AlignedNorm：通过耦合提示场提示视觉–语言模型

马琦[1 2] 王晨洋[1 2] 高德宏[3] 范登平[1 2 4]

摘要

视觉–语言模型（VLMs）的提示词学习主要通过端到端与解耦两种范式来平衡基类与新类上的性能。然而，现有方法面临一个根本瓶颈：在任务特定的特征空间中进行逐样本优化，容易使模型陷入局部最优，难以实现全局最优性。为解决这一问题，本文提出一个关键洞见：**耦合提示场**可以被用来更好地提示 VLMs——在这个共享的场下，基类与新类任务能够相互约束——并提出 **AlignedNorm**来强化场的耦合。通过将提示词范数和 VLMs 的原生尺度动态对齐，AlignedNorm 能够联合优化基类与新类任务。无需复杂设计，该方法在 4 种实验设置下的 15 个数据集上达到与当前领先的解耦方法相当的性能，为提示词学习中的局部最优困境提供了新的建模视角与实用的解决方案。代码见 https://github.com/QByteM/AlignedNorm。

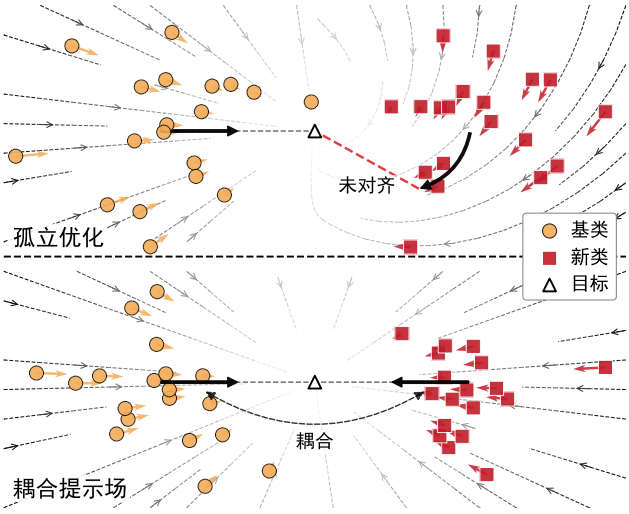


图 *1.* **耦合提示场**。本文通过将提示词诱导的变化视作一个共享场学习，以此耦合基类与新类任务。尽管二种任务之间存在差距，耦合场仍提供二者互相约束的校正，并在统一目标下趋向共同的优化方向。

1. 引言

提示词学习能够高效地将视觉–语言模型（VLMs），例如对比式语言–图像预训练模型（CLIP）（Radford et al., 2021），适配到广泛的下游任务。近期学界进展主要围绕适应性与泛化性的权衡展开。现有方法大体分为两类：端到端提示词学习 (Zhou et al., 2022a; Khattak et al., 2023b;a) 和解耦提示词学习 (Zhang et al., 2024; Li et al., 2025a; Guo & Gu, 2025)。

尽管上述两类范式取得了成功，它们仍难以兼顾基类适应与新类泛化。本文将这个问题归因于孤立的提示词优化，即提示词学到的知识被局限在特定任务的特征空间中。这促使本文将提示词学习视作场的构建：在特征空间上学习一种能在统一的推理规则下被基类和新类两个任务共享的变换。

本文认为提示词诱导的场应当跨越基类与新类任务形成耦合让二者相互约束，而不是被孤立地优化（见图 1）。然而，近期方法 (Zhang et al., 2024; Guo & Gu, 2025; 2026) 仍依赖单一任务的优化或推理机制，这可能削弱任务无关的耦合，并限制可靠的跨任务泛化。

为此，本文提出一种通过**耦合提示场**（CPF）实现提示词学习的框架。这里自然引发一个疑问：**构建这样一个耦合场的关键是什么？**本文观察到范数漂移会破坏 VLMs 所维持的均匀性–容忍度平衡 (Wang & Isola, 2020; Wang & Liu, 2021)，从而削弱任务无关的耦合。这使得动态范数对齐成为构建耦合提示场的必要条件。基于耦合场的分析（见 第 4.1 节），本文发现嵌入范数是关键因素，并由此引入一个简洁的基于范数的正则项 **AlignedNorm**来约束微调动态。

当 Transformer 中间层的提示词交互已经被削弱时，仅在投影后进行范数对齐可能并不够。实验发现，随机初始化的提示词与全局词元之间可能呈现较弱的注意力耦合；这一现象被称为纠缠坍塌，它会导致全局信息交换失败。因此，**AlignedNorm** 在每个引入提示词的编码层中将提示词与类别词元进行范数对齐，使耦合场保持稳定。

在不依赖外部知识蒸馏或解耦推理的情况下，AlignedNorm 在所评估的各类实验设置中展现出极有竞争力的性能和稳健的泛化能力。均匀性和容忍度指标进一步表明，MMRL++ (Guo & Gu, 2026) 中观察

[1] 南开大学视觉计算与智能感知实验室 & 计算机学院 [2] 南开国际先进研究院（深圳福田） [3] 西北工业大学 [4] 深圳河套学院. 通讯作者：范登平 <fdp@nankai.edu.cn>.

本文为 ICML2026 (Ma et al., 2026) 的中文翻译版，由聂博文翻译，范登平、马琦、王晨洋校稿。

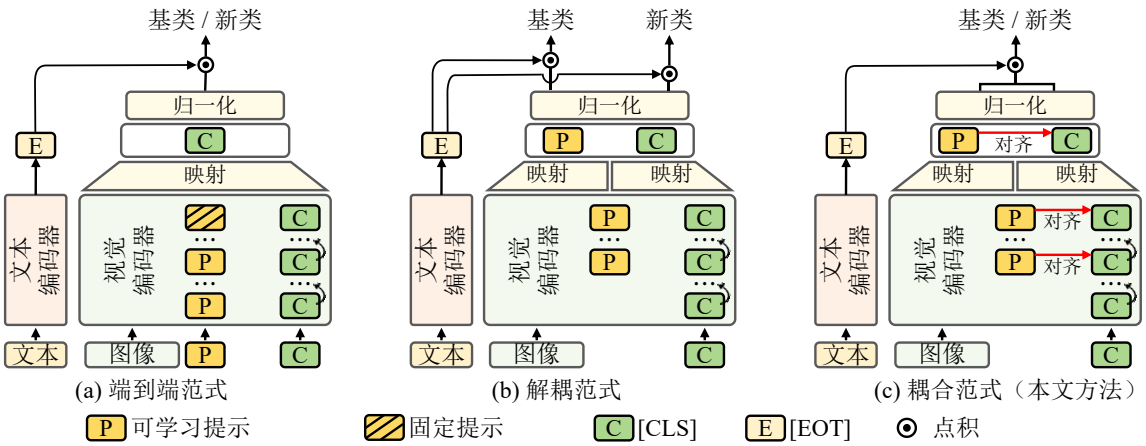


图 2. **VLMs 提示词学习范式比较**。(a) 端到端范式：提示词在基类数据上优化，并直接复用于基类与新类任务，因而常常需要两种任务的权衡。(b) 解耦范式：设计分离的模块分别用于基类学习与新类推理（*e.g.*，基类与新类任务使用不同的网络分支），测试时通常需要预先获得任务标识，即预先知道当前样本属于基类还是新类。(c) 耦合范式（本文）：本文通过在编码器每一层以及投影后的特征上进行范数对齐来耦合不同任务的分支，从而构建耦合提示场。

到的几何差距可由 AlignedNorm缓解，证明了本文方法的有效性。

综上所述，CPF 将 VLM 提示词学习诱导的变换建模为一个基类和新类任务共享的场，在该场中，基类任务与新类任务相互约束。这种建模方式既将预训练知识作为稳定且被两种任务共享的锚点保留下来，又允许在统一的推理规则下进行特定任务适配。本文进一步指出嵌入范数是稳定这种耦合的关键因素。AlignedNorm 将直观的投影后范数对齐与逐层范数对齐相结合：前者稳定最终的提示词诱导的场，后者缓解中间提示词交互中潜在的纠缠坍塌问题。在 15 个数据集和 4 种实验设置上，AlignedNorm 展现出极有竞争力的性能。本文希望 CPF 的建模方式及范数对齐方法能够为开辟高效且可泛化的多模态提示词学习新方向带来启发。

2. 相关工作

端到端提示词学习。早期工作 (Zhou et al., 2022a;b; Khattak et al., 2023a) 通过上下文优化来调整模型。这类方法存在严重的局限，其中最突出的是基类和新类任务的权衡，即微调后的模型在基类上表现越好，往往越难泛化到未见过的新类。

随后，主流方法主要围绕缓解这一权衡展开，大致可分为三类。最常用的是 **(i) 自正则化**：该类方法利用原始 CLIP 模型的内在知识来缓解过拟合 (Yao et al., 2023; Khattak et al., 2023b; Xie et al., 2025)，或者通过元学习策略有效实现正则化 (Park et al., 2024)。当自正则化信号逐渐趋于饱和时，**(ii) 外部参考** 则对其进行补充：其中一类方法通过模仿更强大的教师模型来增强泛化能力 (Li et al., 2024c)，另一类方法 (Ding et al., 2025; Khattak et al., 2025; Li et al., 2025b;c) 则借助大模型构造离散锚点，以此为优化提供引导；这些方法能够带来显著性能提升，但也会带来较大的计算开销。第三类通过 **(iii) 不确定性建模** 对模型最终输出进行松弛化操作，避免过于绝对的单一解导致过拟合 (Lu et al., 2022; Cheng & Han, 2025)。

这些方法试图通过端到端范式逼近预先定义好的理想结果；然而，基类与新类之间固有的差距使得预先定义的结果很难满足理想化的假设。

解耦提示词学习。近期，一些方法注意到，从基类任务中学习到的知识可能会干扰新类任务的性能，因此提出这两类知识应该分开建模。这类方法通常是在单独隔离的一个特征子空间中学习基类知识，以保留模型原有的泛化能力。具体而言，DePT (Zhang et al., 2024) 通过子网络实现知识隔离，DPC (Li et al., 2025a) 通过不同提示词实现这一目标，而 MMRL (Guo & Gu, 2025) 与 MMRL++ (Guo & Gu, 2026) 则通过建立不同的表征空间来实现这一点。这些方法能够保留预训练知识并缓解灾难性遗忘，但也存在一个明显局限：其操作通常需要预先获得任务标识，即事先知道当前输入应按基类还是新类处理；这在开放世界场景中往往并不现实。

现有范式主要关注如何优化最终输出，以同时满足基类与新类任务的要求。与之不同的是，本文从动态视角分析提示词学习的过程：不依赖预先定义的理想结果，而是关注提示词所诱导的有益变化，从而为后续方法提供扩展思路。

3. 前置知识

3.1. 端到端提示词学习

提示词学习 (Zhou et al., 2022a; Khattak et al., 2023a; Guo & Gu, 2025) 通过向视觉编码器、文本编码器或二者中同时引入可学习提示词来微调 CLIP。这里本文

给出形式化定义，给定输入图像 x，CLIP 视觉编码器第 L 层输出类别嵌入 $c_L(x) \in \mathbb{R}^D$，随后该嵌入由针对类别词的视觉投影层 $P_v^c : \mathbb{R}^D \to \mathbb{R}^d$ 映射到共享图文空间：$z_c(x) = P_v^c(c_L(x)) \in \mathbb{R}^d$. 对于 C 个类别，相应的 ℓ_2 归一化文本类别特征记为 $\{w_k\}_{k=1}^C$，其中 $w_k \in \mathbb{R}^d$。图像特征进一步被 ℓ_2 归一化为 $f_c(x) = \frac{z_c(x)}{\|z_c(x)\|}$，并定义温度系数缩放后的分类器 $\pi(\cdot)$，其中 $\tau > 0$：$\pi(f) = \mathrm{Softmax}\big(\frac{1}{\tau}[w_k^\top f]_{k=1}^C\big)$，其中 $\pi_k(f)$ 表示第 k 类的预测概率。给定训练样本 (x, y)，优化过程使用交叉熵损失 $\mathcal{L}_{ce}(\pi(f_c(x)), y)$。

3.2. 解耦提示词学习

最新的解耦提示词学习 (Guo & Gu, 2025; 2026) 通过将可学习提示词注入 CLIP 中更深的编码层，并将平均后的提示词作为额外分类分支，从而扩展提示词学习范式。令 $\mathcal{V} = \{\mathcal{V}_i\}_{i=1}^L$ 表示一个隐藏层宽度为 D 的 L 层 ViT 视觉编码器。记第 i 层，类别词元为 $c_i \in \mathbb{R}^D$，图像切分好的词元为 $E_i \in \mathbb{R}^{M \times D}$，提示词为 $P_i \in \mathbb{R}^{n_p \times D}$。提示词从指定层 J 开始插入：

$$\begin{aligned}[c_i, E_i] &= \mathcal{V}_i([c_{i-1}, E_{i-1}]), \quad i = 1, \ldots, J-1, \\ [c_j, P_j, E_j] &= \mathcal{V}_j([c_{j-1}, \widetilde{P}_j, E_{j-1}]), \quad j = J, \ldots, L,\end{aligned} \tag{1}$$

