arXiv:2512.14574v1 [cs.CV] 16 Dec 2025

FoodLogAthl-218：利用饮食管理应用程序构建真实世界食品图像数据集

Mitsuki Watanabe
东京大学
东京，日本
watanabe@hal.t.u-tokyo.ac.jp

Sosuke Amano
foo.log Inc.
东京，日本
amano@foo-log.co.jp

Kiyoharu Aizawa
东京大学
东京，日本
aizawa@hal.t.u-tokyo.ac.jp

Yoko Yamakata
东京大学
东京，日本
yamakata@hal.t.u-tokyo.ac.jp

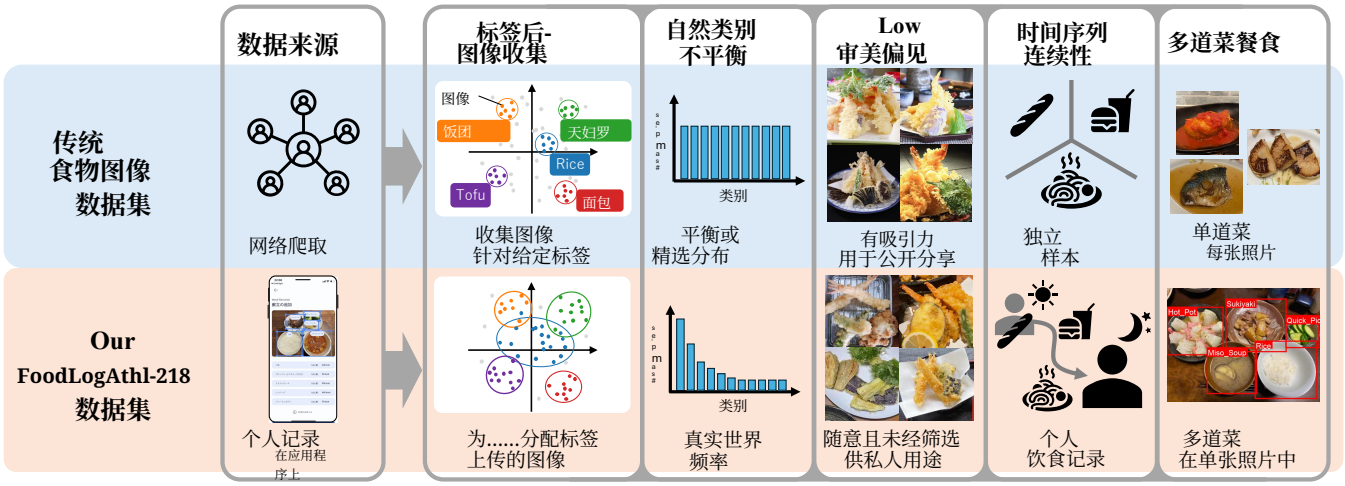


图 1：传统食物图像数据集与 FoodLogAthl-218 的比较。

摘要

食物图像分类模型对于饮食管理应用至关重要，因为它们能够减轻手动记录餐食的负担。然而，大多数公开可用的此类模型训练数据集依赖网络爬取图像，而这些图像往往与用户在现实生活中拍摄的餐食照片有所不同。本文介绍 *FoodLogAthl-218*，这是一个基于饮食管理应用 FoodLog Athl 收集的真实餐食记录构建的食物图像数据集。该数据集包含 6,925 张图像，涵盖 218 个食物类别，共有 14,349 个边界框。每张图像均附有丰富的元数据，包括餐食日期和时间、匿名用户 ID 及餐食级上下文。与传统数据集不同——传统数据集由预定义的类别集合引导网络图像收集——我们的数据从用户提交的照片开始，之后再添加标签。这种做法带来了更高的类内多样性、自然形成的餐食类型频率分布，以及随手拍摄、未经筛选且用于个人用途而非公开分享的图像。除了 (1) 标准分类基准测试外，我们还提出两项 FoodLog 专属任务：(2) 遵循用户日志时间序列的增量微调协议，以及（3）上下文感知分类任务，其中每张图像包含多道菜，模型必须利用整体餐食上下文对每道菜进行分类。我们使用大型多模态模型（大型多模态模型）对这些任务进行评估。该数据集已公开发布于 https://huggingface.co/datasets/FoodLog/FoodLogAthl-218。

CCS 概念

• **计算方法学** → **计算机视觉任务**；• **应用计算** → 医疗保健信息系统；• **信息系统** →数据管理系统。

关键词

食物图像数据集；食物管理应用；个人食物记录；时间序列餐食日志；多菜肴餐食图像

1 引言

大多数公开可用的食物类别分类数据集均由网络来源图像构建 [5, 9, 10, 19, 21, 26–28, 35,36,38–40,42, 45, 50, 53]。尽管有些研究使用专有食物日志，但由于隐私和许可限制，这些数据集通常不公开 [4, 20, 29]。

为弥补这一空白，我们基于用户通过我们开发的食物管理应用提交的个人

这是被 ACM Multimedia 2025（数据集赛道）接收的论文作者预印本版本。正式版本请参见：https://doi.org/10.1145/ 3746027.3758276。

