基于视觉的少样本人类活动识别与基于 MLLM 的视觉强化学习

Wenqi Zheng*, Yutaka Arakawa*
*) Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University, JAPAN
E-mail: zheng.wenqi.005@s.kyushu-u.ac.jp, arakawa@ait.kyushu-u.ac.jp

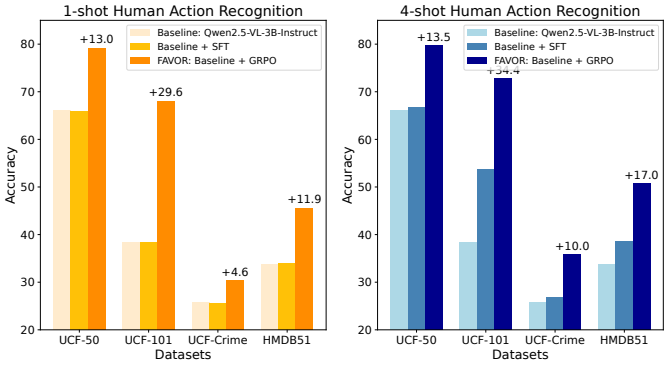

Fig. 1. FAVOR 在 HAR 分类的各种任务中表现优于传统的有监督微调（SFT）。

Abstract—**大型推理模型中的强化学习能够通过对模型输出的反馈来进行学习，这在微调数据有限的情况下特别有价值。然而，它在多模态人体活动识别（HAR）领域的应用尚未得到充分探索。我们的工作将强化学习扩展到了具有多模态大型语言模型的人体活动识别领域。通过在训练过程中引入视觉强化学习，模型在小样本识别上的泛化能力可以大大提升。此外，视觉强化学习还能提高模型的推理能力，并在推理阶段实现可解释的分析。我们将结合视觉强化学习的小样本人体活动识别方法命名为 FAVOR。具体来说，我们的方法首先利用多模态大型语言模型（MLLM）为人体活动图像生成多个候选响应，每个响应都包含推理轨迹和最终答案。然后，使用奖励函数对这些响应进行评估，并使用群组相对政策优化（GRPO）算法对 MLLM 模型进行优化。通过这种方式，MLLM 模型可以仅通过少量样本适应于人体活动识别。在四个人体活动识别数据集和五种不同设置上的大量实验验证了所提出方法的优越性。**

Index Terms—few-shot recognition, human activity recognition, reinforcement learning.

I. 引言

人类活动识别（HAR）专注于通过分析从各种设备收集的数据来识别和分类人类活动，这些设备包括视觉技术和基于物联网的系统 [16], [18], [19] 。目前广泛使用两种主要类型的 HAR 系统：基于视频的系统和基于传感器的系统 [16] 。基于视频的系统依赖于诸如 RGB 视频、骨架数据和深度信息等视觉模态，而基于传感器的系统则利用像陀螺仪和加速度计这样的设备。在这些系统中，由于能够捕捉更丰富的信息并提供对场景的背景理解，视频为主的方法在当前 HAR 研究中占据主导地位。目前，HAR 已被应用于广泛的领域，包括智能家居、医疗保健、社会科学、康复工程、健身等 [2], [6], [12] 。HAR 显著提高了全球人类的安全和福祉。本文主要讨论日常生活场景中的基于视觉的 HAR。

传统的人类活动识别（HAR）系统采用经典的机器学习分类器 [18], [19] 来处理从人工活动数据中手动提取的特征向量。深度学习的出现将这个范式转变为自动特征提取。神经网络结构可以以端到端的方式直接处理原始的人类活动数据流 [6], [12] 。随后，变压器架构的引入进一步革新了深度学习 [4], [21], [22] 。通过采用多头自注意力机制，变压器模型通过注意力头的协作从多个角度学习多样化的表示 [19] 。这种方法优于传统方法，同时通常需要更少的计算资源。然而，大多数这些方法依赖于大规模的标记数据集，并且在以前未见过的人类活动方面显示出有限的泛化能力 [6] 。此外，它们主要产生直接的活动分类输出，缺乏可解释性，这阻碍了它们在实际场景中的部署。此外，在 HAR 领域，它们的全部潜力和局限性仍然很大程度上未被探索，尤其是在利用多模态数据的丰富多样性方面 [8] 。

