arXiv:2602.02633v1 [stat.ML] 2 Feb 2026

重新思考测试时训练：通过倾斜潜在分布实现少样本无源自适应

TahirQasim Syed卡拉奇工商管理学院 tahirqsyed@gmail.com

Behraj Khan卡拉奇工商管理学院 behrajkhan@iba.edu.pk

2026年2月4日

摘要

在部署场景中，常常会遇到各种限制，例如，即便是轻量级的参数更新（如参数高效微调）也可能导致模型偏移或调优不稳定。我们研究了在完全冻结模型设置下，基础模型用于少样本分类的测试时自适应问题；此外，上游数据也无法访问。我们提出了一种据我们所知首创的免训练推理方法，通过对编码器生成的潜在嵌入分布进行测度变换，使预测适应新任务。该方法利用从小型带标签支持集得出的任务相似度分数，以 KL 最优方式对潜在分布进行指数倾斜重加权，同时不修改模型参数。实验结果表明，在多个基准测试和不同样本数设置下，该方法始终能够与基于参数更新的方法相媲美，同时所处的限制严格得多且适用范围更广。这些结果证明，即使在完全冻结的模型流程中，推理层面的分布校正仍可用于测试时自适应。

1 引言

冻结的基础模型编码器正越来越多地被部署为跨下游任务共享的人工智能系统组件。这种部署有时会受到禁止在测试时修改模型参数的约束。这可能出现在受监管环境、共享的部署基础设施中，或仅仅源于主流的模型不可变部署范式。从建模者的角度来看，不允许参数更新也可能是一种方法上的选择。允许测试时优化会使推理与随机优化动态纠缠在一起，从而引入对批次顺序、超参数、初始化以及对稀疏样本暴露的敏感性；这会使理论分析和实验可复现性都变得复杂，例如由于模型崩溃。例如，在对预训练表示进行少样本评估时，线性分类器或余弦分类器通常在评估时被冻结 Braham et al. (2025)。与此同时，预训练与下游使用之间的分布偏移按定义即存在，导致模型在分布外应用时预测性能系统性下降。在无源自适应中，在没有源监督的情况下更新分类器权重是不适定的，因为它将表示偏移与标签/先验概率偏移混为一谈。通过冻结分类头，我们可以分离并研究推理时条件化效应。

我们假设，即使模型参数被冻结，预测行为也可以被改变。预测不仅取决于已学习的权重，还取决于推理时下游数据的表示或嵌入分布。这表明了一种互补的自适应视角：与其修改模型，我们也可以通过根据下游任务的分布来调节推理，从而自适应预测的形成方式。

在这项工作中，我们提出了一种或可称为首创的推理时自适应方法：它完全使用冻结的编码器和分类器，无需访问源数据，也不进行测试时优化。我们将自适应形式化为对潜在表示进行测度变换，并通过吉布斯/指数倾斜来实现，

具体而言，利用评分函数估计潜在表示中样本嵌入的任务相关性，而这些嵌入是通过接触有限的下游数据获得的。通过对嵌入上的概率测度重新加权，我们得到了一个第一性原理机制，可调整不同潜在区域的贡献，使倾斜分布下预测函数的期望更好地符合下游真实标签。随后可依据标准决策规则推断自适应后的预测。

在贡献方面，我们将冻结模型的测试时预测适配重新表述为对潜在表示分布进行指数倾斜，从而得到一种有原则的推理时测度变换，且不依赖于架构、训练目标和优化细节。该方法避免了优化引发的不稳定性、测试批次间的偏移以及不可复现性，同时适用于无源和少样本设定。

2 相关工作

测试时自适应。 测试时自适应（TTA）通过在推理时修改模型行为来应对分布偏移。大多数现代 TTA 方法使用无标签测试数据来调整模型参数，通常采用熵最小化、一致性正则化或辅助自监督目标 Wang 等人（2020）；Niu 等（2022）；Zhang 等（2022）。近期研究进一步探讨了测试时更新的鲁棒性与稳定性，包括测试时增强、批归一化适配和选择性参数更新 Sun 等（2020）；Yuan 等（2023）。尽管这些方法效果显著，但它们需要进行梯度计算、维护优化器状态或修改参数，因此不符合本文所考虑的严格冻结模型约束。近期研究还指出了基于熵的测试时目标的局限性，并探索了无需梯度更新的替代性非参数或基于检索的适配策略 Han 等人（2025）；Hu 等（2025）。

参数高效与基于提示的适配。 为降低自适应成本，近期一些方法提出通过提示、适配器或低秩更新实现参数高效的测试时自适应 Zhou 等（2022）；Khattak 等（2023）。虽然这些方法减少了可训练参数的数量，但仍需要在测试时或任务自适应期间进行优化或修改参数。相比之下，我们的方法无需引入任何可训练组件即可实现自适应。

事后校准与 Logit 调整。 事后方法无需修改模型参数即可调整预测。温度缩放及相关校准技术可以纠正置信度校准失准，但不会直接纳入任务特定结构 Guo 等人（2017）。标签偏移校正方法利用估计的类别先验对后验概率重新加权 Saerens 等（2002）。近期研究重新探讨了分布偏移下的校准问题，但这类方法仍作用于 logit 或概率层面，而非潜在表示层面 Ovadia 等（2019）。

使用冻结编码器的少样本学习。 少样本学习通常假设使用冻结的特征提取器，并采用任务特定的推理规则。基于原型的方法和度量学习方法根据查询样本与由支持集计算得到的类别代表之间的距离进行分类 Snell 等（2017）。近期研究通过改进嵌入几何或采用任务特定度量扩展了这一思路，但仍依赖于修改决策规则本身 Chen 等（2019）。相关研究还探讨了在分布偏移下纳入小规模带标签支持集进行测试时自适应，并表明即使推理时只有有限监督，也能显著提升鲁棒性 Xiao 与 Snoek（2024）。相比之下，我们的方法保留分类器和决策规则，并通过在潜在空间中进行分布重加权来调整预测。

分布偏移与潜在重加权。 重加权方法是应对分布偏移的经典方法，包括协变量偏移下的重要性加权 Shimodaira（2000）。指数倾斜是在矩约束下的 KL 最优测度变换 Csiszár（1975），近期也在鲁棒优化和分布鲁棒学习等机器学习场景中重新受到关注 Duchi 等人（2021）。我们利用这一原理，通过对冻结编码器生成的潜在表示进行重加权来实现测试时自适应，而不更新模型参数。近期研究探索了无需训练或无需反向传播的测试时自适应方法，通过估计或重加权基础模型的测试分布，尤其是在视觉-语言场景中，通常采用贝叶斯或分布更新，而非参数学习 Han 等人（2024）；Kim（2025）。

潜在空间自适应。 近期有多项研究探索了通过操纵潜在空间来调整表示或预测，例如通过原型优化或几何感知评分。然而，这些方法通常会引入额外的优化、学习得到的参数或启发式调整。相比之下，我们将测试时自适应建模为对潜在表示进行严格的测度变换，从而得到一种简单、免训练且具有清晰概率解释的方法。

3 方法

我们提出了一种免训练测试时自适应方法，通过对冻结编码器生成的潜在表示进行重加权来实现。我们首先形式化问题设定与约束，随后将自适应表述为潜在空间中的测度变换，再基于由此得到的倾斜分布构建分数并进行预测。我们始终强调，自适应完全通过分布重加权实现，不会修改模型参数、对数几率或决策规则。

问题设定与约束。 我们考虑一个冻结编码器 f 和一个冻结分类头 $p_0(y \mid z)$。给定输入图像 $x \in \mathcal{X}$，编码器产生一个潜在表示

$$z = f(x), \qquad f \text{ fixed.} \tag{1}$$

分类头定义了给定潜在表示时标签上的条件分布，

$$p_0(y \mid z), \qquad p_0 \text{ fixed.} \tag{2}$$

在测试时，我们获得一个小型带标签支持集

$$\mathcal{D}_{\text{few}} = \{(x_i, y_i)\}_{i=1}^{n}, \qquad n \ll 100, \tag{3}$$

该支持集抽取自可能不同于预训练分布的下游任务分布 .