表 1：食品图像分类任务数据集比较。

Year	数据集名称	#类别	#样本	菜系类别	数据来源
2009	PFID [12]	101	4,545	西方	快餐连锁店
2014	UEC-Food 256 [28]	256	25,088	国际	**网络** + 原始数据库
2014	食物-101 [10]	101	101,000	西方	**网络** (foodspotting.com)
2015	UPMC 食物-101 [50]	101	90,840	西方	**网络** (Google [17])
2016	UNICT-FD1200 [16]	1,200	4,754	国际	实验参与者
2017	VegFRU [21]	25/292*	160,000	蔬菜、水果	**网络**（Google、ImageNet [14], 等）
2019	FoodAI-756 [45]	152/756*	400,000	国际	**网络**（Google、Bing 等）
2021	Food2K [40]	2,000	1,036,564	国际	**网络**（美团 [37]）
2023	Food-500 Cap [35]	494	24,700	国际	**网络**（Google、Bing）
2025	FoodLogAthl-218	218	14,349	国际	**饮食管理应用**

* 层次结构化数据集，其中 n/m 表示 m 个类别划分为 n 个更高层级类别。

食品记录，构建并公开发布了一个新的食物图像数据集 <style id='1'>FoodLogAthl-218</style>。该数据集经机构伦理委员会批准后整理而成；所有参与者均已提供知情同意，并通过两阶段人工流程移除了个人身份信息。由于用户提交的原始数据中存在漏检、误检及标签噪声，我们采用定制的多阶段筛选流程，生成了一个干净、可用于研究的子集，其中包含 6,925 张图像和 14,349 个边界框，涵盖 218 个食物类别，规模与现有分类数据集相当。

FoodLogAthl-218数据集具有以下五项关键特性，使其区别于现有食品图像基准，并更适用于真实世界的饮食记录场景。**图像采集后再标注** 与大多数现有数据集不同，后者会先定义一组固定的类别标签，再收集图像以匹配这些标签；我们的数据集则从用户提交的餐食照片开始，并在事后进行标注。这种自下而上的流程显著提高了类内多样性，也更准确地反映了真实世界饮食记录任务的性质。

自然类别不平衡现有数据集通常会尝试平衡每个类别的样本数量，或依赖网络来源图像的可用性；而我们的数据集保留了日常食品记录中餐食出现的自然频率分布。

低审美偏差 不同于用于公开展示（例如食谱、广告或社交媒体）的网络来源图像，我们的数据集由用户为个人使用而随手拍摄、未经筛选的照片组成。这些图像常常光照不佳、装盘未经修饰，或呈现吃了一部分的菜肴——这些场景在网络收集的数据中鲜少出现，因此其视觉特征截然不同。

用于个性化的时间序列连续性 由于许多用户会在较长时间内记录餐食，该数据集具有部分时间序列结构。这种时间连续性有助于纵向建模，并可用于评估增量学习和个性化。

多菜品餐食构成 数据集中的每张图像都可能包含同一盘或托盘上共现的多道菜，反映了真实的餐食构成，并支持基于单餐内菜品共现情况的上下文感知分类。

图 1 展示了其中几项显著特征，包括视觉多样性、餐食构图和真实世界数据特征。

综合来看，这些方面凸显了传统食物图像数据集在支持实际食物记录应用方面的局限性，也说明了为何需要 FoodLogAthl-218 这类更贴近现实、以用户为中心的资源。

2 相关工作

2.1 基于图像的食物记录

随着智能手机的广泛普及，许多餐食管理应用程序相继推出。在日本，Asken [6], MyFit- nessPal [15], 和 Calomeal [32] 较为知名。这些应用程序利用图像识别技术自动识别餐食内容并估算其营养价值。

本研究使用了本实验室出于学术目的开发的食品记录应用程序“FoodLog Athl” [41] 。FoodLog Athl 的一项重要功能是促进用户与注册营养师之间的沟通。用户在应用程序中关联账户后，即可与指定的营养师分享食品记录；营养师随后可以查看记录并提供个性化的营养指导。为确保营养分析准确，FoodLog Athl 不允许用户自由输入菜品名称，而是要求用户从预设数据库中选择菜肴。该数据库包含约 1,400 种常见的日本料理，以及额外的 14 万条项目（包括日常食品、餐厅菜肴和包装食品）。该数据库汇编自 SARAH [1] 和《日本食品成分标准表》 [47]。

2.2 食物图像数据集

为了使模型能够准确预测现实世界的数据，必须使用能够反映真实环境的丰富数据进行训练。因此，数据集是训练和评估模型的基础，其质量、规模和多样性直接关系到模型的性能。在膳食研究领域，食物图像数据集的研究不断发展，例如 Food-101 [10]包含 101,000 张图像，涵盖 101 个类别；UEC-Food 256 [28]包含 25,088 张图像，涵盖 256 个类别；Food2K [40] 包含 1,036,564 张图像，涵盖 2,000 个类别。这些数据集的规模逐年扩大。

如表1所示，现有的大多数为分类任务构建的食品图像数据集都依赖网络（包括社交网站（SNS））作为其主要数据来源。

表 2：数据筛选流程及图像、边界框（bboxes）和类别的数量。

Step	操作	#图像	(差异)	#边界框	(差异)	#类别	(差异)
0	原始数据	11,460		61,414		5,889	
1	移除没有图像的样本	11,460	0)	26,646	(-34,768)	3,167	(-2,722)
2	移除没有边界框的样本	11,285	(-175)	26,065	-581)	3,071	-96)
3	移除包含 <10 个样本的类别	9,121	(-2,164)	21,944	(-4,121)	1,180	(-1,891)
4	根据图像特征移除离群值	7,537	(-1,584)	17,513	(-4,431)	1,180	0)
5	人工审核并移除无效图像	6,925	(-612)	14,349	(-3,164)	1,180	0)
6	合并相似类别并清理标签	6,925	0)	14,349	0)	218	(-982)