最近，许多研究人员开始探索使用多模态大型语言模型（MLLMs）来解决 HAR 任务 [9], [10], [12], [14], [18]–[20] 。虽然 MLLMs 经过大规模的预训练后可以很好地泛化，但通过微调可以进一步增强其在 HAR 任务中的表现。监督微调（SFT）范式直接模仿高质量数据集中提供的“真实标签”答案，因此依赖于大量的训练数据。另一个值得注意的后训练方法是强化微调（RFT），在该方法中，奖励评分由预定义的规则确定。这些规则评估模型的响应并基于反馈指导优化。RFT 和 SFT 之间的一个关键区别在于数据效率：RFT 可以使用少至数十到数千个样本有效地微调模型，而 SFT 通常需要更多的数据。在本文中，我们介绍了一种使用视觉强化学习（FAVOR）的少量样本人类活动识别方法，该方法无需领域特定的优化即可实现对新任务的零样本和少样本泛化。因此，这种方法使 MLLMs 能够处理各种多模态 HAR 任务（见图 2 ）。然而，如图 2 所示，SFT 直接产生输出，没有经过多步骤推理或上下文推理，这常常导致不准确的响应 [8], [11] 。为了解决这一限制，我们在图 2 中展示了 FAVOR 的实现细节。具体来说，对于每个输入，FAVOR 利用 MLLM 生成多个包含推理标记和最终答案的响应。重要的是，我们定义了基于规则的可验证奖励函数，以使用组相对策略优化 (GRPO) 引导策略优化。例如，我们为基于分类的 HAR 任务设计了量身定制的奖励函数。通过这样做，FAVOR 能够探索各种可能的推理路径，并学习优化以达到这些可验证奖励函数定义的理想结果。因此，我们的方法将训练范式从 SFT 中的数据扩展转变为针对特定多模态 HAR 任务定制的灵活奖励函数的战略设计。

此外，如图 1 所示，在少样本实验中，FAVOR 用最少的训练数据展示了卓越的性能，其少样本学习能力显著强于 SFT。这些多样化的视觉 HAR 任务进一步强调了强化

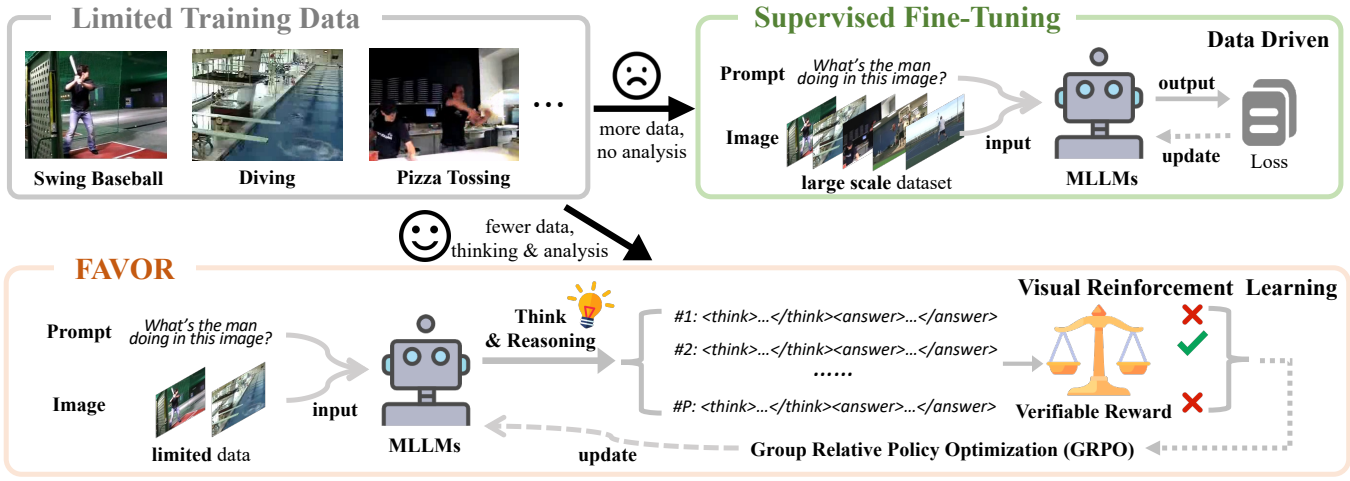

Fig. 2. 使用多模态大型语言模型的人的活动识别（HAR）概述

学习在增强多模态环境中的视觉感知和推理方面的关键作用。总结来说，我们的主要贡献如下：

- 我们提出了 FAVOR，这是一种人类活动识别方法，它将具有可验证奖励的视觉强化学习引入到人类活动识别领域，即使在训练数据稀少的情况下也被证明是有效的。
- 为人体活动识别任务提出了一种可验证的奖励函数，该函数提高了识别准确率并能够进行可解释的分析。
- 在 4 个人类活动识别数据集和 5 种不同设置上进行的大量实验表明，与基线和监督微调相比，所提出的方法具有优越性。