本研究设定了以下约束:

- 编码器参数被冻结，
- 分类头被冻结，
- 测试时不进行梯度更新或优化，
- 可用的测试样本数量较少（低样本/少样本），并且
- 无法获取上游数据或其统计数据。

目标是在满足这些约束的前提下，仅使用 $\mathcal{D}_{\text{few}}$ 对下游查询样本的预测进行自适应。

冻结编码器诱导的潜在分布。 尽管编码器 f 是确定性的，但将其应用于数据集会在潜在表示上诱导出一个分布。给定数据分布 $P(x)$，编码器会诱导出一个推前测度 $P(z)$，定义为 $z = f(x)$。对于有限数据集，这对应于经验分布

$$\hat{P}(z) = \frac{1}{n} \sum_{i=1}^{n} \delta(z - f(x_i)), \tag{4}$$

其中 $\delta(\cdot)$ 表示狄拉克δ函数。
模型预测取决于相对于此类诱导潜在分布所计算的期望。当下游任务与预训练任务不同时，即使模型参数保持不变，潜在空间上的诱导分布也会发生变化。这促使我们通过修改潜在表示上的分布来调整预测，而非修改模型本身。

适应作为KL最优测度变换。 当编码器和分类器都被冻结时，测试时自适应只能通过修改推理时潜在表示的聚合方式来实现。这相当于在潜在空间上选择一个新的概率测度 $P(z)$，同时保持其相对于冻结编码器所诱导的参考测度 $P_0(z)$ 的绝对连续性。

我们将自适应形式化为一个约束推理问题。给定由少样本支持集构造的任务相关评分函数 $s(z)$，我们寻求一个分布 $P(z)$，使其在与下游监督保持一致的同时，最大限度地忠实于冻结推理。由此得到以下变分表述:

$$P^\star = \arg\min_{P} \ \mathrm{KL}(P \,\|\, P_0) \quad \text{s.t.} \quad \mathbb{E}_{z \sim P}[s(z)] = c, \tag{5}$$

其中， c 是由支持集确定的任务相关约束。

式（5）的解由指数倾斜分布给出

$$P_\lambda(z) = \frac{P_0(z)\exp(\lambda s(z))}{Z(\lambda)}, \qquad Z(\lambda) = \int P_0(z)\exp(\lambda s(z))\,dz, \tag{6}$$

其中 $\lambda \geq 0$ 是与期望约束相关联的拉格朗日乘子。该分布是满足任务约束、同时在库尔贝克-莱布勒散度下相对 P_0 引起最小可能偏离的唯一测度 Csiszár (1975); Cover (1999)。

这种KL最优性使指数倾斜在我们的设定中具有明确的操作意义。倾斜并非启发式地对嵌入重新加权，而是实现与下游监督相容的、侵入性最小的推理时自适应。任何施加相同任务约束的替代重加权，都必然相对冻结推理产生更大的散度。

在实践中，我们将 P_0 取为由少样本支持集嵌入诱导的经验分布，从而确保除测试时所观测到的内容外，不对上游数据或潜在几何作任何假设。随后，指数倾斜将概率质量重新分配到潜在空间中与任务一致的区域，而不修改编码器参数、分类器权重或决策规则。

这一视角为冻结模型约束下的测试时自适应提供了统一的解释：自适应并非通过学习新的表示来实现，而是通过在经任务调整的潜在测度下进行 KL 最优推理来实现。

基于少样本数据的分数构造。 指数倾斜提供了重新加权的机制，但这需要一个定义在潜在表示上的标量评分函数。该分数决定了在自适应测度下，哪些嵌入应获得更高或更低的概率质量。

少样本数据为哪些潜在区域与正确预测相关联提供了稀疏约束。评分的作用是将这些约束编码成适用于指数倾斜的形式。评分函数 $s(z)$ 将来自少样本支持集的任务信息注入潜在分布。

我们考虑两种简单且互补的实例化方式。

标签感知分数。 对于一个带标签的支持样本（x_i，y_i），其嵌入为 $z_i = f(x_i)$，我们定义

$$s_{\text{label}}(z_i) = \log p_0(y_i \mid z_i). \tag{7}$$

该分数会提高那些在冻结分类器下已与观测标签一致的嵌入的权重。它可以被视为一种基于似然的重新加权，用于突出潜在空间中的标签一致区域。

几何感知分数。 我们还将这一思路扩展到推理时可能没有标签，或标签过于稀少、无法估计类别条件统计量的场景。众所周知，预训练基础模型会将数据组织成语义上有意义的几何结构，其中与任务相关的信息体现在变化方向、局部邻域和类别对齐子流形中（Assran 等人（2023年））。该分数估计相对于预训练参考测度，下游分布中过度或不足表征的表征空间区域。也就是说，这是一种基于潜在几何而非标签条件得出的似然替代量。

令 $\hat{}\mu$ 并令 $\hat{}\Sigma$ 表示支持集嵌入 $\{z_i\}_{i=1}^n$的经验均值和协方差。我们定义 e

$$s_{\text{geom}}(z) = \log \frac{\mathcal{N}(z; \hat{\mu}, \hat{\Sigma})}{\mathcal{N}(z; 0, I)}. \tag{8}$$

该分数强调与支持集几何结构相符的嵌入，并以简单的参考分布为参照。为确保在低样本场景下的稳定性，可以对协方差估计进行正则化；我们避免复杂的密度估计或高容量建模。

标签感知评分和几何感知评分既可以单独使用，也可以相加组合，从而形成综合任务相关分数。

倾斜分布下的预测。 我们现在将潜在测度倾斜后的预测计算方式形式化，并说明标准预测规则在适配分布下可以自然且确定地扩展。

给定　　倾斜分布 $P_\lambda(z)$ 下的预测通过对冻结分类器取期望来形成　　，

$$p_\lambda(y) = \mathbb{E}_{z\sim P_\lambda(z)}\left[p_0(y \mid z)\right]. \tag{9}$$

查询样本的最终预测通过标准决策规则获得

$$\hat{y} = \arg\max_y p_\lambda(y). \tag{10}$$

分类头和决策规则保持不变。在一个任务回合内，所有查询样本均使用同一个由支持集构造的倾斜分布 P_λ。因此，自适应作用于潜在分布层面，而不是针对每个查询样本进行参数或分数更新。

实际近似。 在实践中，倾斜分布 $P_\lambda(z)$ 下的期望通过有限的经验支持集进行近似，该支持集由可用的测试嵌入或少样本嵌入组成。这会得到对分类器输出的加权聚合，其中权重由指数倾斜分数决定。倾斜分布下的预测通过标准的重要性采样蒙特卡洛方法实现，并以预训练潜在分布作为提议分布。因此，该方法不会引入额外的建模假设。

由此产生的计算开销随支持集嵌入的数量线性增长，因为每增加一个样本，就会进行一次分数评估和一次加权预测。编码器和分类器对每个输入仅评估一次。因此，与标准的冻结模型推理相比，该方法的额外开销可以忽略不计。

4 实验

我们在公开可用的基准数据集上，在少样本图像分类设置下评估我们的方法。我们的实验旨在回答以下问题：

i. 在严格的无自适应约束下，仅凭分布重加权能否提升少样本性能　　nts?
ii. 指数倾斜与标准冻结推理相比表现如何？以及 iii. 归纳式参考测度与传导式参考测度分别有何影响?