其中 $\widetilde{P}_j \in \mathbb{R}^{n_p \times D}$ 表示第 j 层的提示词输入。在最后一层之后，提示词通过平均池化进行聚合，$\bar{P}_L(x) = \mathrm{Mean}(P_L(x))$，并由 $P_v^p : \mathbb{R}^D \to \mathbb{R}^d$ 投影到共享空间，得到 $z_p(x) = P_v^p(\bar{P}_L(x))$；随后进行 ℓ_2 归一化，得到 $f_p(x) = \frac{z_p(x)}{\|z_p(x)\|}$。连同前文定义的类别词元嵌入 $f_c(x)$，两个分支均由同一个分类器 $\pi(\cdot)$ 进行分类。

视觉锚定。视觉锚定通过自正则化 (Khattak et al., 2023b; Roy & Etemad, 2024) 实现，并且仅作用于 f_c，从而将两个分支解耦：f_c 用于保留 CLIP 语义，而 f_p 则侧重于特定任务的适配。

优化目标。给定训练样本 (x, y)，该方法通过 $\mathcal{L}_{\text{task}}$ (Guo & Gu, 2025; 2026) 进行优化。

解耦推理。对于基类任务，两个分支 (Guo & Gu, 2025; 2026) 按如下方式组合：

$$f_b(x) = (1-\alpha)\, f_c(x) + \alpha\, f_p(x), \tag{2}$$

并通过 $\hat{p}_k(x) = \pi_k(f_b(x))$ 得到预测；对于新类，则通过 $\hat{p}_k(x) = \pi_k(f_c(x))$ 得到预测。

4. 方法

4.1. 耦合提示场

动机。现有的提示词学习方法难以预先定义理想的最终结果，其主要原因在于固有的基类—新类差距。受经典观点的启发 (He et al., 2016; Chen et al., 2018; Lipman et al., 2023; Ilievski et al., 2025)，本文认为，由提示词所引发的变化在跨越基类—新类差距时具有更强的可迁移性。基于先进的解耦式深度提示词学习框架 (Guo & Gu, 2026)，本文能够将这种变化显式地建模为特征空间上的一个场，只需一条统一的、任务无关的推理规则。

场的构建。式 (2) 中的解耦推理为基类构造了一个混合嵌入。将其改写为锚定形式，可以更清晰地揭示其结构：

$$f_b(x) = f_c(x) + \alpha\big(f_p(x) - f_c(x)\big). \tag{3}$$

由于视觉锚定的存在，$f_c(x)$ 可被视为特征空间中的稳定锚点。因此，本文将提示场定义为：

$$\mathbf{u}_f(x) \triangleq \alpha\big(f_p(x) - f_c(x)\big) \in \mathbb{R}^d. \tag{4}$$

耦合提示场。解耦学习在实践中的一个严重局限在于：在开放世界评估中通常无法获得任务标识，也就是说，模型并不能预先知道当前样本应按基类还是新类处理。提示场视角由此给出了一种适用于所有任务的统一推理策略：$\hat{p}_k(x) = \pi_k\big(f_c(x) + \mathbf{u}_f(x)\big)$. 这将解耦推理转化为一种任务无关的耦合要求：同一个场 $\mathbf{u}_f(\cdot)$ 必须能够跨越基类和新类的差距实现泛化。

场畸变。在耦合推理下，定义范数比 $r(x) \triangleq \frac{\|z_p(x)\|}{\|z_c(x)\|}$. 于是，耦合场依赖于逐样本的范数比 $r(x)$：

$$\begin{aligned}\mathbf{u}_f(x) &= \alpha\left(\frac{z_p(x)}{\|z_p(x)\|} - \frac{z_c(x)}{\|z_c(x)\|}\right) \\ &= \frac{\alpha}{\|z_c(x)\|}\left(r(x)^{-1} z_p(x) - z_c(x)\right).\end{aligned} \tag{5}$$

如 图 3(a) 所示，MMRL++ 中的范数比 $r(x)$ 在基类和新类任务上存在明显的差距：图中 $\mathcal{D}_{\text{base}}$ 与 $\mathcal{D}_{\text{new}}$ 的取值明显分离。$r(x)$ 的这种偏移意味着，在同一耦合规则下，提示词诱导的偏移会被施加上不一致的尺度，从而使耦合提示场在跨越基类和新类的差距时发生非预期畸变：

$$\mathbb{E}_{x \sim \mathcal{D}_{\text{base}}}\big[r(x)\big] \;\neq\; \mathbb{E}_{x \sim \mathcal{D}_{\text{new}}}\big[r(x)\big], \tag{6}$$

进而破坏任务无关性与全局一致性。

非均匀场更新。仅通过归一化嵌入 $f_p(x)$ 分析任意一个对提示词分支依赖的损失项 $\mathcal{L}$。记 $g(x) \triangleq \partial\mathcal{L}/\partial f(x) \in \mathbb{R}^d$。由反向传播得到：

$$\frac{\partial \mathcal{L}}{\partial z(x)} = \frac{1}{\|z(x)\|}\Big(I - f(x) f(x)^\top\Big)\, g(x). \tag{7}$$

因此，梯度幅值会被 $\|z(x)\|$ 反向缩放。假设随机梯度可分解为 $g(x) = \bar{g}(x) + \xi(x)$，且 $\mathbb{E}[\xi(x)] = 0$。则提示词分支上的更新噪声会被放大为：

$$\mathbb{E}\Big[\big\|\nabla_z \mathcal{L}(x) - \mathbb{E}\big[\nabla_z \mathcal{L}(x)\big]\big\|^2\Big] \;\le\; \mathbb{E}\Big[\frac{\|\xi(x)\|^2}{\|z(x)\|^2}\Big]. \tag{8}$$

因此，当 $\|z_p(x)\|$ 小于 $\|z_c(x)\|$ 时，式 (8) 中的 $1/\|z_p(x)\|$ 因子会放大提示词分支上的有效更新尺度，使其对随机梯度噪声更加敏感（见 附录 B.1）。

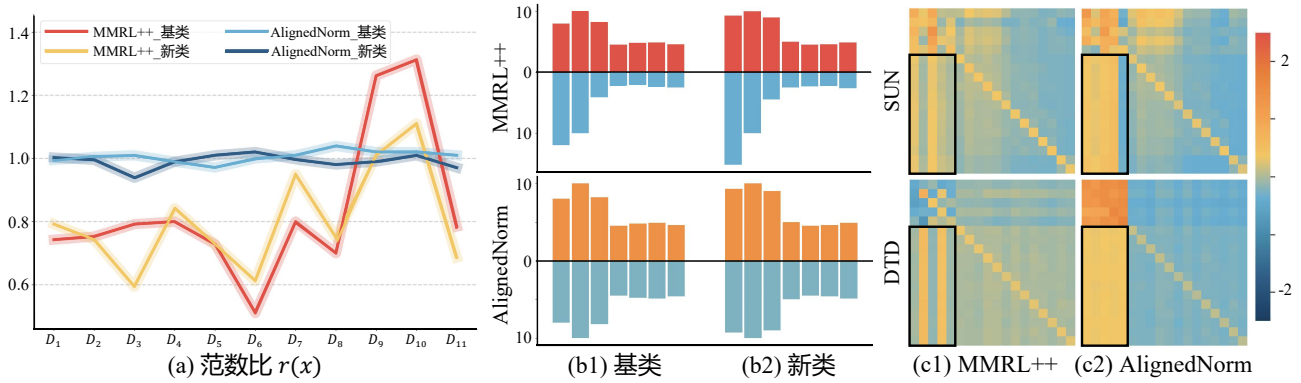


图 3. **实验性观察**。(a) 在 11 个数据集上可视化 MMRL++（基类 与 新类）和 AlignedNorm （基类 与 新类）的 $r(x)$。(b) 在 ImageNet 上统计 7 个层中的词元范数，其中包括 MMRL++（类别词元 与 提示词）和 AlignedNorm （类别词元 与 提示词）。(c) 在两个数据集上可视化注意力分数（前五个词元为提示词元）。MMRL++ 中的提示词元处于孤立状态，受到的注意力较弱；而 AlignedNorm 恢复了提示–图像块耦合，从而促进更充分的全局信息交换。

场对扰动的稳定性。考虑提示词分支上的一个随机扰动：$z_p = z_p^\star + \varepsilon,\quad \mathbb{E}[\varepsilon \mid x] = 0$, 其中 ε 表示由有限数据和随机优化引入的估计噪声。在 $z_p^\star$ 邻域作一阶近似，$f_p(x) - f_p^\star(x) \approx \frac{1}{\|z_p^\star(x)\|}\big(I - f_p^\star(x) f_p^{\star\top}(x)\big)\,\varepsilon$, 得到上界：

$$\mathbb{E}\big[\|\mathbf{u}_f(x) - \mathbf{u}_f^\star(x)\|_2^2 \mid x\big] \approx \alpha^2\,\mathbb{E}\big[\|f_p(x) - f_p^\star(x)\|_2^2 \mid x\big] \le \alpha^2\,\frac{\mathbb{E}[\|\varepsilon\|_2^2 \mid x]}{\|z_p^\star(x)\|^2}. \tag{9}$$

锚定机制使 f_c 保持相对稳定，因此，控制 $\|z_p(x)\|$ 是提升耦合提示场噪声鲁棒性的关键（见 附录 B.2）。

4.2. AlignedNorm

式 (5) 与 式 (6) 表明，基类任务和新类任务之间 $r(x)$ 的不一致会导致耦合提示场发生畸变；式 (8) 与 式 (9) 进一步揭示，提示词分支范数会影响耦合提示场的均匀性与稳定性。综合来看，这些结果说明：稳定投影后的提示词范数，是实现稳健的、任务无关耦合的关键。

投影后范数对齐。因此，本文直接对最终投影后的提示词特征进行正则化，通过 ℓ_1 惩罚使其范数与类别分支的投影范数对齐：

$$\mathcal{L}_{\text{proj}}(x) = \Big|\,\|z_p(x)\| - \text{sg}\big(\|z_c(x)\|\big)\,\Big|, \tag{10}$$

其中 $\text{sg}(\cdot)$ 表示停止梯度。然而，仅在投影后进行对齐的收益有限，因此还需分析编码器内部的范数动态。

纠缠坍塌。由于注意力得分由点积驱动，可写为 $q^\top k = \|q\|\,\|k\|\cos\theta$。在高维空间中，随机初始化的提示词方向以较高概率近似正交于其他词元，因此在训练早期 $\cos\theta$ 接近 0（见 附录 B.3），导致提示词得到其他词元的注意力得分较低且容易失衡。训练过程中，提示词的范数会大于类别词元的范数（见 图 3(b)）；在方向未对齐的情况下，这会进一步放大得分差距，并使 softmax 门控迅速饱和到难以恢复的单边主导状态，造成许多全局词元会给提示词分配近乎为零的注意力（见 图 3(c)），从而造成纠缠坍塌：提示词无法与全局词元进行语义信息交换（见 附录 B.4）。

算法 1 AlignedNorm 损失

输入： 图像 x，类别词元 $\{c_l(x)\}_{l=J}^{L}$，提示词 $\{P_l(x)\}_{l=J}^{L}$，投影输出 $z_c(x), z_p(x)$，权重 β, γ
输出： 对齐损失 $\mathcal{L}_{\text{align}}$
// **1. 投影后范数对齐**
$n_{z_c} \leftarrow \|z_c(x)\|,\quad n_{z_p} \leftarrow \|z_p(x)\|$
$\mathcal{L}_{\text{proj}} \leftarrow \big|\, n_{z_p} - \text{sg}(n_{z_c})\,\big|$
初始化 $\mathcal{L}_{\text{token}} \leftarrow 0$
// **2. 逐层范数对齐**
for $l = J$ **至** L **do**
  $n_c \leftarrow \|c_l(x)\|,\quad n_p \leftarrow \|\text{Mean}(P_l(x))\|$
  $\mathcal{L}_{\text{token}} \leftarrow \mathcal{L}_{\text{token}} + \big|\, n_p - \text{sg}(n_c)\,\big|$
end for
$\mathcal{L}_{\text{align}} \leftarrow \beta\,\mathcal{L}_{\text{proj}} + \gamma\,\mathcal{L}_{\text{token}}$
return $\mathcal{L}_{\text{align}}$