II. 相关工作

A. 强化学习

多模态大型语言模型（MLLM）将多种数据模态——如音频、图像和文本——整合到一个统一的框架中。在这种背景下，研究人员已经证明，后训练阶段可以有效提高模型的响应质量和推理能力，而不需要大量的监督数据或复杂的提示技术。在这些发现的基础上，最近的研究越来越多地集中于通过强化学习（RL）技术增强 MLLM 的推理能力。具体来说，RL 使模型可以通过与环境互动并接收奖励作为反馈来进行决策，从而最大化长期收益。为了进一步促进对齐和推理，开发了群体相对策略优化（GRPO），这将对齐任务重新表述为基于人工标记偏好的排序损失。与传统的监督方法不同，这种方法通过基于群体的奖励比较直接优化策略，同时保持与人类意图的一致性。

B. 在人体活动识别任务中应用强化学习

起初，强化学习（RL）被应用于深度学习模型，例如通过结合深度学习和 RL 来预测智能家居中的人类活动，或控制家用机器人以改善对人类互动的理解。随后，随着大型语言模型（LLM）的兴起，RL 在后训练阶段变得至关重要。特别是，带有人类反馈的强化学习（RLHF）在使模型与人类意图对齐方面发挥了关键作用。此外，奖励模型已证明在推理期间有效提高推理性能。随着研究的不断进展，RL 现在越来越多地被应用于多模态大型语言模型（MLLMs）。值得注意的是，在 HAR 的背景下，RL 有助于在各种多模态 HAR 任务中实现更好的推理、对齐和泛化。

VisualThinker-R1-Zero [3] 表明，应用 R1 于基础 MLLM 可以显著提高性能，并引发所谓的“恍然大悟时刻”的出现。OpenVLThinker 采用了通过监督微调 (SFT) 和强化学习 (RL) 交替进行的迭代自我改进训练策略。具体而言，SFT 用于指导模型获取初始推理结构，而 RL 则用于提升其性能和泛化能力。此外，VLM-R1 [14] 提出了一个跨模态推理管道，并将强化学习应用于视觉感知任务。同样地，R1-Omni [20] 将可验证奖励的强化学习 (RLVR) 应用于用于情感识别的全多模态大型语言模型。Open-Reasoner-Zero [5] 证明即使一个简单的基于规则的奖励策略也能提高回应长度和基准性能。值得注意的是，Visual-RFT [10] 使得一个 MLLM 在物体检测任务中表现优于仅通过 SFT 训练的模型。与这些模型相比，我们的方法采用不同的基础模型，并在一个不同的应用环境中进行评估。

III. 方法论

A. 预备知识

1) 通过增强实现策略梯度优化：强化算法是一种策略梯度方法，通过根据其行动获得的奖励来调整模型的策略（政策），以改善决策。它通过优化可能行动的概率分布来实现，从而增加选择导致更好结果行动的可能性。在每次迭代中，模型根据其过去决策的表现来更新其参数（θ），逐渐优化其行为。对于给定的问题 Q，GRPO 使用当前策略 $\pi_{\theta_{\text{old}}}$ 生成 P 个不同的响应 $\{O_1, O_2, \ldots, O_P\}$。随后对每个响应进行评估，以获得相应的奖励 $\{r_1, r_2, \ldots, r_P\}$。为评估其相对质量，GRPO 通过计算均值和标准差来标准化这些奖励：

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_P\})}{\text{std}(\{r_1, \ldots, r_P\})}, \tag{4}$$

