。在此基础上，我们提出任务-层交互向量，以便更系统地进行分析，从而有效发现模式并揭示层级干扰的一致性。

3.1. 发现任务干扰层

我们的目标是分离并量化各个层对特定任务能力的贡献。为此，我们受 Chen 等人[6], Zhang 等人 [56]的启发，采用参数干预，系统地探测每一层的功能作用。具体而言，对于每一层，我们将自注意力模块的参数替换为零或均匀分布（即将每个参数设为相同的值 $1/N$，其中 N 对应于矩阵的第一维），并评估任务准确率相较于未修改模型的变化。将参数置零实际上会使注意力机制失效，仅保留残差连接，从而使相距较远的层能够直接通信，绕过中间变换。至于均匀缩放，它会将复杂的注意力运算简化为对输入特征进行简单的全局平均，使输出变为秩一矩阵。我们的假设是，干预后性能提升表明该层妨碍了任务表现，因此可将其视为任务干扰层。

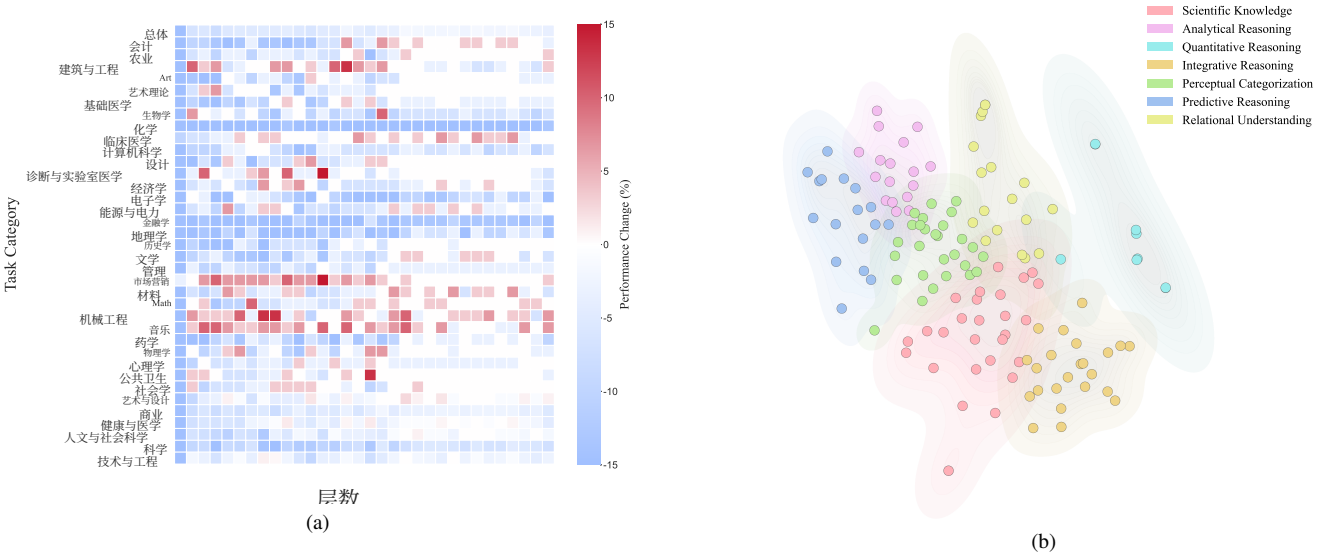


图 2. **任务干扰层的实证验证。(a)** 在 *LLaVA-Next-LLaMA3-8B* 上将每一层置零后，各任务准确率变化百分比的可视化。红色表示相较于基础模型性能有所提升，蓝色表示性能下降。许多任务在层干预下性能有所提升，这表明干扰层在视觉语言模型中普遍存在。 **(b)** 任务簇的 t-SNE 可视化。图中的每个点代表一个以任务-层交互向量 $\mathbf{v}^{(\mathcal{T})}$编码的任务 ρ_{ij} 。根据不同向量之间通过皮尔逊相关系数衡量的两两相似度对任务进行聚类，能力需求相似的任务会形成紧密的簇。不同颜色的簇对应不同类型的任务，表明具有共同认知需求的任务呈现出相似的干预响应，反映了 LLaVA 中结构化的功能布局 [24] (完整的聚类详情见表 11)。

相反，性能下降则表明该层对任务有积极贡献。

随后，我们对 LLaVA-Next [24]，（一个由 32 层组成的模型）实施参数置零干预，并在 MMMU [54] 数据集上评估性能。如图 2a 所示，将单层参数置零后，54.1% 的任务性能提升超过 5%。在其他模型和数据集上也观察到类似趋势：对于 Qwen-VL，这一比例达到 75.6%。为进一步验证研究发现的普遍性，我们在附录 C.4 中展示了针对多种模型、基准测试和干预策略开展的补充实验。这一一致的模式为任务干扰层的存在提供了直接的实证证据：这些层的激活非但无助于任务表现，反而会造成阻碍。

值得注意的是，我们的研究专门针对视觉语言模型的大语言模型主干网络进行干预。视觉编码器（例如基于 CLIP 的视觉变换器）被视为固定的特征提取器，不作修改，因为它已针对视觉表征进行了高度优化。我们假设，任务干扰的主要来源并非感知，而是跨模态推理和任务执行——这些过程主要由大语言模型组件负责。此外，我们有意选择只干预自注意力模块：初步实验表明，对前馈网络模块（即使只在单层中）应用相同的归零干预，也会导致灾难性的性能下降，并且往往会产生语义不连贯的输出。因此，在本文中，“层干预”专指对给定层中的自注意力模块进行修改。

3.2. 刻画任务干扰层

将任务与层的交互建模到向量空间中。为了揭示任务对层干预的系统性响应模式，我们引入了任务-层交互向量。该向量旨在刻画任务性能与施加于模型各层的干预之间的关系。

具体而言，任务-层交互向量是一种表示，用于刻画任务对各个模型层的敏感程度。向量的每个维度对应一个网络层，并捕捉相对于基础模型，对该层进行干预所导致的任务准确率变化。正值表示该层会干扰任务，表现为干预后准确率提升。反之，负值表示该层对任务有积极贡献，表现为修改该层后准确率下降。更正式地，对于任务 $\mathcal{T}$ 和一个具有 L 层的模型，**任务-层交互向量** 定义如下：

$$\mathbf{v}^{(\mathcal{T})} = \left(v_1^{(\mathcal{T})}, v_2^{(\mathcal{T})}, \ldots, v_L^{(\mathcal{T})}\right) \in \mathbb{R}^L, \tag{1}$$

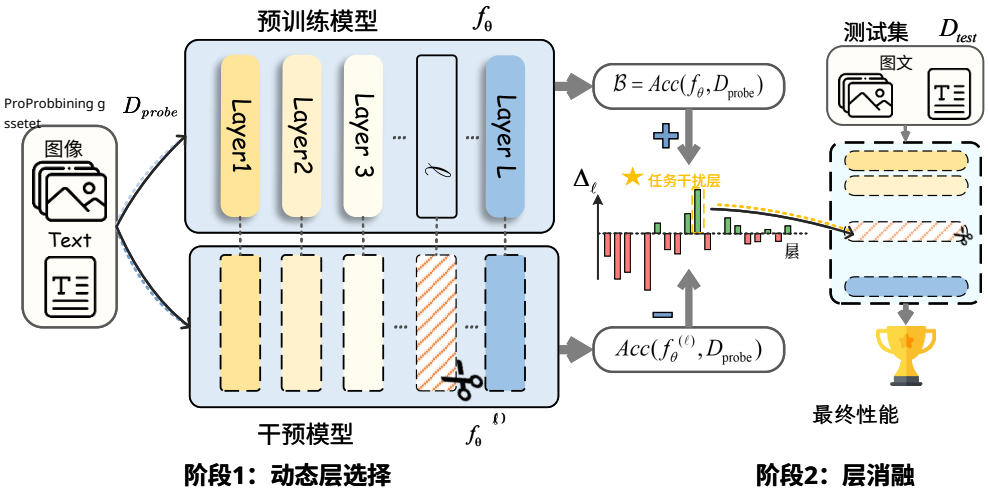


图 3. **TaLo 的框架**。 TaLo 首先针对特定任务动态选择任务干扰层，并在最终评估过程中剔除该层。

其中，每个元素 $v_i^{(\mathcal{T})}$ （称为**层敏感度分数**）用于量化在层 i 进行干预后任务表现的变化。其形式化定义如下：

$$v_i^{(\mathcal{T})} = \text{Acc}\left(\mathcal{M}_{\text{intv}}^{(i)}, \mathcal{T}\right) - \text{Acc}\left(\mathcal{M}_{\text{base}}, \mathcal{T}\right). \quad (2)$$

这里，$\text{Acc}(\cdot, \mathcal{T})$ 表示任务 $\mathcal{T}$ 上的准确率， $\mathcal{M}_{\text{base}}$ 是未经干预的基础模型，而 $\mathcal{M}^{(i)}{}_{\text{intv}}$ 则是对第 i 层的参数进行干预后的模型。

通过这种向量表示，我们将每一层对某项任务的影响抽象为高维向量空间中的结构化表示。这为后续的模式分析提供了可量化且可比较的分析工具。

刻画任务-层交互模式。我们假设，依赖相同底层认知技能（例如数值推理或算术推理）的任务应当会调用模型内部相似的处理路径。由于任务-层交互向量$\mathbf{v}^{(\mathcal{T})}$能够捕捉任务对各层的依赖关系，反映其在模型架构中的传播过程，因此我们假设相关任务会呈现高度相关的交互向量。

为验证这一假设，我们对 6 个基准测试中的近 100 项任务进行了系统分析。对于每对任务 $\mathcal{T}_i$ 和 $\mathcal{T}_j$ （下文中，我们用索引 i 和 j 简写本节所述的对应任务），我们计算它们的皮尔逊相关系数 $\rho_{ij} = \text{Corr}(\mathbf{v}^{(i)}, \mathbf{v}^{(j)})$，并将距离度量定义为 $d_{ij} = 1 - \rho_{ij}$。这确保了越相似的任务之间距离越小，为聚类提供了坚实基础。

如图所示2b，结果证实了我们的假设：依赖共同能力的任务会聚集在一起，分布于任务-层交互空间中，反映了它们共同的内部处理机制。这表明，具有相同底层认知需求或特定领域需求的任务，对层干预表现出高度相似的敏感度模式，揭示了模型建筑学中功能依赖关系的结构化组织。例如，一个显著的簇以**定量推理**任务为主（如数值常识、算术推理和几何）。另一个不同的簇则将**科学任务**等特定领域任务归为一组，例如物理学 和科学推理，体现了它们对形式化科学知识的共同依赖。这表明，模型对层干预的内部响应与任务的底层认知结构高度一致，也揭示了任务干扰层能够泛化到具有相似能力的任务上。这种泛化表明，层敏感度由功能需求决定，而非任务特定属性，因此可以根据少量代表性样本可靠地估计干扰层。

不同干预方法之间的一致性。为验证任务干扰层现象并非方法上的假象，我们考察了两种不同干预策略下研究发现的一致性。对于每项任务 $\mathcal{T}$和层 i，我们分别测量模型在两种干预类型下的准确率，得到两个性能分数：$\text{Acc}(\mathcal{M}_{\text{zero}}^{(i)}, \mathcal{T})$ 和 $\text{Acc}(\mathcal{M}^{(i)}{}_{\text{均匀缩放}}\mathcal{T})$。我们在六个基准测试中进行了这项分析。为了直观呈现两者之间的关系，我们绘制了散点图（图4），其中每个点代表一个任务—层组合，其坐标由均匀缩放（x 轴）和参数置零（y 轴）下的准确率决定。随后，我们计算所有

