通过逻辑推理增强视觉语言模型的态势感知能力

Pavana Pradeep Kumar[1],Krishna Kant[1] 和 Suya You[2]

arXiv:2601.11322v1 [cs.CV] 16 Jan 2026

摘要—**视觉语言模型（VLM）能够根据图像和视频生成复杂活动的高层次、可解释描述，因此在态势感知（SA）应用中具有重要价值。在此类场景中，重点是以高可靠性和准确率识别低频但意义重大的事件，同时提取细粒度细节并评估识别质量。本文提出了一种通过显式逻辑推理将视觉语言模型与传统计算机视觉方法相结合的方法，从三个关键方面增强态势感知能力：（a）提取细粒度的事件细节；（b）采用智能微调（FT）策略，相较于无信息依据的选择实现显著更高的准确率；以及（c）在推理过程中为视觉语言模型的输出生成合理性论证。我们证明，智能微调机制能够提高准确率，并在推理过程中提供一种有价值的方式，用于确认视觉语言模型输出的有效性，或说明其可能存在疑问的原因。**

影响声明——**本文将新兴的视觉语言模型（VLM）应用于信息物理环境中的基于视觉的态势感知，重点关注安全、安保和政策合规。视觉语言模型能够对罕见但关键的事件进行高层次描述，而传统计算机视觉（TCV）方法则更善于捕捉人员、物体、地点和运动等更细微的细节。我们探索通过逻辑推理整合视觉语言模型与传统计算机视觉，以增强态势感知、提高微调（FT）效率并确保输出可靠性。微调视觉语言模型的成本可能很高，尤其是针对罕见事件。我们提出一种技术，不仅能提高微调效率，还能在推理过程中提供一种对输出进行合理性检查的方法。后者可提高关键事件识别的可靠性，这对于态势感知至关重要。**

索引词——**多模态语言模型、逻辑推理、异常活动检测、基于视频的监控、态势感知**

一、引言

许多信息物理系统会定期对场所进行基于视频的监控，并以不同程度的自动化检测需要人类进一步关注的事件。理想情况下，应将这一过程完全自动化，并描述可能值得关注的情况。例如，在道路交通监控中，可能需要检测事故、险情、利用车辆实施的犯罪活动等。这些情况比正常交通活动少见，但却是监控的首要关注对象。尽管值得关注的目标 事件/活动通常具有负面性质（例如，安全、安保、违反政策），但也可能是任意类型的（例如，监控谁进入或离开建筑物或房间）。本文始终假设目标活动/事件有明确描述，以便开展适当的监控。下文将这一组事件称为目标任务。

以大语言模型（LLMs）为基础构建的视觉语言模型（VLMs），在识别和描述复杂场景方面展现出令人瞩目的能力。然而，要获得较高的准确率，视觉语言模型（以及大语言模型）通常需要针对特定任务进行调整，方式包括指令微调（即向模型展示输入和输出样例）或微调（修改神经网络权重）[1]。本文重点关注微调（FT），这通常需要大量资源。因此，对于视觉语言模型表现不佳的情形，智能地选择输入至关重要。在态势感知（SA）应用中，许多目标活动可能相对少见（例如，事故与正常交通），因此，有针对性地选择可能包含所需活动的视频片段就显得尤为重要。如果能在标注所选片段之前完成筛选，我们不仅可以减少处理工作量，也能降低标注成本。

将视觉语言模型用于态势感知的关键吸引力在于，它们能够提供语义丰富的场景描述，其能力远远超出使用传统计算机视觉（TCV）技术通常可以合理实现的范围。在此，TCV是指专门为识别物体、物体属性、特定动作等目的而设计和训练的深度学习网络，例如卷积神经网络或Transformer。[2]。TCV方法仍然十分重要，至少可以通过以下三种方式加以利用：（a）提取视觉语言模型通常难以处理的情境细粒度定量细节；（b）智能选择视频片段，以便高效地针对目标任务对视觉语言模型进行微调；（c）帮助为识别出的活动提供合理性论证（当认为识别结果可靠时），或说明输出可能不可靠的原因。所有这些用途都需要以通用方式整合视觉语言模型和TCV的输出，为此，我们利用高效开展的逻辑推理来满足实时情境感知目标。

据我们所知，这是首项将显式逻辑推理与视觉语言模型和 TCV 相结合的工作，并在基于视觉语言模型的态势感知领域做出了三项重要贡献：

1) 一种为视觉语言模型的输出补充细粒度细节的机制。

[1]计算机与信息科学系，天普大学，美国宾夕法尼亚州费城。（电子邮件：pavana.pradeep@temple.edu，kkant@temple.edu）[2]ARL，美国加利福尼亚州圣何塞。（电子邮件：Suya.you.civ@army.mil）本研究由ARL资助，资助编号为272411

978-1-5386-5541-2/18/$31.00 ©2018 IEEE

2) 一种智能（或有向）微调机制，其准确率远高于无信息依据或随机选择（下文称为无向选择）。即使微调周期的时长保持不变，也能实现这一点；3) 一种在推理阶段为视觉语言模型的输出提供合理性论证的机制（包括指出输出可能不可靠的情况）。

本文其余部分的组织如下。第 II 节介绍我们定向微调机制的详细设计。第 IV 节讨论定向微调与无向微调之间的实验评估。第 II-D 节讨论相关工作。最后，第 VI 节总结全文。

II. 提出的框架

A. VLM/TCV 集成的动机

态势感知（SA）的关键在于快速检测和跟踪异常事件，以及其他对确保受监测基础设施安全运行至关重要的各类事件。视觉语言模型在提供对情境的准确高层描述方面表现出色，但其基于大语言模型的设计并不适用于精细的定量信息，例如精确的地点、距离、朝向、运动等。这些细节不仅通常是 SA应用所必需的，也可以通过其他方式加以利用，包括视觉语言模型微调和输出论证，本文对此进行了详细说明。此外，只要我们将范围限制在物体及其姿态/属性、简单动作和定量信息（例如地点、速度、数量等）等简单方面，这些内容都可以通过成熟的预训练 TCV 算法与简单的视觉和推理操作相结合来轻松处理（例如透视校正、地点/距离估计、运动的帧间跟踪等）。尽管可能需要进行一些额外的数据标注、训练，或对现有 TCV 算法进行调整，但所需工作量与调整视觉语言模型流程及其微调处于相近水平。然而，为了同时捕捉情境的高层特征和细粒度特征，我们需要将视觉语言模型和 TCV 的输出进行匹配并整合起来，为此我们采用逻辑推理——尤其是理论模可满足性（SMT）。事实上，我们还使用辅助视觉语言模型，以进一步协助实现无需标注工作量的高效微调，并提供更深入的论证，后续章节将对此进行介绍。

B. 任务与活动

我们从一个预定义的 目标任务开始，以下称为主 任务， $\mathcal{A}^m$，它由一组我们希望识别和跟踪的“活动”（或情境）构成。[1] 我们还定义了一个辅助任务 $\mathcal{A}^a$，其中包含$\mathcal{A}^m$中每项活动的一个变体。因此，我们有两个视觉语言模型：一个主视觉语言模型 （VLM^m）和一个辅助视觉语言模型（VLM^a），用于识别这些活动。我们还定义了一个代理任务 $\mathcal{A}^p$，其中包含一组简单的 活动