这里，A_i 表示对第 i 个响应的归一化分数。通过对奖励进行归一化，GRPO 鼓励模型在每个组内优先考虑质量更高的响应，帮助它有效地区分强弱输出。

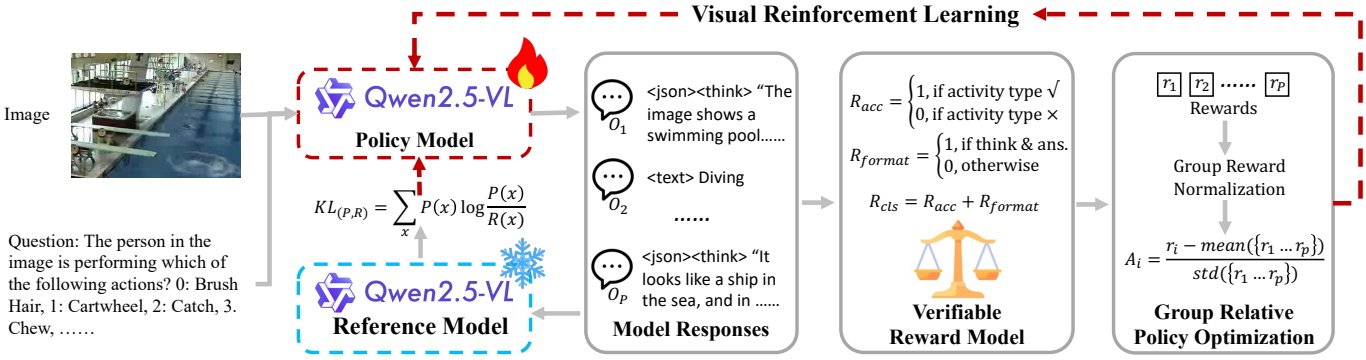

Fig. 3. FAVOR 的框架

Trivial Reasoning Trajectory

Question: <image> The person in the image is performing which of the following activities? 0: brush_hair, 1: cartwheel, 2: catch, *****
Response: First output the thinking process in <think> </think> tags and then output the final answer in <answer> </answer> tags. Output the final answer in JSON format.

Fig. 4. 将 GRPO 应用于基础模型的示例响应。

2) 通过强化学习实现可验证奖励：可验证奖励可以改善 MLLMs 在具有客观正确答案的任务上的表现，这直接采用了一个验证函数来评估输出的正确性。给定一个图像和一个输入问题 Q，多模态大型语言模型 M_θ 生成一系列响应，记作 O，这些响应由一个可验证的奖励函数 $R(Q,O)$ 进行评估。该函数根据输出是否与真实值匹配来分配分数：

$$R(Q,O) = \begin{cases} 1, & \text{if } O = \text{ground truth}, \\ 0, & \text{otherwise}. \end{cases} \tag{1}$$

优化目标定义为：

$$\max_{\pi_\theta} \mathbb{E}_{O \sim \pi_\theta(Q)} \left[R_{\text{RLVR}}(Q,O) \right], \tag{2}$$

其中

$$R_{\text{RLVR}}(Q,O) = R(Q,O) - \alpha \cdot \text{KL}\left[\pi_\theta(O \mid Q) \| \pi_{\text{ref}}(O \mid Q)\right]. \tag{3}$$

在这里，π_{ref} 表示优化前的参考模型，α 是一个超参数，用于在奖励最大化和 KL-散度正则化之间取得平衡，鼓励优化后的模型保持接近原始模型。在这项工作中，我们将 GRPO 与多模态大型语言模型结合，以利用两种方法的优势，增强模型在推理和活动识别方面的能力。

B. 偏好

1) FAVOR 框架：所示的 FAVOR 框架利用 Qwen2.5-VL-3B-Instruct [1] 作为策略模型 π_θ，该模型接受来自用户的多模态输入——即图像和问题——并生成推理过程以及一组候选响应。我们设计了一种提示格式，引导模型在给出最终答案之前生成推理轨迹，如图 4 所示。在训练过程中，奖励函数鼓励模型生成包括推理过程和最终答案的结构化输出。推理轨迹促进在强化微调过程中的自我改进，而基于答案的奖励推动策略优化。通过结合两个奖励，引导模型生成既准确又结构良好的预测。每个响应都通过可验证的奖励函数进行评估，并执行分组奖励计算以评估每个输出的质量。这些评估然后用于更新策略模型。为了确保训练的稳定性，FAVOR 引入了 KL 散度正则化以限制当前策略模型 π_θ 和参考模型 $\pi_{\theta_{\text{old}}}$ 之间的偏差。