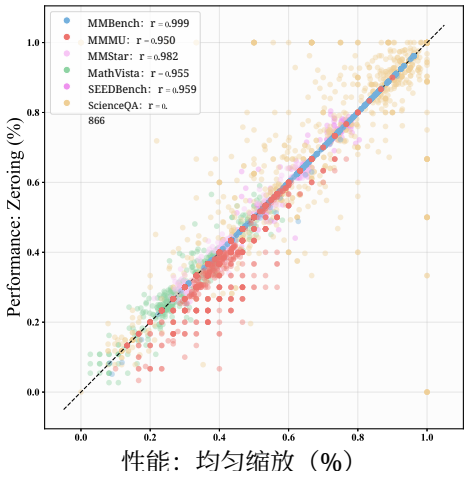


图4. 不同干预的一致性分析。

点在各个基准测试上的皮尔逊相关系数。结果显示，所有基准测试中均存在强烈且具有统计显著性的正相关。这一结果在本质上截然不同的干预策略下均可复现，有力地增强了我们研究发现的可信度。这表明，任务干扰层的存在并非特定干预选择所造成的假象，而是反映了模型在预训练期间由任务冲突引发的内在属性。

4. 任务自适应层敲除

基于在视觉语言模型中发现任务干扰层的结果，并受到动态模型适 [2, 3, 42, 48, 57]，应领域最新进展的启发，我们提出了一种简单而有**效的算法——Task-AdaptiveLayer Knockout**(TaLo)，这是一种无需训练的方法，可在测试时进行任务级模型定制，并作为概念验证，以验证任务干扰层的存在及其实际效用。TaLo 分为两个阶段：（1）针对特定任务动态选择层；（2）从模型中剔除任务干扰层。通过策略性地利用模型中的任务干扰层，该方法能够提升特定任务上的性能。

我们的方法以定向干预为指导原则。目标是找出一个最优层，通过修改该层来释放模型潜在的特定任务能力。搜索首先从确定性能基线开始。对于一个包含 L 层的模型 f_θ（参数为 $\theta = \{\theta_1, ..., \theta_L\}$），以及从给定下游任务中采样的探测集 $D_{\text{probe}} = \{(x_i, y_i)$，在 N 次采样的设置下，我们将未修改模型的基线分数定义为：$\mathcal{BB} = Acc(f_\theta, D_{\text{probe}})$。在我们的实验中，$Acc$ 表示探测集 D_{probe} 上的准确率。随后，我们系统地测试每一层的潜力。对于每一层 ℓ，我们进行干预I（此处，我们采用参数置零，$I(\theta_\ell) = 0$），以创建一个修改后的模型 $f_\theta^{(\ell)}$，并测量由此带来的准确率提升 Δ_ℓ:

$$\Delta_\ell = Acc(f_\theta^{(\ell)}, D_{\text{probe}}) - Acc(f_\theta, D_{\text{probe}}). \quad (3)$$

这一迭代过程揭示了每一层的功能如何为特定任务作出贡献。当我们确定负责带来最大正向性能提升的最优层 ℓ^* 时，搜索便告结束：

$$\ell^* = \underset{\ell \in \{1,...,L\}}{\text{argmax}} \ \Delta_\ell. \quad (4)$$

随后将该层选为任务干扰层。如果没有任何层表现出统计显著的敏感度峰值，我们就保留原始模型，不做任何修改。确定该层后，我们会在推理过程中对其实施干预（即屏蔽）。随后，我们在目标任务的留出测试样本上评估干预模型；这些样本未包含在初始探测集内，从而确保评估结果公正无偏。

5. 实验

5.1. 设置

模型与基准测试。我们在三种不同规模的视觉语言模型上进行了实验：Qwen2-VL-2B [47]，LLaVA-Next-LLaMA3-8B 和 InternVL2-26B [7]。为评估干预的影响，我们在五项重要的选择题基准测试上评估了干预模型和原始模型：MMStar [4],MMBench [27], MMMU [54], ScienceQA [28], 以及 SEED-Bench [23]。有关基准测试和模型配置的更多细节见附录 A。

基线。我们的目标是通过无需训练、即插即用的层干预来提升特定任务的性能。所有比较均针对原始预训练模型进行。在第 5.3 节中，我们还使用与 TaLo 相同的少样本样本对微调方法进行了评估。需要特别指出的是，我们在表格中报告的基线性能反映的是这些特定任务上的准确率。这一性能可能与整个基准测试报告的总体平均准确率存在显著差异。我们使用官方评估工具和标准聊天模板验证了基线实现，以确保比较公平且准确。

任务选择依据。我们没有报告整个基准套件上的平均性能，因为这可能会掩盖任务特定行为，而是选择性地分析各个子任务。我们发现，并非所有任务都会表现出我们所称的“任务干扰层”。如图 2a和附录 C.4中的敏感性可视化结果所示，对于某些任务，热力图主要呈蓝色，表明干预任何一层都会导致性能下降。这证实，在这些特定情形下，不存在需要缓解的显著干扰，因此我们的方法

表1. 将**TaLo** 应用于两个模型（置零）的性能：*LLaVA-Next-LLaMA3-8B*、*Qwen2-VL-2B*在五个基准测试上的结果。MMStar（逻辑推理标注为L-R）、MMBench（物理属性推理标注为PP-R，图像情感标注为I-E）、MMMU（科技与工程标注为T&E，健康与医学标注为H&M）、ScienceQA（自然地理标注为P-G），以及SEEDBench（视觉推理标注为V-R，文本理解标注为T-U）。TaLo在各项任务上的其他结果见表9和表10。样本数表示探针集大小。基础性能可能因任务而异，尤其是在不同样本设置下样本有限的任务。因此，为确保公平评估，我们主要关注相对于基础模型的相对提升。

模型	样本数	MMStar		MMBench		MMMU		ScienceQA		SEEDBench		Avg
		Math	I-R	PP-R	I-E	T&E	H&M	P-G	Maps	V-R	T-U	
LLaVA	10次示例	32.9 ↑2.5	58.3 ↑4.2	55.3 ↑7.8	65.6 ↑0.6	35.2 ↑1.1	42.2 ↑5.1	34.5 ↑6.9	16.7 ↑2.4	70。4 ↑1.3	53.6 ↑7.2	3.91↑
	15样本	33.2 ↑3.0	57.9 ↑4.3	56.0 ↑3.8	69.3 ↑0.7	37.5 ↑8.3	35.3 ↑5.9	**41.4 ↑10.4**	23.8 ↑7.1	70.8 ↑1.1	55.4 ↑9.0	5.36↑
	20次示例	33.9 ↑3.0	58.7 ↑4.3	53.4↑4.8	66.4↓1.5	29.5 ↓3.6	46.2–0.0	34.5 ↑3.5	26.2 ↑9.5	72.5 ↑2.0	55.4 ↑5.4	2.74↑
Qwen-VL	10次示例	44.8 –0.0	54.8 ↓1.9	48.6 ↑8.5	69.4 ↑1.3	28.6 ↑0.6	34.5 ↓1.6	31.0–0.0	31.0 ↑2.4	70.8 ↓1.5	57.1 ↑8.9	1.67↑
	15样本	44.2 ↑0.5	54.2 ↓1.6	55.3 ↑1.2	65.7 ↓0.7	24.4 ↓1.9	35.8 ↓2.9	34.5 ↑3.5	**45.2 ↑16.6**	71.2 ↓2.6	55.4 ↑1.8	1.39↑
	20次示例	38.2 ↓1.2	55.3 ↓0.5	60.3 ↑4.8	67.9 ↑0.7	26.1 ↓0.5	23.5 ↓5.9	37.9–0.0	16.7–0.0	71.3↓0.8	44.6 –0.0	0.34↓

自然不适用。因此，我们特别关注预训练模型表现不佳的任务，因为这些任务往往能揭示模型预训练知识中的偏见或缺失。本研究并非旨在宣称所有任务的性能都能普遍提升，而是希望提供概念验证（如第 6节所述），证明层级干预能够释放模型在这些特定薄弱环节中的潜在能力。

实现细节。所有实验均在单张*80GB A100 GPU* 上、于相同环境下进行，以确保可复现性与公平比较。评估采用标准化的 VLMEvalKit 框架 [12]。更多实验细节请参阅附录 B。

5.2. 实验结果

表格中给出的结果1和2 有力地证明了 TaLo 的有效性。具体而言，对于 LLaVA 模型，TaLo 取得的峰值性能提升达**10.4%**，覆盖五个基准测试和十项不同任务。值得注意的是，在全部三种样本设置下，其平均性能始终高于基线模型。同样，在同一组任务上，Qwen-VL 模型的最大性能提升达**16.6%**。对于InternVL，如表2所示，尽管我们的评估是在较少的八项任务上进行的，但结论一致：TaLo 在所有配置下均带来平均性能提升，峰值增益达到**10.0%**。

5.3. 更多分析

在本小节中，我们使用 MMStar [4]，进行简明而全面的分析。MMStar 是一个综合基准测试，涵盖六种不同的视觉语言模型任务类别，可用于系统地验证方法的有效性并揭示关键模式。

比较研究。 为进一步评估 TaLo，我们将其与模型合并进行比较 [6, 20, 49]，以及旨在提升特定任务性能的微调方法（LoRA [18], OFT [37]）。我们强调，尽管在标注数据充足时，微调可能构成更强的基线，但我们的目标并非声称在数据充足场景下优于微调（FT）。相反，在与 TaLo 相同的极低数据量设置下，我们发现 LoRA 等参数高效微调方法难以有效学习，往往无法超越未经修改的基础模型。对于所有微调实验，我们选取收敛后验证性能最佳的轮次对应的检查点。我们使用 LLaVA-Next-LLaMA3-8B 作为基础视觉语言模型，并为 TaLo 和微调方法采用相同的少样本样本。详细配置见附录B。如表 3所示，在大多数设置下，TaLo 的性能优于模型合并和微调基线，且耗时更短。至关重要的是，与所有基线方法不同，TaLo 既不需要外部模型，也不需要任何形式的训练或任务特定参数更新。相反，它通过在基础模型内部进行最小化的动态干预，实现即时适配，彰显了其高效性、简便性和强大的实际应用价值。

将 TaLo 扩展至多层干预。 为了评估 TaLo 的有效性是否能推广到单层干预之外，我们将该方法扩展为同时干预成对的层。具体而言，在为某项任务确定性能最佳的单层后，我们系统地测试在所有其余层中增加第二次干预，并选择能带来最高性能提升的层对。正如附录C.3所报告的研究结果表明，对于许多任务，增加第二次干预相较于最佳单层结果几乎没有提升，甚至毫无改善；而在若干情况下，根本找不到可行的第二个任务干扰层

表 2. 将 **TaLo** 应用于 *InternVL2-26B*（置零）后，在多个任务上的性能。说服策略 标注为 P-S，基本经济原理 标注为 B-EP，自然地理 标注为 P-G，地理学 标注为 Geo，美洲：地理标注为 A:Geo，*G2T* 标注为从基因到性状。

模型	样本数	任务								Avg
		P-S	B-EP	P-G	解法	Geo	A:Geo	G2T	材料	
InternVL	10次示例	58.3 ↑8.3	27.9 –0.0	34.5 –0.0	40.0 ↑6.7	35.4 ↑2.1	**35.0 ↑10.0**	25.0 ↓3.1	38.5 ↑1.3	3.16↑
	15样本	33.3 ↑8.3	32.6 ↑2.4	31.0 ↑6.9	31.1 –0.0	27.1 –0.0	20.0 –0.0	37。5 ↑3.1	41.0 ↓1.3	2.43↑

表3。TaLo 与其他方法的比较。每个条目报告性能和运行时间（$\times 10^2$ 秒）。TaLo 以更低的适应时间获得了更高的准确率。

Math								
样本数	Base	合并	LoRA微调		OFT		TaLo	
	分数↑	分数↑	分数↑	时间↓	分数↑	时间↓	分数↑	时间↓
10次样本	30.42	32.38	31.25	1.75	30.83	14.67	**32.92**	1.70
15次样本	30.21	32.63	32.17	2.53	32.34	21.91	**33.19**	1.76
20样本	30.87	31.76	33.04	3.46	33.04	29.16	**33.91**	2.84
Avg	30.50	32.26	32.15	2.58	32.07	21.91	**33.34**	2.10

实例推理								
样本数	Base	合并	LoRA微调		OFT		TaLo	
	分数↑	分数↑	分数↑	时间↓	分数↑	时间↓	分数↑	时间↓
10次示例	54.16	53.33	55.83	2.06	55.42	15.34	**58.33**	1.33
15样本	53.61	52.34	55.32	2.87	56.17	25.43	**57.87**	1.81
20次示例	54.35	53.04	55.65	4.06	56.52	33.81	**58.69**	2.89
Avg	54.04	52.90	55.60	2.99	56.04	24.86	**58.30**	2.01