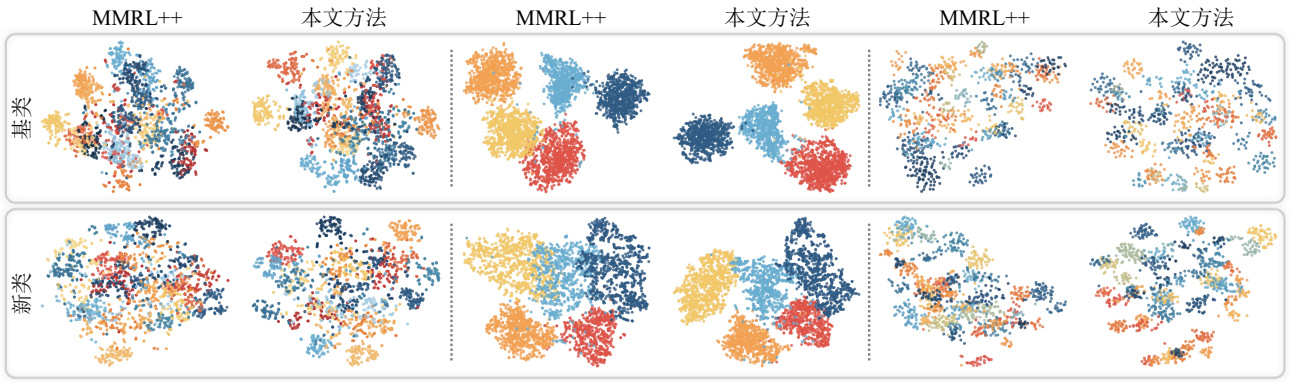
逐层范数对齐。式 (10) 中的对齐能够稳定输出范数 $\|z_p(x)\|$，但并不能直接控制 Transformer 内部提示词的范数。因此，在每个引入提示词的层中，将提示词的范数与类别词元对齐。对于 $l = J, \ldots, L$，定义每层平均池化后的提示词为 $\bar{P}_l(x) \triangleq \text{Mean}(P_l(x))$，并施加如下约束：

$$\mathcal{L}_{\text{token}} = \sum_{l=J}^{L} \Big|\,\|\bar{P}_l(x)\| - \text{sg}\big(\|c_l(x)\|\big)\,\Big|. \tag{11}$$

总体目标。两类范数对齐被合并为一个统一损失（见 Algorithm 1）：

$$\mathcal{L}_{\text{align}} = \beta\,\mathcal{L}_{\text{proj}} + \gamma\,\mathcal{L}_{\text{token}} \tag{12}$$

其中 β 与 γ 分别控制两项损失的相对强度；整体训练目标为 $\mathcal{L} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{align}}$。



(a) DTD (纹理分类) (b) EuroSAT (卫星图像分类) (c) OxfordFlowers (细粒度分类)

图 4. **t-SNE 可视化**。解耦方法 MMRL++ 与 AlignedNorm 在三个不同的图像识别数据集上的嵌入可视化。AlignedNorm 在基类和新类任务上呈现出更清晰的可分性，能在一定程度上反映本文方法的有效性。

5. 实验

5.1. 实验设置

遵循先前工作 (Guo & Gu, 2025; 2026)，本文在以下四个基准上评估 AlignedNorm 的能力：

数据集。 AlignedNorm 在 11 个数据集上进行评估，以考察其在多样化任务场景中的泛化能力，涵盖通用目标识别（ImageNet (Deng et al., 2009)、Caltech101 (Fei-Fei et al., 2004)）、细粒度分类（OxfordPets (Parkhi et al., 2012)、StanfordCars (Krause et al., 2013)、Flowers102 (Nilsback & Zisserman, 2008)、Food101 (Bossard et al., 2014) 和 FGVCAircraft (Maji et al., 2013))、动作识别（UCF101 (Soomro et al., 2012))、纹理分类（DTD (Cimpoi et al., 2014))、场景识别（SUN397 (Xiao et al., 2010)）以及卫星图像分类（EuroSAT (Helber et al., 2019)）。此外，还在 ImageNet-A (Hendrycks et al., 2021b)、ImageNet-R (Hendrycks et al., 2021a)、ImageNet-Sketch (Wang et al., 2019) 和 ImageNetV2 (Recht et al., 2019) 上评估 AlignedNorm 对领域偏移的鲁棒性。

基类到新类泛化。为评估模型在单个数据集内的开放集泛化能力，每个数据集被均匀划分为两个类别互不重叠的子集：基类与新类。模型仅在基类上训练，并同时在基类和新类上测试。该任务要求模型在学习能力与类别级泛化之间取得平衡，并通过基类与新类准确率的调和平均（HM）进行量化。

跨数据集迁移。在该设置下，源域上训练的模型在 ImageNet (Deng et al., 2009) 的全部 1,000 个类别上训练，并在不进行任何微调的情况下，直接在其余 10 个数据集上评估。

少样本学习。该基准旨在考察模型在数据稀缺条件下的学习能力，*i.e.*，使用每个数据集的 K-shot 样本（K = {1, 2, 4, 8, 16}）训练模型，并在完整测试集上进行测试。模型需要在获取特定任务知识的同时，保留 CLIP 的预训练表征。

领域泛化。该设置在专门构造的分布外数据集上评估模型的鲁棒性。源域上训练的模型在 ImageNet (Deng et al., 2009) 的全部类别上训练，并迁移到四个分布偏移领域。

实现细节。 AlignedNorm 基于 MMRL++ 模型构建，报告结果为 3 个随机种子的加权平均。实验设置与 MMRL++ 的默认配置保持一致。更多实现细节见 附录 C。所有实验均在 RTX 3090 GPU 上完成。

5.2. 基类到新类泛化实验

如 表 1 所示，实验选取端到端与解耦两类范式中的若干代表性方法进行比较。在端到端设置下，对比 MaPLe (Khattak et al., 2023a)、PromptSRC (Khattak et al., 2023b) 和 HicroPL (Zheng et al., 2025)。对于解耦策略，选取两个最具竞争力的模型 (Guo & Gu, 2025; 2026)。为保证公平比较，进一步移除这些方法中的解耦推理，并使用相同的基类推理规则进行评估，这会导致泛化性能明显下降。相比之下，AlignedNorm 在不依赖解耦推理的情况下提升了新类准确率，同时没有损害基类性能。如 图 4 所示，AlignedNorm 得到的新类嵌入具有更清晰的可分性。

AlignedNorm 在 FGVCAircraft 的新类上出现轻微下降。一个可能原因是，该数据集的数据分布与 CLIP 的预训练数据存在差异，从而无法匹配 CLIP 的表征空间。由于本方法倾向于保留 CLIP 原始空间结构，模型在该场景下可用于适配的空间可能更有限，进而限制学习并影响泛化。

除经典的准确率外，本文进一步采用 *均匀性* 与 *容忍度* (Wang & Isola, 2020; Wang & Liu, 2021)，用于比较 MMRL++ 与 AlignedNorm 学到的表征的几何特性。这些指标揭示范数对齐如何影响嵌入空间，使 AlignedNorm 的实验性能更具可解释性。

均匀性。令 $\mathcal{D} = \{(f_i, y_i)\}_{i=1}^{N}$ 表示新类的 ℓ_2 归一化嵌

表 1. **基类到新类的泛化实验**。表中报告各数据集在基类和新类上的准确率（%），HM 表示调和平均。Δ 表示 AlignedNorm 相较于不使用解耦策略的 MMRL++ 所取得的提升。

方法	平均			ImageNet			Caltech101			OxfordPets		
	基类	新类	HM	基类	新类	HM	基类	新类	HM	基类	新类	HM
CLIP (Radford et al., 2021)	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
基于 **端到端**范式的方法：												
MaPLe (Khattak et al., 2023a)	81.91	75.09	78.35	76.60	70.77	73.57	97.73	95.30	96.50	95.70	98.07	96.87
PromptSRC (Khattak et al., 2023b)	84.23	75.78	79.78	77.60	70.37	73.81	98.07	94.00	95.99	95.23	97.17	96.19
HicroPL (Zheng et al., 2025)	85.16	76.50	80.60	78.34	71.68	74.86	98.36	95.45	96.88	95.62	97.69	96.64
基于 **解耦**范式的方法：												
MMRL (Guo & Gu, 2025)	85.71	76.28	80.72	77.70	71.20	74.31	98.83	94.33	96.53	95.97	97.50	96.73
不使用解耦策略	85.71	72.60	78.61	77.70	69.13	73.16	98.83	94.83	96.79	95.97	94.70	95.33
MMRL++ (Guo & Gu, 2026)	85.43	77.79	81.43	77.60	71.40	74.37	98.90	94.40	96.60	95.43	96.97	96.19
不使用解耦策略	85.43	76.48	80.71	77.60	71.30	74.32	98.90	94.57	96.69	95.43	96.87	96.14
基于 **耦合提示场**的方法：												
AlignedNorm	85.46	77.79	81.45	77.60	71.47	74.41	98.90	94.77	96.79	95.63	97.43	96.52
Δ	**+0.03**	**+1.31**	**+0.74**	+0.00	**+0.17**	**+0.09**	+0.00	**+0.20**	**+0.10**	**+0.20**	**+0.56**	**+0.38**

方法	StanfordCars			Flowers102			Food101			FGVCAircraft		
	基类	新类	HM	基类	新类	HM	基类	新类	HM	基类	新类	HM
CLIP (Radford et al., 2021)	63.37	74.89	68.65	72.08	77.80	74.83	90.10	91.22	90.66	27.19	36.29	31.09
基于 **端到端**范式的方法：												
MaPLe (Khattak et al., 2023a)	72.30	73.80	73.04	96.03	73.33	83.16	90.70	92.03	91.36	36.07	34.47	35.25
PromptSRC (Khattak et al., 2023b)	78.20	75.47	76.81	98.07	77.37	86.50	90.63	91.50	91.06	43.33	36.27	39.49
HicroPL (Zheng et al., 2025)	81.13	75.04	77.97	98.10	74.75	84.85	90.74	91.72	91.23	46.06	37.61	41.41
基于 **解耦**范式的方法：												
MMRL (Guo & Gu, 2025)	81.30	74.83	77.93	98.97	76.97	86.59	90.57	91.53	91.05	46.13	37.47	41.35
不使用解耦策略	81.30	70.00	75.23	98.97	73.03	84.04	90.57	89.90	90.23	46.13	35.10	39.87
MMRL++ (Guo & Gu, 2026)	81.23	75.23	78.11	98.50	77.47	86.73	90.50	91.70	91.10	46.47	38.50	42.11
不使用解耦策略	81.23	72.43	76.58	98.50	73.80	84.38	90.50	91.57	91.03	46.47	38.63	42.19
基于 **耦合提示场**的方法：												
AlignedNorm	81.70	73.83	77.57	98.40	76.03	85.78	90.57	91.60	91.08	46.20	38.60	42.06
Δ	**+0.47**	**+1.40**	**+0.99**	-0.10	**+2.23**	**+1.40**	**+0.07**	**+0.03**	**+0.05**	-0.27	-0.03	-0.13

方法	SUN397			DTD			EuroSAT			UCF101		
	基类	新类	HM	基类	新类	HM	基类	新类	HM	基类	新类	HM
CLIP (Radford et al., 2021)	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
基于 **端到端**范式的方法：												
MaPLe (Khattak et al., 2023a)	80.80	78.33	79.55	79.87	57.60	66.93	91.70	75.10	82.57	83.53	77.23	80.26
PromptSRC (Khattak et al., 2023b)	82.50	78.87	80.64	83.27	61.50	70.75	92.80	72.20	81.21	86.87	78.83	82.65
HicroPL (Zheng et al., 2025)	83.25	78.99	81.06	83.60	65.30	73.33	94.04	72.28	81.74	87.47	81.02	84.12
基于 **解耦**范式的方法：												
MMRL (Guo & Gu, 2025)	83.07	79.23	81.10	85.87	64.10	73.40	96.10	72.33	82.54	88.30	79.63	83.74
不使用解耦策略	83.07	76.63	79.72	85.87	59.77	70.48	96.10	59.20	73.27	88.30	76.30	81.86
MMRL++ (Guo & Gu, 2026)	82.93	79.57	81.22	85.07	65.83	74.22	95.73	84.17	89.58	87.37	80.43	83.76
不使用解耦策略	82.93	78.43	80.62	85.07	61.70	71.52	95.73	81.63	88.12	87.37	80.40	83.74
基于 **耦合提示场**的方法：												
AlignedNorm	82.93	79.13	80.99	84.63	65.23	73.67	96.10	86.63	91.12	87.43	81.00	84.09
Δ	+0.00	**+0.70**	**+0.37**	-0.44	**+3.53**	**+2.15**	**+0.37**	**+5.00**	**+3.00**	**+0.06**	**+0.60**	**+0.35**

入集合。均匀性通过成对距离上的高斯势来度量：

$$\mathcal{L}_{\text{unif}} = \log \mathbb{E}_{i \neq j} \left[\exp \left(-t \left\| f_i - f_j \right\|^2 \right) \right], \tag{13}$$

其中遵循 CLIP 的默认设置，取 $t = 0.01$。较小的 $\mathcal{L}_{\text{unif}}$ 表明嵌入在超球面上分布得更加均匀，这通常对应于新类之间更好的可分性。