4.1 实验设置

数据集。 我们在标准少样本分类基准上评估我们的方法，包括 CIFAR-FS、CUB 和 mini-ImageNet。为评估分布偏移下的鲁棒性，我们还考察了跨域小样本基准：模型在 mini-ImageNet 上训练，并在 CUB、CARS 和 Places，以及 ImageNet 及其变体（ImageNet-V2、ImageNet-S、ImageNet-A 和 ImageNet-R）和 DomainNet 上进行评估。所有实验均遵循标准的情节式小样本学习协议。在每个任务回合中，从数据集中采样 N 个类别，并为每个类别抽取 K 个带标签的支持样本和 Q 个查询样本。除非另有说明，我们采用 5 路 5样本设置，每个类别使用 15 个查询样本。

编码器。 我们考虑不进行任何微调的冻结视觉编码器：

- **DINO ViT-S/16 Caron 等人（2021年）**，通过在 ImageNet 上进行自监督学习预训练，
- **JEPA ViT-B/16 Assran 等人（2023年）**，一种采用联合嵌入预测架构预训练的视觉模型，通过自监督预测学习结构化且具有语义意义的潜在表示。

评估期间，所有编码器参数均保持冻结。推理在评估模式下进行，并禁用梯度。

分类器。 遵循基于度量的小样本学习领域的既有工作，我们采用使用固定温度参数和余弦相似度的最近原型分类器。类别原型由支持集嵌入的均值计算得出。不训练线性分类器、偏置项或校准参数。

尽管类别原型在每个任务回合中都由支持集嵌入重新计算，但该操作并不涉及学习或优化。分类器形式、相似度度量和温度参数在所有任务和测试回合中均保持固定。原型计算被视为依赖于数据的充分统计量，而非可训练的模型组件，这与标准的小样本评估协议一致。

表1：5分类任务上的少样本分类准确率（%）。所有模型均使用冻结编码器和分类器。对于指数倾斜，我们报告每种样本数下表现最佳的 λ。性能提升完全源自潜在分布重加权。

数据集	方法	1样本	2样本	4样本	8样本	16样本
			DINO			
CIFAR-FS	基线	53.87 ± 9.41	66.61 ± 6.80	74.68 ± 7.76	82.48 ± 5.85	84.68 ± 5.38
	倾斜（我们的方法）	$\mathbf{55.49 \pm 10.32}$	$\mathbf{69.29 \pm 7.45}$	$\mathbf{78.31 \pm 8.33}$	$\mathbf{83.81 \pm 5.80}$	$\mathbf{85.05 \pm 5.55}$
CUB	基线	49.21 ± 8.91	61.75 ± 7.40	70.42 ± 6.94	76.95 ± 6.22	79.83 ± 5.57
	倾斜（我们的方法）	$\mathbf{50.12 \pm 9.50}$	$\mathbf{64.81 \pm 7.88}$	$\mathbf{73.99 \pm 7.30}$	$\mathbf{80.11 \pm 6.08}$	$\mathbf{82.13 \pm 5.42}$
mini-ImageNet	基线	42.64 ± 8.23	54.99 ± 6.56	64.31 ± 6.84	71.82 ± 5.95	74.13 ± 5.25
	倾斜（我们的方法）	$\mathbf{44.93 \pm 8.10}$	$\mathbf{57.45 \pm 7.00}$	$\mathbf{68.10 \pm 6.44}$	$\mathbf{74.52 \pm 5.88}$	$\mathbf{77.04 \pm 5.11}$
			JEPA			
CIFAR-FS	基线	27.37 ± 6.59	27.83 ± 9.20	31.40 ± 7.26	33.79 ± 4.93	35.17 ± 7.24
	倾斜（我们的方法）	$\mathbf{28.17 \pm 9.15}$	$\mathbf{30.29 \pm 6.37}$	$\mathbf{31.92 \pm 7.43}$	$\mathbf{33.80 \pm 4.92}$	$\mathbf{34.81 \pm 6.35}$
CUB	基线	29.84 ± 7.02	34.91 ± 6.48	38.72 ± 6.91	41.95 ± 5.87	44.10 ± 5.43
	倾斜（我们的方法）	$\mathbf{31.12 \pm 7.58}$	$\mathbf{36.88 \pm 6.95}$	$\mathbf{40.21 \pm 6.77}$	$\mathbf{43.02 \pm 5.61}$	$\mathbf{45.36 \pm 5.28}$
mini-ImageNet	基线	25.46 ± 6.84	30.27 ± 6.11	34.85 ± 6.32	38.41 ± 5.76	40.92 ± 5.19
	倾斜（我们的方法）	$\mathbf{26.71 \pm 7.20}$	$\mathbf{32.09 \pm 6.54}$	$\mathbf{36.43 \pm 6.28}$	$\mathbf{40.05 \pm 5.63}$	$\mathbf{42.31 \pm 5.06}$

4.2 基线与自适应变体

我们比较以下推理策略：

冻结基线。 使用由支持集计算得到的均匀加权类别原型进行标准冻结推理。不进行任何自适应。

倾斜（归纳式）。 仅对由带标签支持集嵌入诱导的经验潜在分布应用指数倾斜。随后，根据倾斜分布下的加权平均值重新计算类别原型。

倾斜（传导式）。 将无标签查询嵌入纳入参考潜在分布。使用冻结基线分类器获取伪标签，随后对合并后的支持集和查询嵌入应用指数倾斜。此变体不涉及基于梯度的优化或参数更新，而是通过分布重加权利用无标签测试数据。

在所有变体中，查询嵌入均保持不变，并使用相同的冻结分类器进行预测。

对比方法。 我们将本方法与在冻结模型约束下运行的免训练基线进行比较，如表2所示。这些基线包括：

i. 使用余弦相似度的冻结原型网络基线（Snell 等（2017））。

ii. 在冻结嵌入空间中进行 k 近邻分类（Avdiukhin 等人（2024））。

iii. 对冻结分类器的对数几率应用事后温度缩放（Guo 等人（2017））。

iv. 利用无标签查询嵌入、无需基于梯度的优化的标签传播和原型细化方法（Ren 等人（2018）；Zhu 与 Koniusz（2023））。

v. 假设能够访问上游数据或统计数据的分布式或基于检索的免训练适应方法 Han 等人（2025，2024）。

.