[1]虽然我们只考虑预定义的活动，但其中一些实际上可能是通过过去的视觉语言模型描述自动学习得到的。

这些活动将对 $\mathcal{A}^m$ 和 $\mathcal{A}^a$提供较为宽泛的描述，具体如下文所述。辅助活动主要是为了能够通过标准的预训练 TCV 算法识别，不过在某些情况下，可能还需要调整算法和/或增加标注训练数据。有了这些设定，我们的微调和论证方法便围绕这三个任务输出之间的一致性展开。请注意，虽然给定了 $\mathcal{A}^m$，但 $\mathcal{A}^a$ 和 $\mathcal{A}^p$ 在很大程度上可以自动构建。表 I 列出了这 3 个任务及其对应活动的符号表示。这里， A_i^a，$i=1..I$ 的定义确保了 A_i^m与 A_i^a之间存在一一对应关系，因此 i可被视为用于识别的“类别”。然而，代理活动的解释方式有所不同，如图 1 所示。图中展示了任务$\mathcal{A}^m$ 的每项活动 A_i^m都与代理活动的某个子集 $\mathbb{S}_i^m$（A^p）相关联。例如， A_2^m关联了三个代理活动 $A_1^p..A_3^p$。请注意，辅助任务也适用类似的表示方式，只是其子集 $\mathbb{S}_i^a$（A^p）有所不同。

任务类型	**定义方式**
主任务	主活动集（$\mathcal{A}^m=\{A_i^m,i=1..I\}$）
辅助任务	辅助活动集 ($\mathcal{A}^a=\{A_i^a,i=1..I\}$)
代理任务	代理活动集 ($\mathcal{A}^p=\{A_k^p,k=1..K\}$)

表 I：任务与活动

现在，我们提出要求：$A_i^m \Longrightarrow \mathbb{S}_i^m$（$A^p$）。此处，$\Longrightarrow$ 表示如果观察到活动 A_i^m ，那么集合 $\mathbb{S}_i^m$ (A^p) 中的所有（更简单的）活动也应当被观察到。

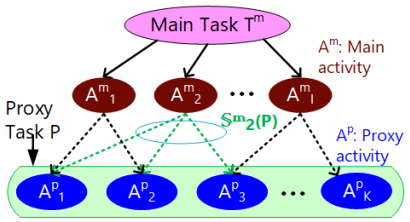


图 1：任务与活动关系

也就是说，集合 $\mathbb{S}_i^m(A^p)$ 是活动 A_i^m发生的必要条件，但可能不是充分条件。此处的说明较为非正式；在实际实现中，这类要求及其底层活动会表示为逻辑断言，后文将对此进行讨论。

我们的意图是使用 TCV 识别 $\mathbb{S}_i^m(A^p)$，并利用单向蕴含，将其作为一种虽较弱但高效的一致性检查，与视觉语言模型的输出进行核对。对于 TCV，我们使用 YOLO 变体等标准算法检测物体/姿态；如有需要，还可通过为 YOLO 系列目标检测器增加颈部/头部层，来识别其他属性 [3]。运动跟踪可通过逐帧监测物体、进行透视变换，以及推理物体的位置/运动来实现。特别是，当人/机器暂时从视野中消失后又重新出现时，在大多数情况下，我们都可以轻松且准确地对其进行跟踪。为简单起见，我们也对辅助视觉语言模型使用同一组代理活动。也就是说，$A_j^a \Longrightarrow \mathbb{S}_j^a(A^p)$，其中 $\mathbb{S}_j^a(A^p)$ 是 A^p 的某个非空子集。

由于我们主要将 TCV 用于简单且经过充分研究的任务，因此我们预期 TCV 准确率相当高；不过，如果 TCV、视觉语言模型或两者都出错，它们仍然保持一致的可能性极低。因此，

这些错误只意味着可能需要进一步微调，而这或许并无必要。还需注意，在连续视频监控环境中，我们可以轻松利用时间一致性来即时纠正暂时性的错误。例如，如果卡车在连续几帧中被识别为汽车或完全未被检测到，而后又在后续帧中重新出现，那么纠正起来就很简单，也很容易通过我们的逻辑推理机制实现。不过，这确实需要保留最近 K 帧的识别结果缓冲区，其中 K 的取值需要在输出延迟与时间纠正能力之间权衡。

C. 一致性驱动的微调

根据上述设置，我们可以通过 VLM^m、VLM^a 和 TCV 处理视频帧，并通过时空逻辑推理确定以下事项：(A) VLM^m 的输出类别（例如 i）是否与 VLM^a 的输出类别（例如 i'）相同；(B) 集合 $\mathbb{S}_i^m(A^p)$中的所有活动是否都被观察到；以及 (C) 集合 $\mathbb{S}_{i'}^a(A^p)$中的所有活动是否也都被观察到。如果其中任何一项一致性条件 不成立，我们就会智能地或**有向地** 选择“输入”（或视频片段），以便进一步微调 VLM^m 和 VLM^a。

这一过程可以循环重复，直到微调性能趋于饱和，或因其他原因（如微调数据 (FTD) 用尽）而终止。此后，我们将用于评估微调需求的数据称为评估数据 (ED)。请注意，条件 (A)–(C) 不需要任何带标签数据。也就是说，无需先对数据进行标注，就能确定是否需要进一步微调。一旦确定需要进一步微调，微调本身仍像往常一样需要带标签数据。这有助于更有针对性地开展数据标注工作。当然，如果 ED 已有标签，我们只需检查预测类别 i 是否与标签相同。下文将展示一个有趣的结果：纯粹一致性驱动的 评估方法所获得的微调结果质量，几乎与准确率驱动的方法（即使用带标签 ED 的方法）相同。

D. 相关工作

据我们所知，现有文献中尚无研究利用直接逻辑推理和/或 TCV 来提升微调效果及可辩护性；不过，针对这些单独主题的研究非常丰富。

我们的工作并不关注微调的具体实现方式，因此无论采用何种底层方法均适用。例如，参数高效微调（PEFT）旨在尽量减少被修改的参数（权重） [4]。典型的微调方法包括在视觉骨干网络上添加分类器 [5], 、添加特征适配器 [6], , 或利用提示结合视觉语言模型的思维链（CoT）推理来调整分类概率。

本文采用的逻辑推理，是通过显式的推理规则 (RoIs) 扩展一阶演绎推理。这与近期其他利用视觉语言模型/大语言模型输出进行推理的概念不同 [7]–[9]。这类推理会微调视觉语言模型/大语言模型，使其直接提出问题并生成答案，也可能结合通过网络搜索获取的外部信息。尽管大语言模型的推理能力持续提升，但大语言模型推理这一概念本身也受到质疑 [7]。

在 TCV 领域以及更广泛的深度学习领域中，推理问题通常被称为神经符号人工智能 [10]–[12]。这种人工智能尝试通过神经网络的损失函数间接施加约束。例如，广受关注的逻辑张量网络（LTN） [13]通过使用布尔运算的可微扩展 [14], 隐式且近似地施加逻辑约束，从而避免梯度爆炸或消失的问题。显式逻辑推理方法受到的探索相对较少 [15], [16]。参考文献 [17], [18]尝试运用显式推理来表征事故和驾驶员行为。

可解释人工智能领域的相关文献日益增多 [19]。其中许多研究都关注解释人工智能的决策，但如今研究重点已扩展至一个更重要的问题：这些决策是否具有可辩护性 [20]。我们的方法不仅以简单的方式支持可辩护性，还能在输出可能不可靠时发出提示。