2) 在 HAR 任务中的可验证奖励：奖励模型在强化学习中起着关键作用，它检查模型的预测是否与真实答案完全匹配。本文侧重于对图像进行推理，以根据给定的问题对人所执行的活动进行分类。因此，奖励函数 R_{cls} 包含两个部分：准确性奖励 R_{acc} 和格式奖励 R_{format}。总体奖励定义如下：

$$R_{\text{cls}} = R_{\text{acc}} + R_{\text{format}}. \tag{5}$$

我们的强化学习策略避免了复杂的奖励建模，而是选择了一种基于简单规则的奖励函数，以鼓励具有正确内容和适当格式的响应：

TABLE I
FAVOR 训练中使用的超参数。

Setting	Value
Batch size per device	8
Data_seed	4
Gradient accumulation steps	2
Training steps	20
Learning rate	5×10^{-5}
Temperature	1.0
Maximum response length	2048
Responses per GRPO step	16
KL coefficient	0.04

- 准确性奖励（+1）：在回复提供正确最终答案时分配。
- 格式奖励（+1）：在响应使用 <think> 标签包裹其推理并使用 <answer> 标签包裹最终答案时给予。
- 否则：分配奖励 0。

IV. 实验

A. 数据集

大多数先进的活动识别方法是在具有有限活动类别和控制环境下拍摄的数据集上进行评估的。相比之下，UCF50

[13] 数据集由现实世界的 YouTube 视频组成，涵盖 50 种不同的活动类别。HMDB51 [7] 数据集包括 51 个类别，既包含面部表情（如微笑、笑声、咀嚼和讲话），又包含 47 个全身活动。UCF101 [15] 提供 101 个人类活动类别，拥有超过 13,000 个视频片段，总计 27 小时的录像。UCF-Crime [17] 数据集包含 1,900 个长且未经修剪的监控视频，描绘了 13 种现实中的异常—例如打斗、交通事故、入室盗窃和抢劫—以及正常的日常活动。这些视频在现实条件下录制，通常涉及相机运动和杂乱的背景，使得该数据集对活动识别任务具有特别的挑战性。许多现有的方法在这些数据集上表现欠佳，这些数据集中的视频来自网络，通常面临诸如相机运动、光线不佳、场景杂乱、尺度和视角变化，以及对人类活动本身不一致的关注等挑战。在本研究中，我们证明了我们的 FAVOR 模型能够处理 HAR 推理和分类任务。我们在四个数据集上训练了我们的模型：UCF50 [13] 、UCF101 [15] 、UCF-Crime [17] 和 HMDB51 [7] 。

B. 评估

为了评估我们方法的推理和分类能力，我们在四个数据集上进行了实验。我们首先从每个数据集中获取视频，并从每个类别中的几个视频中随机提取一个帧来构建一个新的图像数据集。这些提取的帧根据它们在数据集中原始视频的类别进行分类。随后，从每个类别中随机选择少量图像构成训练集。剩余的图像用作后续评估和测试的测试集。我们采用小样本设置来评估模型的识别能力，应用强化学习于有限的数据上。具体而言，我们使用强化学习（RL）和监督微调（SFT）对从训练集中随机抽样的 1-shot、2-shot、4-shot、8-shot 和 16-shot 示例训练 Qwen2.5-VL-3B-instruct。剩余的数据被用作评估的测试集。

C. 实现细节

该模型使用学习率 5×10^{-5} 和温度 1.0 训练了 20 个周期。最大响应长度设置为 2048 个标记。在 GRPO 优化中，我们每步采样 8 个响应，并使用 0.04 的 KL 系数，如表 I 所示。为了展示直接将 GRPO 应用于基础模型的有效性，我们将我们的方法与两个基线进行比较：未进行 SFT 预训练的模型和在相同数据集上使用 SFT 微调的基础模型。

D. 主要结果

我们评估了多种模型在多个 HAR 数据集上的少样本识别准确率，包括 UCF-50 [13] 、UCF-101 [15] 、UCF-Crime [17] 和 HMDB51 [7] 。我们的骨干模型是 Qwen2.5-VL-3B-Instruct，我们在不同的学习范式下对其进行微调。如表 II 和表 III 所示，我们比较了三种设置：基线、监督微调 (SFT) 和我们提出的 FAVOR 方法。