。这表明任务干扰层可能较为稀疏，而且复杂的层间依赖关系可能会掩盖各层的独立作用，进一步证明 TaLo 针对单层的方法是合理的。（更多结果和分析见附录 C.3 。）

6. 讨论

局限性与未来工作。 我们的分析依赖于现有基准测试，其中预定义的任务类别可能会影响我们所观察到的层敏感性模式。因此，未来工作可以在更细粒度的任务分解上验证我们的发现，这可能有助于确立该现象的普遍性。就我们的方法 TaLo 而言，我们承认它是一个极简框架。我们的主要目标并非实现最先进的性能，而是提供一种简单、即插即用的解决方案，作为任务干扰层现象实际效用的概念验证。我们还探索了扩展方案，例如多层干预，结果表明识别多个干扰层并非易事。这一发现支持我们将重点放在极简的单层方法上，将其作为一种稳健且实用的起点。还有若干有前景的方向值得探索，例如开发更精细的动态层选择机制、研究更好的多层调制策略，或更细粒度的干预操作。此外，这种逐层诊断信号有望指导模型训练；例如，通过对已识别的干扰层施加定向正则化，或许可以从源头缓解任务冲突。我们乐观地认为，未来研究能够扩展 TaLo 的能力，并将其原则应用于更广泛、更复杂的场景。

假设与解释。 我们进一步提出一种假设，以解释为何某些层可能成为任务干扰层。现代大型模型在多样化的多任务数据上进行预训练，其中每一层都会学习一种折中表征，以近似所有任务的全局最优解。然而，这一全局最优解可能偏离任何特定任务的局部最优解。我们推测，任务干扰层捕获的特征虽能带来平均收益，但应用于特定任务时会引入噪声或失配。通过将这些层置零或统一缩放，TaLo 能有效抑制或重新平衡它们的影响，从而避免传播与任务无关甚至有害的信息。我们假设，这种干预能够引导模型的内部计算转向参数空间中更有利的区域，使其更好地契合目标任务的局部最优解，进而在不进行任何参数更新的情况下有效提升性能。

7. 结论

通过广泛的实证分析，我们发现大规模预训练视觉-语言模型中存在特定层，这些层会主动抑制某些下游任务上的性能。我们将这些层称为任务干扰层，因为策略性地绕过它们可以带来显著的性能提升。我们的进一步研究发现了一个关键模式：需要相似功能能力的任务对层干预表现出高度一致的响应模式。这表明干扰现象是围绕模型的功能能力系统组织的，从而使任务干扰层的影响能够泛化到相关任务。

基于这些发现，我们提出了TaLo，一种免训练适配方法，可在推理时识别并绕过这些干扰层。TaLo在多种模型上的性能表明，简单且有针对性的层干预可以成为模型适配的高效策略，从而无需任何参数更新。

参考文献

[1] Ameen Ali Ali、Shahar Katz、Lior Wolf 和 Ivan Titov。《检测并剪除大型语言模型中显著但有害的神经元》。载于第二届语言建模会议论文集，2025年。3[2] Xiaofeng Cao、Weiyang Liu 和 Ivor W Tsang。《通过最小化超球面能量实现数据高效学习》。IEEE 模式分析与机器智能汇刊，45(11):13422–13437，2023年。6[3] Xiaofeng Cao、Ya ming Guo、Heng Tao Shen、Ivor W Tsang 和 James T Kwok。《导师式学习：提升学生学习者的泛化能力与收敛速度》。机器学习研究期刊，25(325):1–45，2024年。6[4] Lin Chen、Jinsong Li、Xiaoyi Dong、Pan Zhang、Yuhang Zang、Zehui Chen、Haodong Duan、Jiaqi Wang、Yu Qiao、Dahua Lin 和 Feng Zhao。《我们评估大型视觉语言模型的方法正确吗？》。载于 NeurIPS 2024，。6、7、2[5] Qizhou Chen、Taolin Zhang、Chengyu Wang、Xiaofeng He、Dakan Wang 和 Tingting Liu。《归因分析与模型编辑相结合：借助 visedit 推进视觉语言模型中的知识纠正》。载于 AAAI-25，第 2168–2176 页，。3[6] Shiqi Chen、Jinghan Zhang、Tongyao Zhu、Wei Liu、Siyang Gao、Miao Xiong、Manling Li 和 Junxian He。《让视觉具备推理能力：通过模型合并理解感知与推理》，2025年。2、3、7[7] Zhe Chen、Jiannan Wu、Wenhai Wang、Weijie Su、Guo Chen、Sen Xing、Muyan Zhong、Qinglong Zhang、Xizhou Zhu、Lewei Lu 等。《InternVL：扩展视觉基础模型并使其适配通用视觉语言任务》。载于 CVPR，第 24185–24198 页，2024年。6、2[8] Hongrong Cheng、Miao Zhang 和 Javen Qinfeng Shi。《深度神经网络剪枝综述：分类体系、比较、分析与建议》。IEEE 模式分析与机器智能汇刊，46(12):10558–10578，2024年。3[9] Damai Dai、Li Dong、Yaru Hao、Zhifang Sui、Baobao Chang 和 Furu Wei。《预训练 Transformer 中的知识神经元》。载于 ACL，2022年，第 8493–8502 页。3[10] Nicola De Cao 、 Wilker Aziz 和 Ivan Titov。《编辑语言模型中的事实知识》。载于 2021年自然语言处理实证方法会议论文集，2021年。3[11] DeepSeek-AI。《DeepSeek-R1：通过强化学习激励大语言模型的推理能力》，2025年。1[12] Haodong Duan、Junming Yang、Yuxuan Qiao、Xinyu Fang、Lin Chen、Yuan Liu、Xiaoyi Dong、Yuhang Zang、Pan Zhang、Jiaqi Wang 等。《VL MEvalKit：用于评估大型多模态模型的开源工具包》。载于 ACM MM，2024年，第 11198–11201 页，2024年。7

[13] Razvan-Gabriel Dumitru、Vikas Yadav、Rishabh Mahesh- wary、Paul-Ioan Clotan、Sathwik Tejaswi Madhusudhan 和 Mihai Surdeanu。逐层量化：一种超越整数位宽量化大语言模型的实用有效方法。arXiv 预印本 arXiv:2406.17415，2024年。3[14] Chun-Mei Feng、Kai Yu、Yong Liu、Salman Khan 和 Wangmeng Zuo。利用扩散模型进行多样化数据增强，以有效实现测试时提示调优。载于 I CCV 2023，第 2704–2714 页。3[15] Andrey Gromov、Kushal Tirumala、Hassan Shapourian、Paolo Glorioso 和 Daniel A. Roberts。深层的效果为何出人意料地有限。载于第十三届国际学习表征会议，ICLR 2025。1[16] Ziang Guo、Zakhar Yagudin、Artem Lykov、Mikhail Ko- nenkov 和 Dzmitry Tsetserukou。VLM-Auto：基于视觉语言模型的自动驾驶助手，具备类人行为，并能理解复杂道路场景。载于第二届基础模型与大语言模型国际会议，FLLM 2024，第 501–507 页。1[17] Yihui He、Xiangyu Zhang 和 Jian Sun。通过通道剪枝加速超深度神经网络。载于 ICCV 2017，第 1398–1406 页。3 [18] Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen- Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适配。载于第十届国际学习表征会议，ICLR 2022。7[19]Shaohan Huang、Li Dong、Wenhui Wang、Yaru Hao、Saksham Singhal、Shuming Ma、Tengchao Lv、Lei Cui、Owais Khan Mohammed、Barun Patra、Qiang Liu、Kriti Aggarwal、Zewen Chi、Nils Johan Bertil Bjorck、Vishrav Chaudhary、Subhojit Som、Xia Song 和 Furu Wei。语言并非你所需的一切：让感知与语言模型对齐。载于 NeurIPS 2023。1[20] Gabriel Ilharco、Marco T´ulio Ribeiro、Mitchell Wortsman、Ludwig Schmidt、Hannaneh Hajishirzi 和 Ali Farhadi。通过任务算术编辑模型。载于第十一届国际学习表征会议，ICLR 2023。7[21] Yusuke Iwasawa 和 Yutaka Matsuo。用于模型无关领域泛化的测试时分类器调整模块。载于 NeurIPS，2021年。3[22] Adilbek Karmanov、Dayan Guan、Shijian Lu、Abdulmotaleb El-Saddik 和 Eric P. Xing。视觉语言模型的高效测试时适应。载于 CVPR 2024，第 14162–14171 页。3[23] Bohao Li、Rui Wang、Guangzhi Wang、Yuying Ge、Yix- iao Ge 和 Ying Shan。SEED-Bench：对具备生成式理解能力的多模态大语言模型进行基准测试。arXiv 预印本 arXiv:2307.16125，2023年。6, 2 [24] Bo Li、Kaichen Zhang、Hao Zhang、Dong Guo、Renrui Zhang、Feng Li、Yuanhan Zhang、Ziwei Liu 和 Chunyuan Li。LLaVA-Next：更强大的大语言模型为开放环境中的多模态能力赋能，2024年。2, 4[25] Chunyuan Li、Cliff Wong、Sheng Zhang、Naoto Usuyama、Haotian Liu、Jianwei Yang、Tristan Naumann、Hoifung Poon,