表2：在严格的冻结模型约束下，CIFAR-FS 数据集上免训练测试时自适应方法的比较。我们的方法无需参数更新、梯度或访问上游数据，却始终优于非自适应基线。标记有 ✓ 的方法违反了一项或多项约束，作为上界。

方法	参数?	梯度?	Opt?	流式?	上游?	准确率（%）		
						1样本	5样本	10样本
免训练且符合约束的方法								
冻结基线	×	×	×	×	×	53.87	74.68	82.48
k-NN（嵌入）Avdiukhin 等人（2024）	×	×	×	×	×	58.21	76.34	83.15
温度缩放（TS）Guo 等人（2017）	×	✓	✓	×	×	54.92	75.81	82.90
标签传播 Zhu 与 Koniusz（2023）	×	×	×	✓	×	59.83	77.42	84.06
排序 Han 等人（2025）	×	×	×	×	✓	59.45	77.18	83.92
DOTA（改编版）Han 等人（2024）	×	×	×	✓	×	58.76	76.89	83.54
我们的方法（归纳式）	×	×	×	×	×	**60.49**	**78.31**	**83.81**
我们的方法（传导式）	×	×	×	✓	×	**62.17**	**79.85**	**85.22**
轻量级参数更新方法（上界）								
Tent Wang 等人（2020年）	✓	✓	✓	✓	×	61.34	79.12	85.45
Tip-Adapter-F Zhang 等人（2021年）	✓	✓	✓	×	×	62.89	80.47	86.31
线性微调 Radford 等人（2021年）	✓	✓	✓	×	×	63.72	81.25	87.08
我们的方法（线性头微调）	✓	✓	✓	×	×	**64.73**	**83.59**	**88.89**

4.3 评分函数与倾斜

除非另有说明，我们使用基于置信度的评分，其定义为冻结分类器下的最大预测类别概率。倾斜权重计算如下：

$$w_i \propto \exp(\lambda s(z_i)),$$

其中 λ的 范围为（0.25 - 2.00），适用于所有实验。 λ 的值不会针对每个任务回合或每个数据集进行调优 .

4.4 评估指标

我们报告从 100 个随机采样的任务回合中得到的平均分类准确率。结果以平均准确率± 标准差的形式呈现。由于未进行概率校准学习，准确率是主要评估指标。

4.5 免训练测试时自适应协议

所有实验均严格遵循免训练测试时自适应协议。评估过程中，编码器参数始终保持固定，且任何阶段都不进行基于梯度的优化。推理过程不进行梯度跟踪，也不使用反向传播或优化器。查询嵌入被视为不可变对象，在推理或自适应过程中绝不会被修改。

这些约束确保所有报告的改进都仅来自对经验潜在分布的重加权，而无需更新模型参数或修改底层推理过程。

总体而言，结果表明，指数倾斜是在严格约束下实现少样本自适应的一种简单而有效的机制。该方法完全在潜在分布层面运作，既能实现自适应行为，又能保留冻结推理。

4.6 结果

表1报告了不同数据集、编码器和样本数下的小样本分类准确率。对于 DINO 和 JEPA 编码器，指数倾斜始终优于冻结基线。归纳式变体在所有样本数设置下均带来稳定提升，而传导式变体取得的提升最大，凸显了通过分布重加权纳入无标签测试嵌入的优势。

重要的是，所有性能提升都无需基于梯度的优化、参数更新，也不需要修改分类器或决策规则。这些提升完全源于对冻结编码器所诱导的经验潜在分布进行重新加权。

4.7 倾斜强度消融

我们分析了在冻结模型、免训练测试时自适应设置下，指数倾斜强度 λ 对少样本分类性能的影响。不同样本数设置下的结果见附录表3。
在所有设置下，指数倾斜的性能始终优于基线（$\lambda = 0$），这表明仅通过对潜在空间重新加权，就足以实现有效的测试时自适应，无需进行参数更新。在低样本场景中，相对收益最大，尤其是在 1样本和 2样本分类中；此时基线性能受分布失配的影响最为显著。

一个重要观察是，性能对 λ 具有鲁棒性。当样本数不超过 8 时，只要 λ 超过一个较小的正值，准确率便会迅速趋于平稳，并在较宽的倾斜强度范围内保持稳定。即使对于较大的 λ，也未观察到明显的性能下降。在 16样本设置中，较小的 λ 表现最佳，而更强的倾斜带来的额外收益有限，这表明随着带标签支持集的增加，激进的重新加权已无必要。

总的来说，这些结果表明，指数倾斜是一种条件良好且鲁棒的测试时调整机制。与可能对优化超参数敏感的基于梯度的自适应方法不同，此处的性能基本不受 λ 具体取值的影响，从而减少了对任务特定调优的需求，并有助于在实践中实现可靠部署。

5 讨论

本节提供实证结果，以验证我们对方法的主张：

> i. 在严格的冻结模型约束下，仍可实现有意义的测试时自适应；ii. 指数倾斜通过纠正潜在分布失配来提升性能，而非学习新参数；iii. 由此得到的自适应稳定、可解释，且能与更强的基线相媲美。

5.1 潜在分布重加权的有效性

表1表明，在所有数据集、编码器和样本数设置下，指数倾斜都能持续提升少样本分类准确率。对于 DINO 编码器，在 1–4 个样本数设置下，倾斜可带来 $+1.6$ 至 $+3.8$ 个百分点的绝对提升，即使在 16 个样本时仍有益处。尽管 JEPA 编码器的基线准确率整体较低，也观察到了类似的提升，这表明该效果并非特定于某种架构。

这些提升是在不修改编码器、分类器或决策规则的情况下实现的，因此能够单独体现潜在分布重加权的效果。这些改进证实，在冻结推理下，由少样本数据诱导的经验潜在分布与下游任务之间的失配是误差的重要来源。指数倾斜通过将概率质量重新分配至与任务一致的嵌入，在推理时纠正了这种失配。

重要的是，随着提示样例数增加，提升依然持续，但幅度逐渐减小。这一现象符合预期：随着可用的标注样本增多，经验潜在分布更能代表该任务，从而降低所需校正的程度。这种单调但趋于饱和的提升模式直接支持我们将自适应表述为测度变换，而非参数学习。

5.2 严格冻结模型约束下的比较

表2将指数倾斜置于遵守相同部署约束的免训练测试时自适应方法之中。我们的归纳式变体优于所有符合约束的基线，包括冻结原型、 k-NN分类和温度缩放。传导式变体进一步提升了性能，在不使用梯度、参数更新或上游数据的方法中取得了最高准确率。

线性头微调下的性能。在这组实验中，我们放宽了冻结头这一假设。大型编码器仍保持冻结。在采用线性头微调时，我们的方法在轻量级自适应方法中性能最佳。如表2所示，将我们的方法与线性头结合后，在所有样本数设置下均优于以往基于优化的方法（例如 TENT、Tip-Adapter-F）。

值得注意的是，受冻结约束的指数倾斜与线性头微调之间的差距，远小于它们与冻结推理之间的差距。例如，在5样本设置下，无需梯度或参数更新，指数倾斜便能取得线性头微调带来的大部分提升。

这些结果表明，大部分测试时改进源自潜在空间分布校正，而非表征学习；无论是在冻结设置还是仅具备最少可训练参数的设置下，指数倾斜都提供了强劲且有竞争力的适应基线。

5.3 归纳式与传导式参考测度

归纳式变体与传导式变体的区别仅在于倾斜所用参考测度的选择。归纳式变体仅依赖带标签支持集嵌入，而传导式变体则将无标签查询嵌入纳入参考分布。

如表2所示，传导式变体始终优于归纳式变体。这一改进直接源于测度变换解释：纳入查询嵌入后，得到的参考分布能够更贴近真实的测试时潜在分布。至关重要的是，这一优势无需修改查询表征、迭代更新伪标签，也无需引入优化循环即可实现。

