在线动作堆叠提升强化学习在空中交通管制中的性能

Ben Carvell*NATS，怀特利，汉普郡，PO15 7FL，英国

George De Ath†埃克塞特大学，埃克塞特，英格兰 EX4 4QJ，英国

EseogheneBenjamin‡ 和 Richard Everson§艾伦·图灵研究所，伦敦，英格兰 NW1 2DB，英国

arXiv:2601.04287v1 [cs.LG] 7 Jan 2026

我们提出了在线动作堆叠，这是一种用于强化学习策略的推理时封装器，能够生成符合实际的空中交通管制指令，同时允许在更小的离散动作空间上进行训练。策略通过简单的航向或飞行高度层增量调整进行训练，并结合动作阻尼惩罚，以降低指令频率，使智能体以短时突发的方式发布指令。在推理阶段，在线动作堆叠会将这些原始动作突发编译为符合领域要求的复合许可指令。我们使用近端策略优化和 BluebirdDT 数字孪生平台，训练智能体沿横向航路引导航空器，管理爬升和下降至目标飞行高度层，并在最小间隔约束下执行双航空器防撞。在横向导航实验中，与阻尼基线相比，动作堆叠大幅减少了发布的指令数量；尽管仅使用五个动作，其性能仍可与在 37 维动作空间上训练的策略相当。这些结果表明，在线动作堆叠有助于弥合标准强化学习形式与空中交通管制运行要求之间的关键差距，并为扩展到更复杂的控制场景提供了一种简单机制。

I. 引言

A 背景MATION

空中交通管制是航空领域面临的最严峻挑战之一。国际航空运输协会（IATA）估计，到 2025 [1], 将有52亿名乘客乘坐飞机出行，预计到 2050 [2]这一数字将增加至多达124亿。英国政府响应IATA提出的“Fly Net Zero”决议，作出了自己的“Jet Zero”承诺，力争到 2050 [3]实现航空业净零碳排放。在这一背景下，对能够彻底提升运行效率的先进决策支持系统的需求日益迫切。在维持严格安全标准的同时满足这些需求，需要开展与领域密切相关的研究，为空中交通管制员（ATCO）的自动化支持奠定坚实基础。

英国的空中交通管制员使用一系列支持系统来协助执行管制任务，但这些系统的复杂程度各不相同。其中最先进的是iFACTS [4] 和iTEC [5], 等预测系统，它们利用Eurocontrol开发的航空器数据基础（BADA）航空器性能模型 [6]，提供由航空器行为驱动的预测结果。随后，这些工具会对预测结果进行比较，以帮助识别航空器之间潜在的冲突，并支持对潜在管制许可开展“假设情景”探查，从而评估其影响。系统不会建议应发布何种管制许可，相关决策及其安全性完全由空中交通管制员负责。

自 Wesson等人于20世纪70年代开展研究以来，推动空中交通管制实现全面决策自动化的研究一直在持续 [7], ，期间涌现并经过原型验证的方法丰富多样。基于规则的方法和启发式方法已有详尽记录，其中最成熟的示例是 ARGOS——由 Eurocontrol 开发的原型系统 [8]。

*共同研究者，Project Bluebird，NATS

†讲师，计算机科学系，埃克塞特大学

‡研究软件工程师，Project Bluebird，艾伦·图灵研究所 §首席研究员，Project Bluebird，艾伦·图灵研究所和埃克塞特大学

SESAR（单一欧洲天空空中交通管理研究）计划开展了大量行业工作，包括 AGENT [9]，其采用基于规则的方法进行冲突消解；近期的 HYPERSOLVER [10] 则利用强化学习，同时管理流量管理和战术空中交通管制。

强化学习（RL）已成为应对这一挑战的一种颇具前景的方法，许多研究都在探索其在空中交通管制任务中的应用 [11–18]。与传统的基于规则的系统不同，强化学习能够从经验中学习，并为复杂场景发现创新解决方案，DeepMind 在 AlphaStar 和 AlphaGo 方面的开创性工作便证明了这一点 [19, 20], OpenAI 在《Dota》中的研究 2 [21], 以及其在机器人学等领域的其他应用也同样如此 [22]。这些特性使强化学习尤其适合空中交通管制所要求的动态且细致的决策。