在各种少样本设置中，FAVOR 始终优于 SFT 和基线。在低样本状态下，如 1-shot 时，FAVOR 相比基线平均准确率提高了 14.78 % ，而 SFT 稍逊一筹（-0.04 %）。这种增益在 UCF-101（+29.62 %）和 UCF-50（+13.01 %）上尤为显著。在 2-shot 和 4-shot 设置下，这一趋势仍在继续，FAVOR 分别带来了平均增益 15.32 % 和 18.75 % 。在 UCF-101（最高 +34.43 %）和 HMDB51（最高 +17.04 %）上的性能提升最为显著，显示了模型从有限例子中良好泛化的能力。

在高次尝试（higher-shot）场景中，FAVOR 的优势变得更为明显。在 8 次尝试设置下，FAVOR 实现了 61.22 % 的平均准确率，优于基线模型（41.00 % ）和 SFT（53.06 % ）。它在所有数据集上都表现出持续的改进，包括在具有挑战性的 UCF-Crime 数据集上比 SFT 高出显著的 +11.05 % ，而 SFT 在该数据集上则表现出轻微的性能下降（-1.33 % ）。在 16 次尝试设置下，FAVOR 保持领先，平均准确率达到 63.55 % ，相比基线模型提高了 +22.55 % ，相较于 SFT 提高了 +4.92 % 。同样，最高的增益出现在 UCF-101（+41.18 % ）和 HMDB51（+20.64 % ）数据集上。

这些结果明确指出，FAVOR 显著提升了 HAR 任务中的少样本识别性能。它不仅在低数据环境中优于传统的 SFT，而且还表现出更好的可扩展性和鲁棒性。重要的是，在某些情况下，SFT 对基线的改进较少甚至为负，而 FAVOR 则始终提供一致且通常是较大的提升——这突显了其作为数据稀缺环境中现实世界活动识别的可推广解决方案的潜力。

为研究我们方法中不同组件对性能的贡献，我们在 UCF-50 的 2-shot 数据集上进行了一个消融研究，如表 IV 所示。从基础的 Qwen2.5-VL-3B-Instruct 模型开始，该模型达到了 66.16 % 的准确率，我们检查了几项关键设计选择的影响。特别是，使用冻结视觉模块的 FAVOR 导致显著提升，准确率达到 79.69 %。这表明保持视觉编码器不变使得模型能够更好地将其能力分配用于在强化学习中学习有效的推理策略。此外，我们探索了训练期间采样响应数量（P ）变化的效果。随着 P 从 2 增加到 16，模型性能持续提升——从 $P = 2$ 的 75.34 % 提高到 $P = 16$ 的 80.14 %。这突出了在优化过程中更丰富的轨迹探索的好处，使模型能够更好地通过多样的推理路径来完善其策略。总体而言，结果显示即使没有微调视觉基础网络，通过适当配置强化学习设置也可以实现显著的增益。这强调了在多模态学习中基于奖励的推理探索的重要性，并指出视觉和语言模块之间复杂的动态关系需要进一步研究。

V. 结论

本论文中，我们提出了一种名为 FAVOR 的方法，用于调整基于 GRPO 的强化学习，以提高 MLLMs 的模型推理能力。FAVOR 采用一种可验证的奖励系统，减少了对人工标注的依赖，并简化了奖励计算。在各种 HAR 任务中表现出色。大量实验表明，FAVOR 在推理和小样本学习方面表现优异，持续超越使用最少数据的监督微调（SFT）。我们的研究结果强调了视觉强化学习在提高 MLLMs 效率和效果方面的潜力，以优化 HAR 任务。

VI.

致谢 此项工作在博士期间得到日本🄵原基金会第五次研究资助和中国国家留学基金委的部分支持。

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

[2] Marwa R. M. Bastwesy, Hyuckjin Choi, and Yutaka Arakawa. Tracking on-desk gestures based on wi-fi csi on low-cost micro-controller. In International Conference on Mobile Computing and Ubiquitous Network, pages 1–6, 2023.