结果表明，在这一设定中，传导式增益的关键驱动因素是获得更好的参考测度，而非增加监督或学习能力。

5.4 通过潜在分布重加权实现跨域泛化

附录表5报告了跨域少样本性能，其中模型在 mini-ImageNet 上训练，并在未见过的目标数据集上进行评估。在所有目标域（CUB、CARS、STL-10、CIFAR-10/100）和样本数设置下，指数倾斜始终优于冻结推理，且在低样本设置以及 STL-10 和 CARS 等语义距离较远的域中提升最大。在 1样本设置下，倾斜使各数据集的准确率提高约 $+1.6$、$+2.1$ 个百分点；随着样本数增加，提升幅度逐渐减小，这与域内饱和行为一致。

我们进一步评估了模型在 ImageNet 分布偏移下的鲁棒性，包括 ImageNet-V2、ImageNet-R 和 ImageNet-A。如附录表6所示，指数倾斜在所有偏移类型下均能提升性能，包括 ImageNet-A 等严重偏移，且无需进行参数更新。总的来说，这些结果表明，潜在分布重加权能够有效泛化至数据集层面和基于偏移的域变化。更多分析见附录C。

5.5 性能提升的机制解释

对倾斜权重的分析揭示了一种一致的模式：分类器置信度较高或与支持集几何对齐程度较强的嵌入会获得更大的权重，而模糊或置信度较低的嵌入则会被降低权重。这一现象在不同数据集和样本数设置下均可观察到。

指数倾斜并不改变表示或决策边界，而是重塑了对预测进行平均时所依据的有效潜在总体。这为观察到的性能提升提供了具体的机制解释，也与指数倾斜的 KL 最优性解释相一致。因此，实证结果既支持这一概率化表述，也证明了所提方法在实践中的有效性。

总体而言，实验确认，在冻结模型约束下，指数倾斜是一种有理论依据且有效的测试时自适应方法，仅通过推理级潜在分布重加权便实现了稳定的性能提升。

5.6 局限性

所提出的方法存在若干局限。首先，指数倾斜作用于由冻结编码器诱导的潜在分布，无法弥补表示级缺陷；如果编码器无法区分下游类别，仅靠重加权可能并不足够。其次，该方法依赖规模较小的带标签支持集来确定任务相关性，而在噪声极大或对抗性少样本设置中，这种相关性可能不可靠。最后，传导式变体假定可以访问一组静态的无标签测试点，因此可能无法自然地扩展到流式或非平稳测试分布。

6 结论

我们提出了指数倾斜，将其作为一种有理论依据的免训练测试时自适应方法，用于少样本分类。通过将自适应重新表述为对冻结编码器所诱导的潜在分布进行测度变换，所提出的方法无需梯度更新、参数修改或访问上游数据，即可实现任务特定推理。我们的理论表述将测试时自适应与任务诱导约束下的KL最优分布重加权联系起来，并为标签感知和几何感知校正提供了统一的解释。实证结果表明，在多种样本数设置下，该方法相较于冻结推理和其他免训练基线均取得了稳定提升，验证了推理级自适应的有效性。这项工作凸显了潜在分布推理作为基础模型基于参数的自适应之外一种尚未得到充分探索的替代方案。未来工作可以将该方法扩展到流式测试分布、多模态设置，或纳入结构化任务信息的更丰富评分函数，同时保持该方法免训练的特性。

影响声明

本研究提出了一种面向冻结模型约束下小样本图像分类的免训练测试时自适应方法，通过一种用于潜在空间重加权的概率方法，推进了对分布自适应的理解。该方法基于预训练模型表征运行，无需引入新的学习目标或数据收集流程，因此不会带来额外的伦理问题。它无需梯度或模型更新即可实现自适应，支持在受限场景中部署，同时提升对分布偏移的鲁棒性，并有望降低失效率。在实际应用中，负责任地使用预训练模型仍至关重要。

参考文献

Assran, M.、Duval, Q.、Misra, I.、Bojanowski, P.、Vincent, P.、Rabbat, M.、LeCun, Y. 和 Ballas, N. 基于联合嵌入预测架构从图像进行自监督学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第15619–15629页，2023年。Avdiukhin, D.、Chatziafratis, V.、Fischer, O. 和 Yaroslavtsev, G. 对比学习的嵌入维度与k近邻。神经信息处理系统进展，37:41359–41393，2024年。Braham, N. A. A.、Albrecht, C. M.、Mairal, J.、Chanussot, J.、Wang, Y. 和 Zhu, X. X. Spectralearth：大规模训练高光谱基础模型。*IEEE*应用地球观测与遥感精选主题期刊，2025年。Caron, M.、Touvron, H.、Misra, I.、Jégou, H.、Mairal, J.、Bojanowski, P. 和 Joulin, A. 自监督视觉Transformer中的新兴特性。载于*IEEE/CVF*国际计算机视觉大会，第9650–9660页，2021年。Chen, W.-Y.、Liu, Y.-C.、Kira, Z.、Wang, Y.-C. F. 和 Huang, J.-B. 深入探讨少样本分类。*arXiv*预印本 *arXiv:1904.04232*，2019年。Cover, T. M. 信息论基础。John Wiley & Sons，1999年。Csiszár, I. 概率分布的I散度几何与最小化问题。概率年刊，第146–158页，1975年。Duchi, J. C.、Glynn, P. W. 和 Namkoong, H. 鲁棒优化的统计数据：一种广义经验似然方法。运筹研究数学，46(3):946–969，2021年。Guo, C.、Pleiss, G.、Sun, Y. 和 Weinberger, K. Q. 现代神经网络的校准。载于国际机器学习大会，第1321–1330页。PMLR，2017年。Han, J.、Na, J. 和 Hwang, W. 持续测试时自适应的排序熵最小化。*arXiv*预印本 *arXiv:2505.16441*，2025年。Han, Z.、Yang, J.、Wang, G.、Li, J.、Xu, Q.、Shou, M. Z. 和 Zhang, C. Dota：视觉语言模型的分布式测试时自适应。*arXiv*预印本 *arXiv:2409.19375*，2024年。Hu, Z.、Hu, Y.、Li, X.、Tang, S. 和 Duan, L.-Y. 超越熵：野外测试时自适应的区域置信度代理。*arXiv*预印本 *arXiv:2505.20704*，2025年。Khattak, M. U.、Rasheed, H.、Maaz, M.、Khan, S. 和 Khan, F. S. Maple：多模态提示学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第19113–19122页，2023年。Kim, B. 面向视觉语言模型的超轻量测试时自适应。*arXiv*预印本 *arXiv:2511.09101*，2025年。