B. 强化学习与 PPO

在强化学习中，智能体通过在给定环境中选择要执行的动作来最大化所获得的累积奖励 [23], ，这一交互过程被表述为马尔可夫决策过程 [24]。智能体在一系列离散时间步 t = 0、1、2、.....中与环境交互，并且在每个时间步，根据当前观测到的环境状态 $S_t \in S$，从可用动作集合 A 中选择一个动作 A_t。状态的表示方式必须经过选择，使得 $Pr[S_{t+1}|S_t, A_t] = Pr[S_{t+1}|S_1, A_1, ..., S_t, A_t]$ 满足马尔可夫性质，即转移到后继状态 S_{t+1} 的概率与所有先前的状态和动作无关，只取决于当前状态 S_t和动作 A_t。如图 1 所示，环境随后前进一个时间步，智能体获得奖励 R_{t+1}以及状态 S_{t+1}的更新观测，从而形成由状态、动作、所得奖励和下一状态组成的元组 S_t，A_t，R_{t+1}，S_{t+1}。

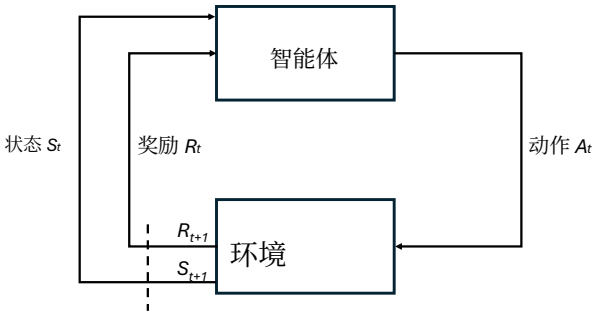

图 1 马尔可夫决策过程中智能体与环境的交互循环。

通过在多个重复的训练回合中，随时间处理这些元组构成的轨迹，智能体可以学会在给定特定状态时选择哪些动作，以最大化其获得的累积奖励。智能体的行为由折扣因子 $\gamma \in (0, 1]$，调节，该因子决定智能体在多大程度上倾向于获得即时奖励，而非预测的未来奖励。γ 值较低时，智能体会偏好即时奖励；γ 值较高时，智能体则可能牺牲即时奖励，以换取之后获得更高总奖励的机会。训练智能体主要有两种方法：基于策略的方法和基于价值的方法。基于策略的方法学习一种策略 $\pi(s)$，将任意给定状态映射为相应动作（或选择各动作的概率）；而基于价值的方法则学习价值函数 $V^{\pi}(s)$，用于预测在遵循策略 [23] 时，给定特定状态下的期望回报（从当前时间步直到回合结束所获得的总奖励）。

在本研究中，我们采用近端策略优化（PPO） [25], ——一种先进的强化学习方法，它以演员-评论家配置结合了上述两类方法的要素。在这种配置中，策略 π 和价值函数V^{π} 都会接受训练，随后价值函数将用作“评论家”。通过记录策略下实际获得的总奖励，并将其与价值函数给出的期望回报进行比较，可以衡量智能体相较于以往经验的表现是更好还是更差，并据此指导策略更新。PPO还采用了一种高效方法来限制策略更新的幅度，以防止训练不稳定。凭借这些优势，PPO已被广泛视为一流方法，并被用于DeepMind的 AlphaStar [19] 以及近期突破性大语言模型的输出优化 [26]。

C. 先前工作：空中交通管制的强化学习形式化

强化学习适用于解决不确定条件下的序贯决策问题，因此已被应用于多种管制任务。Brittain and Wei [12] 的研究是新一波尝试通过强化学习实现空中交通管制任务自动化的论文中较早的一篇。作者将空中交通管制建模为多智能体强化学习问题，并使用Bluesky模拟器 [27] 构建环境。该研究展示了强化学习作为空中交通管制自动化可行方案的潜力。尽管取得了成功，作者采用了简化的场景设置（相较于完整的三维空中交通管制任务），将智能体可用的动作A限制为仅能提高、降低或保持航空器当前速度。随后，他们利用这一动作空间解决这样一种场景：航空器以相同飞行高度层沿多条不同航线飞行，通过调整速度，确保航空器在同一航线上前后飞行时，以及在交叉点与其他航线上的航空器相遇时，均保持所需的3海里间隔。

在文献中，其他尝试将强化学习应用于空中交通管制各个环节的方法，也普遍存在对环境设计和动作空间加以限制的特点。Wang等人 [14] 同样只考虑横向冲突消解，但引入了航向调整的连续动作空间，而非速度调整。Dalmau和Allard [17] 通过同时纳入速度和横向导航动作，扩展了横向动作空间的功能，但仍然省略了垂直控制这一环节。当前文献中普遍缺少垂直控制。垂直控制是控制任务的关键组成部分，既要使航空器达到巡航高度层、下降着陆，也要满足各类协调协议，并利用垂直维度进行冲突消解。