图2：错误数据示例。

2.3 食品领域的视觉-语言模型

2.3.1 CLIP。 得益于较高的零样本准确率，对比式语言-图像预训练（CLIP） [43]可作为食品图像分类的基线模型。例如，CLIP 的零样本分类器在 Food-101 [43]上取得了超过90%的 Top-1准确率——比基于 ResNet-50 的分类器 [18] 高出15个百分点以上。此外，在成分估计任务中，基于 CLIP 的模型在每种成分的召回率和 F1分数方面均高于基于 ResNet 的模型 [48]。

2.3.2 大型多模态模型（*LMM*）。除 CLIP 外，近期多模态学习的进展主要集中在通用型大型多模态模型上，这类模型将图像编码器与大型语言模型整合到端到端框架中。LLaVA [33], MiniGPT- 4 [54], OpenFlamingo [7], Phi-3.5-Vision [2], 和 Qwen-VL [8]等模型在一系列食物相关任务中均展现出强大的零样本和少样本性能。例如，FoodMLLM-JP 利用这些架构生成日语食谱，表现出出色的流畅度和烹饪准确性 [23]。经过领域适配的变体——如 FoodLMM [51], ChefFusion [31],和 RoDE [24]——进一步提升了细粒度分类、食材定位和营养估计的效果。这些成果证实，CLIP 和大型多模态模型可作为食品图像基准的可靠基线。

3 FoodLogAthl-218 数据集

如引言所述，用户通过日常饮食记录应用程序提交的食品记录往往存在大量噪声，如图2所示。为解决这一问题，我们开发了一套定制的多阶段过滤流水线，以提取一个质量可与现有食物图像数据集相媲美的干净子集。

3.1 数据收集

本研究使用的数据集由 2023 年 5 月至 2024 年 10 月期间通过 FoodLog Athl 应用收集的食物记录组成，涵盖 632 名用户。由于许多用户仅上传一次，或提交的并非食物图像，我们采用了以下筛选步骤，以获得可靠的约 60,000 条记录的子集：(i) 选取至少关联三个用户账户的营养师；(ii) 确定与这些营养师关联的所有用户账户；(iii) 获取这些用户的每条餐食记录。每条记录都包含匿名用户 ID、餐食日期和时间、菜品名称、份量、一张带有边界框标注的图像，以及营养值（最初均由应用程序内置的检测与识别模型推断，之后根据需要由用户进行修正）。由于图像中通常包含多道菜，因此将每道菜都视为一个独立样本，这使得一张图像可以对应多个数据集样本。

3.2 数据筛选

表 2 展示了数据清理步骤及其对样本数量和类别数量的影响。步骤 1 和 2 会移除图像缺失或图像及边界框分配错误的记录。步骤 3 会剔除样本数量不足、不适合用于训练的类别。

步骤 4 使用基于 CLIP 的异常值检测来识别较隐蔽的错误，例如裁剪图标签错误或非食物图像。每张菜品裁剪图都由预训练 CLIP 模型（ViT-B/16）进行编码，并计算其特征与类别质心之间的距离。为确定合适的阈值，我们人工检查了不同距离范围内的样本，发现将高于类别均值 1.5 个标准差的样本标记出来，既能以较高精度剔除明显错误，又能保留有效但少见的样本。因此，任何距离超过该阈值的样本都会被标记为异常值。

步骤 5：人工审核所有剩余样本，修正边界框和标签，或移除无效条目，以确保最终数据集的质量。

为保留真实世界记录中食品记录的自然分布，我们将每张图像视为一个原子单位。

由于一张图像可能包含多道菜，移除单个样本可能会扭曲餐食层级模式。

因此，如果某张图像中的任何样本被标记，则会丢弃所有相关样本。

这一全有或全无策略保留了多菜品餐食的原始频率，并支持整餐任务，例如第 4 节中的描述引导分类。经过筛选至步骤5后，最终保留了来自513位用户的6,925张餐食图像和14,349个边界框。

表 3：使用特征提取器对 FoodLogAthl-218 和 UEC-Food 256 的分布指标进行比较。较高的迹 和 戴维斯–博尔丁指数 表示分布更分散，而较低的轮廓系数 和 卡林斯基–哈拉巴斯指数 表示聚类不够紧密。

	迹 (↑)			轮廓系数 (↓)			戴维斯–博尔丁指数 (↑)			卡林斯基–哈拉巴斯指数 (↓)		
数据集	**CLIP**	**EVA-CLIP**	**Swin**	**CLIP**	**EVA-CLIP**	**Swin**	**CLIP**	**EVA-CLIP**	**Swin**	**CLIP**	**EVA-CLIP**	**Swin**
UEC-Food 256	71.21	45.09	264.00	0.0335	0.0577	-0.0064	3.64	3.02	4.11	74.05	115.80	49.17
FoodLogAthl-218（我们）	**81.70**	**49.92**	**282.57**	**-0.0210**	**0.0036**	**-0.0288**	**4.52**	**3.68**	**4.51**	**27.51**	**51.51**	**21.29**