E. 为推理提供合理性论证

在许多 SA应用中，监测目标活动的可信度至关重要。因此，提供结果的合理性论证（当结果被认为可靠时），或指出结果不可靠，都是必不可少的。只需在一致性驱动的 微调机制中，于推理阶段继续进行一致性检查，即可使其适用于这一用途。在这种情况下，如果有可用资源，我们可以在两块不同的图形处理器上并行运行两个视觉语言模型；也可以在推理过程中舍弃 VLM^a ，从而降低验证力度。第三种选择是仅在“困难场景”中使用 VLM^a 。我们可以根据以往观察到的不一致情况来学习识别此类场景，但本文不探讨这一方面。之后我们将量化合理性论证所需的开销；随着硬件速度提升，有望实现实时处理。

III. 实施定向微调

A. 识别代理活动和辅助活动

由于感兴趣的主要活动可能是任意的，最好假定这些活动是预先给定的（尽管其中一些可以通过学习获得）。不过，自动生成相应的代理活动和辅助活动会很方便。代理活动可视为将主要活动分解为其由哪些更简单的低层活动构成；而辅助活动则可视为对主要活动或非常相似活动的另一种描述。由于这些识别工作只需在设计阶段进行一次，最简单的方法是请能力出众的多模态大语言模型完成层次分解和相似活动识别。还可以利用大语言模型剔除有缺陷的描述。例如，Video-chatGPT [21] 不仅可以用于生成

另一种描述，还可以用于检查主要描述与辅助描述之间的语义对齐。Video-chatGPT 会在 4 个不同方面给出分数（采用 0-5 分制）：正确性、一致性、朝向、情境理解以及对时间变化的理解。

在大多数情况下，层次分解最终涉及场景中的主要物体及其属性（姿态、属性、运动、与另一物体的时空关系等）。例如，驾驶车辆这一相当复杂的活动，可以用人、车辆及二者的相对朝向／位置／运动来描述。如前所述，这种描述是驾驶活动的必要条件，但并非充分条件。

接下来，我们自动确定视觉语言模型识别出的活动 A_i所隐含的代理活动集合，并用映射 $\mathbb{S}_i(A^p)$表示。（此处省略了上标 m或 a，因为两种情况适用相同的流程。）首先，我们将每个代理活动 A^p 表示为一个逻辑断言，例如 ψ_{A^p}。令 D 表示可用于微调的整个输入数据集，它是由带标签的短视频片段构成的集合。然后，对于每个类别 i，我们将每个输入送入TCV，并识别 ψ_{A^p} 中的哪些断言对该输入成立。最后，对于每个类别 i，我们选取在大多数情况（例如90%）下成立的断言。对于每个 i，这一集合便刻画了集合 $\mathbb{S}_i$ 。在现实场景中，对断言进行模糊处理（即采用90%的阈值）是必要的。另一种允许模糊性的方式（本文未采用）是为每个断言附加一个强度度量，用以表示对该断言的置信程度，并可通过适当的相似度度量进行量化。

B. 使用显式逻辑推理

显式逻辑推理已成功应用于众多领域 [22] 和场景 [17], [18], [23], [24]。最基本的用法是采用扩展了一些附加“理论”的一阶逻辑，以便对整数/实数算术、线性/角运动定律、空间关系等相关主题进行推理。可以使用基于模理论可满足性（SMT）的工具进行推理，例如广受欢迎的 Z3 [25] 和 YICES [26]，，这些工具通常能够解决规模非常庞大的实际问题。SMT 模型包含三个关键部分：(a) 推理规则（RoIs）；(b) 各种现实世界约束（例如，步行速度为 <2 m/s）；以及 (c) 根据环境确定的空间定位或事实（例如，通过分析各个帧中的物体并进行透视校正）。我们还可以将更高层次的概念定义为适用于“逻辑程序”的可复用函数。例如，考虑函数“following(V1, V2)”，它断言车辆 V1 正在跟随车辆 V2。如有需要，该函数可以分解为一系列帧中 V1 与 V2 之间的相对距离。它也可以定义为一个未解释函数，其真值通过 TCV，甚至 VLM 来确定。需要注意的是，本文讨论的显式逻辑推理不同于相关工作（第 II-D 节）中讨论的其他推理形式，既不涉及损失函数，也不涉及任何训练。

虽然对于本文而言，简单的基于 SMT 的推理已足够，但很容易将其扩展以涵盖时序推理（例如，某种情境持续成立，直到某个事件使其终止），或超时等实时关系。这种适应性通过若干时序扩展示例得以展现，例如 [27]–[29]，如 [17], [18] 中所用。具体而言，推理可以同时包含“硬”子句（即必须成立）和“软”子句（根据权重设定、最好成立的子句），如 [17] 所探讨。

C. 微调算法

所有部分都准备就绪后，我们按照图 2 所示描述整体微调算法。为便于理解，这里的流程图和说明做了一定简化，实际实现时也可以采用不同方案。紫色方框表示输入/输出数据，蓝色/绿色方框表示操作。我们从两个数据集开始：微调数据（FTD）和评估数据（ED）。如前所述，ED 无需标注，但在我们的案例中对其进行了标注，以便研究上文介绍的基于一致性和基于准确率的推理。理想情况下，ED 应为每个视觉语言模型^m 类别至少包含一个批次，而每个 FTD 批次则聚焦于在一致性检查中未通过的特定视觉语言模型^m 类别。请注意，*ED* 在微调过程中使用，因此不同于最终用于评估性能的测试数据集。

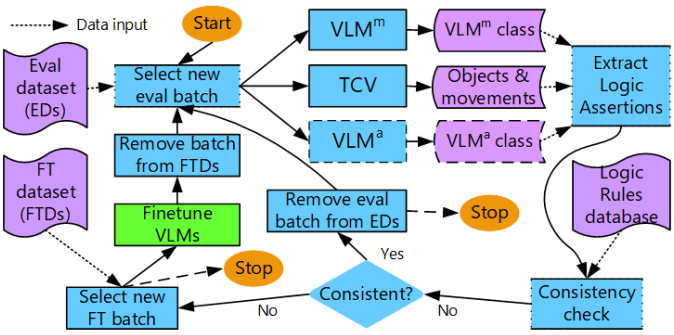


图 2：视觉语言模型的应用有向微调 Ms

图 2 所示微调循环的“开始”阶段，在每次评估时都会使用一个来自 ED 的输入批次，如图 2 中的 评估批次 所示。该批次会依次通过视觉语言模型^m、视觉语言模型^a（若使用）以及 TCV 推理步骤，并收集结果。输出为检测类别和代理活动的逻辑表示中相关断言提供“空间定位”，并借助预先构建的逻辑规则数据库进行一致性检查。接下来，如图所示，我们使用可满足性模理论进行一致性检查。如果某个评估批次未发现不一致，我们就将其从 ED 中移除，因为它对发现不一致没有帮助。如果发现不一致，可满足性模理论框架会给出导致问题的断言（即未通过检查的断言子集），据此可以确定哪些视觉语言模型类别需要进一步微调。因此，下一步是从 FTD 中选择一个批次，如图 2 中的 微调批次所示。

举例来说，考虑 TU DAT 数据集中的图 3 所示视频帧,– 其中呈现了一起追尾事故。TCV（YoLovx）给出所有汽车的边界框，我们可以为它们分配一些伪标识符，例如汽车1、汽车2等。此处识别出的 VLM^m 类别为“1”，如表 III 左栏所定义的主要任务活动所示。