TABLE II
HAR 数据集上的少样本识别精度（1-shot, 2-shot 和 4-shot）。

Dataset	Baseline	1-shot				2-shot				4-shot			
		+SFT	Δ SFT	+FAVOR	Δ FAVOR	+SFT	Δ SFT	+FAVOR	Δ FAVOR	+SFT	Δ SFT	+FAVOR	Δ F
Average	41.00	40.96	-0.04	55.78	+14.78	42.51	+1.51	56.32	+15.32	46.47	+5.47	59.75	+1
UCF-50 [13]	66.16	65.96	-0.20	79.17	+13.01	67.06	+0.90	80.14	+13.98	66.78	+0.62	79.65	+1
UCF-101 [15]	38.41	38.31	-0.10	68.03	+29.62	42.94	+4.53	72.02	+33.61	53.65	+15.24	72.84	+3
UCF-Crime [17]	25.77	25.56	-0.21	30.32	+4.55	25.82	+0.05	26.75	+0.98	26.80	+1.03	35.81	+1
HMDB51 [7]	33.67	34.02	+0.35	45.59	+11.92	34.20	+0.53	46.38	+12.71	38.66	+4.99	50.71	+1

TABLE III
HAR 数据集上的少样本识别准确率。

Models	Avg.	UCF-50 [13]	UCF-101 [15]	UCF-Crime [17]	HMDB51 [7]
Baseline	41.00	66.16	38.41	25.77	33.67
8-shot					
+ SFT	53.06	78.54	67.87	24.44	41.38
Δ SFT	+12.06	+12.38	+29.46	-1.33	+7.71
+ FAVOR	61.22	80.94	75.27	36.82	51.83
Δ FAVOR	+20.22	+14.78	+36.86	+11.05	+18.16
16-shot					
+ SFT	58.63	82.93	75.79	27.84	47.96
Δ SFT	+17.63	+16.77	+37.38	+2.07	+14.29
+ FAVOR	63.55	83.44	79.59	36.86	54.31
Δ FAVOR	+22.55	+17.28	+41.18	+11.09	+20.64

TABLE IV
UCF-50 2-shot 数据集上的消融结果。

Ablation Setup	Accuracy (%)
Base model: Qwen2.5-VL-3B-Instruct	66.16
FAVOR with frozen vision modules	79.69
FAVOR with responses num $P = 2$	75.34
FAVOR with responses num $P = 4$	78.32
FAVOR with responses num $P = 8$	79.93
FAVOR: trainable vision modules, responses num $P = 16$	80.14

[3] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. arXiv preprint arXiv:2503.17352, 2025.

[4] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

[5] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. arXiv preprint arXiv:2503.24290, 2025.

[6] Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, et al. Efficient multimodal large language models: A survey. arXiv preprint arXiv:2405.10739, 2024.

[7] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: A large video database for human motion recognition. In International Conference on Computer Vision, pages 2556–2563, 2011.

[8] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. arXiv preprint arXiv:2502.21321, 2025.

[9] Ming Li, Keyu Chen, Ziqian Bi, Ming Liu, Benji Peng, Qian Niu, Junyu Liu, Jinlang Wang, Sen Zhang, Xuanhe Pan, et al. Surveying the mllm landscape: A meta-review of current surveys. arXiv preprint arXiv:2409.18991, 2024.

[10] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785, 2025.

[11] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435, 2023.

[12] Wen Qi, Xiangmin Xu, Kun Qian, Björn W Schuller, Giancarlo Fortino, and Andrea Aliverti. A review of aiot-based human activity recognition: From application to technique. IEEE Journal of Biomedical and Health Informatics, 29(4):2425–2438, 2024.

[13] Kishore K Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. Machine Vision and Applications, 24(5):971–981, 2013.

[14] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615, 2025.

[15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[16] K. Suda, S. Ishida, and H. Inamura. User estimation with touch panel buttons toward in-home activity recognition. In 2023 Fourteenth International Conference on Mobile Computing and Ubiquitous Network, pages 1–6, Kyoto, Japan, 2023.

[17] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6479–6488, 2018.

[18] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3):3200–3225, 2022.

[19] Yafeng Yin, Lei Xie, Zhiwei Jiang, Fu Xiao, Jiannong Cao, and Sanglu Lu. A systematic review of human activity recognition based on mobile devices: overview, progress and trends. IEEE Communications Surveys & Tutorials, 26(2):890–929, 2024.

[20] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. arXiv preprint arXiv:2503.05379, 2025.

[21] Xiaoqiang Zhou, Chaoyou Fu, Huaibo Huang, and Ran He. Dynamic graph memory bank for video inpainting. IEEE Transactions on Circuits and Systems for Video Technology, 34(11):10831–10844, 2024.

[22] Xiaoqiang Zhou, Huaibo Huang, Zilei Wang, and Ran He. Ristra: Recursive image super-resolution transformer with relativistic assessment. IEEE Transactions on Multimedia, 26:6475–6487, 2024.