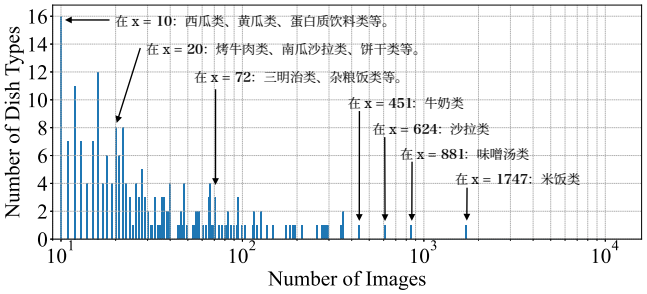


图 3：FoodLogAthl-218 中各类别条目数量的对数－线性图（横轴为对数刻度，纵轴为线性刻度），并附有示例图。



图 4：FoodLogAthl-218 中所含数据示例。

3.3 聚类

如第2.1节所述，食物名称（即类别标签）从超过14万个选项中选取，筛选后仍存在各种标注不一致的问题。为了将我们的数据集与现有食物图像数据集进行比较，我们在步骤6采用了两阶段聚类流程，将指向同一道菜的标签合并。首先，我们向 GPT-4o [3] 提示："请将以下菜品名称归入合适的类别"，利用其预训练语言知识和领域知识，自动聚类同义词、拼写变体和细微的措辞差异。这一模型辅助步骤大幅减少了统一庞大且嘈杂的标签集所需的人工工作。其次，我们人工审核并完善这些聚类，以纠正错误归类，并确保每个类别的定义清晰且彼此有别。通过将可扩展的自动分组与有针对性的人为整理相结合，我们最终的类别定义兼具效率和较高的标注准确性。

3.4 数据集概述与类别分布

图 3 展示了 FoodLogAthl-218 中各类别条目数量的对数－线性分布图（横轴为对数刻度，纵轴为线性刻度）。各类别的样本数量高度不平衡，真实反映了日本 FoodLog Athl 用户记录的日常餐食的实际频率分布。

图 4 展示了来自 FoodLogAthl-218 的餐食图像示例。该数据集包含网络上很少见、但经常出现在个人食物日志中的图像，例如盒装餐食（红框）和在便利店购买的包装食品（蓝框），以及电饭煲中吃剩的部分米饭（紫框）。另一个显著特征是构图和光照的广泛变化。例如，在黄色框标出的图像中，拍摄者的智能手机在食物上投下了阴影——这是随手拍摄、未经编辑的餐食快照的常见特征。

为了评估类内变异性和类间可分性，我们使用三种不同的编码器，从 FoodLogAthl-218 和 UEC-Food 256——另一个日本食物图像分类数据集——的每个样本中提取图像特征：CLIP (ViT-L/14) [43],EVA-CLIP (eva_giant_patch14_clip_224) [46], 以及 Swin Transformer (swin_large_patch4_window12_384.ms_in22k) [34]。对于每个编码器及其对应的特征集，我们计算了四种离散度指标，以评估类别可分性和变异性：

- **迹(↑)** [25]：所有特征维度上的方差总和。数值越高，表示类内离散度越大。
- **轮廓系数(↓)** [44]：衡量类别的类内凝聚度和分离度；分数越低，表示重叠越多。
- **Davies–Bouldin 指数(↑)** [13]：类内散度与类间分离度之比。数值越大，表示类别越不紧凑、重叠越多。
- **Calinski–Harabasz 指数(↓)** [11]：类间方差与类内方差之比。数值越低，表示分离度越弱。

表 3 总结了各编码器在这四项指标上的结果。列标题中的箭头表示方向（↑越大，↓越小）对应着越高的类内方差和类间重叠，也就是更高的分类难度。在所有指标和编码器上，与 UEC-Food 256 相比，FoodLogAthl-218 的类内方差和类间重叠始终更高，这表明用户提交的自下而上的数据集比通过传统的基于网络的方法收集的数据集具有更多样的视觉表现（也伴随更多噪声）。

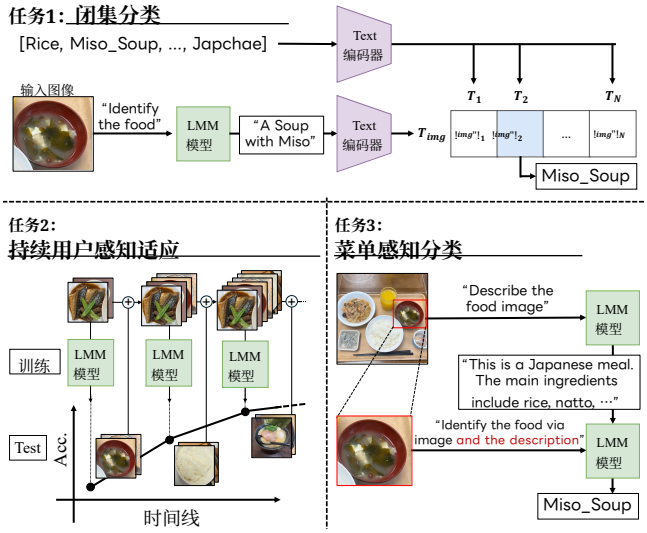


图 5：三个评估任务概览。

4 基准任务与实验设置

本节定义了三项任务，用于评估 FoodLogAthl-218 数据集的复杂性，并为每项任务提供基于大型多模态模型（LMM）的基线方法，如图 5 所示。

4.1 任务 1：标准闭集分类

4.1.1 任务定义与基线方法。与传统的食物图像基准测试一样，每张菜品裁剪图恰好对应 218 个类别标签中的一个。我们使用 lmms-eval 对各种大型多模态模型进行该任务的基准测试；lmms-eval 是一个统一且标准化的多模态评估框架，包含 10 多个模型[30, 52]。在基线方法中，我们首先为每张菜品图像裁剪图生成自由形式的描述，然后使用 intfloat/e5-large-v2 文本编码器 [49]分别对该描述和全部 218 个类别标签进行嵌入。我们选择与描述嵌入具有最高余弦相似度的标签作为预测标签。该整体流程示于图 5 顶部。对于 EVA-CLIP 和 CLIP，我们通过直接计算图像特征与标签文本嵌入之间的余弦相似度来执行分类。为应对不确定性和标签歧义，我们报告 Top-1/3/5 准确率。