Langley, P. 撰写机器学习论文。载于 Langley, P.（编），第*17*届国际机器学习会议论文集（*ICML 2000*），第1207–1216页，加利福尼亚州斯坦福，2000年。Morgan Kaufmann。Niu, S.、Wu, J.、Zhang, Y.、Chen, Y.、Zheng, S.、Zhao, P. 和 Tan, M. 无遗忘的高效测试时模型适应。载于国际机器学习大会，第16888–16905页。PMLR，2022年。Ovadia, Y. 等。你能信任模型的不确定性吗？评估数据集偏移下的预测不确定性。神经信息处理系统进展，32，2019年。Radford, A.、Kim, J. W.、Hallacy, C.、Ramesh, A.、Goh, G.、Agarwal, S.、Sastry, G.、Askell, A.、Mishkin, P.、Clark, J. 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习大会，第8748–8763页。PMLR，2021年。Ren, M.、Triantafillou, E.、Ravi, S.、Snell, J.、Swersky, K.、Tenenbaum, J. B.、Larochelle, H. 和 Zemel, R. S. 面向半监督少样本分类的元学习。*arXiv*预印本 *arXiv:1803.00676*，2018年。Saerens, M. 等。将分类器输出调整为新的先验概率。*Neural Computation*，2002年。Shimodaira, H. 通过对数似然函数加权改进协变量偏移下的预测推理。载于*Journal of Statistical Planning and Inference*，2000年。Snell, J.、Swersky, K. 和 Zemel, R. 面向少样本学习的原型网络。神经信息处理系统进展，30，2017年。Sun, Y.、Wang, X.、Liu, Z.、Miller, J.、Efros, A. 和 Hardt, M. 通过自监督进行测试时训练，以实现分布偏移下的泛化。载于国际机器学习大会，第9229–9248页。PMLR，2020年。Wang, D.、Shelhamer, E.、Liu, S.、Olshausen, B. 和 Darrell, T. Tent：通过熵最小化实现完全测试时自适应。*arXiv*预印本 *arXiv:2006.10726*，2020年。Xiao, Z. 和 Snoek, C. G. 超越测试时的模型适应：综述。*arXiv*预印本 *arXiv:2411.03687*，2024年。Yuan, L.、Xie, B. 和 Li, S. 动态场景下的鲁棒测试时自适应。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第15922–15932页，2023年。Zhang, M.、Levine, S. 和 Finn, C. Memo：通过适应与增强实现测试时鲁棒性。神经信息处理系统进展，35:38629–38642，2022年。Zhang, R.、Fang, R.、Zhang, W.、Gao, P.、Li, K.、Dai, J.、Qiao, Y. 和 Li, H. Tip-adapter：无需训练的CLIP适配器，用于更好的视觉语言建模。*arXiv*预印本 *arXiv:2111.03930*，2021年。Zhou, K.、Yang, J.、Loy, C. C. 和 Liu, Z. 学习为视觉语言模型生成提示。*International Journal ofComputer Vision*，130(9):2337–2348，2022年。Zhu, H. 和 Koniusz, P. 通过迭代图细化实现基于原型的标签传播的传导式少样本学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第23996–24006页，2023年。

附录 A

A.1 指数潜在倾斜的理论性质

我们给出理论结果，将指数倾斜形式化为一种有原则的、无需训练的机制，用于在冻结模型约束下进行测试时自适应。

定理 A.1（KL 最优潜在测度适应）。设 P_0 为潜在空间 $\mathcal{Z}$ 上的参考概率测度，并且令 $s:\ \mathcal{Z}\to\mathbb{R}$ 为一个可测评分函数。对于任意 $c\in\mathbb{R}$，考虑约束优化问题

$$\min_{P\ll P_0}\ \mathrm{KL}(P\,\|\,P_0)\quad \textit{subject to}\quad \mathbb{E}_{z\sim P}[s(z)]=c. \tag{11}$$

则唯一解由指数倾斜分布给出

$$P_\lambda(z)=\frac{P_0(z)\exp(\lambda s(z))}{Z(\lambda)}, \tag{12}$$

其中 $\lambda\in\mathbb{R}$ 的选取满足 $\mathbb{E}_{z\sim P_\lambda}[s(z)]=c$，且 $Z(\lambda)=\int P_0(z)\exp(\lambda s(z))\,dz$。

证明。该结果可由凸对偶得到。定义拉格朗日函数

$$\mathcal{L}(P,\lambda,\nu)=\mathrm{KL}(P\,\|\,P_0)+\lambda\big(\mathbb{E}_P[s(z)]-c\big)+\nu\Big(\int dP-1\Big). \tag{13}$$

对 P 求泛函导数并令其为零，可得

$$\log\frac{dP}{dP_0}(z)+1+\lambda s(z)+\nu=0, \tag{14}$$

这意味着

$$\frac{dP}{dP_0}(z)\propto\exp(\lambda s(z)). \tag{15}$$

归一化即可得到所述形式。由 KL 散度的严格凸性可知其唯一性. □

推论 A.2（经验潜在倾斜）。令参考测度 P_0 为由有限支撑的潜在表示集合诱导的经验分布

$$P_0=\frac{1}{n}\sum_{i=1}^{n}\delta_{z_i},$$

其中 $z_i=f(x_i)$ 是由冻结编码器生成的嵌入。令 $s:\ \mathcal{Z}\to\mathbb{R}$ 为定义在潜在空间上的实值评分函数，并假设对于相关范围内的 λ，$\sum_{i=1}^{n}\exp(\lambda s(z_i))<\infty$。
那么，解 定理 *A.1* 中 *KL* 约束优化问题的解是离散分布 n

$$P_\lambda=\sum_{i=1}^{n}w_i\,\delta_{z_i},\qquad w_i=\frac{\exp(\lambda s(z_i))}{\sum_{j=1}^{n}\exp(\lambda s(z_j))}. \tag{16}$$

此外，对于任意可测函数 $g:\ \mathcal{Z}\to\mathbb{R}$，倾斜分布下的期望可化为加权经验平均，

$$\mathbb{E}_{z\sim P_\lambda}[g(z)]=\sum_{i=1}^{n}w_i\,g(z_i). \tag{17}$$

定理 A.3（倾斜潜在测度下的贝叶斯最优边缘预测）。令 $p_0(y\mid z)$ 为固定的条件分类器，令 P_λ 为潜在空间 $\mathcal{Z}$上的概率测度。考虑输出边缘分布 $q(y)\in\Delta(\mathcal{Y})$ 且与 z无关的预测器类别。

则预测分布

$$p_\lambda(y)=\mathbb{E}_{z\sim P_\lambda}\left[p_0(y\mid z)\right] \tag{18}$$

唯一地最小化期望对数损失

$$\mathbb{E}_{z\sim P_\lambda}\left[\mathbb{E}_{y\sim p_0(\cdot|z)}[-\log q(y)]\right] \tag{19}$$

对所有边缘预测器 $q(y)$。

证明。 目标函数可写为

$$\mathbb{E}_{y\sim p_\lambda}[-\log q(y)],$$

其中 $p_\lambda(y) = \mathbb{E}_{z\sim P_\lambda}[p_0(y \mid z)]$。这是 p_λ 与 q之间的交叉熵，其唯一最小值出现在$q = p_\lambda$时。

□

命题 A.4(倾斜测度的经验形式)。 设

$$P_0(z) = \frac{1}{n}\sum_{i=1}^{n}\delta(z - z_i) \tag{20}$$

为由潜在嵌入 $\{z_i\}_{i=1}^n$诱导的经验分布。则倾斜分布赋予权重

$$w_i = \frac{\exp(\lambda s(z_i))}{\sum_{j=1}^{n}\exp(\lambda s(z_j))}, \tag{21}$$

且 P_λ 下的期望可化为支持集上的加权和。

证明。 将 P_0 的经验形式代入 P_λ 的定义可得

$$P_\lambda(z) = \frac{\sum_i \delta(z - z_i)\exp(\lambda s(z_i))}{\sum_j \exp(\lambda s(z_j))}, \tag{22}$$

权重由此直接得出。 □

推论 A.5 （标签偏移校正作为特例）。 假设评分函数仅依赖于预测标签，

$$s(z) = \log \pi\big(y(z)\big), \tag{23}$$

其中 π 表示类别先验比，且 y（z） $= \mathrm{argmax}_y\, p_0$（$y \mid z$）。则对 P_0（z） 进行指数倾斜会诱导出等价于经典标签偏移校正的后验重加权。