这些形式化方法大幅简化了训练时的奖励设计，因为它们没有考虑完整三维控制任务中涉及的复杂成功判据。仅在横向平面上判断间隔，可以简化对航空器间相互作用的预测，无须考虑不同的高度层和垂直剖面。

最小化每架航空器所需管制许可数量这一目标也常被忽略，而这构成了强化学习中更普遍的动作稀疏性挑战；Biedenkapp等人 [28]在非空中交通管制领域的研究探讨了这一问题。出于多种运行原因，限制管制许可数量十分必要，包括管理飞行员工作负荷、管理空中交通管制员工作负荷，以及管理用于发布指令的无线电频率负载。尽管未来的技术发展或许能够缓解其中一些因素，但由于安全挑战以及难以强制要求航空器配备相关设备，空中交通管制领域的技术发展自然较为缓慢。因此，忽视这一关键约束会严重削弱研究对近期应用的相关性。

近期的一些论文对完整管制任务的刻画更为全面。Sui等人 [15]采用了横向和垂直方向的动作，但将问题限定为一对航空器之间的冲突探测与解脱，忽略了扇区层面的更广泛目标。Vouros等人 [18] 的研究从领域角度来看设计较为完善，其动作空间涵盖垂直、横向和速度调整，并采用了包含多架航空器及更广泛扇区目标的空中交通管制场景设计。然而，该研究的安全奖励函数有所简化，因为它只对间隔丧失进行惩罚，而没有像现实运行所要求的那样，激励采用故障安全的运行方式。这显著扩大了可行解空间，使其中包含在现实世界中会被判定为极不安全的决策，因此从领域应用角度看，研究结果的相关性仍然有限。Machine Basic Training框架 [29]提出了一套针对早期成熟阶段空中交通管制智能体的明确性能目标，这些目标改编自真实的空中交通管制员培训课程。该框架明确指出，即使对于处于早期成熟阶段的空中交通管制自动化方案，广阔的动作空间和严格的安全定义也至关重要。

总而言之，目前关于强化学习在空中交通管制中应用的大多数文献都采用了受限的问题设定，掩盖了在该领域尝试利用强化学习实现决策自动化时所面临的挑战。已有研究探讨了状态空间的可变性，但对可变且庞大的动作空间、动作稀疏性、符合领域实际的安全性衡量标准，以及垂直控制这一学科的关注不足，甚至完全未予涉及。这为相关研究提供了契机：综合应对这些挑战，并开发能够在多种此类困难条件下训练智能体的技术。

D. 研究动机与贡献

空中交通管制的一项关键目标，是在有效管制航空器的同时，尽可能减少所需的管制许可。这样可以减轻飞行员的工作负荷，并减少扇区无线电频率的拥塞，确保所有必要的管制许可都能及时发出。这一操作要求与空中交通管制中常见的强化学习形式相冲突：在这类形式中，航向和高度层通过小幅渐进调整来改变，导致需要发出大量指令，才能完成实际操作中一条指令就能实现的任务。在实践中，空中交通管制员（ATCO）会发出一条指令（“左转*30*度”），而不是连续发出三条指令（“左转*10*度”）；并且只会在必要时才给出指令，而不是持续微调航空器的航迹。尽管状态信息

会定期更新（在英国，每6秒刷新一次雷达），与轮流行动的游戏相同，但这些操作限制使空中交通管制有别于围棋、国际象棋或星际争霸等其他备受瞩目的强化学习应用。空中交通管制中的成功不仅取决于最终结果，也取决于执行的效率和方式。

在本研究中，我们引入了在线动作堆叠，这是一种在线策略驱动方法，使仅使用有限离散动作训练的智能体能够生成完整的空中交通管制指令，同时尽量减少所发布指令的总数。这项技术在概念上与Sutton等人的*options*框架相关。 [30],但它专门针对管制员式指令进行了设计。然而，与options、FiGAR [31], 和TempoRL [28], 不同，我们不改变训练目标，也不引入半马尔可夫决策过程。相反，我们在标准PPO策略外封装一个推理时过程，将原子指令编译为宏指令，同时保留底层马尔可夫决策过程结构。通过解决动作频率这一挑战，动作堆叠显著简化了训练智能体执行基本空中交通管制行为的过程，并为实现更符合领域实际的方案提供了关键基础。

II. 方法

A. 场景规范

空中交通管制运行是一项复杂且多方面的任务，要求空中交通管制员在确保航空器之间安全间隔的同时，提供高效的横向和垂直剖面，并满足扇区之间的移交约束（称为协调）。在此，我们描述如何构建场景，以训练和测试这一挑战的相关方面。