4.1.2 实验结果。 表 4 显示，GPT-4o 的 recall@1 最高（≈42%），在所有开源模型中，Qwen2.5-VL-7B-Instruct 表现最强，但即使是表现最好的大型多模态模型，recall@1 也停留在 40% 左右，凸显了 FoodLogAthl- 218 的挑战。有趣的是，与大型多模态模型相比，规模较小的分类模型 CLIP (ViT-L/14) 达到了最高的 recall@5（60%）。这是因为它直接进行图文相似度评分，自然能够生成完整的排序列表；而基于嵌入的大型多模态模型描述往往将概率集中在首个猜测上，因此生成多样化 Top-N 候选项的效果较差。

4.2 任务 2：持续的用户感知适应

4.2.1 任务定义和基线方法 *1*。用户通常会表现出重复的饮食模式，经常在一段时间内食用相似的餐食。这表明，持续使用新

表 4：FoodLogAthl-218 上的分类性能。

模型	F1	R@1	R@3	R@5
GPT-4o	**0.3732**	**0.4243**	**0.4991**	0.5221
Qwen2.5-VL-7B-Instruct	0.2349	0.3684	0.4609	0.4859
Qwen2-VL-7B-Instruct	0.1594	0.3656	0.4590	0.4781
Qwen2-VL-3B-Instruct	0.1386	0.2663	0.3250	0.3517
llava-onevision-qwen2-7b-ov	0.1546	0.3629	0.4497	0.4639
EVA-CLIP (ViT-g)	0.2465	0.2654	0.4897	0.5944
CLIP (ViT-L/14)	0.2393	0.2567	0.4885	**0.6000**
llava-v1.6-34b	0.1354	0.2432	0.3779	0.4063
llava-v1.5-7b	0.0991	0.2140	0.2721	0.2973
llava-v1.6-mistral-7b	0.0944	0.1932	0.2595	0.2831

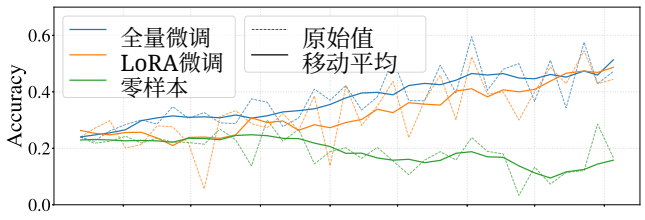


图 6：所有用户的增量准确率：零样本、全量微调和 LoRA 微调。

记录的餐食更新模型可能会提升分类性能，因为此前见过的餐食会变得更容易识别。我们利用"用餐日期"元数据，对截至第 $n-1$ 周的所有记录逐步微调模型，并对第 n 周记录的餐食进行测试，如图 5 左下方所示。

使用 Qwen2.5-VL-3B-Instruct，我们在留出数据划分上评估了三种模型变体：(1) 全量微调（更新所有模型参数）、(2) LoRA 微调（仅调整低秩适配器权重） [22], 以及 (3) 不进行微调的零样本基线。

4.2.2 实验结果 *1*。图 6 展示了三种变体的增量准确率。实线表示加权一个月移动平均（± 两周），权重与每周记录数成正比；虚线表示每周原始准确率。

图 6 显示，通过对真实膳食日志持续进行微调，准确率从约 20% 提升至 50% 以上，相比零样本基线提高了 30 多个百分点，最终甚至超过了大型商业大语言多模态模型。LoRA 微调在最初几周略逊于全量微调，当时只有几十条记录；但随着数据不断积累，它很快迎头赶上，仅使用一小部分可训练参数，就达到了接近全量微调的表现。相比之下，零样本曲线缓慢下滑，这凸显了用户饮食历史更新对于抵消现实世界膳食记录数据分布偏移的重要性。