证明。 被分配到同一预测类别的所有潜在表示获得相同的评分。因此，指数倾斜在每个类别内均匀地重新分配概率质量，从而按类别先验比进行后验重缩放，如同经典标签偏移适应中那样。

□

命题 A.6 （小 λ 的一阶稳定性）。 对于足够小的 λ，倾斜预测分布可作如下展开

$$p_\lambda(y) = p_0(y) + \lambda\,\mathrm{Cov}_{P_0}\big(p_0(y \mid z), s(z)\big) + O(\lambda^2). \tag{24}$$

证明。 该结果由指数权重与归一化常数$Z(\lambda)$的一阶泰勒展开得出。

□

B 实验与结果

B.1 消融实验与进一步分析

我们针对指数倾斜强度 λ 进行了详细的消融研究，以更好地了解所提出的适应机制在严格的冻结模型、免训练设置下的行为、稳健性和局限性。所有实验均遵循与主要结果相同的评估方案。

表3显示，在所有样本数设置下，指数倾斜的性能均稳定优于基线。即使倾斜强度较小（$\lambda = 0.25$），在1样本和2样本设置下也能带来显著提升，这凸显了在标注数据极其有限时，概率重加权的有效性。

一个关键发现是，性能在较大范围的 λ 取值内都很稳定。对于1样本、2样本和4样本设置，应用倾斜后，准确率变化很小； $\lambda = 0.25$ 与 $\lambda = 2.0$之间的差异远在标准差范围之内。这表明该方法对精确的超参数调优并不敏感，也不太可能因 λ的适度设定偏差而失效。

随着提示样例数增加，增强倾斜所带来的相对收益会逐渐减小。在 16样本设置中，较小的 λ 可取得最佳性能，而较大的倾斜强度则会略微降低准确率。这种现象是可以预期的：

表3：对指数倾斜强度 λ 进行消融实验，评估其在5路少样本分类中的表现。所有方法均使用冻结编码器和分类器。粗体表示每种样本数设置下的最佳性能。

提示样例数	λ	基线	倾斜准确率
1样本	–	53.87 ± 9.41	–
	0.25	–	$\mathbf{60.49 \pm 10.32}$
	0.50	–	60.44 ± 10.35
	0.75	–	60.39 ± 10.40
	1.00	–	60.35 ± 10.45
	1.25	–	60.35 ± 10.40
	1.50	–	60.40 ± 10.39
	1.75	–	60.41 ± 10.38
	2.00	–	60.37 ± 10.37
2样本	–	66.61 ± 6.80	–
	0.25	–	72.16 ± 7.56
	0.50	–	72.12 ± 7.50
	0.75	–	72.24 ± 7.41
	1.00	–	72.25 ± 7.45
	1.25	–	72.27 ± 7.43
	1.50	–	$\mathbf{72.29 \pm 7.45}$
	1.75	–	72.29 ± 7.56
	2.00	–	72.27 ± 7.51
4样本	–	74.68 ± 7.76	–
	0.25	–	78.04 ± 8.39
	0.50	–	78.09 ± 8.31
	0.75	–	78.13 ± 8.31
	1.00	–	78.17 ± 8.26
	1.25	–	78.24 ± 8.29
	1.50	–	78.20 ± 8.32
	1.75	–	78.21 ± 8.33
	2.00	–	$\mathbf{78.31 \pm 8.33}$
8样本	–	82.48 ± 5.85	–
	0.25	–	83.75 ± 5.81
	0.50	–	83.67 ± 5.80
	0.75	–	83.72 ± 5.79
	1.00	–	83.80 ± 5.76
	1.25	–	83.80 ± 5.76
	1.50	–	$\mathbf{83.81 \pm 5.80}$
	1.75	–	83.73 ± 5.85
	2.00	–	83.75 ± 5.89
16样本	–	84.68 ± 5.38	–
	0.25	–	$\mathbf{85.05 \pm 5.55}$
	0.50	–	84.97 ± 5.62
	0.75	–	84.91 ± 5.65
	1.00	–	84.85 ± 5.70
	1.25	–	84.81 ± 5.73
	1.50	–	84.73 ± 5.73
	1.75	–	84.75 ± 5.70
	2.00	–	84.73 ± 5.70

当有充足的带标签支持集可用时，对潜在表示进行激进的重新加权就不再那么必要，还可能过度强调细微的分布变化。

总体而言，这项消融实验支持将指数倾斜理解为一种稳健且表现良好的测试时调整。它在低数据情形下带来稳定收益，在广泛的参数范围内保持稳健，并且随着监督增加而自然减弱，与轻量级、免训练适应机制的预期作用高度一致。

表4：指数倾斜对 λ选择的敏感性。在较宽的 λ 取值范围内，准确率变化极小，表明无需逐任务调优即可保持鲁棒性。

提示样例数	λ 范围	准确率范围	最大 Δ
1样本	$[0.25, 2.0]$	60.35–60.49	$+6.63$
2样本	$[0.25, 2.0]$	72。16–72。29	$+5.68$
4样本	$[0.25, 2.0]$	78。04–78。31	$+3.63$
8样本	$[0.25, 2.0]$	83。67–83。81	$+1.33$
16样本	$[0.25, 2.0]$	84。73–85。05	$+0.37$

表5：5分类任务上的跨域少样本分类准确率（%）。模型在 mini-ImageNet 上训练，并在目标数据集上评估。所有编码器和分类器均被冻结。性能提升完全源于潜在分布重加权。

数据集	方法	1样本	2样本	4样本	8样本	16样本
			DINO ViT-S/16			
CUB	基线	51.42 ± 7.96	61.87 ± 6.88	72.14 ± 6.21	79.02 ± 5.54	82.63 ± 5.01
	倾斜（我们的方法）	$\mathbf{53.06 \pm 7.89}$	$\mathbf{63.44 \pm 6.74}$	$\mathbf{74.01 \pm 6.18}$	$\mathbf{80.21 \pm 5.48}$	$\mathbf{83.11 \pm 4.96}$
CARS	基线	47.35 ± 8.22	58.41 ± 7.05	69.08 ± 6.47	75.66 ± 5.92	78.94 ± 5.36
	倾斜（我们的方法）	$\mathbf{48.98 \pm 8.14}$	$\mathbf{60.02 \pm 6.91}$	$\mathbf{70.91 \pm 6.39}$	$\mathbf{76.88 \pm 5.87}$	$\mathbf{79.42 \pm 5.31}$
STL-10	基线	40.73 ± 8.51	52.19 ± 7.24	63.02 ± 6.66	69.88 ± 6.02	73.41 ± 5.48
	倾斜（我们的方法）	$\mathbf{42.81 \pm 8.39}$	$\mathbf{53.96 \pm 7.10}$	$\mathbf{64.87 \pm 6.58}$	$\mathbf{71.02 \pm 5.96}$	$\mathbf{74.18 \pm 5.42}$
CIFAR-10	基线	55.63 ± 6.84	64.92 ± 6.21	77.71 ± 5.47	81.94 ± 4.88	84.67 ± 4.32
	倾斜（我们的方法）	$\mathbf{57.33 \pm 6.79}$	$\mathbf{66.01 \pm 6.05}$	$\mathbf{77.71 \pm 5.44}$	$\mathbf{82.02 \pm 4.85}$	$\mathbf{84.71 \pm 4.30}$
CIFAR-100	基线	66.47 ± 6.12	73.85 ± 5.63	87.47 ± 4.91	90.26 ± 4.35	92.11 ± 3.98
	倾斜（我们的方法）	$\mathbf{68.64 \pm 6.08}$	$\mathbf{75.02 \pm 5.58}$	$\mathbf{87.47 \pm 4.89}$	$\mathbf{90.31 \pm 4.33}$	$\mathbf{92.14 \pm 3.96}$