和 Jianfeng Gao。Llava-med：用一天时间训练用于生物医学的大型语言与视觉助手。载于 NeurIPS 2023。1[26] Tianwei Lin、Wenqiao Zhang、Sijing Li、Yuqian Yuan、Binhe Yu、Haoyuan Li、Wanggui He、Hao Jiang、Mengze Li、Xiaohui Song、Siliang Tang、Jun Xiao、Hui Lin、Yueting Zhuang 和 Beng Chin Ooi。Healthgpt：通过异构知识适配统一理解与生成的医学大型视觉语言模型。CoRR，abs/2502.09838，2025年。1[27] Yuan Liu、Haodong Duan、Yuanhan Zhang、Bo Li、Songyang Zhang、Wangbo Zhao、Yike Yuan、Jiaqi Wang、Conghui He、Ziwei Liu 等。MMBench：你的多模态模型是否面面俱到？载于 ECCV，2024年，第 216–233 页。6, 2[28] Pan Lu、Swaroop Mishra、Tony Xia、Liang Qiu、Kai-Wei Chang、Song-Chun Zhu、Oyvind Tafjord、Peter Clark 和 Ashwin Kalyan。学会解释：通过思维链进行多模态推理，以解答科学问题。载于 NeurIPS，2022年。2, 6[29] Pan Lu、Hritik Bansal、Tony Xia、Jiacheng Liu、Chunyuan Li、Hannaneh Hajishirzi、Hao Cheng、Kai-Wei Chang、Michel Galley 和 Jianfeng Gao。MathVista：在视觉情境中评估基础模型的数学推理能力。载于 ICLR，2024 年。2[30] Xiaolong Ma 、 Wei Niu 、 Tianyun Zhang、Sijia Liu、Sheng Lin、Hongjia Li、Wujie Wen、Xiang Chen、Jian Tang、Kaisheng Ma、Bin Ren 和 Yanzhi Wang。一种用于移动设备实时推理的基于图像增强模式的稀疏性方法。载于 ECCV，2020年，第 629–645 页。3[31] Xin Men、Mingyu Xu、Qingyu Zhang、Qianhao Yuan、Bingn- ing Wang、Hongyu Lin、Yaojie Lu、Xianpei Han 和 Weipeng Chen。Shortgpt：大语言模型中的层比你预想的更加冗余。载于计算语言学协会论文集，ACL 2025，第 20192–20204 页，2025年。1[32] Kevin Meng、David Bau、Alex Andonian 和 Yonatan Be- linkov。在 GPT 中定位并编辑事实关联。载于 NeurIPS 2022。3[33] Kevin Meng、Arnab Sen Sharma、Alex J. Andonian、Yonatan Belinkov 和 David Bau。在 Transformer 中大规模编辑记忆。载于第十一届国际学习表征会议，ICLR 2023。3[34] Eric Mitchell、Charles Lin、Antoine Bosselut、Chelsea Finn 和 Christopher D. Manning。大规模快速模型编辑。载于第十届国际学习表征会议，ICLR 2022。3[35] Eric Mitchell、Charles Lin、Antoine Bosselut、Christopher D. Manning 和 Chelsea Finn。大规模基于记忆的模型编辑。载于国际机器学习会议，ICML 2022，第 15817–15831 页。3[36] Saurav Muralidharan、Sharath Turuvekere Sreenivas、Raviraj Joshi、Marcin Chochowski、Mostofa Patwary、Mohammad Shoeybi、Bryan Catanzaro、Jan Kautz 和 Pavlo Molchanov。通过剪枝和知识蒸馏构建紧凑型语言模型。载于 NeurIPS，第 41076–41102 页，2024年。3[37] Zeju Qiu、Weiyang Liu、Haiwen Feng、Yuxuan Xue、Yao Feng、Zhen Liu、Dan Zhang、Adrian Weller 和 Bernhard Sch¨olkopf。通过正交微调控制文本到图像扩散。在 NeurIPS 2023。 7[38] Steffen Schneider、Evgenia Rusak、Luisa Eck、Oliver Bringmann、Wieland Brendel 和 Matthias Bethge。通过协变量偏移适应提高对常见损坏的鲁棒性。在 NeurIPS 2020。 3[39] Manli Shu、Weili Nie、De-An Huang、Zhiding Yu、Tom Goldstein、Anima Anandkumar 和 Chaowei Xiao。用于视觉语言模型零样本泛化的测试时提示调优。在 NeurIPS 2022。 3[40] Shoaib Ahmed Siddiqui、Xin Dong、Greg Heinrich、Thomas Breuel、Jan Kautz、David Krueger 和 Pavlo Molchanov。深入探究大语言模型的深度剪枝。arXiv 预印本 arXiv:2407.16286，2024年。 3[41] Chonghao Sima、Katrin Renz、Kashyap Chitta、Li Chen、Hanxue Zhang、Chengen Xie、Jens Beißwenger、Ping Luo、Andreas Geiger 和 Hongyang Li。DriveLM：通过图视觉问答实现驾驶。在 ECCV 2024，第 256–274 页。 1[42] Qi Sun、Edoardo Cetin 和 Yujin Tang。Transformer平方：自适应大语言模型。在第 13 届国际学习表征会议（ICLR 2025）。 6[43] Wenfang Sun、Xinyuan Song、Pengxiang Li、Lu Yin、Yefeng Zheng 和 Shiwei Liu。大语言模型的深度诅咒。arXiv 预印本 arXiv:2502.05795，2025年。 1[44] Yu Sun、Xiaolong Wang、Zhuang Liu、John Miller、Alexei A. Efros 和 Moritz Hardt。通过自监督进行测试时训练，以实现分布偏移下的泛化。在第 37 届国际机器学习会议（ICML 2020），第 9229–9248 页。 3[45] Dongsheng Wang、Natraj Raman、Mathieu Sibue、Zhiqiang Ma、Petr Babkin、Simerjot Kaur、Yulong Pei、Armineh Nourbakhsh 和 Xiaomo Liu。DocLLM：面向多模态文档理解的布局感知生成式语言模型。在第 62 届计算语言学协会年会（ACL 2024），第 8529–8548 页。 1[46] Dequan Wang、Evan Shelhamer、Shaoteng Liu、Bruno A. Olshausen 和 Trevor Darrell。Tent：通过熵最小化实现完全测试时适应。在第 9 届国际学习表征会议（ICLR 2021）。 3[47] Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhihao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge 等。Qwen2-VL：在任意分辨率下增强视觉语言模型对世界的感知。arXiv 预印本 arXiv:2409.12191，2024年。2, 6[48] Xin Wang、Fisher Yu、Zi-Yi Dou、Trevor Darrell 和 Joseph E Gonzalez。SkipNet：学习卷积网络中的动态路由。在 ECCV，第 409–424 页，2018年。 6[49] Enneng Yang、Li Shen、Guibing Guo、Xingwei Wang、Xiaochun Cao、Jie Zhang 和 Dacheng Tao。大语言模型、多模态大语言模型及其他模型中的模型合并：方法、理论、应用与机遇。CoRR，abs/2408.07666，2024年。 7[50] Xikai Yang、Juzheng Miao、Yuchen Yuan、Jiaze Wang、Qi Dou、Jinpeng Li 和 Pheng-Ann Heng。具备多图像视觉能力的医学大型视觉语言模型。在医学图像计算与计算机辅助干预会议（MICCAI 2025），第 402–412 页。 1

[51] Chenyu Yi、Siyuan Yang、Yufei Wang、Haoliang Li、Yap-Peng Tan 和 Alex C. Kot。用于鲁棒视频分类的时间一致性测试时优化。在第十一届国际学习表征会议（ICLR 2023），2023年。 3[52] Lu Yin、You Wu、Zhenyu Zhang、Cheng-Yu Hsieh、Yaqing Wang、Yiling Jia、Gen Li、Ajay Kumar Jaiswal、Mykola Pechenizkiy、Yi Liang、Michael Bendersky、Zhangyang Wang 和 Shiwei Liu。异常值加权的逐层稀疏性（OWL）：将大语言模型剪枝至高稀疏度所缺失的关键要素。在第四十一届国际机器学习会议（ICML 2024）。 3, 1[53] Shukang Yin、Chaoyou Fu、Sirui Zhao、Ke Li、Xing Sun、Tong Xu 和 Enhong Chen。多模态大语言模型综述。《国家科学评论》，11(12)，2024年。 1[54] Xiang Yue、Yuansheng Ni、Kai Zhang、Tianyu Zheng、Ruoqi Liu、Ge Zhang、Samuel Stevens、Dongfu Jiang、Weiming Ren、Yuxuan Sun、Cong Wei、Botao Yu、Ruibin Yuan、Renliang Sun、Ming Yin、Boyuan Zheng、Zhenzhu Yang、Yibo Liu、Wenhao Huang、Huan Sun、Yu Su 和 Wenhu Chen。MMMU：面向专家级通用人工智能的大规模多学科多模态理解与推理基准测试。在 CVPR，2024年，第 9556–9567 页。 2, 4, 6[55] Marvin Zhang、Sergey Levine 和 Chelsea Finn。MEMO：通过适应与数据增强实现测试时鲁棒性。在 NeurIPS 2022。 3[56] Yang Zhang、Yanfei Dong 和 Kenji Kawaguchi。研究大语言模型中的层重要性。在第七届 BlackboxNLP研讨会：分析与解释自然语言处理神经网络，第 469–479 页，2024年。 2, 3[57] Anhao Zhao、Fanghua Ye、Yingqi Fan、Junlong Tong、Zhiwei Fei、Hui Su 和 Xiaoyu Shen。SkipGPT：结合词元感知与模块解耦，重新设计动态层剪枝。CoRR，abs/2506.04179，2025年。 6[58] Kai Zheng、Wei Wu、Rui Feng 等。正则化掩码调优：揭示预训练视觉语言模型中的隐藏知识。在 CVPR 2023 论文集，第 11663–11673 页。 3[59] Chen Zhu、Ankit Singh Rawat、Manzil Zaheer、Srinadh Bhojanapalli、Daliang Li、Felix X. Yu 和 Sanjiv Kumar。修改 Transformer模型中的记忆。CoRR，abs/2012.00363，2020年。 3

所有单独层都有帮助吗？视觉语言模型中任务干扰层的实证研究

补充材料

A. 模型与基准测试

我们在表4中展示了实验所用的全部模型，并在表5中列出了我们使用的全部基准测试。

在本研究中，所有数据集均仅以准确率作为唯一指标进行评估。大多数数据集，即MMStar、MMMU、SEEDBench、MMBench和ScienceQA，都属于选择题基准测试；模型预测的选项会从其输出中提取，并与真实标签进行匹配。MathVista虽然被构建为视觉问答任务，但其官方评估同样采用生成回答与参考答案之间的直接字符串匹配，从而确保所有任务在指标上保持一致。

具体而言，MMStar是一个综合性基准测试，包含在六项核心能力上均衡分布的250个样本：粗粒度感知、细粒度感知、实例推理、逻辑推理、数学，以及科学与技术。MMBench包含2,974道多项选择题，用于评估广泛的能力，包括粗粒度感知、细粒度感知（单实例与跨实例）、实例推理、逻辑推理、属性推理和关系推理。MMMU涵盖30个学科，包括艺术与设计、商业、科学、健康与医学、人文与社会科学以及工程学，覆盖183个子领域，并包含30类异构图像（例如图表、示意图、地图、表格、乐谱、化学结构），重点考察结合领域专属知识的高级感知与推理。SEED-Bench包含19,000道人工标注的多项选择题，覆盖12个评估维度，其中包括图像理解。MathVista是一个具有挑战性的基准测试，融合了多样的数学与视觉推理任务，共有6,141个示例，来自28个现有的多模态数学相关数据集以及三个新整理的数据集。最后，ScienceQA包含21,208道从中小学课程中收集的多模态科学问题。这一多样且严格的基准测试选择，使我们能够在统一的准确率指标下，对任务特定能力进行全面评估。

B. 实验细节

在我们的 TaLo 实验中，识别最佳干预层的流程始于任务定义和样本准备。我们首先根据数据集的元数据确定目标任务，随后从一个与最终测试进行设置以防止任何数据重叠。

我们的识别流程采用迭代式管线。首先，我们使用初始探测样本集评估未修改模型，以确定基线性能。如果基线准确率达到 100%，则认为该样本集缺乏信息，并将其弃用；随后从探测样本池中抽取新样本集，再次评估基线。确定基线后，我们按层系统地逐一干预参数，并测量每层带来的性能提升。如果某一层带来的正向提升独一无二且最大，则将其指定为最优目标。如果多个层的性能并列最佳，或没有任何层带来正向提升，我们将启动多轮增强采样策略以消除歧义。具体而言，我们会额外补充 $shot/2$ 个样本进行重新评估；如果仍然并列，则进一步补充 $shot/4$ 个样本。如果经过这两轮后仍无法确定唯一的最优层，我们会从最终候选层中选择索引最高的层，以确保鲁棒性[15, 31, 43, 52]。

在微调实验中，我们采用两种参数高效微调（PEFT）方法：LoRA（低秩适配）及其变体 OFT（正交微调）。所有实验均在 LLaVA- Next-8B 模型上进行。对于 LoRA，我们将秩设为 $r = 8$，缩放因子设为 $\alpha = 16$，并将适配器模块应用于语言通路和视觉通路中的所有线性投影。这种全架构适配策略可确保在微调过程中，视觉和文本表征都得到全面对齐。为了确保模型在微调过程中真正理解问题所涉及的知识，而非仅仅记住选项，我们将答案格式设为“选项 +选项内容”。这种方法有助于模型学习每个选项的具体含义及其与问题之间的关系。

在模型合并实验中，用于合并的大语言模型是 DeepSeek-R1-Distill-Llama-8B[11], , 融合系数 λ为 0.9。然而，如表3所示，模型合并的效果对外部大语言模型和目标任务的选择都高度敏感，这表明其性能在不同配置下并不稳健，需要针对具体任务进行仔细调优。

Name	Size	Hugging Face 检查点
LLaVA-Next-LLaMA3 [24]	8B	llava-hf/llama3-llava-next-8b-hf
Qwen2-VL [47]	2B	Qwen/Qwen2-VL-2B-Instruct
InternVL2 [7]	26B 和 40B	OpenGVLab/InternVL2-40B

表4. 我们实验中使用的模型详情。

基准测试	类别	Hugging Face网址
MMStar [4]	MCQ	Lin-Chen/MMStar
MMBench-EN [27]	MCQ	lmms-lab/MMBench
MMMU-VAL [54]	MCQ	MMMU/MMMU
ScienceQA-VAL [28]	MCQ	derek-thomas/ScienceQA
MathVista-MINI [29]	VQA	AI4Math/MathVista
SEEDBench-IMG [23]	MCQ	lmms-lab/SEED-Bench

表 5. 我们实验所用基准测试的详情。

图 5. 随机噪声干预的定性案例研究。



C. 补充实验结果与分析

C.1. 具体聚类详情

表11 列出了七个集群中包含的所有任务。从表中可以看出，集群划分似乎能根据任务在功能上的相似性进行合理分组。例如，集群 1 将数值常识、算术推理和数学应用题归为一组——这些任务都明显围绕数值理解与计算展开。这表明，该方法成功识别并分离出了定量推理这一连贯的能力。