VLM^a 类别也将为“1”（见表 III 右栏）。代理活动集 $\mathbb{S}_1^m(A^p)$ 包含表 VI 中以抽象方式描述的活动（ #1 、 #3 ）。这些活动可根据所示帧以及视频中此前的若干帧进行空间定位。因此，对 #1（“同一车道上，一辆汽车在另一辆汽车后方”）进行空间定位，需要确定两辆汽车的伪标识符以及该陈述的真值。完成空间定位后，我们会得到类似这样的陈述：“汽车 2在同一车道上跟在汽车1后方行驶”。这种空间定位将在逻辑域中进行（而非自然语言），并将相应断言标记为真。因此，可以使用 SMT 轻松检查各种一致性条件。

为避免图示过于繁杂，图 2 仅在两处标出转至“停止”的状态，以表示流程终止；当 FTD 或 ED 数据耗尽、超出时间限制或准确率几乎不再提升时，流程便会终止。我们先使用少量已标注输入进行初步微调，之后再进行定向微调。微调时，我们会将较长的视频切分成较短的视频，尽可能使每段视频只聚焦于一种交互类别。[2] 我们根据所用特定视觉语言模型的要求，为这些视频片段添加标签（或说明文字）。对于基于图像的视觉语言模型，我们会为视频片段中的每一帧添加相同的标签。

IV. 实验设置详情

A. 用于评估的视觉语言模型

虽然我们主要关注基于视频的视觉语言模型，但也在基于图像的视觉语言模型上评估了我们的微调方法 [30]。我们还同时考察了由大语言模型支持的视觉语言模型，以及仅进行图文匹配的视觉语言模型。这些模型采用不同的策略来对齐视觉与语言模型。通过分析这些多样化的模型，我们旨在得出具有较强普适性的结论。

MiniGPT-4 [31] 是一种基于图像的视觉语言模型，它定义了两个可训练层，用于将冻结的视觉 Transformer 模型与冻结的大语言模型对齐。MiniGPT4-Video [32] 是 MiniGPT-4 的扩展，可处理视频和文本对话。X-CLIP [33] 是一种流行的、基于视频文本匹配的视觉语言模型，用于视频文本检索，并生成多粒度的视觉和文本表征。Video-LLaMA [34] 是一种先进的多模态框架，使大语言模型能够理解音频和视频信号。还有许多其他视觉语言模型，并且仍在不断涌现，但我们认为所选模型集很好地代表了这一领域的广泛范围。

近年来，选择性状态空间模型（SSM） [35]已成为基于 Transformer 的视觉语言模型的一种颇具吸引力的替代方案。例如，VideoMamba [36] 是一种专为视频理解设计的纯 SSM 模型，融合了卷积和注意力的优势。它采用线性复杂度方法，从实时情境理解的角度来看颇具吸引力。

[2]即使视频片段非常短，其中也可能包含不止一种活动。对于经常同时出现的状况，可以通过定义额外的复合类别来识别此类情况。预计此类组合的数量会很少。

B. 所用数据集的描述

我们使用三个截然不同的数据集评估了 FT 方法。前两个数据集涉及的活动有限，而第三个则更加开放。我们的第一个数据集名为 TU DAT [37]，聚焦道路交通，包含在驾驶条件和天气条件都颇具挑战性的情况下采集的视频。图 3（将在评估部分中展示）呈现了该数据集中的一个场景，其中发生了一起追尾事故。

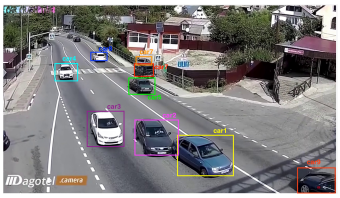


图 3：查找场景中的关键对象



图 4：跆拳道中的绿带动作



(a)



(b)

图 5：Kinetics 数据集中的部分活动

我们的第二个数据集——跆拳道数据集，记录了跆拳道运动员所做的动作。正如 Taekwondo America 学生手册 [38]所述，理解动作模式是跆拳道训练的重要组成部分。该数据集包含 35 个视频，每个视频中有一名或多名学生依次演示各腰带套路中的动作。图 4(a) 展示了一名身穿深绿色腰带的学生所做的行走姿势和下段格挡；图 4(b) 展示了该学生所做的行走姿势和反击拳。

表 II：数据集详情

数据集 type	FT 数据	评估 数据	测试 数据
TU DAT	200	75	45
跆拳道	105	65	30
Kinetics	2000	800	350

我们的第三个数据集是 Kinetics-100 [39]，，这是一个专注于人体动作的大型视频数据集。动作类别列表包括单人动作（例如绘画、饮酒）、人与人互动动作（例如拥抱、握手）以及人与物体互动动作（例如拆礼物、割草、洗碗）。Kinetics-100 包含 100 个人体动作类别，每种动作有 400–1150 个视频片段，每个片段时长约为 10 秒。图 5 展示了 Kinetics-100 数据集中的一些活动，例如 (a) 修剪草坪和 (b) 洗碗。

我们采用多种数据增强方法，为 TU DAT 数据集和跆拳道数据集获取足够的数据量。具体而言，我们使用 Keras 提供的多种实时图像增强方法。所采用的数据增强方法包括翻转、平移、错切和旋转。表 II 展示了全部三个数据集的详细信息。

No.	视觉语言模型中的类别^m	视觉语言模型中的类别^a
1	汽车被另一辆车从后方撞击	汽车同向前后行驶
2	汽车被另一辆车从侧面撞击	汽车与另一辆车反向且垂直行驶
3	汽车被另一辆车从前方撞击	汽车反向行驶
4	汽车撞上静止物体	汽车紧贴静止物体行驶
5	摩托车撞上行人	摩托车紧贴步行行人行驶
6	交通视频	同向或反向行驶的汽车在其他汽车旁边、横穿其前方或位于其后方
7	未定义	汽车和摩托车彼此并排行驶
8	未定义	行人步行

分数范围（低 - 高） **0-5**

平均性能	分数
正确性分数	4.09
一致性分数	4.02
详细方向分数	3.91
情境理解得分	3.97
时序理解得分	4.01

表 III：(a) TU DAT 的类别描述，(b) VLM^m 和 VLM^a 类别的相似度得分

表 IV：跆拳道数据集使用的类别

No.	VLM^{m}中的类别	VLM^{a}中的类别
1	左腿静止，右腿静止	左臂和右臂伸出
2	左腿静止，右腿向前	左臂伸出，右臂弯曲
3	左腿静止，右腿向后	左臂弯曲，右臂伸出
4	右腿静止，左腿向前	左臂和右臂交叉抱于胸前
5	右腿静止，左腿向后，左臂放在头上，右臂折叠的	
6	左腿向前，右腿向后-向后	右臂放在头上，左臂交叉折叠
7	右腿向前，左腿向后—方	未定义

表 V：Kinetics 数据集所用类别的描述

编号。	视觉语言模型^m 类别	视觉语言模型^a 类别
1	Arm 摔跤	用一只手抓住对手的手，并施加压力，双腿稳住身体，并调整借力姿势
2	烘焙 Cookie（小型文本文件）	混合食材、擀开面团、塑形饼干用手操作，借助腿部够到烤箱并将其打开以便将烤盘放入其中
3	刷洗牙齿	用一只手拿着牙刷，挤上牙膏另一只手拿着牙膏，并以画圈的方式移动牙刷来清洁牙齿和牙龈
4	Cart 滚动	双手推离地面并支撑身体同时双腿向上踢起并在空中旋转，以完成侧手翻
5	欢呼领先	双手做出拍手、挥手等动作，挥——.....或手持啦啦队彩球，同时双腿做跳跃动作、踢腿，和特技动作，以丰富啦啦队动作编排
6	移动 Lawn	用一只手启动割草机、调整转向和切割高度，并用双腿推动或引导割草机并绕开障碍物
7	洗碗	用双手刮洗餐具，并将肥皂涂抹在餐具上，再用水冲洗餐具