1. 空中交通管制问题陈述

We designed 场景 ios to train and evaluat代理 t 竞争 tence in four fundamenta空中交通管制边界 ehavi:

- 通过发布航向指令，引导航空器沿航线穿越扇区，并确保其从指定航路点离开
- 通过发布垂直指令，引导航空器爬升或下降至与下一扇区协调一致的出口高度层
- 在横向平面内为航空器实施冲突调配，确保航空器轨迹始终保持5海里间隔标准，即使不再发布任何指令也是如此
- 在有效维持管制的同时，尽量减少发布的指令数量

这些行为与CAP493对战术管制员的要求直接吻合，即提供安全、有序且高效的管制 [32]。场景运行了300步（每步代表6秒仿真时间），为航空器穿越空域提供了充足时间。根据具体用例，场景使用一架或两架航空器生成。我们使用两架航空器训练横向导航与避障，使用一架航空器训练垂直导航。训练场景和测试场景均通过相同的随机过程生成，起始位置和航线存在随机变化，详见第II.A.3节。

2. 空域

本研究中的所有场景均使用作为***ProjectBluebird**** [33]一部分构建的测试空域“X-Plus扇区”，如图2所示。该空域的航路宽度为20海里，范围受限于一个120海里的边界框内，代表标准的航路扇区配置。空域中包含用于界定航空器可飞行路线的导航定位点，最外围的定位点则作为创建航空器的“种子点”。该空域旨在为航空器导航带来挑战，并在航路交叉点呈现潜在的航空器交互。穿越该空域的可行航线如图3所示。

3. 航空器

对于横向控制场景，创建了两架航空器，其参数如表1所示。对于垂直控制场景，创建了一架航空器，其参数如表2所示。

B. 问题形式化

我们的仿真使用 BluebirdDT [34]数字孪生平台† 。航空器的性能采用基于 BADA 物理模型的简化实现进行建模 [6]。模拟器更新航空器

*Project Bluebird 是 NATS、艾伦·图灵研究所与埃克塞特大学之间由 EPSRC 资助的 Prosperity Partnership。†该平台正在作为 Project Bluebird 的一部分开发， [33], 并将纳入即将发布的开源版本。