同样，集群 3 将天文学、化学和科学推理等领域特定的科学任务归为一个统一主题，凸显了它们对正式科学知识的共同依赖。

C.2. 干预组件的消融

在正文中，我们将层干预定义为对自注意力模块的修改。为证明这一选择的合理性，我们进行了全面的消融实验，将相同的干预应用于 MLP模块。

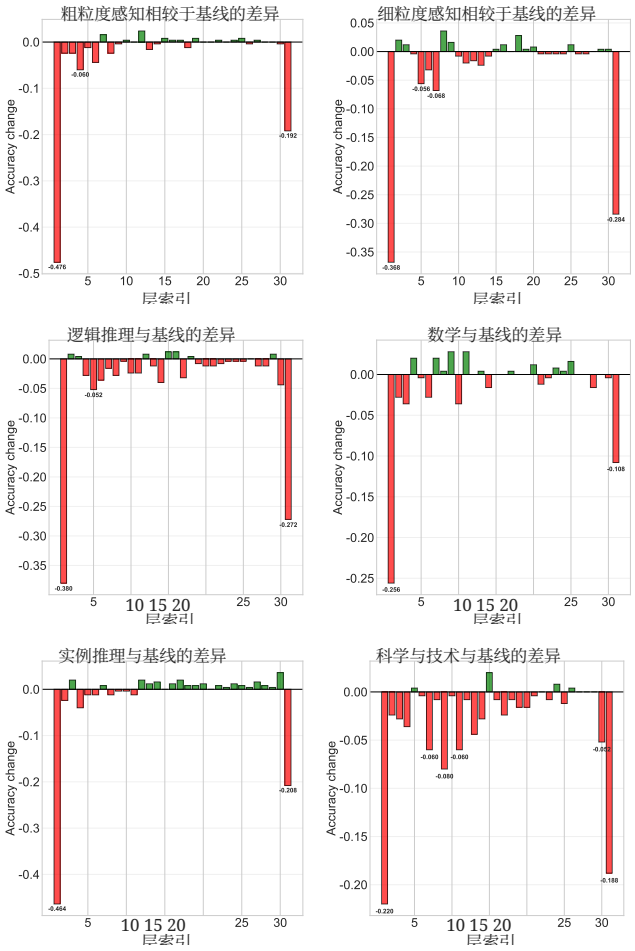
如图6所示，对 MLP模块进行干预表现出明显的位置依赖敏感性，其模式与自注意力模块有根本不同。我们发现，对模型早期层或最后一层进行干预，会导致所有任务的性能灾难性崩溃。至于绝大多数中间层，尽管干预 MLP模块不会导致模型崩溃，表现出一定的鲁棒性，但也无法带来显著的性能提升。这一发现证明了我们将本研究重点放在自注意力模块上的合理性，因为它们明显呈现出我们旨在研究的、能显著提升性能的任务干扰层现象。因此，后续所有分析及所提出的 TaLo 方法都仅针对这些自注意力模块展开。

C.3. TaLo 的补充分析

干预方法的消融研究 尽管早期结果（第 3.2节）表明，不同干预类型总体上会引发相似的逐层趋势，但它们对各项任务的影响可能存在细微差异。为了在受控且具有代表性的设置下考察这些影响，我们选取了 MMStar 和 MMBench：这两个均衡的基准测试汇集了多个来源的数据，任务多样性适中且覆盖面广。它们避免了高度专业化或过于狭窄的数据集，因此非常适合用于研究这些因素如何与 TaLo 的表现相互作用。

除了参数置零和均匀缩放之外，

图6. 对MLP层进行干预后的准确率变化



我们还尝试将层权重替换为其均值和随机噪声。然而，实验表明，注入随机噪声会严重破坏模型的稳定性，实际上抹去了学习到的表征（如图 5所示）。模型输出会变得不连贯，往往还会退化，与输入几乎毫无关联。这表明，预训练参数即使对特定任务而言并非最优，仍编码了对模型功能至关重要的基本结构和语义先验。

鉴于基本能力出现这种下降，我们将分析重点放在影响较温和、控制更精细的干预上：置零、均匀缩放和均值替换。这些方法保留模型的基础知识，同时允许有针对性的调节。它们在扰动与稳定性之间取得了更好的平衡，使我们能够有意义地研究逐层任务适配，而不会导致总体性能崩溃。

为了进一步考察三种干预类型的影响，我们在两个基准数据集上，针对范围广泛的任务开展消融研究，并采用一致的 10-shot 设置。

如表 6、7所示，我们观察到，置零和均匀缩放产生的效果相近，而置零在各项任务上的平均性能更好。相比之下，均值替换始终会导致较差的结果。

这一观察与我们在第3.2节中的早期发现一致：缩放和置零都呈现出相似的层敏感性模式。尽管在特定情境下可能存在细微差异，这些差异取决于模型架构或任务性质。

在实践中，缩放和置零之间的选择可能取决于任务特定行为或实现的简便性。两者都支持有效的即插即用适配。TaLo 的优势似乎并不在于干预本身，而在于策略性地选择何时以及何处应用干预。起决定性作用的似乎是层，而非操作。

多层干预 为补充基于单层干预的主要结果，我们对双层 TaLo 干预进行了系统研究。对于每项任务，我们首先使用标准 TaLo 流程确定最优单层。然后固定该层，并在大语言模型主干网络的其他每一层上迭代应用第二次归零干预，同时测量由此产生的性能变化，并保持所有其他组件不变。这样便为每项任务生成完整的成对干预矩阵，据此选出表现最佳的双层组合。

由于计算成本随模型深度呈二次增长，我们将探索范围限制在双层组合上，以此作为高阶交互的一种可行近似。结果一致表明，增加第二次干预带来的收益有限或毫无收益，这进一步印证了主论文中观察到的任务干扰层具有稀疏性。

层选择的鲁棒性 为证明我们的层选择具有鲁棒性，并非仅限于特定领域，我们将自助法分析（N=50）扩展到了三个少样本设置下的两个不同任务类别。如图 7所示，在所有任务中，层选择分布始终高度集中，这证实了无论任务类型或采样方式如何变化，TaLo 都能持续定位到稳定的干扰区域。尤为重要的是，不同任务中的干扰区域位置各不相同，这有力地验证了 TaLo 捕捉到了模型内部表示中固有的、有意义的任务特定功能冲突，而非仅仅识别普遍冗余的层。

TaLo 的更多实验结果如表9 10所示的综合结果所证明的那样，评估涵盖了 MMBench 和 ScienceQA，TaLo

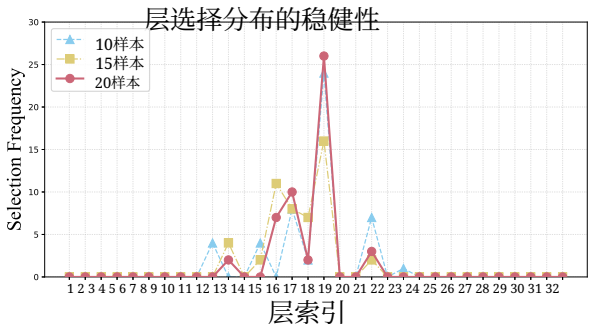
表6. 不同干预方法下 TaLo 在 *LLaVA-Next-Llama3-8B* 上的结果。最后一列（**平均值**）报告了各任务的平均变化。 表示该方法未能找到任务干扰层。

干预	MMStar						Avg
	粗粒度感知	细粒度感知	实例推理	科学与技术	逻辑推理	Math	
置零	63.8 ↓2.9	41.9 ↑0.9	57.6 ↑3.8	32.9 ↑1.0	38.6 ↓3.8	31.0 ↑2.9	0.32↑
均匀缩放	66.2 ↑2.4	41.0 ↓0.4	51.0 ↓2.8	24.8 ↓5.2	40.0 ↓1.0	34.3 ↑3.8	0.53↓
均值替换	✗	34.8 ↓6.2	52。9 ↓0.9	27.6 ↓3.4	41.9 ↑1.4	30.5 ↑1.0	1.14↓

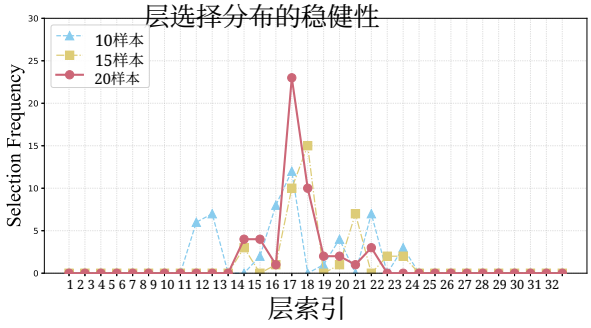
表7. TaLo 在*L*上的结果 *LaVA-Next-LLaMA3-8B* 采用不同干预方法时（结构化图文理解意为 r
结构化图像文本理解 意）。

干预	MMBench					Avg
	物理属性推理	结构化图文理解	属性识别	名人识别	图像情感	
置零	55.3 ↑7.8	55.8 ↑0.8	65.2 ↓3.1	68.5 ↓1.4	65.6 ↑0.6	0.94↑
均匀缩放	57。0 ↑5.0	55.8 ↑0.8	71.0 ↑0.5	67.7 ↓2.8	65.6 ↓0.6	0.58↑

图7. 层选择的鲁棒性分析。



(a) 数学任务分析



(b) 实例推理任务分析

展现出在广泛任务中稳健提升性能的能力。不过，对于特别具有挑战性的类别，提升幅度相对有限，在某些情况下甚至有所下降，例如涉及复杂多步推理和精细视觉属性辨别的任务。这一模式表明，尽管我们提出的层级干预 tion 为任务适应提供了一种有效机制，但其效力受预训练模型底层能力的限制。在这些高难度场景中，性能停滞或下降可能表明存在架构或数据层面的局限，未来工作或可通过整合更强的归纳偏置或辅助知识源来解决这些问题。

C.4. 任务干扰层的实证验证

多个模型（*LLaVA-Next-Llama3-8B*、*Qwen2-VL-2B*和 *InternVL2-40B*）在不同干预策略下的准确率变化热力图如图9 至20所示。在多种基准测试上进行评估后，这些热力图揭示了层特定性能提升的一致模式，构成了任务干扰层现象的核心实证依据。这些结果并非孤立的异常，而是表明单个层在支持相互竞争的任务需求时存在系统性权衡，这种现象可在不同模型规模和架构中观察到。

我们观察到，即使面对相同的任务，任务-层交互向量在不同视觉语言模型中的形态也存在显著差异。我们认为，这并非局限，而是一个关键且符合预期的发现。由于基础模型在架构、预训练数据和目标方面各不相同，它们会形成独特的内部功能专门化路径。因此，任务干扰本质上是模型特有的现象；在一个模型中妨碍性能的层，在另一个模型中未必如此。这种差异使通用且静态的移除策略难以实施，也充分凸显了我们采用自适应方法 TaLo 的必要性。TaLo 利用小型探测集，有效应对这种异质性，动态识别并绕过各模型表征空间中特有的干扰层。

表8. 双层干预（10样本）下 *LLaVA* 上的 TaLo 结果。*'✗'*表示无法识别出第二个任务干扰层的情况。MMStar 的详细信息见附录 A。

指标	MMStar					
	CP	FP	IR	S&T	LR	Math
任务干扰层	L1、L6	L15、L29	L11,	L31,	L1，L8	L6，L13
性能（两层）	61.9 ↓3.8	40.0 ↓0.5	56.2 ↑1.0	31.0 ↑0.5	37.1 ↓3.8	25.7 ↓3.8
性能（单层）	63.8 ↓2.9	41.9 ↑0.9	57.6 ↑3.8	38.6 ↓3.8	31.0 ↓2.9	32.9 ↑2.5