容忍度。容忍度刻画同一新类内部嵌入的局部聚合程度。对于归一化嵌入，本文将同类样本对上余弦相似度的期望作为计算结果：

$$\text{Tol} = \mathbb{E}_{i \neq j} \left[f_i^\top f_j \cdot \mathbb{I}(y_i = y_j) \right], \tag{14}$$

其中 $\mathbb{I}(\cdot)$ 为指示函数。更高的 Tol 表明更强的类内紧致性；但若过度优化均匀性，语义相近的样本可能被推得过远，从而降低容忍度。

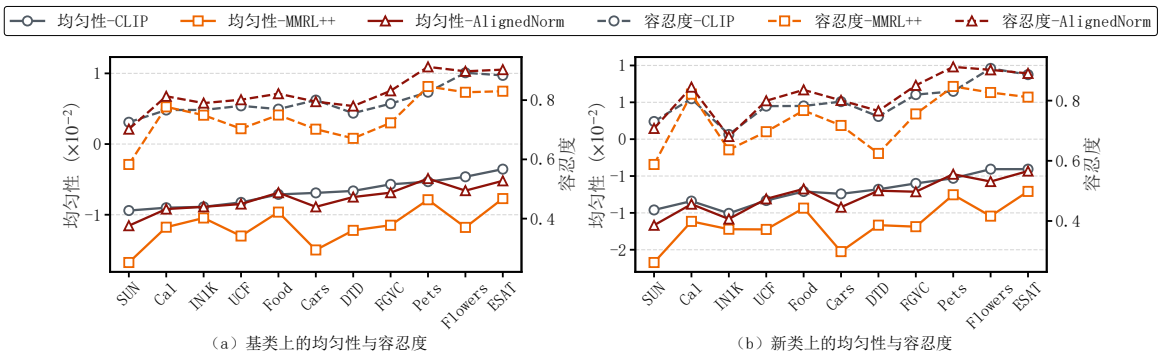
如 图 5 所示，MMRL++ 与 AlignedNorm 的均匀性和容忍度被进一步可视化。对于新类，AlignedNorm 取得了更接近原始 CLIP 的均匀性，说明耦合提示场以全局一致的方式重塑了嵌入空间。此外，AlignedNorm 获得了与 CLIP 相当甚至更高的容忍度，而 MMRL++ 的容忍度显著更低，这表明基类知识得到了更有效的迁移，并能更好地泛化到新类。

5.3. 跨数据集实验

如 表 2 所示，AlignedNorm 在不依赖任何解耦策略的情况下提升了跨数据集泛化能力。其中AlignedNorm 在 10 个数据集上取得了一致增益，体现出其对跨数据集分布偏移更强的鲁棒性。

5.4. 小样本实验

在小样本实验下，MMRL++ 在所有样本数设置中均采用统一的推理策略。如 表 4 所示，在 AlignedNorm 中引入范数对齐并不会损害性能；相反，其收益在样本数极少的场景下更为明显。这一现象表明，当监督信号稀缺时，范数对齐主要通过提升优化稳定性发挥



（a）基类上的均匀性与容忍度　　（b）新类上的均匀性与容忍度

图 5. **均匀性与容忍度可视化**。MMRL++ 在两项指标上始终落后于 CLIP，而AlignedNorm 达到了与 CLIP 相当的均匀性，并进一步提升了容忍度。这些结果表明 AlignedNorm 能在不同数据集、架构和评估设置下有效平衡两项关键性质。

表 2. **跨数据集基准评估**。值得注意的是，在测试时，AlignedNorm 在没有任何额外线索的情况下取得了有竞争力的性能，体现出其在多样化领域中的稳健泛化能力。

	源	目标										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	平均
PromptSRC	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	**46.87**	45.50	68.75	65.81
MMRL	71.93	94.40	91.20	65.67	72.90	86.23	26.43	67.40	46.50	52.27	69.03	67.20
不使用解耦	**71.93**	94.17	89.53	63.40	71.77	84.87	24.80	65.80	46.40	53.10	68.40	66.22
MMRL++	71.83	94.57	91.30	66.53	73.30	86.63	26.23	67.90	46.47	53.50	69.60	67.60
不使用解耦	71.83	94.40	91.17	65.87	72.27	86.07	26.27	67.50	46.50	**55.50**	69.10	67.47
AlignedNorm	71.83	**94.50**	**91.57**	**66.50**	**73.00**	**86.67**	**26.30**	**67.97**	46.63	53.47	**69.50**	**67.61**

表 3. **领域泛化**。所有方法均在 ImageNet 上训练，并在 4 个存在领域偏移的数据集上评估。

	源域	目标域				
	ImageNet	-V2	-S	-A	-R	平均
CLIP	66.73	60.83	46.15	47.77	73.96	57.17
MMRL	**71.93**	64.57	49.13	50.70	77.20	60.40
不使用解耦	**71.93**	**65.37**	49.10	49.57	77.00	60.26
MMRL++	71.83	64.37	49.23	**51.10**	77.63	60.58
不使用解耦	71.83	65.27	**49.53**	50.80	**78.10**	60.93
AlignedNorm	71.83	65.13	**49.53**	**51.10**	78.07	**60.96**

作用；此时模型更容易受到过拟合与表征漂移的影响。总之，这些结果说明范数对齐有助于保留可迁移知识。

5.5. 域泛化实验

在存在领域偏移的情形下，本文进一步评估领域泛化能力。按照既有做法，模型在 ImageNet 上训练，并在不进行任何额外微调的情况下直接在目标数据集上评估。如 表 3 所示，AlignedNorm 在整体上保持了稳定且具有竞争力的表现，说明其在领域偏移下具有稳健的泛化能力，尤其适用于从 ImageNet 迁移到分布不匹配的目标数据集的场景。

表 4. **少样本学习**。AlignedNorm 在低样本设置下展现出高度竞争力。详细结果见 附录 11。

方法	1-shot	2-shot	4-shot	8-shot	16-shot
线性探测 CLIP	45.83	57.98	68.01	74.47	78.79
PromptSRC	72.27	75.24	78.20	80.55	82.76
MMRL	72.62	**75.74**	79.10	**81.40**	**84.28**
MMRL++	72.57	75.35	78.95	81.30	84.07
AlignedNorm	**72.93**	75.57	**79.11**	81.35	84.08

6. 讨论与未来展望

范式差距。已有研究分别从通道分布 (Zhang et al., 2024; 2026) 与优化 (Li et al., 2025a) 视角说明了解耦提示词学习相较端到端提示词学习的优势，表明将面向基类适应与泛化知识保持分离是有益的。然而，解耦方法通常需要在测试时获得任务标识，由此造成推理差距。受基于均匀性与容忍度的表征几何视角启发，如 图 5 所示，AlignedNorm 通过耦合提示场弥合了这一推理差距，同时仍遵循解耦思想的核心：保护预训练知识，并避免基类适应对泛化能力的破坏。CPF 并不试图建模一个跨基类与新类的理想最终结果，而是建模更具可迁移性的变化。

计算开销。AlignedNorm 仅在训练阶段额外加入一个范数对齐损失并不引入过多的计算量。如 表 5 所示，

表 5. **计算开销**。在 batch size 为 32 时，于 ImageNet 上统计的平均训练与推理时间。

方法	训练 (s)	测试 (s)	参数量 (M)	HM
PromptSRC	**0.350**	0.124	0.046	79.78
HiCroPL	0.362	0.163	0.246	80.60
MMRL++	0.357	**0.087**	**0.045**	80.71
AlignedNorm	0.357	**0.087**	**0.045**	**81.45**

表 6. **不同组件的影响**。结果为 11 个数据集上的平均值。HM 表示调和平均。

方法	基类	新类	HM
1: MMRL++	85.43	76.48	80.71
2: + $\mathcal{L}_{\text{token}}$	85.16	76.37	80.52
3: + $\mathcal{L}_{\text{proj}}$	85.42	76.96	80.97
4: + $\mathcal{L}_{\text{token}}$ + $\mathcal{L}_{\text{proj}}$	**85.46**	**77.79**	**81.45**

相比 MMRL++，这一设计几乎不会带来额外训练时间开销，也不会增加可学习参数量。由于对齐损失在推理阶段并不使用，AlignedNorm 与 MMRL++ 具有完全相同的测试开销。因此，AlignedNorm 在提升其任务无关耦合行为的同时，保留了 MMRL++ 相较其他提示词学习方法的效率优势。

范数对齐消融。表 6 评估了本文的范数对齐策略。投影后对齐能够带来稳定增益，而进一步加入逐层对齐可以继续提升性能。然而，单独使用层间或投影后对齐都会不同程度影响基类性能，尤其是仅使用层间对齐时更为明显。这表明两个层级需要以互补方式协同工作，才能取得最佳结果。

损失函数消融。表 7 比较了范数对齐的不同损失函数。铰链损失旨在调整提示词与类别词元之间的相对范数关系，但并未提升 HM，说明仅施加不等式形式的约束并不足够。平滑 L_1 损失与环形损失 (Zheng et al., 2018) 能够产生更强的范数收缩或范数匹配效果，但二者都没有达到 L_1 对齐的性能。这表明，过于激进的范数收敛可能造成数值捷径，使模型转而专注于满足范数约束，而不是学习语义特征。还需要注意的是，环形损失通常将特征范数对齐到一个固定目标，而本文的对齐则动态使用类别词元范数作为样本相关锚点。这种动态锚定更契合耦合提示场的目标。

对齐权重选择。表 8 报告了 11 个数据集上使用的对齐权重 (β, γ)。其中，β 与 γ 分别控制投影后范数对齐和逐层范数对齐的强度。在实践中，本文遵循两条经验原则。首先，它们的取值应使对齐损失与交叉熵损失保持在相近量级；过大的权重可能导致模型过度优化范数匹配，从而抑制语义学习。其次，本文设置 $\gamma \leq \beta$，使投影表示的范数优先得到稳定，再对中间提示词施加约束。这种层级安排为逐层对齐提供优化信号，实验结果表明其有助于缓解纠缠坍塌。

局限性与未来工作。当前的耦合提示场采用单步优化方案，这一设计相对高效，但也可能限制其在更具挑战性的实验设置下的进一步增益。未来工作将探索多步优化，以进一步提升鲁棒性。对于范数控制而言，更自适应的策略可能带来更高效的训练。此外，本文主要研究 CLIP 等对比式 VLMs 中的范数动态；在这类模型中，图像与文本表征会在归一化嵌入空间中对齐。类似效应在生成式多模态架构中是否成立有待研究。

表 7. **损失函数消融**。结果显示，$L1$ 损失带来更好的性能。HM 表示调和平均。

方法	基类	新类	HM
1: MMRL++	85.43	76.48	80.71
2: + 铰链损失	**85.44**	76.19	80.55
3: + 平滑 $L1$ 损失	85.32	76.83	80.85
4: + 环形损失	85.29	76.94	80.90
5: + $L1$ 损失	85.42	**76.96**	**80.97**

表 8. **对齐权重选择**。基类到新类泛化实验中各数据集使用的 β 与 γ 的取值。

数据集	β	γ	数据集	β	γ
ImageNet	0.005	0.005	UCF	0.150	0.150
EuroSAT	0.100	0.050	DTD	0.200	0.150
Caltech	0.050	0.050	SUN	0.100	0.010
Aircraft	0.100	0.010	Cars	0.150	0.150
Flowers	0.150	0.100	Pets	0.100	0.010
Food	0.150	0.001			

7. 结论

本文提出了一种耦合提示场建模范式，为理解基类到新类泛化权衡提供了新的视角，并进一步识别出此前被忽视的嵌入范数作用，通过 AlignedNorm予以处理；该方法简单而有效，同时施加逐层范数对齐与投影后范数对齐。大量实验表明，在不依赖解耦推理或外部蒸馏的情况下，AlignedNorm 能够在多种任务设置中取得高度竞争性的性能。本文希望这项研究能够为理解提示词调优的内在机制提供新的启发。

致谢

本工作部分受到国家自然科学基金（编号 62476143）和中央高校基本科研业务费专项资金（南开大学，编号 63263250）的支持。感谢邹先予、陈龙、杜泽文、聂博文（NKU）、季葛鹏（ANU）、张煜坤（BIT）以及沈冠翔（RUC）富有启发的讨论。

影响声明

本文旨在推动视觉–语言模型中高效且可泛化的提示词学习。本文的分析表明，嵌入范数并非单纯的数值副产物，而是可能影响表征几何的重要因素。这一观察或许与未来关于鲁棒性、隐私泄露或对抗性滥用的研究相关，尽管本文并未开发或评估此类攻击。实际部署中仍需进行细致的数据审计、鲁棒性评估，并采取具备隐私意识的防护措施。

References

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *ECCV*, 2014.