C. 用于微调视觉语言模型的类别

TU DAT 数据集包含道路交通中的多个事故场景，这些场景构成了用于微调视觉语言模型的类别。由于我们提出的方法需要微调两个视觉语言模型，因此 TU DAT 数据集中的视频被划分为两类：用于视觉语言模型^m 建模事故场景的视频，以及用于视觉语言模型^{a}识别车辆相对位置和运动的视频。表 III(a) 展示了在 TU DAT 数据集上用于微调视觉语言模型^m 和视觉语言模型^{a}的类别说明。表中将编号相同的类别并列展示，体现了二者之间的一一对应关系。正如前文所述，视觉语言模型^a 的描述是使用 ChatGPT 根据视觉语言模型^m 的描述生成的（为便于在表中展示，略作缩短）。

为了展示 VLM^m 和 VLM^a 的描述契合程度，我们获取了 Video-chatGPT 的相似度得分，并将其列于表 III(b)中。所有得分都接近 4.0（满分 5.0），显然二者在各个方面都高度一致。对于下文讨论的另外两个数据集，我们也发现了类似的匹配结果，但为简洁起见，此处不列出相似度得分。

对于跆拳道数据集，微调 VLM^m 以识别学生的腿部动作，同时微调 VLM^a 以识别学生的手臂动作。表 IV 展示了在跆拳道数据集上微调 VLM^m 和 VLM^a 时所用类别的说明。理解跆拳道规则的视觉语言模型能够自动生成与腿部动作相对应的手臂动作。

对于 Kinetics 数据集，我们从总共 100 项活动中精心选取了约 50 项作为“目标活动”。之所以选择这么多目标活动，是为了进行压力测试。如前所述，在现实中，人们通常关注的是某种程度上不寻常或异常的目标活动。由于篇幅所限，我们不对这 50 项活动逐一详述，而是在表 V 中介绍其中选定的几个类别。

D. 代理活动的识别

这里以 TU-DAT 为例说明代理活动的推导过程；由于篇幅所限，其他数据集不再赘述。我们首先列举一些简单活动；在道路交通场景中，这些活动体现为关键对象之间的空间关系。这些关系在表 VI(a) 中以非正式方式列出，且很容易表示为逻辑断言。请注意，我们以通用形式表达这些关系，因为我们希望能够检查任意一对汽车之间，或汽车与行人之间的关系。对象 ID 将通过处理实际视频获得，具体如下文所述。

表 VI：说明代理活动识别

(a) 可能的空间关系（非正式）	**(b) 空间定位 of 空间关系（非正式）**
同一车道内一辆汽车位于另一辆汽车后方 同一车道内一辆汽车正对着另一辆汽车 一辆汽车正在接近前方的汽车 一辆汽车正沿同一方向驶入相邻车道 一辆汽车正驶入相邻的反向车道 一名行人位于车道内。一辆汽车正在接近一名行人	汽车1和汽车2一前一后行驶；汽车1、汽车2、汽车3和汽车4朝同一方向行驶；汽车1紧跟汽车2，距离非常近；汽车1和汽车9在相反方向的车道上行驶；汽车7停放着，没有移动；汽车5、汽车8和汽车9朝同一方向行驶

表 VII：推理断言

变量： 汽车1、汽车2.....汽车9均为整数

函数： 布尔值，每个函数都接受一个整数参数：在后方行驶()、距离非常近()、沿相反方向行驶()、沿相同方向行驶()、汽车正在行驶()、汽车被后方撞击()

基元：
在后方行驶(汽车1,汽车2) ∧距离非常近(汽车1,汽车2) ∧沿相同方向行驶(汽车1,汽车2)∧ 沿相同方向行驶(汽车2, 汽车3) ∧沿相同方向行驶(汽车3, 汽车4) ∧ 沿相同方向行驶(汽车5, 汽车8) ∧沿相同方向行驶(汽车8, 汽车9) ∧沿相反方向行驶(汽车1,汽车9) ∧¬汽车正在行驶(汽车7)

代理断言示例：
汽车被后方撞击 (汽车1,汽车2) ⟹ 在后方行驶(汽车1,汽车2) ∧距离非常近(汽车1,汽车2)

我们使用当时最新版本的 YOLO 来识别输入图像中的关键对象，并在各帧之间跟踪这些物体（对于视频），以通过物体、伪对象 ID、位置、距离和相对运动对场景进行“空间定位”。例如，对于图 3 中的场景，识别出的（已定位）关系以非正式方式列于表 VI(b) 中。同样，在实际处理中，我们需要将这些关系转换为适当的逻辑断言，如表 VII 所示。该表展示了图 3 中的基本函数及其空间定位结果。

根据这些空间定位结果，我们可以确定表 VI(a) 中哪些断言成立，以及它们适用于哪些汽车。进而，我们可以据此识别出对于视频标签所指示的场景几乎总是成立的断言，从而得到所需的代理断言。例如，要判断一辆汽车是否从后方被撞，成立的断言应为：汽车一辆在另一辆后方行驶，并且彼此距离很近。表 VII 中也以单向蕴含的形式展示了这一点。

E. 微调细节

我们首先通过表 VIII 中的示例说明微调机制的运作方式。我们使用 TU DAT 数据集中的追尾事故场景，并采用 XCLIP 视觉语言模型。初始微调阶段结束后，视觉语言模型^m 和视觉语言模型^a 生成的字幕如表 VIII 前两行所示。两个输出并不一致，这一点可从视觉语言模型^m 和视觉语言模型^{a}之间缺少预期的类别对应关系得到确认。因此，我们按照图示标签筛选并检索视频，以对两个视觉语言模型进行额外微调。

表 VIII：微调中的一致性检查示例

视觉语言模型^m 输出： 匹配字幕：汽车被另一辆车从后方撞击**视觉语言模型^a 输出：** 匹配字幕：汽车反向行驶**一致性检查：** 否 - 视觉语言模型^m 的输出与视觉语言模型^{a}的输出不一致**检索带有以下标签的视频：** 汽车被另一辆车从后方撞击、汽车沿相同方向行驶且一辆在另一辆后方**微调视觉语言模型^m 和视觉语言模型^a**
选择新的评估批次： 用于确定视觉语言模型^{m}和视觉语言模型^{a}的输出是否一致**一致性检查：** 是 - 视觉语言模型^m 的输出与视觉语言模型^{a}的输出一致

对于我们的实验结果，我们将微调批次大小固定为 20 个视频。这在一定程度上是任意选择的，也可以调整数量。为进行公平比较，我们分别对有向情形和无向情形执行四次循环，每次使用 20 个视频。再次说明，此处的无向情形是指随机选择视频，不考虑微调薄弱的领域。我们在四次迭代后停止，因为一致性的提升似乎已陷入停滞。所有实验均在一台配备两块 48GB NVIDIA RTX A6000 图形处理器的工作站上完成。（推理时仅使用一块图形处理器。）我们的结果不假设采用并行推理（见第 III-C 节）。

V. 实验结果

我们从以下三个方面表征所提出的微调框架的性能：准确率、一致性和开销：

1) 准确率：分类结果与真实标签相符的测试案例所占比例。2) 一致性改进因子 (CIF)：定义为 $(n_b - n_e)/n_b$ ，其中 n_b 表示微调前记录的不一致数量， n_e 表示微调过程结束时的不一致数量。3) 开销：我们测量微调期间有向输入选择所增加的时间，以及推理期间合理性论证所花费的时间。理想情况下，这应以相对于基线时间的比率报告，但我们也会报告绝对时间。