4.2.3 任务定义与基线方法 *2*。为分析长期记录对个性化的影响，我们聚焦于最活跃的单个用户（用户 A）。该用户从 5 月

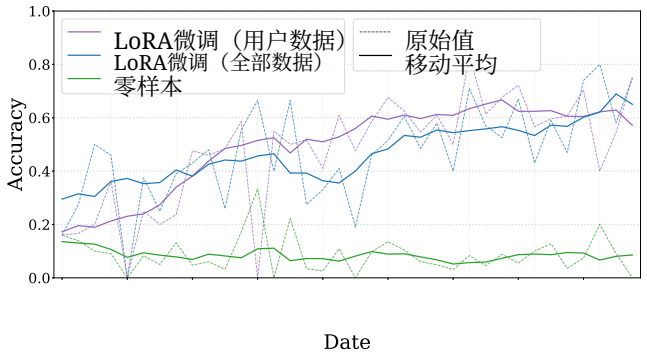


图7：用户 A 的准确率随时间变化：零样本、在所有用户数据上进行 LoRA 微调，以及仅在用户 A 数据上进行 LoRA 微调。

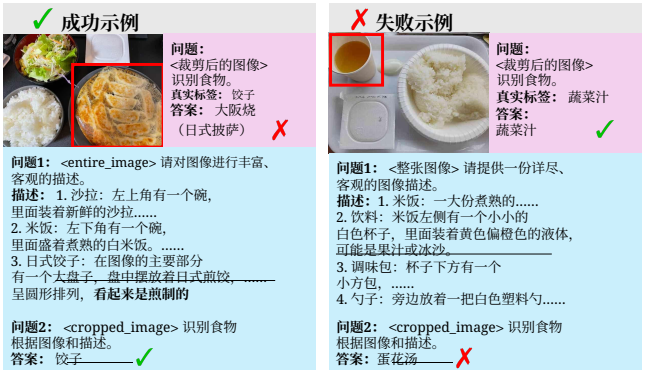


图8：描述引导的分类任务示例。粉色阴影区域表示基线的“无描述”设置，蓝色阴影区域表示“有描述”设置。

表5：Qwen2.5-VL-7B-Instruct 在描述引导分类任务上的 F1分数、Precision@1（P@1）和 Recall@K（R@1、R@3、R@5）。

	F1	P@1	R@1	R@3	R@5
不使用描述	0.1594	0.1019	**0.3656**	**0.4590**	**0.4781**
使用描述	**0.2304**	**0.1788**	0.3240	0.4032	0.4336

12 日持续记录至 2024 年 1 月 12 日，共提交了 274 张图像，涵盖 1,003 条菜品记录。基于该用户的记录，我们比较了三种变体：（1）仅使用用户 A 数据进行 LoRA 微调；（2）使用所有用户数据进行 LoRA 微调；（3）不进行微调的零样本基线。

4.2.4 实验结果 2。 图7展示了它们在一段时间内对用户 A 记录的分类准确率。实线表示一个月移动平均（± 两周），虚线表示每周原始准确率。

从图7可以看出，两种经 LoRA 微调的模型最终都比零样本基线提高了 50 个百分点。全数据模型起初优于用户专属模型，但随着用户 A 的记录不断增加，个性化模型逐渐缩小差距，最终达到或超过全数据模型，展现出有效的用户级适应能力。

4.3 任务 3：菜单感知分类

4.3.1 任务定义与基线方法。 单张照片中往往包含多道菜肴。为利用这一上下文，我们首先将整张图像展示给大型多模态模型，并要求其用简短的“菜单”描述概括所有可见食物。随后，使用每个裁剪图像及菜单描述对其进行重新评估。这种两阶段推理利用餐食级线索来区分视觉上相似的菜肴（例如主菜与配菜，或不同地区的变体），旨在提高单道菜肴的准确率。图 5 右下方展示了这一整体流程。

4.3.2 实验结果。 我们使用 Qwen2.5-VL-7B-Instruct 生成菜单描述。表5展示了模型在有无高层次餐食描述的情况下对每张菜品裁剪图进行分类时的 F1、Precision@1 (P@1) 和 Recall@K (R@1、R@3、R@5)。

加入菜单描述后，F1 从 0.159 提升至 0.230，Precision@1 也提高了 7.5% 以上（从 0.102 提升至 0.179），代价是召回率略有下降。举例来说，模型不再默认预测为“米饭”（占样本的 11.9%），而是能够正确预测大麦_米饭、糙米_米饭和炒_米饭等较少见的变体。重要的是，这并非仅限于米饭类：在整个数据集中，我们都观察到模型对其他主导标签的偏向有所减弱，并转而预测其较少见的子类别，这表明标签多样性得到了普遍提升。

尽管有所提升，加入菜单描述也带来了取舍：它正确地重新分类了此前被误分类的 687 个实例，但同时也将原本正确标注的 1,222 个实例错误分类。图8展示了成功纠正和新增错误的代表性示例。这些结果表明，大型多模态模型能够有效利用整餐的上下文信息。例如，单独评估时，一张日式煎饺的裁剪图会被错误分类为大阪烧 。然而，加入整餐描述后，描述中提到了米饭和配菜——这些通常与日式煎饺 一起供应、但在日本通常不会与大阪烧 一起供应——模型便能正确预测其标签为饺子。相反，一些在没有上下文时分类正确的样本，在加入餐食描述后却被错误分类，这表明上下文虽能提高预测准确率，但有时也会造成歧义或误导模型。

5 结论

我们介绍了*FoodLogAthl-218*，这是一个源自真实膳食日志的数据集，包含 6,925 张图像、14,349 个边界框和 218 个类别，并具有丰富的元数据和自然类别不平衡。通过三项任务——封闭集分类、持续用户感知适应和描述引导分类——我们证明，最先进的大型多模态模型难以处理未经加工的食物日志数据，但通过增量式和上下文感知微调，其性能可显著提升。

可用性。 FoodLogAthl-218 面向非商业学术研究发布。访问需要使用机构电子邮箱。请通过“Apply”在 `https://huggingface.co/datasets/FoodLog/FoodLogAthl-218`申请；经简短的人工审核后，将发放访问凭证。

伦理声明 本研究已获得东京大学信息科学与技术研究生院伦理审查委员会批准（方案编号 **UT-IST-RE-250415-2**）。

6 致谢

本研究得到 JSPS KAKENHI Grant Number 23K25247 和日本 JST NEXUS Grant Number JPMJNX25C9 的资助。