Chen, G., Yao, W., Song, X., Li, X., Rao, Y., and Zhang, K. Plot: Prompt learning with optimal transport for vision-language models. In *ICLR*, 2023.

Chen, L., Sun, R., Wang, X., Huang, G., Wu, J., and Jia, W. Learning corruption-invariant components and cross-modal correspondence for unsupervised visible-infrared person re-identification. *IEEE TIFS*, 21:1712–1724, 2026.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *NeurIPS*, 2018.

Chen, Y., Feng, J., Zhang, H., GONG, L., Zhu, F., Zhao, R., Hou, Q., Cheng, M.-M., and Song, Y. Realigning language to visual objects with an agentic workflow. In *ICLR*, 2025.

Cheng, S. and Han, K. Vamp: Variational multi-modal prompt learning for vision-language models. In *NeurIPS*, 2025.

Chou, J. C.-C. and Alam, N. Embedding geometries of contrastive language-image pre-training. In *ECCV*, 2024.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *IEEE CVPR*, 2014.

Dai, T., Han, M., Du, T., Liu, Z., Li, Z., Khan, S., Yu, J., and Chang, X. See, plan, rewind: Progress-aware vision-language-action models for robust robotic manipulation. *arXiv preprint arXiv:2603.09292*, 2026.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. In *ICLR*, 2024.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, 2009.

Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., and Vedantam, S. R. Hyperbolic image-text representations. In *ICML*, 2023.

Ding, T., Li, W., Miao, Z., and Pfister, H. Tree of attributes prompt learning for vision-language models. In *ICLR*, 2025.

Draganov, A., Vadgama, S., Damrich, S., Böhm, J. N., Maes, L., Kobak, D., and Bekkers, E. J. On the importance of embedding norms in self-supervised learning. In *ICML*, 2025.

Du, Z., Hu, Z., Zhao, G., Jin, Y., and Ma, H. Cross-layer feature pyramid transformer for small object detection in aerial images. *IEEE TGRS*, 63:1–14, 2025a.

Du, Z., Hu, Z., Zhao, G., Jin, Y., and Ma, H. Enhanced head: Exploring strong detection heads with vision transformer. *IEEE TMM*, 27:7834–7848, 2025b.

Fan, D.-P., Ji, G.-P., Sun, G., Cheng, M.-M., Shen, J., and Shao, L. Camouflaged object detection. In *IEEE CVPR*, 2020.

Fan, D.-P., Ji, G.-P., Cheng, M.-M., and Shao, L. Concealed object detection. *IEEE TPAMI*, 44(10):6024–6042, 2021.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE CVPR Workshop*, 2004.

Gao, S., Jia, X., Huang, Y., Duan, R., Gu, J., Bai, Y., Liu, Y., and Guo, Q. Hts-attack: Heuristic token search for jailbreaking text-to-image models. *arXiv preprint arXiv:2408.13896*, 2024a.

Gao, S., Jia, X., Ren, X., Tsang, I., and Guo, Q. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. In *ECCV*, 2024b.

Gao, S., Zhao, S., Jiang, X., Duan, L., Chng, Y. X., Chen, Q.-G., Luo, W., Zhang, K., Bian, J.-W., and Gong, M. Scaling beyond context: A survey of multimodal retrieval-augmented generation for document understanding. *arXiv preprint arXiv:2510.15253*, 2025.

Guo, Y. and Gu, X. Mmrl: Multi-modal representation learning for vision-language models. In *IEEE CVPR*, 2025.

Guo, Y. and Gu, X. Mmrl++: Parameter-efficient and interaction-aware representation learning for vision-language models. *IJCV*, 134(1):11, 2026.

Hariharan, B. and Girshick, R. Low-shot visual recognition by shrinking and hallucinating features. In *IEEE ICCV*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE CVPR*, 2016.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE JSTARS*, 12(7):2217–2226, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE ICCV*, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *IEEE CVPR*, 2021b.

Hu, B.-C., Ji, G.-P., Shao, D., and Fan, D.-P. Pranet-v2: Dual-supervised reverse attention for medical image segmentation. *Computational Visual Media*, 12(2):493–500, 2026.

Hu, H., Lin, T., Wang, J., Sun, Z., and Xu, Y. Context-aware prompt tuning for vision-language model with dual-alignment. *arXiv preprint arXiv:2309.04158*, 2023.

Huang, L., Cao, X., Lu, H., Meng, Y., Yang, F., and Liu, X. Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. In *IEEE ICCV*, 2025.

Ilievski, F., Hammer, B., van Harmelen, F., Paassen, B., Saralajew, S., Schmid, U., Biehl, M., Bolognesi, M., Dong, X. L., Gashteovski, K., et al. Aligning generalization between humans and machines. *Nature Machine Intelligence*, 7(9):1378–1389, 2025.

Ji, G.-P., Zhuge, M., Gao, D., Fan, D.-P., Sakaridis, C., and Van Gool, L. Masked vision-language transformer in fashion. *Machine Intelligence Research*, 20 (3):421–434, 2023.

Ji, G.-P., Liu, J., Fan, D.-P., and Barnes, N. Colon-x: Advancing intelligent colonoscopy from multimodal understanding to clinical reasoning. *arXiv preprint arXiv:2512.03667*, 2025.

Ji, G.-P., Liu, J., Xu, P., Barnes, N., Khan, F. S., Khan, S., and Fan, D.-P. Frontiers in intelligent colonoscopy. *Machine Intelligence Research*, 23(1): 70–114, 2026.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

Jiang, N., Dravid, A., Efros, A., and Gandelsman, Y. Vision transformers don't need trained registers. In *NeurIPS*, 2025.

Jiang, Y., Yan, X., Ji, G.-P., Fu, K., Sun, M., Xiong, H., Fan, D.-P., and Khan, F. S. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2(1):17, 2024.

Jiang, Z., Xu, F. F., Araki, J., and Neubig, G. How can we know what language models know? *TACL*, 8:423–438, 2020.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. Maple: Multi-modal prompt learning. In *IEEE CVPR*, 2023a.

Khattak, M. U., Wasim, S. T., Naseer, M., Khan, S., Yang, M.-H., and Khan, F. S. Self-regulating prompts: Foundational model adaptation without forgetting. In *IEEE ICCV*, 2023b.

Khattak, M. U., Naeem, M. F., Naseer, M., Van Gool, L., and Tombari, F. Learning to prompt with text only supervision for vision-language models. In *AAAI*, 2025.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *IEEE ICCV Workshop*, 2013.

Levi, M. Y. and Gilboa, G. The double-ellipsoid geometry of clip. In *ICML*, 2025.

Li, H., Wang, L., Wang, C., Jiang, J., Peng, Y., and Long, G. Dpc: Dual-prompt collaboration for tuning vision-language models. In *IEEE CVPR*, 2025a.

Li, J., Gao, M., Wei, L., Tang, S., Zhang, W., Li, M., Ji, W., Tian, Q., Chua, T.-S., and Zhuang, Y. Gradient-regulated meta-prompt learning for generalizable vision-language models. In *IEEE ICCV*, 2023.

Li, Y., Li, Z.-Y., Zeng, Q.-S., Hou, Q., and Cheng, M.-M. Cascade-clip: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. In *ICML*, 2024a.

Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.-M., and Shan, Y. Photomaker: Customizing realistic human photos via stacked id embedding. In *IEEE CVPR*, 2024b.

Li, Z., Li, X., Fu, X., Zhang, X., Wang, W., Chen, S., and Yang, J. Promptkd: Unsupervised prompt distillation for vision-language models. In *IEEE CVPR*, 2024c.

Li, Z., Song, Y., Cheng, M.-M., Li, X., and Yang, J. Advancing textual prompt learning with anchored attributes. In *IEEE ICCV*, 2025b.

Li, Z., Song, Y., Zhang, X., Luo, L., Li, X., and Yang, J. Anchoropt: Towards optimizing dynamic anchors for adaptive prompt learning. *arXiv preprint arXiv:2511.21188*, 2025c.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *ICLR*, 2023.

Liu, L., Wang, N., Yang, X., Gao, X., and Liu, T. Surrogate prompt learning: Towards efficient and diverse prompt learning for vision-language models. In *ICML*, 2025a.

Liu, L., Wang, N., Zhou, D., Liu, D., Yang, X., Gao, X., and Liu, T. Generalizable prompt learning via gradient constrained sharpness-aware minimization. *IEEE TMM*, 27:1100–1113, 2025b.

Lu, Y., Liu, J., Zhang, Y., Liu, Y., and Tian, X. Prompt distribution learning. In *IEEE CVPR*, 2022.

Luo, Z., Liu, N., Yang, X., Zhang, D., Fan, D.-P., Khan, F. S., and Han, J. Vscode-v2: Dynamic prompt learning for general visual salient and camouflaged object detection with two-stage optimization. *IEEE TPAMI*, 2025.

Ma, Q., Wang, C.-Y., Gao, D., and Fan, D.-P. Alignednorm: Prompting vision–language models via coupled prompt field. In *ICML*, 2026.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Nakamura, K., Nozawa, Y., Lin, Y.-C., Nakata, K., and Ng, Y. Improving image clustering with artifacts attenuation via inference-time attention engineering. In *ACCV*, 2024.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

Park, J., Ko, J., and Kim, H. J. Prompt learning via meta-regularization. In *IEEE CVPR*, 2024.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. In *IEEE CVPR*, 2012.

Petrov, A., Torr, P., and Bibi, A. When do prompting and prefix-tuning work? a theory of capabilities and limitations. In *ICLR*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Ranzato, M. *Unsupervised learning of feature hierarchies*. PhD thesis, New York University, 2009.

Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., and Lu, J. Denseclip: Language-guided dense prediction with context-aware prompting. In *IEEE CVPR*, 2022.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

Roy, S. and Etemad, A. Consistency-guided prompt learning for vision-language models. In *ICLR*, 2024.

Schrodi, S., Hoffmann, D. T., Argus, M., Fischer, V., and Brox, T. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *ICLR*, 2025.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, 2020.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Sun, R., Chen, L., Zhang, L., Xie, R., and Gao, J. Robust visible-infrared person re-identification based on polymorphic mask and wavelet graph convolutional network. *IEEE TIFS*, 19:2800–2813, 2024.

Tian, X., Zou, S., Yang, Z., and Zhang, J. Argue: Attribute-guided prompt tuning for vision-language models. In *IEEE CVPR*, 2024.

Wang, C.-Y., Ji, G., Shao, S., Cheng, M.-M., and Fan, D.-P. Context-measure: Contextualizing metric for camouflage. *arXiv preprint arXiv:2512.07076*, 2025a.

Wang, E., Peng, Z., Xie, Z., Yang, F., Liu, X., and Cheng, M.-M. Get: Unlocking the multi-modal potential of clip for generalized category discovery. In *IEEE CVPR*, 2025b.

Wang, E., Wang, Q., Wu, Y., Yan, K., Yuan, X., Ding, S., Liu, X., and Cheng, M.-M. Predictive regularization against visual representation degradation in multimodal large language models. *arXiv preprint arXiv:2603.20808*, 2026.

Wang, F. and Liu, H. Understanding the behaviour of contrastive loss. In *IEEE CVPR*, 2021.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

Wang, H., Zhang, T., and Salzmann, M. Sinder: Repairing the singular defects of dinov2. In *ECCV*, 2024.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

Wu, S., Zhang, J., Zeng, P., Gao, L., Song, J., and Shen, H. T. Skip tuning: Pre-trained vision-language models are effective and efficient adapters themselves. In *IEEE CVPR*, 2025a.

Wu, Y.-H., Zhu, Z.-X., Wang, Y., Zhen, L., and Fan, D.-P. Refonce: Distilling references into a prototype memory for referring camouflaged object detection. *arXiv preprint arXiv:2511.20989*, 2025b.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. In *ICLR*, 2024.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE CVPR*, 2010.

Xie, J., Zhang, Y., Peng, J., Huang, Z., and Cao, L. Textrefiner: Internal visual feature as efficient refiner for vision-language models prompt tuning. In *AAAI*, 2025.

Yan, X., Sun, M., Ji, G.-P., Khan, F. S., Khan, S., and Fan, D.-P. Lawdis: Language-window-based controllable dichotomous image segmentation. In *IEEE ICCV*, 2025.

Yao, H., Zhang, R., and Xu, C. Visual-language prompt tuning with knowledge-guided context optimization. In *IEEE CVPR*, 2023.

Yin, D., Zhao, T.-F., Fan, D.-P., Li, S., Du, B., Sun, X., and Hu, S.-M. Remote sensing tuning: A survey. *Computational Visual Media*, 11(5):897–937, 2025.

Yona, I., Shumailov, I., Hayes, J., Barbero, F., and Gandelsman, Y. Interpreting the repeated token phenomenon in large language models. In *ICML*, 2025.

Zang, Y., Li, W., Zhou, K., Huang, C., and Loy, C. C. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *IEEE ICCV*, 2023.