图8. 层置零对 LLaVA-Next 推理影响的定性案例研究。该图*展示了三个对比示例，呈现模型在特定任务干扰层干预前（基础模型）和干预后的推理过程。

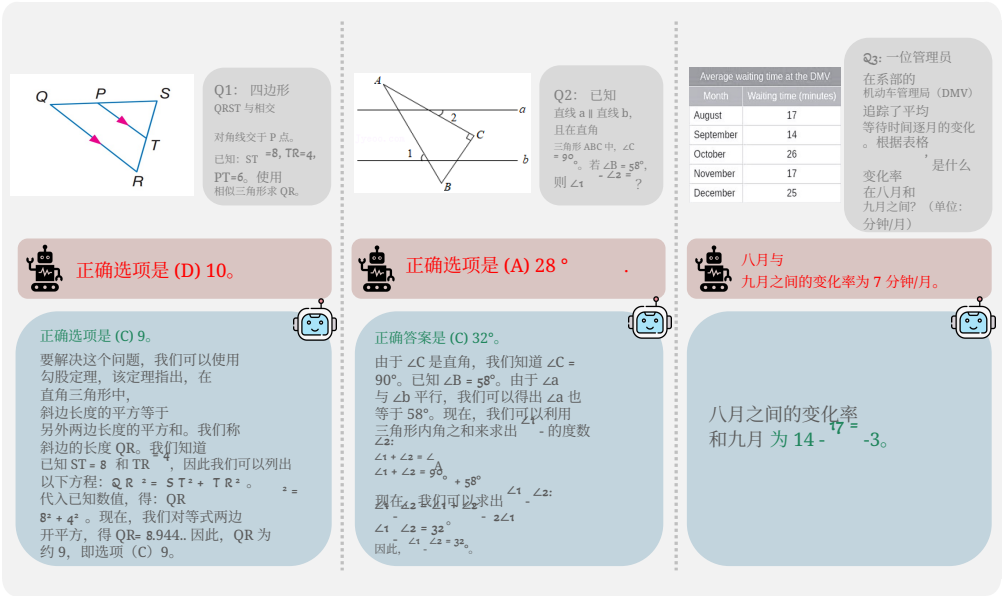


表9. TaLo 在 *LLaVA-Next-LLaMA3-8B* 上的其他结果。此处，未来预测 标注为 F-P，函数推理标注为 F-R。

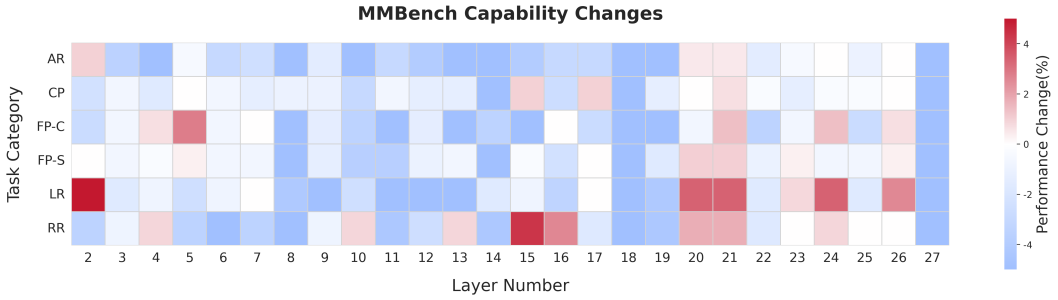
模型	样本数	MMBench			ScienceQA			
		F-P	OCR	F-R	生态相互作用	美洲：地理学	大洋洲：地理学	地理学
LLaVA	10次示例	43。3 ↓3.3	69.8 ↓2.6	68.6 –0.0	17。6 –0.0	30。0 ↑20.0	21.9 –0.0	41.7 ↑8.4
	15样本	48.3 ↓1.1	71.2 –0.0	68.0 ↓1.3	29.4 ↑11.8	25.0 ↑5.0	25.0 ↑3.1	39.6 ↑4.2
	20次示例	58.6 ↑6.9	78.8 ↑3.8	71.0↑2.2	29.4↑11.8	35.0 ↑20.0	15.6–0.0	33.3 ↑2.1

表10。补充结果 TaLo 在 *Qwen2-VL-2B* 中，其中 结构化图文理解标注为 S-I-U，归因 识别标注为 AR， 物理关系标注为 PR，名人识别标注为 CR。 *te*

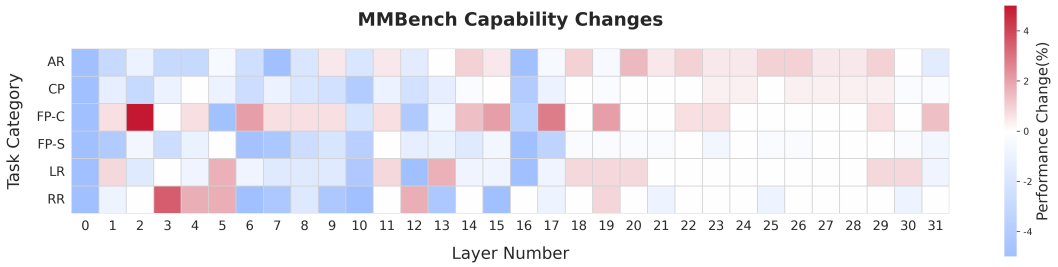
模型	样本数	MMBench				ScienceQA				
		S-I-U	AR	PR	CR	天文学	美洲：地理学	从基因到性状	解法	力与运动
Qwen-VL	10次示例	52.5 ↑3.3	72.3 ↑0.9	47.6 ↑1.6	75.0 ↑2.0	35。5 ↑3.2	25.0 ↑15.0	21.9 ↑12.5	24.4 ↑2.2	35。3 ↑5.9
	15样本	52.3 ↑3.2	72.1 ↑2.0	50.8 ↑3.2	75.3 ↑1.5	38.7 ↑6.4	10.0 –0.0	25.0 ↑3.1	28.9 ↑8.9	29.4 ↓5.9
	20样本	48.5 ↑2.5	73.4 ↑1.1	55.6 ↑3.2	76.6 ↑1.0	32.3 ↑6.5	15.0 ↑5.0	28。1 ↑6.2	35。6 ↑15.6	52。9 ↑17.6

集群类别与包含的任务	
集群 1（定量推理）	数值常识、算术推理、几何推理， 代数推理、几何问题求解、数学应用题，图形问答、统计推理， 城市，信息类文本：第 1 级，粒子运动与能量
集群 2（分析推理）	图像情感，生物学、工程学、公共卫生，实例推理、数学、地理学、视觉推理、建筑学与工程学，诊断学与检验医学、电子学， 心理学、地图、磁体、植物繁殖、领域专用词汇、从基因到性状
集群 3（科学知识）	科学推理、教科书问答、化学， 医学、经济学、物理学、社会学、艺术与设计、科学与技术、天文学、植物、天气，化石、热能， 自然资源、药学，人文与社会科学， 文学
第4簇（综合推理）	逻辑推理、会计，历史学、药学、工程学 实践、生态学、世界宗教、说服策略， 适应和自然选择、经济学、社会学， 人文学科，设计，文学、自然科学理论， 政治史
簇5（知觉分类）	动作识别、属性识别，图像质量、粗粒度感知，细粒度感知、目标定位， 分类、生态系统，力与运动、解法， 物质状态、场景理解、光学字符识别，图像场景、视觉元素、结构化图文理解、属性 比较、名人识别
簇6（预测推理）	未来预测，身份推理、空间关系， 电子学、心理学、数学、音乐、植物繁殖， 速度与加速，实例交互、物理 地理学、分类和学名、语境线索， 文本理解
第7簇（关系理解）	自然关系、物理关系、社会关系、艺术理论， 南北战争与重建、大航海时代、古代 美索不达米亚、板块构造、地质学、动物、农业， 生态系统、文化史、世界宗教、州府