参考文献

[1] SARAH。2025。https://corporate.sarah30.com/。（访问日期：29/05/2025）。[2] Marah Abdin、Jyoti Aneja、Hany Awadalla、Ahmed Awadallah、Ammar Ahmad Awan、Nguyen Bach、Amit Bahree、Arash Bakhtiari、Jianmin Bao、Harkirat Behl 等。2024。Phi-3 技术报告：一款可在手机本地运行的高性能语言模型。arXiv 预印本 arXiv:2404.14219（2024）。[3] Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。2023。GPT-4 技术报告。arXiv 预印本 arXiv:2303.08774（2023）。[4] Kiyoharu Aizawa、Yuto Maruyama、He Li 和 Chamin Morikawa。2013。利用多媒体食物日志中的个人饮食倾向进行食物平衡估算。IEEE Trans. Multimedia 15, 8（2013），2176–2185。[5] Marios M Anthimopoulos、Lauro Gianola、Luca Scarnato、Peter Diem 和 Stavroula G Mougiakakou。2014。一种基于优化词袋特征模型、面向糖尿病患者的食物识别系统。JBHI 18, 4（2014），1261–1271。[6] asken, Inc. 2007。Asken（原为日语）。https://www.asken.jp/。（访问日期：16/03/2024）。[7] Anas Awadalla、Irena Gao、Josh Gardner、Jack Hessel、Yusuf Hanafy、Wanrong Zhu、Kalyani Marathe、Yonatan Bitton、Samir Gadre、Shiori Sagawa 等。2023。OpenFlamingo：用于训练大型自回归视觉语言模型的开源框架。arXiv 预印本 arXiv:2308.01390（2023）。[8] Jinze Bai、Shuai Bai、Yunfei Chu、Zeyu Cui、Kai Dang、Xiaodong Deng、Yang Fan、Wenbin Ge、Yu Han、Fei Huang 等。2023。Qwen 技术报告。arXiv 预印本 arXiv:2309.16609（2023）。[9] Thoranna Bender、Simon Sørensen、Alireza Kashani、Kristjan Eldjarn Hjorleifsson、Grethe Hyldig、Søren Hauberg、Serge Belongie 和 Frederik Warburg。2024。学会品尝：一个多模态葡萄酒数据集。NeurIPS 36（2024）。[10] Lukas Bossard、Matthieu Guillaumin 和 Luc Van Gool。2014。Food-101——利用随机森林挖掘判别性成分。载于 ECCV。[11] T. Caliński 和 J. Harabasz。1974。一种用于聚类分析的树突法。《统计学通讯——理论与方法》3, 1（1974），1–27。[12] Mei Chen、Kapil Dhingra、Wen Wu、Lei Yang、Rahul Sukthankar 和 Jie Yang。2009。PFID：匹兹堡快餐图像数据集。载于 ICIP。[13] D. ˇL. Davies 和 D. ˇW. Bouldin。1979。聚类分离度量。《IEEE模式分析与机器智能汇刊》PAMI-1, 2（1979），224–227。[14] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。2009。ImageNet：大型分层图像数据库。载于 CVPR。[15] Daniel Evans。2017。MyFitnessPal。BJSM 51, 14（2017），1101–1102。[16] Giovanni Maria Farinella、Dario Allegra、Marco Moltisanti、Filippo Stanco 和 Sebastiano Battiato。2016。食物图像的检索与分类。《计算机与生物医学》77（2016），23–39。[17] Google LLC。2024。Google 搜索。https://www.google.com/。（访问日期：16/03/2024）。[18] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016。用于图像识别的深度残差学习。载于 CVPR。[19] Hajime Hoashi、Taichi Joutou 和 Keiji Yanai。2010。通过特征融合识别 85 种食物类别的图像。载于 ISM。[20] Shinji Horiguchi、Sosuke Amano、Masahiko Ogawa 和 Kiyoharu Aizawa。2018。用于食物图像识别的个性化分类器。《IEEE Trans. Multimedia》20, 10（2018），2836–2848。[21] Saihui Hou、Yushan Feng 和 Zilei Wang。2017。VegFRU：用于细粒度视觉分类的领域专用数据集。载于 ICCV。[22] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 等。2022。LoRA：大型语言模型的低秩适配。载于 ICLR，第 1 卷，3。[23] Yuki Imajuku、Yoko Yamakata 和 Kiyoharu Aizawa。2025。FoodMLLM-JP：利用多模态大语言模型生成日语食谱。载于 MMM *2025* 401 414 10 1007 978 981 96 2054 8_30。–。doi: 。/ - - - - [24] Pengkun Jiao 、 Xinlan Wu 、 Bin Zhu 、 Jingjing Chen 、 Chong-Wah Ngo 和 Yugang Jiang。2024。RoDE：用于食物大型多模态模型的线性整流多样化专家混合。arXiv 预印本 arXiv:2407.12730（2024）。[25] I. ˇT. Jolliffe。2002。主成分分析。Springer。[26] Taichi Joutou 和 Keiji Yanai。2009。一种采用多核学习的食物图像识别系统。载于 ICIP。[27] Parneet Kaur、Karan Sikka、Weijun Wang、Serge Belongie 和 Ajay Divakaran。2019。FoodX-251：一个用于细粒度食物分类的数据集。载于 CVPR 细粒度视觉分类研讨会。[28] Y. Kawano 和 K. Yanai。2014。利用领域自适应和现有类别自动扩展食物图像数据集。载于 ECCV Workshops（TASK-CV）。[29] Seum Kim、Yoko Yamakata 和 Kiyoharu Aizawa。2021。通过共享食品记录提升个性化食品图像分类器性能。