为确保公平，对于目标活动数量较少的数据集（TU DAT 和跆拳道），我们对有向和无向情形采用相同的微调迭代次数。对于包含许多（实际上为 50 个）目标活动的 Kinetics 数据集，我们对有向和无向情形进行相同时间的微调。

以下结果表明，在所有情况下，有向方法的表现始终优于无向方法。此外，这种差异在基于图像的视觉语言模型（例如 MiniGPT4 [40] ）及其基于视频的版本 MiniGPT4-video [32] 中同样存在。我们还在基于 Transformer 的视觉语言模型上展示了我们的微调机制，这类模型以大语言模型作为后端，例如 Video-Llama [34], ；此外，也在较新的、基于状态空间模型（SSM）的非 Transformer 视觉语言模型上进行了展示，例如 Video-Mamba [36]。对于不以大语言模型为后端的模型，情况也相同，例如 XClip [33] 和 Video-MAE [41]。我们还表明，在两个截然不同的数据集上，这种提升均能保持：一个涉及道路交通，另一个涉及跆拳道课堂。

A. 实现的分类准确率

对于我们以一致性为驱动的微调方法而言，最重要的结果是其在测试数据集上达到的分类准确率。表 IX 展示了三个数据集以及两种视觉语言模型^m 和视觉语言模型^{a}的结果。可以看到，在所有情况下，有向微调的表现均显著优于无向微调，且这一结论适用于两种视觉语言模型。不过，结果取决于视觉语言模型和数据集。

接下来，我们在测试数据上比较我们的一致性驱动微调方法与准确率驱动方法。回想一下

表 IX：一致性驱动微调的准确率

VLMs	数据集	无向		有向	
		视觉语言模型^m	视觉语言模型^a	视觉语言模型^m	视觉语言模型^a
MiniGPT4	TU_DAT	73.14	72.5	82.14	82.5
	跆拳道	72.1	71.7	81.1	81.45
	Kinetics	79.6	78.4	84.15	84.78
MiniGPT4-V	TU_DAT	75.4	74.15	83.3	83.55
	跆拳道 73.41		73.75	82.81	82.6
	Kinetics	80.15	79.85	84.4	84.2
Video-LLaMa	TU_DAT	78.25	78.35	85.12	85.84
	跆拳道	77.5	77.85	85.22	85.1
	Kinetics	82.15	82.85	88.54	88.25
Video-Mamba	TU_DAT	76.6	75.55	81.3	81.42
	跆拳道 76.41		76.8	80.85	80.8
	Kinetics 80.15		79.35	83.85	83.14

准确率驱动微调会根据评估数据中观察到的不准确之处来选择输入。准确率驱动微调不需要任何 TCV 或视觉语言模型^a，但其缺点是需要带标签的评估数据，且无法在推理时间提供可辩护性。表 X 展示了三个数据集的准确率驱动结果。可以看出，这里的准确率几乎与表 IX 中相同；因此，采用一致性驱动方法并不会让我们有所损失。[3]

表 X：准确率驱动微调的准确率

VLMs	数据集	无向		有向	
		视觉语言模型^m	视觉语言模型^a	视觉语言模型^m	视觉语言模型^a
MiniGPT4-V	TU_DAT 74.5 75.5 82.0 82.52				
	跆拳道 72.8 72.75 82.1 81.5				
	Kinetics	81.25	79.5	85.0	85.25
Video-Mamba	TU_DAT	75.8	76.25	80.25	80.0
	跆拳道 77.15		75.5	80.12	80.35
	Kinetics 79.4		80.55	84.6	84.1

B. TU DAT/跆拳道数据集的 *CIF* 分数

表 XI 展示了分别使用 X-CLIP、Video-MAE、MiniGPT4（基于图像）、MiniGPT4-Video、Video-Llama 和 Video-Mamba 在 TU DAT 和跆拳道数据集上取得的 CIF 分数。

显然，在所有情况下，我们的定向微调都以显著优势胜过无向微调。值得注意的是，对于两种截然不同的视频类型（道路交通与跆拳道），一致性都大幅提升，这证实了这种提升并非取决于视频特征。

表 XI：TU DAT 和跆拳道数据集上的 CIF 结果

VLM 模型	Task 数据集	无向		有向	
		VLM^m	VLM^a	VLM^m	VLM^a
X-CLIP	TU_DAT	54.5	55.15	74.25	73.65
X-CLIP	跆拳道	48.71	48.15	69.85	70.05
VideoMAE	TU_DAT	52.04	52.41	72.65	73.25
VideoMAE	跆拳道	48.5	49.06	69.8	69.31
MiniGPT4	TU_DAT	59.78	60.41	75.51	74.35
MiniGPT4	跆拳道	59.78	60.41	75.71	75.41
MiniGPT4-Video	TU Dat	71.45	71.8	86.35	85.125
MiniGPT4-Video	跆拳道	68.40	68.1	83.10	83.08
Video-Llama	TU_DAT	72.16	72.41	86.85	87.32
Video-Llama	跆拳道	42.57	42.05	84.60	85.08
VideoMamba	TU_DAT	61.95	61.41	80.85	80.4
VideoMamba	跆拳道 61.60	61.75	76.52	75.45	

[3]无向结果与表 IX 略有不同，可能是随机性导致的；理想情况下，两者应当相同。

我们还考察了不使用 VLM^{a}的影响。由于这两个数据集中的 VLM^a 任务很简单，因此可以不使用它。表 XII 展示了使用和不使用 VLM^{a}时，VLM^{m}在测试数据上的准确率 。在后一种情况下，一致性检查是在与 VLM^m 相关联的代理活动和与 VLM^{a}相关联的代理活动之间进行的。本例中，我们使用 MiniGPT4-Video 作为主视觉语言模型。此时准确率有所降低，但仍显著高于纯粹的无向微调。

C. Kinetics 数据集上的 *CIF* 结果

与前两个数据集不同，Kinetics 数据集包含许多相当复杂的活动。这通常也意味着需要为 VLM^a 提供更复杂的描述，因为复杂的活动最好通过其细节来描述，如表 V 所示。例如，很难用 2～3 个词对“割草”进行替代描述。对于定向微调，我们的方法仍然能够达到与前两个较简单数据集大致相同的准确率水平。

表 XII：使用和不使用视觉语言模型时的微调准确率^a

Case	TU_DAT	Taek.
无向	75.4	73.41
使用视觉语言模型^a	83.3	82.81
无 VLM^a	81.0	80.65

表 XIII 比较了 MiniGPT4、MiniGPT4-Video 和 Videomamba 的 CIF 度量。（其他情况类似，故未报告。）这里我们使用批次大小为 50。请注意，我们对有向和无向微调采用相同的总时长。因此，总时间中有很大一部分消耗在视觉语言模型^{a}的微调上。尽管如此，有向方法的 CIF 仍高出约十个百分点。这有力地验证了我们先前的主张：在规模大、复杂的数据集中，使用视觉语言模型^a 很容易证明其合理性；这也应当让人确信我们的结论是可靠的。

表 XIII：Kinetics 数据集上的 CIF 结果

VLMs	无向 有向	
MiniGPT4	74.25%	83.7 %
MiniGPT4-V 78.5%		87.75%
Video-Mamba	72.15%	84.5%