Zhang, J., Wu, S., Gao, L., Shen, H. T., and Song, J. Dept: Decoupled prompt tuning. In *IEEE CVPR*, 2024.

Zhang, J., Luo, X., Gao, L., Zou, D., Shen, H. T., and Song, J. From channel bias to feature redundancy: Uncovering the "less is more" principle in few-shot learning. *IEEE TPAMI*, 2026.

Zhang, X., Yang, X., Li, Y., Yang, J., Cheng, M.-M., and Li, X. Rsar: Restricted state angle resolver and rotated sar benchmark. In *IEEE CVPR*, 2025.

Zhao, K., Yuan, W., Lin, Y., Ruan, L., Lu, X., Fan, D.-P., Cheng, M.-M., and Zeng, D. Attention debiasing for token pruning in vision language models. *arXiv preprint arXiv:2508.17807*, 2025.

Zhao, K., Yuan, W., Wang, Z., Li, G., Zhu, X., Fan, D.-P., and Zeng, D. Open-vocabulary camouflaged object segmentation with cascaded vision language models. *Computational Visual Media*, 12(2):473–492, 2026.

Zheng, H., Yang, S., He, Z., Yang, J., and Huang, Z. Hierarchical cross-modal prompt learning for vision-language models. In *IEEE ICCV*, 2025.

Zheng, P., Gao, D., Fan, D.-P., Liu, L., Laaksonen, J., Ouyang, W., and Sebe, N. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:1–12, 2024a.

Zheng, Y., Pal, D. K., and Savvides, M. Ring loss: Convex feature normalization for face recognition. In *IEEE CVPR*, 2018.

Zheng, Z., Wei, J., Hu, X., Zhu, H., and Nevatia, R. Large language models are good prompt learners for low-shot image classification. In *IEEE CVPR*, 2024b.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *IJCV*, 130 (9):2337–2348, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *IEEE CVPR*, 2022b.

Zhu, B., Niu, Y., Han, Y., Wu, Y., and Zhang, H. Prompt-aligned gradient for prompt tuning. In *IEEE ICCV*, 2023.

Zhuge, M., Gao, D., Fan, D.-P., Jin, L., Chen, B., Zhou, H., Qiu, M., and Shao, L. Kaleido-bert: Vision-language pre-training on fashion domain. In *IEEE CVPR*, 2021.

AlignedNorm：通过耦合提示场提示视觉–语言模型

补充材料

A. 补充相关工作

视觉–语言模型（VLMs），例如 CLIP (Radford et al., 2021) 与 ALIGN (Jia et al., 2021)，通过双塔结构为图像与文本学习共享嵌入空间。它们在大规模图文对上进行预训练，因而能够在广泛下游任务实现强大的开放词表迁移 (Zhuge et al., 2021; Rao et al., 2022; Ji et al., 2023; Gao et al., 2024b;a; Zheng et al., 2024a; Jiang et al., 2024; Li et al., 2024a;b; Chen et al., 2025; Luo et al., 2025; Yan et al., 2025; Zhao et al., 2026; Wu et al., 2025b; Ji et al., 2025; 2026; Wang et al., 2025b; Huang et al., 2025; Dai et al., 2026; Chen et al., 2026; Wang et al., 2026)。在实践中，下游推理与评估通常基于 ℓ_2 归一化嵌入上的余弦相似度进行，强调模态对齐超球面上的角度对齐 (Wang & Isola, 2020; Wang & Liu, 2021)。相比之下，未归一化嵌入空间的几何结构在下游适配与分析中受到的关注较少；而模态差距 (Liang et al., 2022) 与锥形效应 (Liang et al., 2022; Schrodi et al., 2025) 等现象正是在这一空间中产生。近期研究进一步表明，包括嵌入范数在内的此类几何结构与语义属性存在耦合关系 (Desai et al., 2023; Chou & Alam, 2024; Levi & Gilboa, 2025)，这也促使本文进一步审视范数相关结构。

提示学习起源于自然语言处理 (Jiang et al., 2020; Shin et al., 2020)，旨在通过向序列中加入连续且可学习的词元，将 CLIP 等视觉–语言模型适配到下游任务 (Yin et al., 2025)。CoOp (Zhou et al., 2022a;b) 率先采用语言诱导提示，并利用任务特定数据优化文本上下文。然而，将适配限制在 VLMs 的语言分支会显著损害模型在未见类别上的零样本性能。后续工作分别采用 **(i) 视觉–语言分层提示** (Zang et al., 2022; Khattak et al., 2023a; Zheng et al., 2025; Liu et al., 2025a)，由内部参考 (Yao et al., 2023; Khattak et al., 2023b; Tian et al., 2024; Roy & Etemad, 2024; Xie et al., 2025)、外部参考 (Hu et al., 2023; Zheng et al., 2024b; Ding et al., 2025; Li et al., 2025b)、分布多样性 (Lu et al., 2022; Chen et al., 2023; Wu et al., 2025a; Cheng & Han, 2025)、梯度对齐 (Zhu et al., 2023; Li et al., 2023; Park et al., 2024; Liu et al., 2025b)、基类–新类解耦 (Zhang et al., 2024; Li et al., 2025a) 驱动的 **(ii) 网络正则化**，或 **(iii) 蒸馏 (Li et al., 2024c; Khattak et al., 2025)**，以保护 VLMs 的泛化能力。鉴于深层提示相较浅层提示具有更强表达能力 (Petrov et al., 2024)，先进方法 (Guo & Gu, 2025; 2026) 已将表征学习引入更深层。

嵌入范数。嵌入可分解为方向（语义角度）与范数（径向幅值）。方向通常被优先关注，而嵌入范数却常被视为次要或不重要因素。早期无监督学习方法 (Ranzato, 2009; Hariharan & Girshick, 2017; Zheng et al., 2018) 施加范数约束以避免平凡解坍塌，而现代对比式 VLMs (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023; Chou & Alam, 2024) 则通过将嵌入投影到单位超球面并使用余弦相似度，从而绕过范数信息。然而，近期研究 (Levi & Gilboa, 2025; Draganov et al., 2025; Zhang et al., 2025) 表明，嵌入范数与语义特征存在相关性。

高范数词元。高范数词元可能作为 Transformer 训练中的非预期副产物出现 (Xiao et al., 2024; Darcet et al., 2024)。在因果掩码下，它们可能吸引过多注意力并形成 “attention sinks” (Xiao et al., 2024)，这一现象可在推理时通过让相关神经元失活来缓解 (Yona et al., 2025)。在没有因果掩码的情况下，高范数词元通常出现在低熵区域，并聚合全局上下文 (Darcet et al., 2024)。已有工作通过寄存器词元 (Darcet et al., 2024)、轻量微调 (Wang et al., 2024) 或测试时修正 (Nakamura et al., 2024; Jiang et al., 2025; Zhao et al., 2025) 来缓解这些影响。

B. 数学分析推导

B.1. 非均匀场更新

式 (7) 的推导。 假设损失仅通过 ℓ_2 归一化嵌入 $f(z) = \frac{z}{\|z\|} \in \mathbb{R}^d$ 依赖于提示分支（其中 $\|z\| > 0$）。记关于归一化嵌入的梯度为 $g \triangleq \frac{\partial \mathcal{L}}{\partial f} \in \mathbb{R}^d$. 下面推导归一化映射的雅可比矩阵。

令 $s = \|z\| = (z^\top z)^{1/2}$。利用微分可得：

$$ds = d\|z\| = d(z^\top z)^{1/2} = \frac{1}{2}(z^\top z)^{-1/2}\, d(z^\top z) = \frac{1}{\|z\|}\, z^\top dz.$$

注意到 $f = z/\|z\|$，因此 $z^\top dz = \|z\|\, f^\top dz$，于是

$$ds = f^\top dz.$$

现在计算 $f(z) = z/s$ 的微分：

$$df = d\Big(\frac{z}{s}\Big) = \frac{1}{s}\,dz - \frac{z}{s^2}\,ds = \frac{1}{\|z\|}\,dz - \frac{z}{\|z\|^2}\,(f^\top dz).$$

代入 $z = \|z\| f$，得到

$$df = \frac{1}{\|z\|}\Big(I - ff^\top\Big)\,dz.$$

因此，归一化映射的雅可比矩阵为

$$\frac{\partial f}{\partial z} = \frac{1}{\|z\|}\Big(I - ff^\top\Big),$$

其中 $P \triangleq I - ff^\top$ 是单位球面在 f 处切空间上的正交投影矩阵。

由链式法则可得：

$$\frac{\partial \mathcal{L}}{\partial z} = \left(\frac{\partial f}{\partial z}\right)^\top \frac{\partial \mathcal{L}}{\partial f} = \frac{1}{\|z\|}\Big(I - ff^\top\Big)\,g,$$

因为 $I - ff^\top$ 是对称矩阵。这就得到式 (7)。此外，由于 P 是投影矩阵，且其算子范数满足 $\|P\|_2 \le 1$，因此

$$\Big\|\frac{\partial \mathcal{L}}{\partial z}\Big\| \le \frac{1}{\|z\|}\,\|g\|,$$

这表明有效梯度幅值会被 $\|z\|$ 反向缩放。

式 (8) 的推导。 假设随机梯度可分解为

$$g = \bar{g} + \xi, \qquad \mathbb{E}[\xi] = 0,$$

其中期望是针对小批量/优化器中的随机性而取的；在给定输入 x 与当前参数的条件下，z 以及 f 固定。利用式 (7)，并记 $P = I - ff^\top$，有

$$\nabla_z \mathcal{L} = \frac{1}{\|z\|}P(\bar{g} + \xi) = \underbrace{\frac{1}{\|z\|}P\bar{g}}_{\mathbb{E}[\nabla_z \mathcal{L}]} + \frac{1}{\|z\|}P\xi,$$

其中 $\mathbb{E}[\nabla_z \mathcal{L}] = \frac{1}{\|z\|}P\bar{g}$ 由 $\mathbb{E}[\xi] = 0$ 得到。因此，

$$\nabla_z \mathcal{L} - \mathbb{E}[\nabla_z \mathcal{L}] = \frac{1}{\|z\|}P\xi.$$

对其取平方范数，并利用 P 是正交投影矩阵这一事实（即对任意 v，有 $\|Pv\| \le \|v\|$），可得

$$\big\|\nabla_z \mathcal{L} - \mathbb{E}[\nabla_z \mathcal{L}]\big\|^2 = \frac{1}{\|z\|^2}\,\|P\xi\|^2 \le \frac{1}{\|z\|^2}\,\|\xi\|^2.$$

最后取期望，得到

$$\mathbb{E}\Big[\big\|\nabla_z \mathcal{L} - \mathbb{E}\big[\nabla_z \mathcal{L}\big]\big\|^2\Big] \;\le\; \mathbb{E}\Big[\frac{\|\xi\|^2}{\|z\|^2}\Big],$$

这就是式 (8)。特别地，当 $\|z_p(x)\|$ 较小时，因子 $1/\|z_p(x)\|$ 会增大有效更新尺度，并放大随机梯度噪声对提示词分支的影响。

B.2. 场对扰动的稳定性。

式 (9) 的推导。 考虑提示分支归一化前特征受到随机扰动的情形：

$$z_p = z_p^\star + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid x] = 0,$$

其中 $z_p^\star$ 表示无噪声特征，ε 表示由有限数据与随机优化带来的估计噪声。记归一化提示嵌入为 $f_p(z) = \frac{z}{\|z\|}$.