载于 ACM ICMR Workshops。29–32。[30] Bo Li、Peiyuan Zhang、Kaichen Zhang、Fanyi Pu、Xinrun Du、Yuhao Dong、Haotian Liu、Yuanhan Zhang、Ge Zhang、Chunyuan Li 和 Ziwei Liu。2024。LMMs-Eval：加速大型多模态模型的开发。https://github.com/EvolvingLMMs-Lab/lmms-eval [31] Peiyu Li、Xiaobao Huang、Yijun Tian 和 Nitesh V Chawla。2024。ChefFusion：融合食谱与食品图像生成的多模态基础模型。载于 CIKM。3872–3876。[32] Life Log Technology, Inc. 2016。Calomeal（原为日语）。https://www.calomeal.com/about-calomeal/。（访问日期：16/03/2024）。[33] Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。2024。视觉指令微调。NeurIPS 36（2024）。[34] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。2021。Swin Transformer：使用移位窗口的层次化视觉Transformer。载于 ICCV。10012–10022。[35] Zheng Ma、Mianzhi Pan、Wenhan Wu、Kanzhi Cheng、Jianbing Zhang、Shujian Huang 和 Jiajun Chen。2023。Food-500 Cap：用于评估视觉语言模型的细粒度食品描述基准。载于 ACMMM。[36] Yuji Matsuda 和 Keiji Yanai。2012。利用流形排序并考虑共现关系的多食品识别。载于 ICPR。[37] 美团。2024。美团。https://www.meituan.com/。（访问日期：16/03/2024）。[38] Weiqing Min、Linhu Liu、Zhengdong Luo 和 Shuqiang Jiang。2019。用于食物识别的成分引导级联多注意力网络。载于 ACMMM。[39] Weiqing Min、Linhu Liu、Zhiling Wang、Zhengdong Luo、Xiaoming Wei、Xiaolin Wei 和 Shuqiang Jiang。2020。ISIA Food-500：通过堆叠式全局—局部注意力网络实现大规模食物识别的数据集。载于 ACMMM。[40] Weiqing Min、Zhiling Wang、Yuxin Liu、Mengjiang Luo、Liping Kang、Xiaoming Wei、Xiaolin Wei 和 Shuqiang Jiang。2023。大规模视觉食物识别。TPAMI（2023）。[41] Kei Nakamoto、Kohei Kumazawa、Hiroaki Karasawa、Sosuke Amano、Yoko Yamakata 和 Kiyoharu Aizawa。2022。FoodLog Athl：用于饮食指导与食品监测的多媒体食品记录平台。载于 ACMMM Asia。[42] Jianing Qiu、Frank P-W Lo、Yingnan Sun、Siyao Wang 和 Benny Lo。2019。挖掘判别性食品区域，实现精准食物识别。BMVC（2019）。[43] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021。通过自然语言监督学习可迁移的视觉模型。载于 ICML。[44] P. ˇJ. Rousseeuw。1987。轮廓系数：一种用于解释和验证聚类分析的图形辅助工具。《计算与应用数学杂志》20（1987），53–65。[45] Doyen Sahoo、Wang Hao、Shu Ke、Wu Xiongwei、Hung Le、Palakorn Achananuparp、Ee-Peng Lim 和 Steven CH Hoi。2019。FoodAI：通过深度学习实现食物图像识别，助力智能食品记录。载于 KDD。[46] Quan Sun、Yuxin Fang、Ledell Wu、Xinlong Wang 和 Yue Cao。2023。EVA-CLIP：改进 CLIP 的大规模训练技术。arXiv 预印本 arXiv:2303.15389（2023 2303 15389）。https://arxiv.org/abs/。日 [47] 本文部科学省。2015。《日本食品成分标准表——2015——（第七次修订版）》。https://www.mext.go.jp/en/policy/science_technology/policy/title01/detail01/1374030.htm（访问日期：21/07/2024）。[48] Liangyu Wang、Yoko Yamakata、Ryoma Maeda 和 Kiyoharu Aizawa。2024。测量并改善您的饮食：基于成分估计的营养计算器。载于 ACM MM '24。11273–11275。doi:10.1145/3664647.3684997[49] Liang Wang、Nan Yang、Xiaolong Huang、Binxing Jiao、Linjun Yang、Daxin Jiang、Rangan Majumder 和 Furu Wei。2022。通过弱监督对比预训练获得文本嵌入。arXiv 预印本 arXiv:2212.03533（2022）。https://arxiv.org/abs/ 2212 03533。Xin Wang、Devinder Kumar、Nicolas Thome、Matthieu Cord 和 Frederic Precioso。2015。[50] 利用大型多模态食品数据集进行食谱识别。载于 ICMEW。[51] Yuehao Yin、Huiyan Qi、Bin Zhu、Jingjing Chen、Yu-Gang Jiang 和 Chong-Wah Ngo。2023。FoodLMM：基于大型多模态模型的多功能食品助手。arXiv 预印本 arXiv:2312.14991（2023）。[52] Kaichen Zhang、Bo Li、Peiyuan Zhang、Fanyi Pu、Joshua Adrian Cahyono、Kairui Hu、Shuai Liu、Yuanhan Zhang、Jingkang Yang、Chunyuan Li 和 Ziwei Liu。2024。LMMs-Eval：对大型多模态模型评估的现实检验。arXiv: 2407 12772 [ ] 2407 12772。[53] cs.CL https://arxiv.org/abs/。Feng Zhou 和 Yuanqing Lin。未注明年份。通过探索二部图标签实现细粒度图像分类。载于 CVPR。[54] Deyao Zhu、Jun Chen、Xiaoqian Shen、Xiang Li 和 Mohamed Elhoseiny。2023。MiniGPT-4：利用先进的大型语言模型增强视觉语言理解。arXiv 预印本 arXiv:2304.10592（2023）。2016