D. 微调时间分析

本节比较了针对所考虑的全部三个数据集，有向和无向微调方法各自所需的时间。如前所述，我们对两种方法都进行四轮微调有向和无向情形，每种情形均使用 20 个视频。每次迭代的耗时分为两部分，报告的是 500 轮次内每轮的平均值：(a) 实际微调时间和 (b) 准备时间。对于无向情形，准备时间包括从磁盘随机检索 20 个视频。对于有向情形，准备时间也 包括运行 YOLO、查询 VLM^m 和 VLM^a、生成断言以及利用这些断言进行一致性检查的开销。

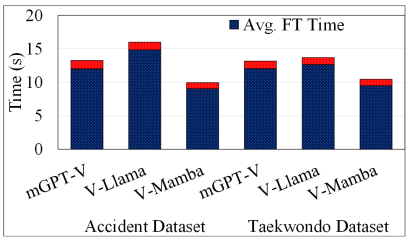


图6：无向情形下每轮的准备与微调时间

图 6 展示了无向情形下每轮次的平均微调时间和微调准备时间。图 7 展示了有向情形下的相同指标。在这两种情形中，微调时间均包含两个视觉语言模型。正如预期，

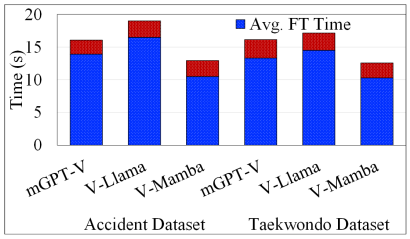


图7：有向情形下每轮次的准备与微调时间

两种情形下的微调时间几乎相同，约为 ~10- 12 秒（或每个视觉语言模型 5-6 秒）。即使在有向情形下，准备时间仍然很短，且比无向情形下长不了多少；这是因为其中大部分时间都用于从磁盘检索和加载视频。

E. 推理与合理性论证时间

如前文所述，我们可以轻松地将微调机制转变为由可辩护性支持的推理机制。为此，我们实际上移除图 2 中的微调部分，从而打破该循环。在这种情况下，每次推理还将附带

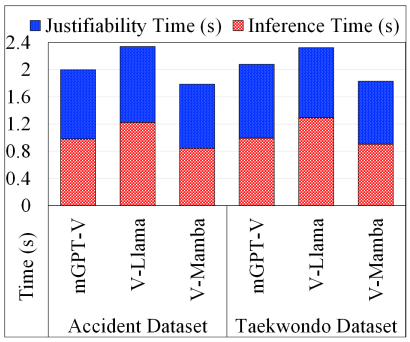


图 8：合理性论证/推理时间

一致性检查的结果。也就是说，如果所有一致性测试均通过，我们就能获得输出结果的额外支持（或合理性论证）；但是，如果其中任何一项测试失败，则表明该输出不值得信赖。

图8展示了 MiniGPT4-Video、Video-Llama 和 Video-Mamba 在两个数据集上的推理时间与可辩护性时间。请注意，可辩护性时间不同于上文报告的微调准备时间，因为我们不再需要承担为微调而从磁盘读取视频所产生的大量开销。可以看出，可辩护性时间与推理时间大致相同。对于准确率可能造成严重后果的关键应用而言，这应该是合理的。对于其他应用，我们可以降低可辩护性的运行频率（例如，定期运行，或仅在视觉语言模型^m 预测出已知存在可靠性问题的类别时运行）。

需要强调的是，我们的目标是检测可能需要干预的目标活动，而非逐帧讲述事件。因此，约 *2* 秒的演示推理时间相当高效，而且只使用一块图形处理器。虽然实验中使用的 NVIDIA A6000 图形处理器属于相对高端的硬件，且主要用于微调，但在部署场景中，关键因素是推理；这些模型也可以在较低端的硬件上运行，只是推理时间会相应增加。

不过，我们认为，在大语言模型和人工智能处理硬件都快速发展的领域，关注本研究中获得的某一时点的推理时间并无意义。例如，我们的许多结果都是使用搭载70亿参数模型的 MiniGPT4-Video 得出的。最近，阿里云发布了同样拥有70亿参数的 Qwen2.5-VL 模型。我们使用 Qwen2.5-VL 进行了若干案例测试以作比较，发现 MiniGPT4-Video 平均1.9秒的运行时间，在 Qwen2.5-VL 上缩短至仅0.9秒。因此，在现有机器上实现平均推理时间低于1秒这一目标，已经可以达成。

在硬件方面，为高端台式机和工作站中央处理器集成并增强神经网络处理器的竞赛仍在持续，这已经使它们能够以较低的功耗运行70亿参数乃至更大的模型。例如，英特尔即将推出的 Panther Lake 工作站处理器（预计于2026年第一季度量产）采用英特尔的1.8纳米制程打造。其机器学习性能最高可达180万亿次运算（每秒万亿次运算），比早前采用台积电3.0纳米制程的 Lunar Lake 处理器提升了50%，同时热设计功耗（热设计功耗）仍为45 W。相比之下，我们的 NVIDIA A6000 每秒万亿次运算指标为310（高出不到 2× ），但其热设计功耗为300 W（高出约 7× ）。因此，Qwen2.5-VL 在 Panther Lake 上运行时，预计推理时间约为1.55秒，功耗为45 W，设备属于从笔记本电脑到工作站级别。

我们还想指出，尽管当前有一种趋势，致力于在移动设备和低端物联网设备上运行精简版大语言模型，但这未必适用于要求响应时间极短的关键应用。在这种情况下，视觉语言模型很可能运行在边缘控制器上，而不是各个物联网设备上。精简版模型往往也会受到准确率低和幻觉问题的困扰，因此不适合关键应用。同样，将推理过程的一部分放在云端而非本地部署，会带来访问延迟和网络不确定性——对于时效性要求高的安全关键用例而言，这同样不可取。

VI. 结论

本文提出了一种面向视觉语言模型的新型一致性驱动微调（FT）方法，将传统计算机视觉（TCV）用于识别细节，并结合显式逻辑推理，以提升视觉语言模型的性能。该机制能够显著减少微调所需的带标签数据，并且与不智能选择输入的机制相比，可实现高得多的准确率。此外，该机制还提供了一种可在推理时继续使用的合理性论证机制，以及一种对态势感知应用至关重要的合理性检查机制。所提出的机制具有很强的通用性，因为我们可以自行决定在一致性检查中引入的复杂程度。

该机制的主要缺点是需要运行另一个视觉语言模型，这会增加资源需求，尤其是在推理过程中生成合理性论证时。另一个局限是，如果使用 TCV 技术来实现超越基本目标/姿态检测和运动跟踪的功能，就需要专用训练数据和相应训练。