表6：5分类任务上的跨域少样本分类准确率（%）。模型使用在 ImageNet 上预训练的 DINOViT-S/16，并在分布偏移数据集上进行评估。所有编码器和分类器均被冻结。指数倾斜在各种分布偏移下均带来了稳定的改进。

数据集	方法	1样本	2样本	4样本	8样本	16样本
			DINO ViT-S/16			
ImageNet-V2	基线	58.23 ± 8.72	68.41 ± 7.65	76.82 ± 6.93	81.54 ± 6.21	84.37 ± 5.68
	倾斜（我们的方法）	$\mathbf{61.07 \pm 8.65}$	$\mathbf{70.98 \pm 7.58}$	$\mathbf{78.65 \pm 6.86}$	$\mathbf{82.89 \pm 6.15}$	$\mathbf{85.21 \pm 5.62}$
ImageNet-R	基线	41.56 ± 9.15	53.28 ± 8.22	65.74 ± 7.41	72.31 ± 6.68	76.42 ± 6.05
	倾斜（我们的方法）	$\mathbf{45.19 \pm 9.08}$	$\mathbf{56.33 \pm 8.14}$	$\mathbf{68.02 \pm 7.35}$	$\mathbf{74.25 \pm 6.61}$	$\mathbf{77.86 \pm 5.98}$
ImageNet-S	基线	38.72 ± 9.37	50.81 ± 8.44	63.05 ± 7.62	69.88 ± 6.89	74.15 ± 6.24
	倾斜（我们的方法）	$\mathbf{42.65 \pm 9.29}$	$\mathbf{54.26 \pm 8.36}$	$\mathbf{65.78 \pm 7.55}$	$\mathbf{71.94 \pm 6.82}$	$\mathbf{75.69 \pm 6.17}$
ImageNet-A	基线	35.41 ± 9.52	47.23 ± 8.68	59.84 ± 7.86	66.92 ± 7.12	71.38 ± 6.46
	倾斜（我们的方法）	$\mathbf{38.94 \pm 9.43}$	$\mathbf{50.67 \pm 8.59}$	$\mathbf{62.51 \pm 7.78}$	$\mathbf{69.03 \pm 7.04}$	$\mathbf{73.02 \pm 6.39}$

C 通过潜在分布重加权实现跨域泛化

这一现象直接支持将指数倾斜理解为纠正域诱导的潜在分布失配的一种机制。当从 mini-ImageNet 迁移到新的视觉领域时，冻结编码器生成的嵌入，其经验分布与下游任务结构并不匹配。指数倾斜通过对潜在空间中与任务一致的区域重新加权来补偿这种失配，而无需进行特定领域再训练、适应数据流或参数更新。

值得注意的是，即使在基线准确率相对较高的数据集（例如 CIFAR-100）上也能观察到改进，这表明该方法并非仅仅从失败案例中恢复，而是对冻结

推理进行持续优化。在更高的样本数下，增益变小或趋于饱和，这是符合预期的，因为经验支持集的分布会越来越能够代表目标域。
至关重要的是，所有改进均仅源于潜在分布重加权。未使用任何目标域数据进行训练，未计算梯度，也未修改模型参数。这将分布校正分离为实际起作用的适应机制，并验证了核心主张：跨域的测试时泛化仅通过推理层面的测度变换即可实现。

总体而言，跨域结果表明，指数倾斜并不局限于域内少样本自适应，而是在严格的冻结模型约束下，为域迁移提供了一种有原则且有效的机制。

对 ImageNet 变体的跨域鲁棒性分析。我们在源自 ImageNet 的具有挑战性的分布偏移上评估指数倾斜。结果见表6。我们的方法在所有偏移类型上均持续提升性能，从较轻微的 ImageNet-V2（1样本为 +2.84%）到严重的 ImageNet-A（1样本为 +3.53%）。这些改进在分布式失配最为严重的低样本场景中最为显著，表明潜在重加权能够在不进行参数更新的情况下有效适应多样的域偏移。

跨少样本设定的鲁棒性。 我们在一系列少样本设置（包括样本数极低的场景）下评估指数倾斜。在考察的所有样本数下，指数倾斜均稳定优于冻结推理。相对提升幅度最大的是低样本场景，此时带标签支持集有限，简单平均尤其容易变得脆弱。随着支持样本数量增加，提升幅度逐渐减小，这表明当分布不确定性最高时，该方法的自适应效果最为显著。这一现象与如下解释相符：指数倾斜会有选择地强调潜在空间中与任务一致的区域。

跨回合采样的稳定性。 为评估统计稳定性，我们考察了测试回合数量不同时的性能。随着回合数量增加，指数倾斜带来的提升始终保持稳定，其波动程度与冻结基线相当。这表明观察到的增益并非有限采样造成的假象，而是反映了预测性能的系统性提升。

与线性头微调的比较。 尽管我们通过允许线性分类器微调放宽了冻结分类器约束，但仍将其作为允许优化时可达到性能的近似上界。我们发现，尽管无需进行梯度计算、参数更新或额外训练，指数倾斜仍能获得微调所带来提升的相当一部分。当优化不可行或不适宜时，这一比较凸显了潜在分布重加权作为替代适应机制的有效性。

骨干网络敏感性。 我们考察了指数倾斜能否适用于不同类型的冻结视觉编码器，包括自监督模型和视觉语言模型。在不同骨干网络上，倾斜相较于冻结推理均带来了稳定的性能提升。尽管绝对性能因编码器而异，分布式重加权带来的相对收益仍然稳定，这表明该方法在很大程度上与架构无关，不依赖编码器特有的属性。

消融实验：评分函数组件。 我们分析了倾斜所用不同评分函数的贡献。标签感知评分带来了主要的性能提升，因为它们会强调与已观察到的支持集标签相一致的嵌入。几何感知评分则通过纳入从支持集中提取的潜在结构，带来互补的收益。两者结合时性能最佳，这证实了基于似然的信号与几何信号为自适应提供了不同且可叠加的信息。

消融实验：对倾斜强度的敏感性。 我们研究了不同样本数设置下倾斜参数的影响。在较宽的取值范围内，性能保持稳定，变化幅度很小。这种鲁棒性消除了针对每个任务进行超参数调优的必要，也使该方法无需验证数据即可实际部署。

归纳式与传导式变体。当有无标签查询嵌入可用时，我们会考虑一种将其纳入参考潜在分布的传导式变体。与归纳式设置相比，该变体能带来进一步改进，同时仍然不需要梯度、优化或参数更新。重要的是，自适应仍然完全基于分布，从而使这种方法有别于基于优化的传导式方法。

对支持集噪声的鲁棒性。 我们通过引入受控的标签损坏来评估对噪声支持集标签的敏感性。指数倾斜相对于冻结推理表现出更高的鲁棒性，因为低置信度或不一致

的样本会在倾斜测度下被自动降权。这种行为自然地从加权机制中产生，并且不需要显式的噪声建模。

计算开销。 指数倾斜在推理时仅产生极少的额外成本。除冻结的前向传播外，自适应只需计算支持集上的标量权重和加权平均值。由此产生的开销随支持样本数量线性增长，与编码器推理相比仍可忽略不计，因此该方法适用于资源受限部署。