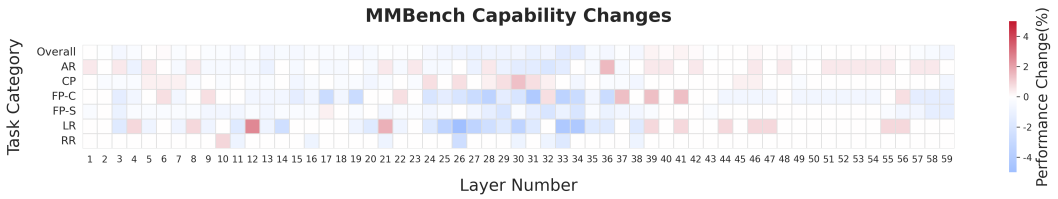
表 11。来自多个基准测试的类别及其包含的任务。



(a) Qwen2-VL 上的准确率变化热力图。



(b) LLaVA-Next 上的准确率变化热力图。



(c) InternVL2 上的准确率变化热力图。

图9. MMBench 上的准确率变化热力图（均匀缩放）。

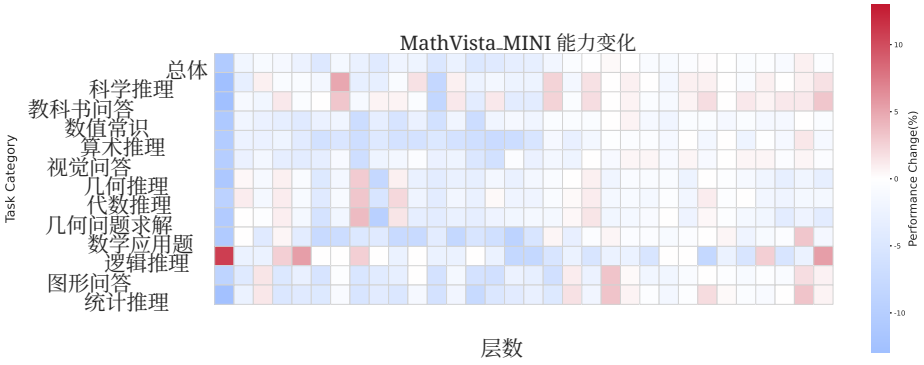


图10. LLaVA-Next 在 MathVista-MINI 上的准确率变化热力图（均匀缩放）。

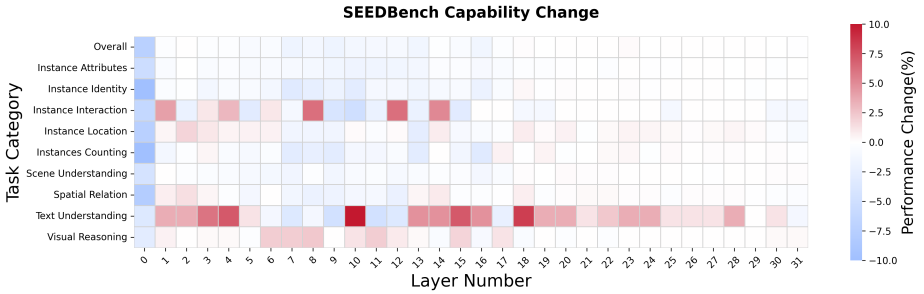


图11. LLaVA-Next 在 SEEDBench 上的准确率变化热力图（均匀缩放）。

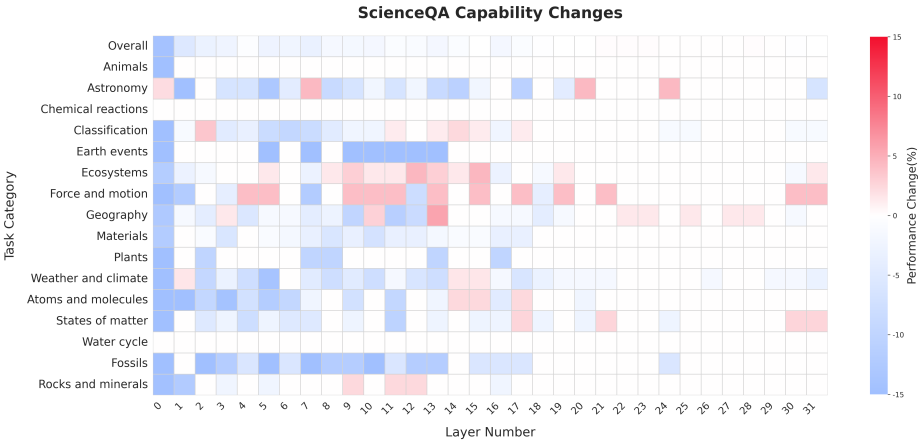
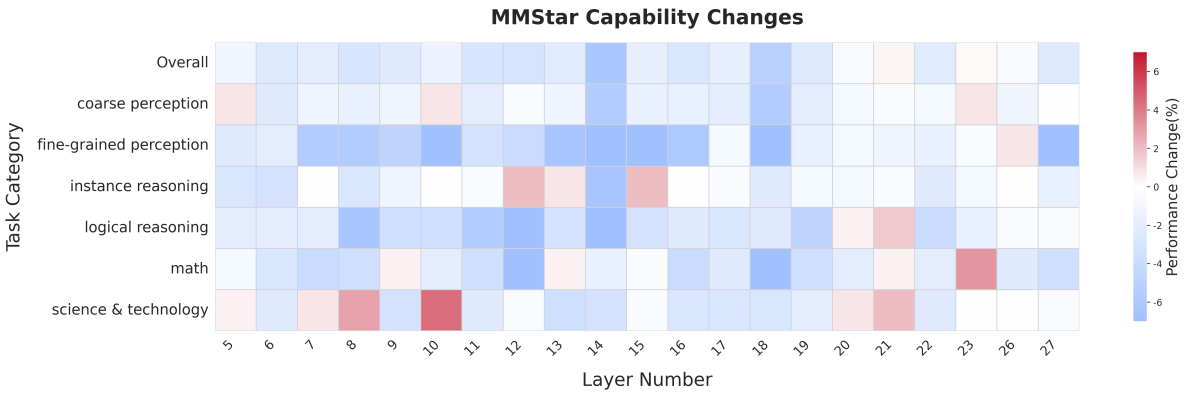
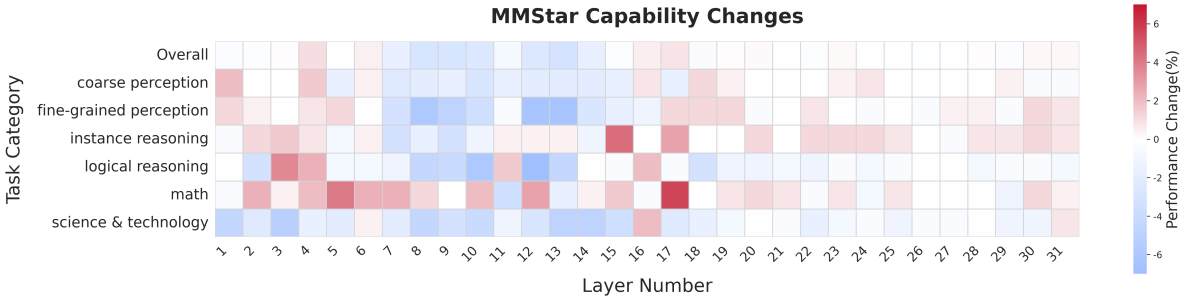


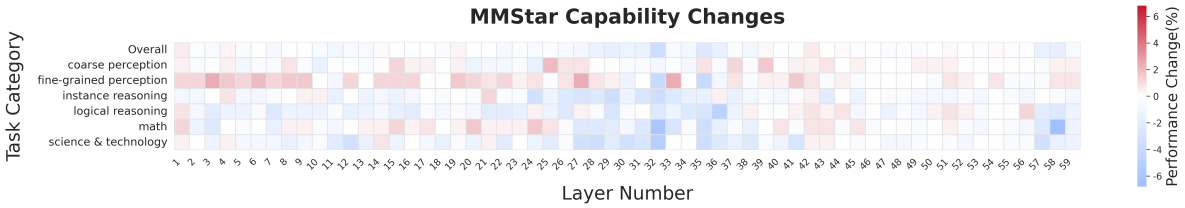
图12. LLaVA-Next 在 ScienceQA 上的准确率变化热力图（均匀缩放）。