参考文献

[1] Xiao-Kun Wu, Min Chen, Wanyi Li, Rui Wang, 等。大语言模型微调：概念、机遇与挑战。Big Data and Cognitive Computing, 9(4):87，2025年。1[2] Nikoleta Manakitsa, George S Maraslidis, Lazaros Moysis, 和 George F Fragulis。机器与机器人视觉中的目标检测、语义分割和人体动作识别：机器学习与深度学习综述。Technologies，12(2):15，2024年。1[3] Nidhal Jegham, Chan Young Koh, Marwan Abdelatti, 和 Abdeltawab Hendawi。YOLO演进：YOLOv12、YOLO11及其早期版本的综合基准与架构综述。arXiv预印本 arXiv:2411.00201，2024年。2[4] Dan Zhang, Tao Feng, Lilong Xue, Yuandong Wang, 等。基础模型的参数高效微调。arXiv预印本 arXiv:2501.13787，2025年。3[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, 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。3[6] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, 等。CLIP-Adapter：通过特征适配器改进视觉语言模型。International Journal of Computer Vision，132(2):581–595，2024年。3[7] Avinash Patil 和 Aryan Jadon。推进大语言模型的推理能力：有前景的方法与途径。arXiv预印本 arXiv:2502.03671，2025年。3[8] Dibyanayan Bandyopadhyay, Soham Bhattacharjee, 和 Asif Ekbal。思考型机器：基于大语言模型的推理策略综述。arXiv预印本 arXiv:2503.10814，2025年。3[9] Yueting Yang, Xintong Zhang, 和 Wenjuan Han。借助大语言模型增强视觉语言模型的推理能力。arXiv 预印本 arXiv:2305.13267，2023年。3[10] Bikram Pratim Bhuyan, Amar Ramdane-Cherif, Ravi Tomar, 和 TP Singh。神经符号人工智能：综述。Neural Computing and Applications，36(21):12809–12844，2024年。3[11] Brandon C Colelough 和 William Regli。2024年的神经符号人工智能：系统综述。arXiv预印本 arXiv:2501.05435，2025年。3[12] Jae Hee Lee, Michael Sioutis, Kyra Ahrens, Marjan Alirezaie, 等。神经符号时空推理。arXiv预印本 arXiv:2211.15566，2022年。3[13] Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, 和 Michael Spranger。逻辑张量网络（LTN）。Artificial Intelligence, 303:103649，2022年。3[14] Emile van Krieken, Erman Acar, 和 Frank van Harmelen。分析可微模糊逻辑算子。Artificial Intelligence，302，103602，2022年。3[15] Alexander Artikis, Marek Sergot, 和 Georgios Paliouras。一种用于活动识别的逻辑编程方法。载于《第28届ACM多媒体事件国际研讨会论文集》，第–页，2010年。3[16] Karinne Ramirez-Amaro, Eun-Sol Kim, Jiseob Kim, Byoung-Tak Zhang, 等。通过时空特征学习和语义规则增强人体动作识别。载于《2013年IEEE-RAS国际人形机器人会议（Humanoids）》，第456–461页。IEEE，2013年。3[17] Pavana Pradeep, Krishna Kant, 和 Amitanghsu Pal。C-FAR：智能交通系统中异常消解的组合式框架。IEEE智能交通系统汇刊，2022年8月。3，4[18] Pavana Pradeep, Krishna Kant, 和 Amitangshu Pal。基于路侧摄像头的非侵入式驾驶员行为表征。IEEE IoT Journal，2023年9月。3，4[19] Md Tanzib Hosain, Jamin Rahman Jim, MF Mridha, 和 Md Mohsin Kabir。深度学习中的可解释人工智能方法：进展、应用与挑战。Computers and Electrical Engineering，117:109246，2024年。3[20] Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, 等。XXAI——超越可解释人工智能。载于《国际研讨会：将可解释人工智能扩展至深度模型和分类器之外》，第3–10页。Springer，2020年。3[21] Muhammad Maaz、Hanoona Rasheed、Salman Khan 和 Fahad Shahbaz Khan。Video-chatgpt：通过大规模视觉与语言模型实现详细的视频理解。arXiv:2306.05424，2024年。3[22] Erika ´ Abrah´am 和 Gereon Kremer。可满足性检查：理论与应用。载于《软件工程与形式化方法》：第14届国际会议，SEFM 2016，作为 STAF 2016 的组成部分于奥地利维也纳举行，2016年7月4日至8日，会议论文集14，第9–23页。Springer，2016年。4[23] Pavana Pradeep、Amitanghsu Pal 和 Krishna Kant。大规模物联网系统中的冲突检测与缓解自动化。CCGrid 会议论文集，2021年5月。4[24] Pavana Pradeep、Krishna Kant、Francesco Di Rienzo、Carlo Vallati 和 Amitangshu Pal。从 RGB-D 图像中识别人类姿态错误。IEEE《人工智能汇刊》，2025年5月。4[25] Leonardo de Moura 和 Nikolaj Bjørner。Z3：高效的 SMT 求解器。载于 C. R. Ramakrishnan 和 Jakob Rehof 编，《TACAS》，第337–340页，2008年 4 – 。[26] Bruno Dutertre。Yices 2.2。《计算机辅助验证》（CAV'2014），8559：737–744，2014年7月。4[27] Savas Konur。用于规定和验证实时系统的时序逻辑综述。《计算机科学前沿》，7(3)：370–403，2013年。4[28] Alexander Artikis、Marek Sergot 和 Georgios Paliouras。用于事件识别的事件演算。IEEE TKDE，27(4)：895–908，2014年。4[29] Christos Vlassopoulos 和 Alexander Artikis。迈向一种用于运行时推理的简单事件演算（RTEC）。载于 Commonsense，2017年。4[30] Jingyi Zhang、Jiaxing Huang、Sheng Jin 和 Shijian Lu。面向视觉任务的视觉与语言模型：综述。IEEE《模式分析与机器智能汇刊》，2024年。5[31] Deyao Zhu、Jun Chen、Xiaoqian Shen、Xiang Li 等。MiniGPT-4：借助先进的大语言模型增强视觉语言理解。arXiv:2304.10592，2023年。5[32] Kirolos Ataallah、Xiaoqian Shen、Eslam Abdelrahman、Essam Sleiman 等。MiniGPT4-Video：利用交错视觉文本词元推进视频多模态大语言模型的视频理解。arXiv 预印本 arXiv:2404.03413，2024年。5，7[33] Yiwei Ma、Guohai Xu、Xiaoshuai Sun、Ming Yan 等。X-CLIP：面向视频文本检索的端到端多粒度对比学习。载于《第30届 ACM 多媒体国际会议论文集》，第638–647页，2022年。5，7[34] Hang Zhang、Xin Li 和 Lidong Bing。Video-Llama：用于视频理解的指令微调音视频语言模型。arXiv 预印本 arXiv:2306.02858，2023年。5，7[35] Xiao Wang、Shiao Wang、Yuhe Ding、Yuehang Li 等。新一代网络中替代 Transformer 的状态空间模型：综述。arXiv 预印本 arXiv:2404.09516，2024年。5[36] Kunchang Li、Xinhao Li、Yi Wang、Yinan He 等。VideoMamba：用于高效视频理解的状态空间模型。arXiv 预印本 arXiv:2403.06977，2024年。5，7[37] Pavana Pradeep Kumar。TU-DAT 数据集。https://github.com/pavana27/TU-DAT，2021年。5[38] Taekwando America。《跆拳道学生手册》。https://taekwondoamerica.org/wp-content/uploads/2017/09/Student-Manual-2012.pdf，2011年。5[39] Will Kay、Joao Carreira、Karen Simonyan、Brian Zhang 等。Kinetics 人类动作视频数据集。arXiv 预印本 arXiv:1705 06950，2017年 5 . . [40] Deyao Zhu、Jun Chen、Xiaoqian Shen、Xiang Li 等。MiniGPT-4：借助先进的大语言模型增强视觉语言理解。arXiv 预印本 arXiv:2304.10592，2023年。7[41] Limin Wang、Bingkun Huang、Zhiyu Zhao、Zhan Tong 等。VideoMAE V2：通过双重掩码扩展视频掩码自编码器。载于《IEEE/CVF 计算机视觉与模式识别会议论文集》

，第14549–14560页，2023年7月