步骤 1：ℓ_2 归一化的一阶近似。归一化映射 $f_p(z) = z/\|z\|$ 在 $z = z_p^\star$ 处的雅可比矩阵为

$$J(z_p^\star) \triangleq \left.\frac{\partial f_p}{\partial z}\right|_{z=z_p^\star} = \frac{1}{\|z_p^\star\|}\Big(I - f_p^\star f_p^{\star\top}\Big),$$

其中 $f_p^\star \triangleq f_p(z_p^\star) = z_p^\star/\|z_p^\star\|$。因此，一阶泰勒展开给出

$$f_p - f_p^\star \;\approx\; J(z_p^\star)\,\varepsilon = \frac{1}{\|z_p^\star\|}\Big(I - f_p^\star f_p^{\star\top}\Big)\varepsilon.$$

步骤 2：给出归一化空间中扰动能量的上界。令 $P \triangleq I - f_p^\star f_p^{\star\top}$。注意，$P$ 是单位球面在 $f_p^\star$ 处切空间上的正交投影，因此它是对称且幂等的（$P^2 = P$），并满足对任意 v 都有 $\|Pv\| \le \|v\|$。利用上述一阶近似，

$$\|f_p - f_p^\star\|^2 \;\approx\; \left\|\frac{1}{\|z_p^\star\|}P\varepsilon\right\|^2 = \frac{1}{\|z_p^\star\|^2}\,\|P\varepsilon\|^2 \le \frac{1}{\|z_p^\star\|^2}\,\|\varepsilon\|^2.$$

取条件期望可得

$$\mathbb{E}\big[\|f_p - f_p^\star\|^2 \mid x\big] \;\lesssim\; \frac{\mathbb{E}[\|\varepsilon\|^2 \mid x]}{\|z_p^\star(x)\|^2}.$$

步骤 3：传播到耦合提示场。回顾耦合提示场定义为 $\mathbf{u}_f(x) = \alpha(f_p(x) - f_c(x))$。令其无噪声对应项为 $\mathbf{u}_f^\star(x) = \alpha(f_p^\star(x) - f_c(x))$，其中 $f_c(x)$ 被视为在锚定机制下保持稳定；更一般地，也可认为其扰动相较 f_p 的扰动可以忽略。于是

$$\mathbf{u}_f(x) - \mathbf{u}_f^\star(x) = \alpha\big(f_p(x) - f_p^\star(x)\big),$$

因此

$$\|\mathbf{u}_f(x) - \mathbf{u}_f^\star(x)\|^2 = \alpha^2\|f_p(x) - f_p^\star(x)\|^2.$$

取条件期望并代入步骤 2 的界，可得

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{u}_f(x) - \mathbf{u}_f^\star(x)\|^2 \mid x\big] &\approx \alpha^2\,\mathbb{E}\big[\|f_p(x) - f_p^\star(x)\|^2 \mid x\big] \\ &\le \alpha^2\,\frac{\mathbb{E}[\|\varepsilon\|^2 \mid x]}{\|z_p^\star(x)\|^2},\end{aligned} \tag{15}$$

这与式 (9) 一致。

注。该界揭示了 $1/\|z_p^\star(x)\|^2$ 形式的敏感性：当提示分支归一化前范数较小时，同等扰动能量 $\mathbb{E}[\|\varepsilon\|^2 \mid x]$ 会在耦合场中造成更大的偏移。由于锚定机制使 f_c 保持相对稳定，控制 z_p 的尺度能够提升耦合提示场的扰动鲁棒性。

B.3. 高维错位

令 u, v 为从 $\mathbb{R}^D$ 中单位球面上均匀采样的独立随机单位向量，并令 $\theta = \arccos(u^\top v) \in [0, \pi]$ 表示二者夹角。由旋转对称性，不妨固定 $u = e_1$，于是 $u^\top v = v_1$，即 v 的第一个坐标。因此，

$$\mathbb{E}[u^\top v] = 0, \qquad \mathbb{E}\big[(u^\top v)^2\big] = \mathbb{E}[v_1^2] = \frac{1}{D}. \tag{16}$$

由此可得 $\mathbb{E}[|u^\top v|] \le \sqrt{\mathbb{E}[(u^\top v)^2]} = D^{-1/2}$，也就是说，点积通常接近 0。由于 $\theta = \arccos(u^\top v)$，这意味着随机方向近似正交：$\theta$ 集中在 $\pi/2$ 附近。特别地，θ 的分布关于 $\pi/2$ 对称，因此

$$\mathbb{E}[\theta] = \frac{\pi}{2}. \tag{17}$$

进一步地，对任意 $\delta \in (0, \pi/2)$，

$$\mathbb{P}\big(|\theta - \tfrac{\pi}{2}| \geq \delta\big) = \mathbb{P}\big(|u^\top v| \geq \sin\delta\big) \;\leq\; 2\exp\!\left(-\frac{(D-1)\sin^2\delta}{2}\right), \tag{18}$$

这表明，随着 D 增大，出现显著对齐（即 θ 明显偏离 $\pi/2$）的概率会指数级降低。

B.4. 纠缠坍塌

全局信息交换在多种场景中都很重要 (Fan et al., 2020; 2021; Sun et al., 2024; Du et al., 2025a;b; Wang et al., 2025a; Gao et al., 2025; Hu et al., 2026)。下面给出一个简单分析，说明控制词元范数能够通过使注意力门控保持在非饱和区间内，防止并且在实践中常常逆转提示词元–patch 纠缠坍塌，并有助于保持信息交换。

设置。 考虑某一层中的一个自注意力头。对于一个固定查询词元（patch 词元或类别词元），令其查询向量为 $q \in \mathbb{R}^D$，并令 $\{k_i\}_{i=1}^N$ 表示同一层中所有词元的键向量。其中，k_p 表示某个提示词元的键，其余向量对应非提示词元。定义注意力权重为

$$a_i \;=\; \frac{\exp(\ell_i)}{\sum_{j=1}^N \exp(\ell_j)}, \qquad \ell_i \;=\; \frac{1}{s}\, q^\top k_i, \tag{19}$$

其中 $s > 0$ 是常用缩放因子（例如 $s = \sqrt{D}$）。注意力输出为 $o = \sum_{i=1}^N a_i v_i$，其中 $\{v_i\}$ 为值向量。

引理 1。 令 $\ell_{\max} = \max_j \ell_j$。则对于任意词元 i，有

$$a_i \;\leq\; \exp\big(\ell_i - \ell_{\max}\big), \qquad a_i \;\geq\; \frac{1}{N}\exp\big(\ell_i - \ell_{\max}\big). \tag{20}$$

证明。上界由 $\sum_j \exp(\ell_j) \geq \exp(\ell_{\max})$ 得到。下界由 $\sum_j \exp(\ell_j) \leq N\exp(\ell_{\max})$ 得到。 □

引理 2。 令 $\mathcal{L}$ 为任意依赖注意力输出 o 的损失。则关于提示键 k_p 的梯度满足

$$\left\|\frac{\partial \mathcal{L}}{\partial k_p}\right\| \;\leq\; \frac{1}{s}\, a_p \,\|q\| \left\|\frac{\partial \mathcal{L}}{\partial o}\right\| \big(\|v_p\| + \|o\|\big). \tag{21}$$

证明。对 $o = \sum_i a_i v_i$ 且 $a = \mathrm{softmax}(\ell)$ 作标准求导，可得

$$\frac{\partial \mathcal{L}}{\partial \ell_p} = \left(\frac{\partial \mathcal{L}}{\partial o}\right)^\top \frac{\partial o}{\partial \ell_p} = \left(\frac{\partial \mathcal{L}}{\partial o}\right)^\top \Big(a_p (v_p - o)\Big).$$

由于 $\ell_p = \frac{1}{s} q^\top k_p$，有 $\frac{\partial \ell_p}{\partial k_p} = \frac{1}{s} q$ 因此

$$\frac{\partial \mathcal{L}}{\partial k_p} = \frac{\partial \mathcal{L}}{\partial \ell_p} \cdot \frac{\partial \ell_p}{\partial k_p} = \frac{1}{s}\, a_p \Big(\Big(\frac{\partial \mathcal{L}}{\partial o}\Big)^\top (v_p - o)\Big) q.$$

取范数并应用 Cauchy–Schwarz 不等式，即得 (21)。 □

含义：为什么孤立状态会自我强化。 由式 (21) 可知，一旦 a_p 变得非常小，提示键接收到的梯度就会趋近于消失，从而难以旋转并与其他词元对齐。因此，提示词元上的低注意力是一种自我强化状态：弱注意力 ⇒ 弱梯度 ⇒ 提示持续弱耦合，这正是纠缠坍塌。

命题。 将提示 logit 写作 $\ell_p = \frac{1}{s}\|q\|\,\|k_p\|\cos\theta_p$ 并类似地令 $\ell_j = \frac{1}{s}\|q\|\,\|k_j\|\cos\theta_j$。假设 (i) 在训练早期，角度尚未得到充分学习，因此某些 $\cos\theta_p$ 可能为负；并且 (ii) 提示范数可能发生漂移，使得 $\|k_p\|$ 远大于典型非提示词元的范数。则 logit 差距 $\Delta \triangleq \ell_{\max} - \ell_p$ 可能随 $\|k_p\|$ 成比例增长，而式 (20) 表明

$$a_p \;\leq\; \exp(-\Delta), \tag{22}$$

也就是说，提示注意力可能受到指数级抑制。结合式 (21) 可知，提示梯度也会被指数级削弱，使坍塌状态难以通过梯度更新恢复。

相比之下，施加词元范数对齐（使跨词元/层的 $\|k_p\|$ 与 $\|k_j\|$ 保持可比）能够控制 logits 的尺度，从而为可能出现的差距 Δ 给出上界。因此，a_p 不会变得指数级小，提示词元的梯度流得以保留，并维持提示–patch 纠缠。

要点：词元范数在点积注意力中充当门控尺度。范数对齐使注意力 logits 保持在可学习范围内，从而让提示词元对其他词元保持可见，并持续接收有意义的梯度；这解释了直接对齐词元范数为何能在实践中缓解纠缠坍塌。

表 *9.* 15 个数据集概览。

数据集	类别数	训练集	验证集	测试集	描述	提示模板
ImageNet	1000	1.28M	∼	50000	大规模目标分类基准	“a photo of a [CLASS].”
Caltech101	100	4128	1649	2465	经典目标类别识别基准	“a photo of a [CLASS].”
OxfordPets	37	2944	736	3669	猫狗品种细粒度识别	“a photo of a [CLASS], a type of pet.”
StanfordCars	196	6509	1635	8041	汽车品牌、型号与年份的细粒度识别	“a photo of a [CLASS].”
Flowers102	102	4093	1633	2463	花卉物种的细粒度分类	“a photo of a [CLASS], a type of flower.”
Food101	101	50500	20200	30300	多样化菜品的细粒度识别	“a photo of [CLASS], a type of food.”
FGVCAircraft	100	3334	3333	3333	飞机型号变体的细粒度分类	“a photo of a [CLASS], a type of aircraft.”
SUN397	397	15880	3970	19850	大规模场景与环境理解	“a photo of a [CLASS].”
DTD	47	2820	1128	1692	纹理与视觉模式分类	“[CLASS] texture.”
EuroSAT	10	13500	5400	8100	基于卫星图像的土地利用与覆盖分类	“a centered satellite photo of [CLASS].”
UCF101	101	7639	1898	3783	真实场景视频帧中的动作识别	“a photo of a person doing [CLASS].”
ImageNetV2	1,000	∼	∼	10,000	用于评估 ImageNet 分布泛化能力的测试集	“a photo of a [CLASS].”
ImageNet-Sketch	1,000	∼	∼	50,889	草图风格领域偏移下的目标识别	“a photo of a [CLASS].”
ImageNet-A	200	∼	∼	7,500	自然对抗 ImageNet 样本上的鲁棒目标识别	“a photo of a [CLASS].”
ImageNet-R	200	∼	∼	30,000	ImageNet 目标类别的多样艺术化呈现	“a photo of a [CLASS].”

C. 补充实现细节

15 个数据集的详细信息见 表 9。为与基线方法进行公平比较，除非特别说明，实验沿用 MMRL++ 的超参数设置。具体而言，α 设为 0.7，并从第 $J = 6$ 层开始启用深层提示。各数据集使用的视觉锚定权重与 MMRL++ 保持一致。除 SUN397 外，所有数据集均沿用 MMRL++ 的训练轮数；对于 SUN397，额外训练 3 个 epoch。所有实验均采用 CLIP ViT-B/16 作为骨干网络，并使用 AdamW 优化器，学习率为 1×10^{-3}，采用余弦学习率调度器，并设置 1 个 warm-up epoch，warm-up 阶段的学习率恒定为 1×10^{-5}。可学习提示词元的数量设为 5。

D. 补充实验结果

所有基类到新类消融实验见 表 10，11 个数据集上少样本学习的详细结果见 表 11。图 6 与 图 7 展示了 MMRL++ 和 AlignedNorm 在不同层中的注意力图。

表 *10.* **基类到新类泛化消融实验**。各数据集上基类与新类的准确率（%）。

方法	平均			ImageNet			Caltech101			OxfordPets		
	基类	新类	HM	基类	新类	HM	基类	新类	HM	基类	新类	HM
baseline	85.43	76.48	80.71	77.60	71.30	74.32	98.90	94.57	96.69	95.43	96.87	96.14
+ $\mathcal{L}_{\text{token}}$	85.16	76.37	80.52	77.43	71.27	74.22	98.83	**94.87**	**96.81**	95.03	96.83	95.92
+ $\mathcal{L}_{\text{proj}}$(Hingeloss)	85.44	76.19	80.55	**77.67**	71.23	74.31	98.90	94.57	96.69	95.63	97.07	96.34
+ $\mathcal{L}_{\text{proj}}$(smooth$L1$)	85.32	76.83	80.85	77.47	71.33	74.27	98.73	94.63	96.64	95.17	96.83	95.99
+ $\mathcal{L}_{\text{proj}}$(Ring loss)	85.29	76.94	80.90	77.60	71.30	74.32	**98.93**	94.70	96.77	95.63	97.37	95.99
+ $\mathcal{L}_{\text{proj}}(L1)$	85.42	76.96	80.97	77.57	71.27	74.29	98.87	94.40	96.58	**95.67**	97.23	96.44
AlignedNorm	**85.46**	**77.79**	**81.45**	77.60	**71.47**	**74.41**	98.90	94.77	96.79	95.63	**97.43**	**96.52**