（a）Qwen2-VL 上的准确率变化热力图。

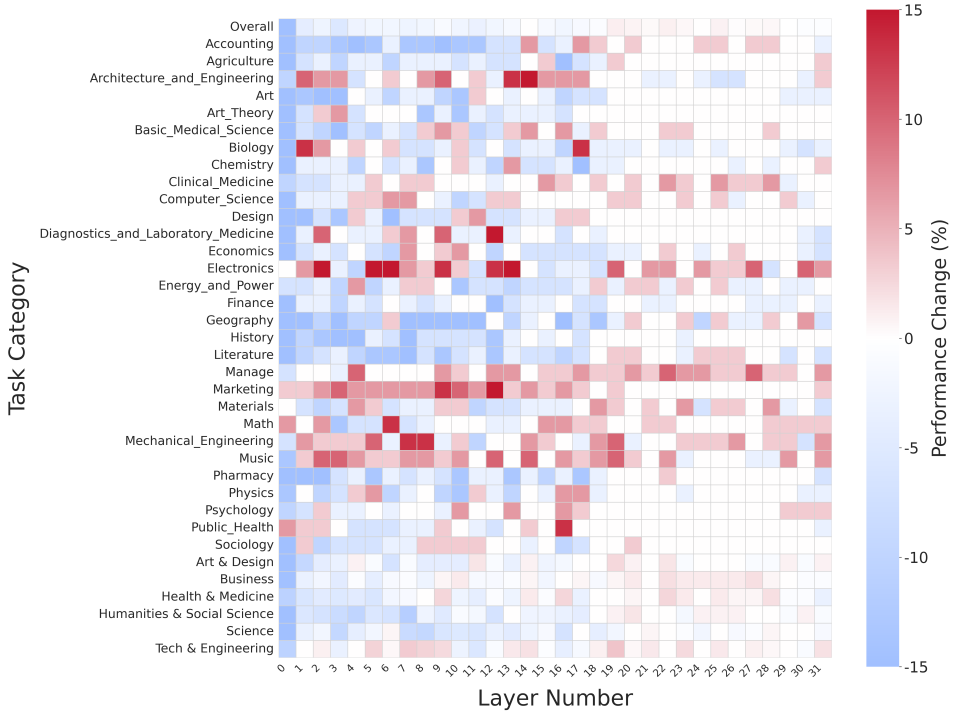


（b）LLaVA-Next 上的准确率变化热力图。

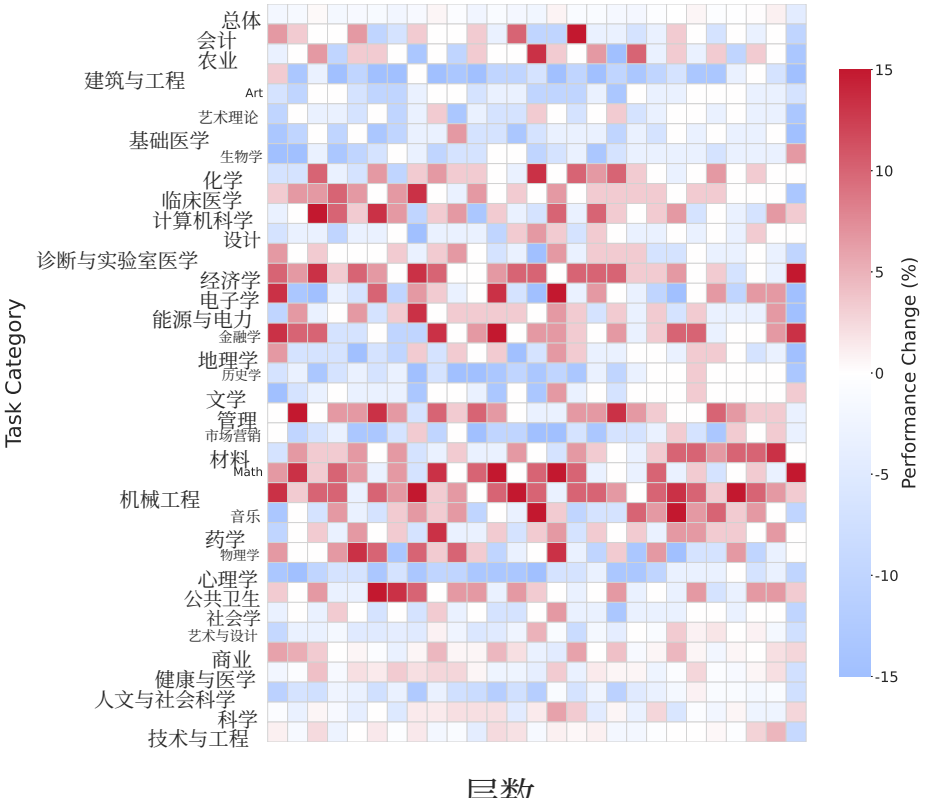


（c）InternVL2 上的准确率变化热力图。

图13。MMStar 上的准确率变化热力图（均匀缩放）。

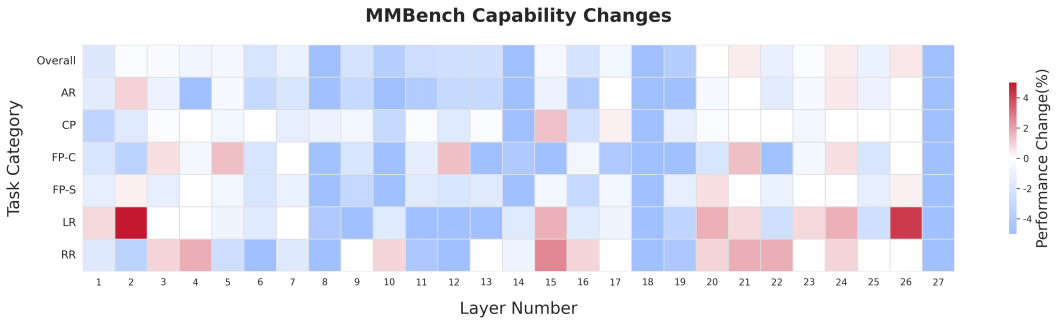


(a) LLaVA-Next上的准确率变化热力图。

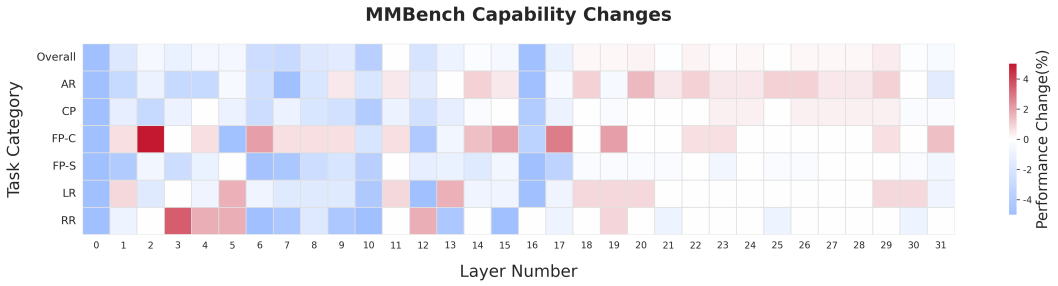


(b) Qwen2-VL上的准确率变化热力图。

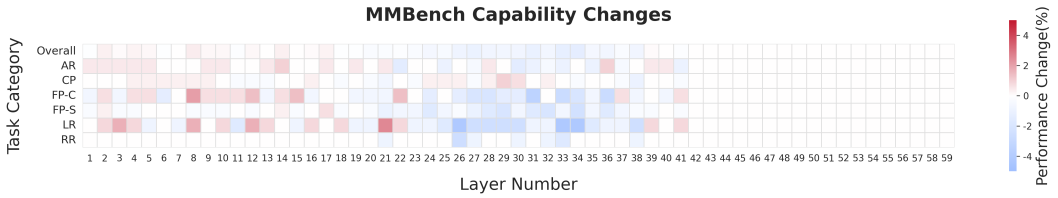
图14. MMMU上的准确率变化热力图（均匀缩放）。



(a) Qwen2-VL 上的准确率变化热力图。



(b) LLaVA-Next 上的准确率变化热力图。



(c) InternVL2 上的准确率变化热力图。

图15. MMBench 上的准确率变化热力图（置零）。

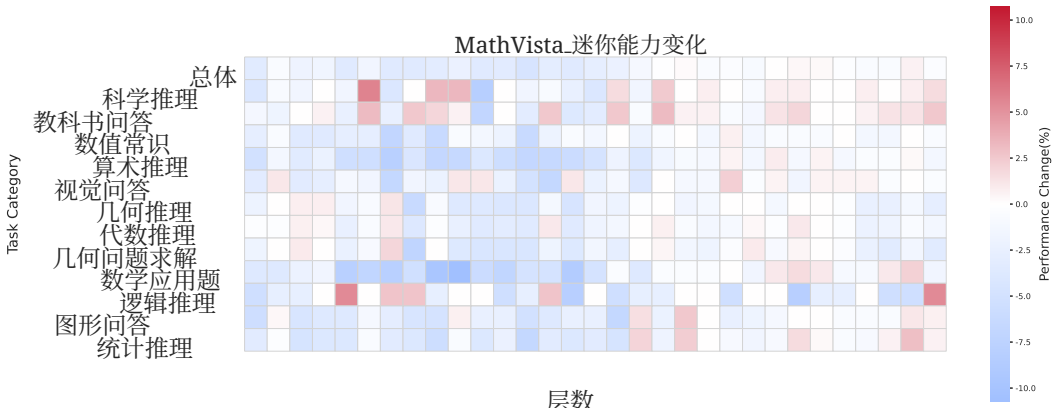


图16. LLaVA-Next 在 MathVista-MINI 上的准确率变化热力图（置零）。

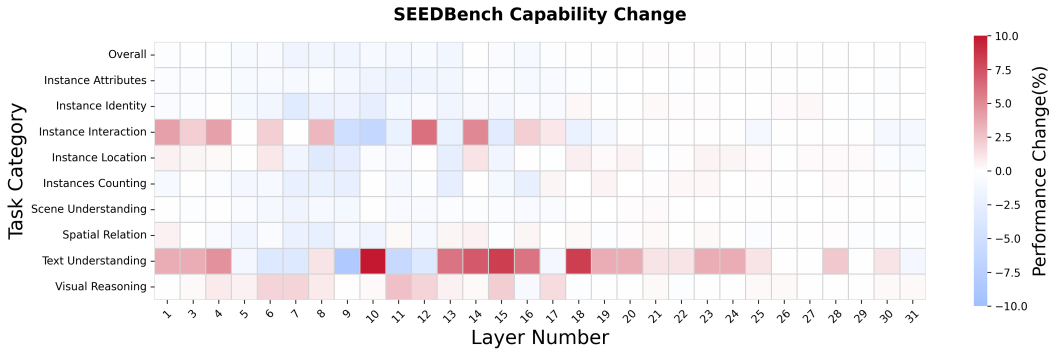


图17. LLaVA-Next 在 SEEDBench 上的准确率变化热力图（置零）。

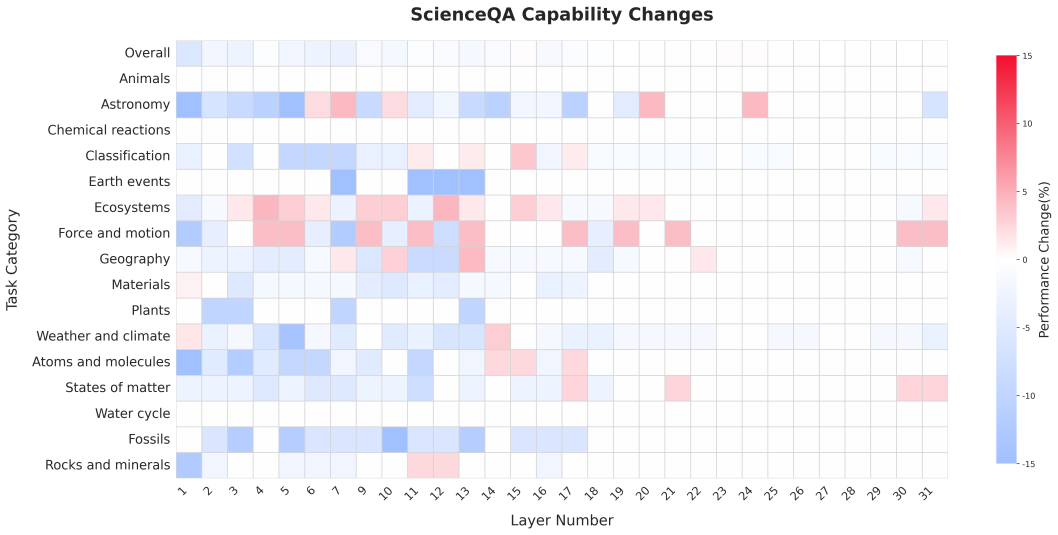
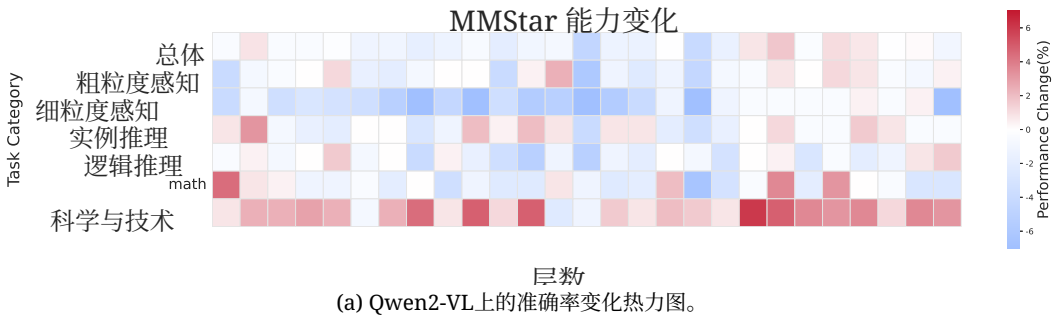
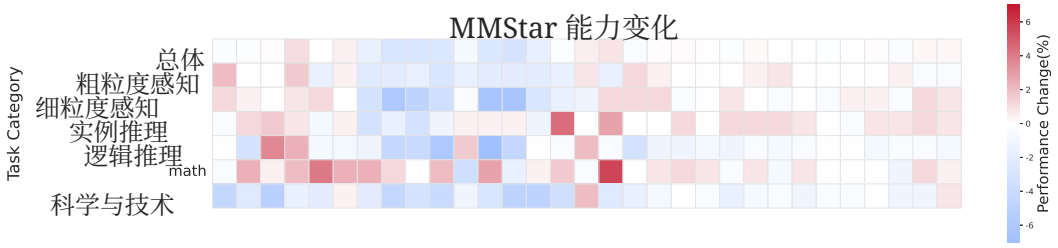


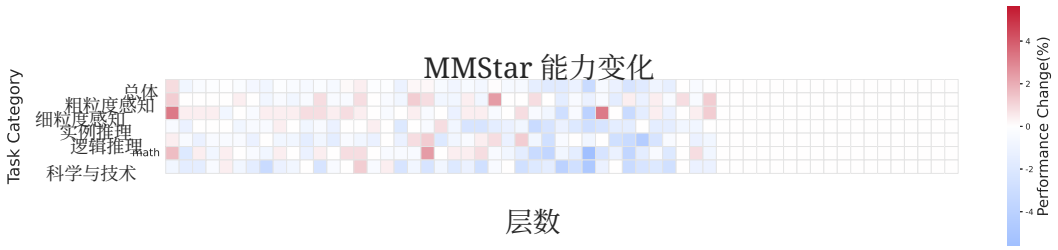
图18. LLaVA-Next 在 ScienceQA 上的准确率变化热力图（置零）。



(a) Qwen2-VL上的准确率变化热力图。

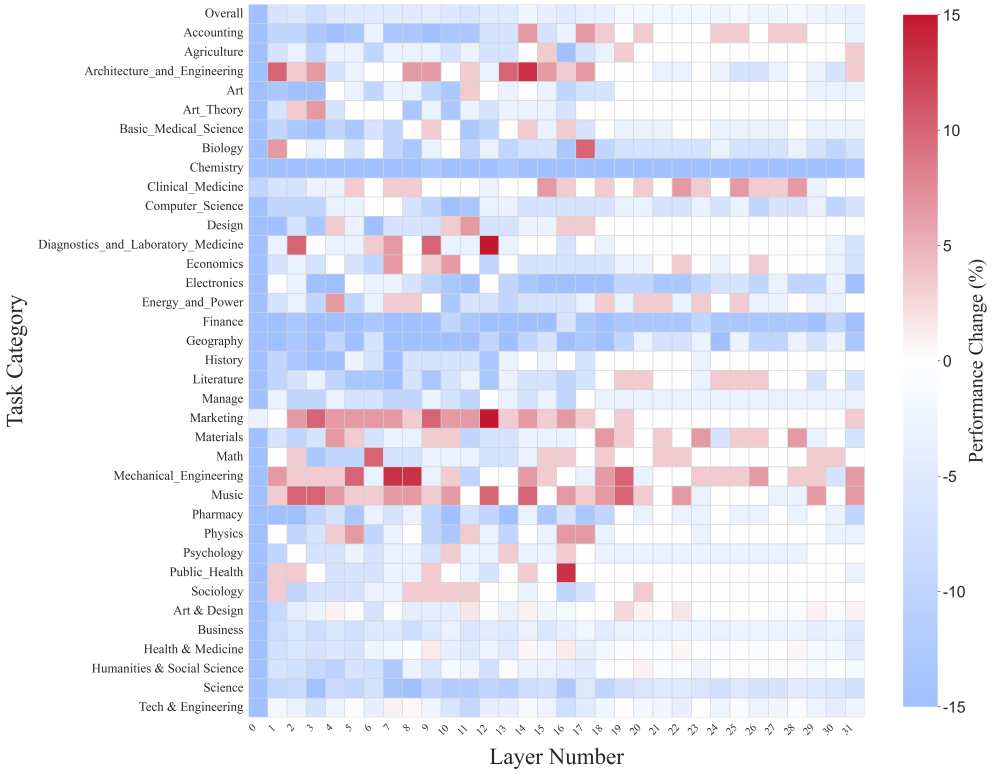


(b) LLaVA-Next上的准确率变化热力图。

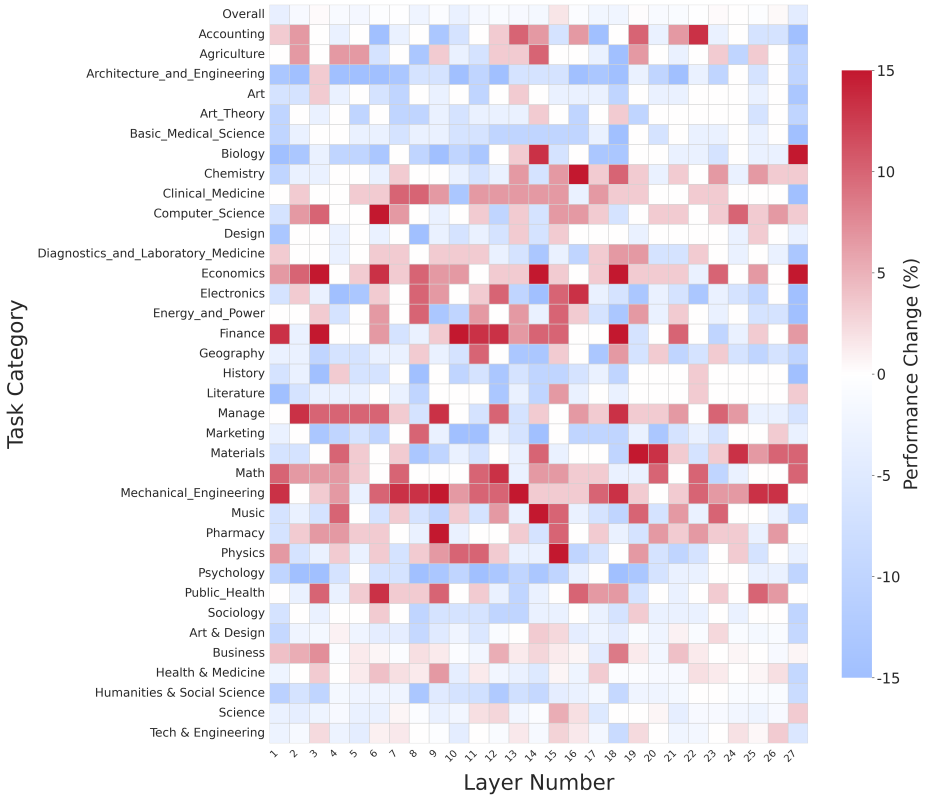


(c) InternVL2上的准确率变化热力图。

图19. MMStar上的准确率变化热力图（置零）。



(a) LLaVA-Next 上的准确率变化热力图。



(b) Qwen2-VL 上的准确率变化热力图。

图20. MMMU 上的准确率变化热力图（置零）。