方法	StanfordCars			Flowers102			Food101			FGVCAircraft		
	基类	新类	HM	基类	新类	HM	基类	新类	HM	基类	新类	HM
baseline	81.23	72.43	76.58	**98.50**	73.80	84.38	90.50	91.57	91.03	**46.47**	**38.63**	**42.19**
+ $\mathcal{L}_{\text{token}}$	80.87	72.23	76.31	98.30	73.83	84.33	90.50	91.30	90.90	45.50	36.97	40.79
+ $\mathcal{L}_{\text{proj}}$(Hingeloss)	80.97	71.17	75.75	98.20	73.47	84.05	90.53	91.30	90.91	45.87	37.20	41.08
+ $\mathcal{L}_{\text{proj}}$(smooth$L1$)	80.67	73.23	76.77	98.27	75.03	85.09	90.57	91.37	90.97	46.43	37.83	41.69
+ $\mathcal{L}_{\text{proj}}$(Ring loss)	80.53	73.53	76.87	98.07	75.90	85.57	90.67	91.50	91.08	46.10	37.37	41.28
+ $\mathcal{L}_{\text{proj}}(L1)$	80.70	73.53	76.95	98.27	75.90	85.65	**90.73**	**91.60**	**91.16**	46.10	38.47	41.94
AlignedNorm	**81.70**	**73.83**	**77.57**	98.40	**76.03**	85.78	90.57	**91.60**	91.08	46.20	38.60	42.06

方法	SUN397			DTD			EuroSAT			UCF101		
	基类	新类	HM	基类	新类	HM	基类	新类	HM	基类	新类	HM
baseline	82.93	78.43	80.62	85.07	61.70	71.52	95.73	81.63	88.12	87.37	80.40	83.74
+ $\mathcal{L}_{\text{token}}$	83.03	78.43	80.66	84.80	63.53	72.64	95.63	80.80	87.59	86.80	80.00	83.26
+ $\mathcal{L}_{\text{proj}}$(Hingeloss)	82.93	78.37	80.59	**85.13**	61.40	71.34	95.93	82.17	88.52	**88.03**	80.10	83.88
+ $\mathcal{L}_{\text{proj}}$(smooth$L1$)	82.87	78.83	80.80	84.77	63.47	72.59	**96.30**	82.77	89.02	87.23	79.80	83.35
+ $\mathcal{L}_{\text{proj}}$(Ring loss)	82.87	78.97	80.87	84.37	62.50	71.81	95.80	83.07	88.98	87.57	80.13	83.68
+ $\mathcal{L}_{\text{proj}}(L1)$	**83.10**	**79.17**	**81.09**	84.93	61.97	71.66	96.10	82.80	88.96	87.53	80.23	83.72
AlignedNorm	82.93	79.13	80.99	84.63	**65.23**	**73.67**	96.10	**86.63**	**91.12**	87.43	**81.00**	**84.09**

表 *11.* AlignedNorm 与 MMRL++ (Guo & Gu, 2026) 在 11 个数据集上少样本学习的比较。

数据集	**方法**	**1-shot**	**2-shot**	**4-shot**	**8-shot**	**16-shot**
ImageNet	MMRL++	70.03	70.80	71.43	72.30	**73.17**
	AlignedNorm	**70.07**	**70.87**	**71.47**	**72.33**	73.07
Caltech101	MMRL++	94.13	94.87	**95.90**	**96.13**	**96.83**
	AlignedNorm	**94.33**	**94.97**	95.83	96.03	**96.83**
OxfordPets	MMRL++	**91.30**	**91.17**	92.67	92.33	**93.53**
	AlignedNorm	91.17	90.93	**92.87**	**92.53**	93.50
StanfordCars	MMRL++	68.30	72.57	**77.97**	**82.43**	**86.20**
	AlignedNorm	**70.17**	**73.87**	77.80	81.83	85.40
OxfordFlowers	MMRL++	83.87	**89.67**	**93.80**	96.23	**98.30**
	AlignedNorm	**84.53**	89.63	93.73	**96.33**	97.87
Food101	MMRL++	82.93	84.03	84.63	85.43	86.13
	AlignedNorm	**84.70**	**85.33**	**85.90**	**86.37**	**87.00**
FGVCAircraft	MMRL++	28.13	32.50	40.70	**49.27**	**58.20**
	AlignedNorm	**28.30**	**33.83**	**40.80**	48.80	58.03
SUN397	MMRL++	69.00	71.13	73.40	75.40	77.47
	AlignedNorm	**69.50**	**71.43**	**73.97**	**75.87**	**77.63**
DTD	MMRL++	**57.13**	60.80	**67.10**	70.73	74.43
	AlignedNorm	56.80	**60.90**	66.80	**71.00**	**74.80**
EuroSAT	MMRL++	**78.00**	**82.40**	88.33	**89.30**	**93.33**
	AlignedNorm	77.17	80.93	**88.67**	89.13	93.03
UCF101	MMRL++	75.50	**78.90**	**82.50**	**84.73**	87.23
	AlignedNorm	**75.53**	78.60	82.40	84.60	**87.67**

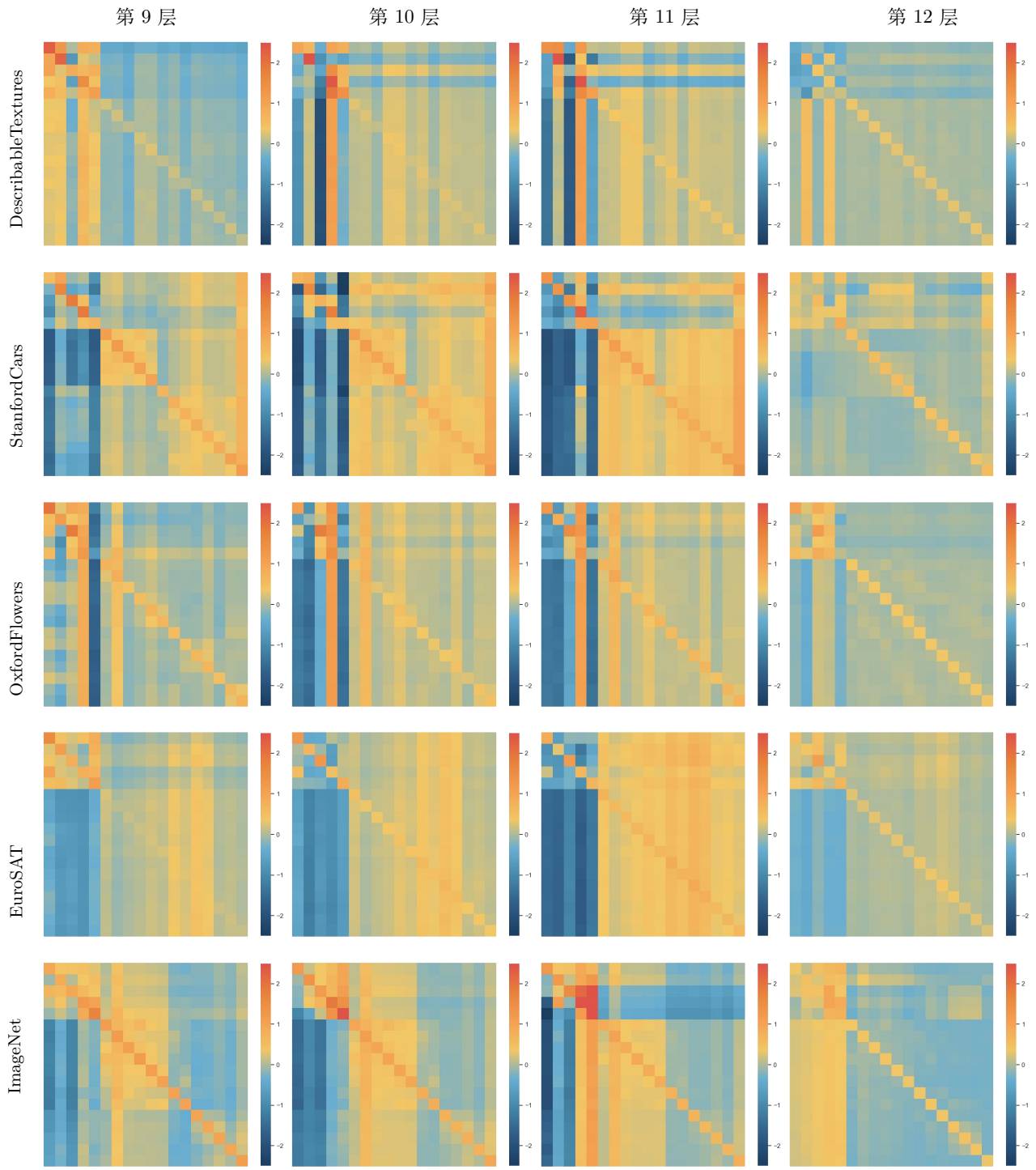


图 *6.* **MMRL++ 最后 4 层的注意力图。**

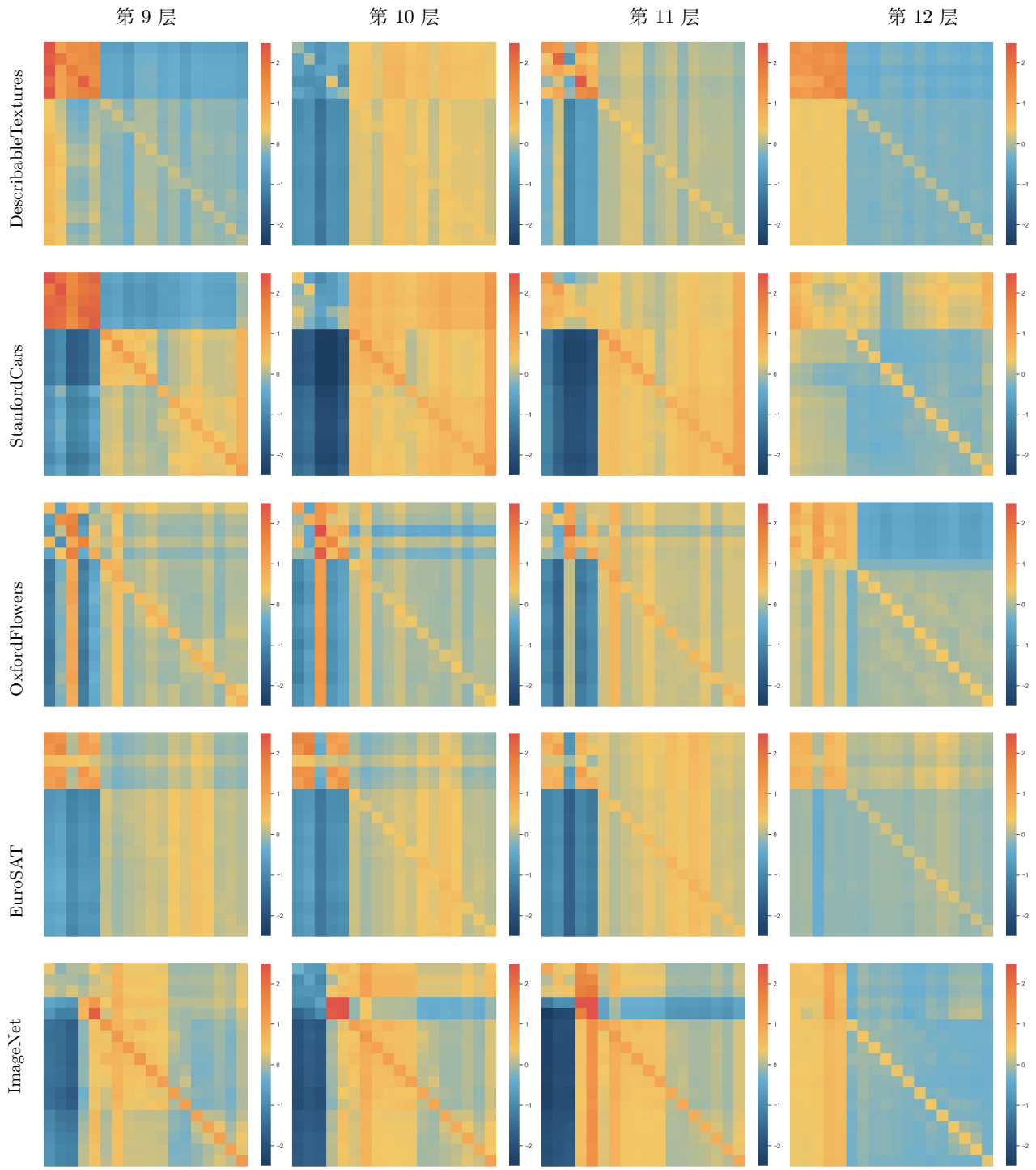


图 7. **AlignedNorm 最后 4 层的注意力图。**