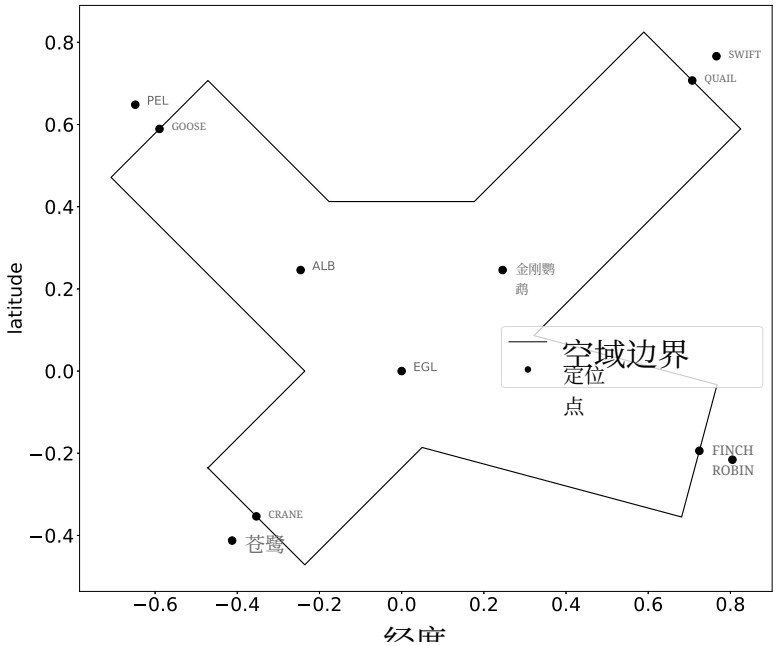

图2 “X-Plus 扇区”：一个人工构建的航路空域表示，包含20条海里航路和命名导航定位点。

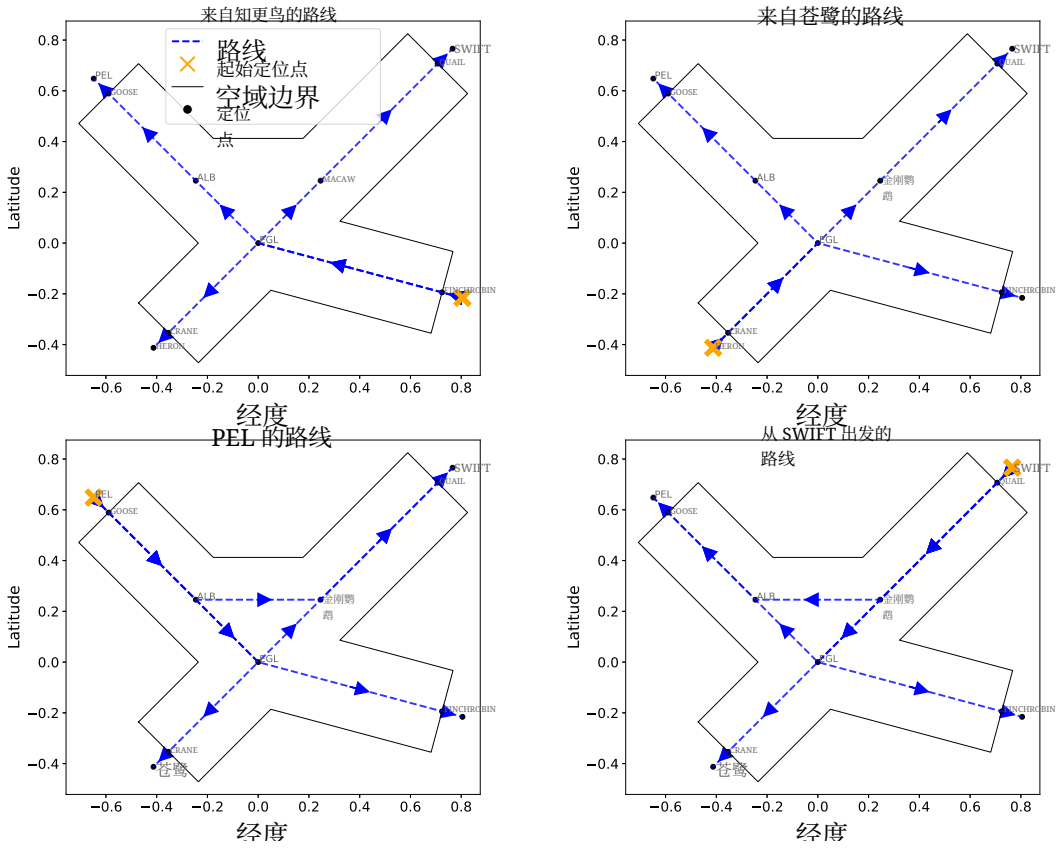

图3 从各个起点穿过 X-Plus 扇区的可行航线。

表1 横向控制场景的航空器参数

参数	值
航空器类型	波音757-300
航线	从可行扇区航线中随机选择*
飞行高度层	飞行高度层300（30,000英尺）
起始位置	随机分布在以第一个航路定位点为中心、半径4海里的圆内
航向	与初始航段平行

*航线起 始点经过冲突消解，以避免场景初始化时已存在冲突 .

表2 垂直控制场景的航空器参数

参数	值
航空器类型	波音757-300
航线	从可行扇区航线中随机选择*
初始高度	在飞行高度层100和飞行高度层300之间随机选择
目标高度	在飞行高度层100和飞行高度层300之间随机选择
起始位置	位于第一个航路定位点
航向	沿航线自主导航

每六秒更新一次状态，与实际运行中的雷达更新率一致。

这一问题设置自然符合马尔可夫决策过程的结构。在任意时刻t，模拟器的当前状态S_t与当前动作A_t共同提供了预测模拟器下一状态所需的信息；预测结果具有不确定性，且无需依赖历史信息，体现了马尔可夫性质。在每个时间步，智能体根据策略选择动作$A_t \in \mathrm{A}$，以调节环境中航空器的行为（*no action* 也属于这一动作集合）。最后，在执行动作A_t后，可以分析更新后的环境状态 S_{t+1}，计算奖励R_{t+1}，以此评估动作A_t所引起的环境变化是否符合良好战术管制的目标。我们采用集中式强化学习（而非多智能体）建模，因此将多架飞机的状态向量拼接成一个向量，动作空间也采用类似方法。下文将详细介绍这一建模方式。

1. 状态

正如 Asadi 和 Huber 所述， [35], 从状态空间中移除冗余变量，同时保留预测下一状态所需的全部信息，可以加快强化学习配置的训练速度。为最大限度地提升未来的可扩展性，我们希望这些简单的初始示例利用领域知识，仅选取状态空间中的必要变量，从而提高训练速度，同时保留有效训练的能力。

横向控制场景中单架航空器的状态向量 S 如表 3 所示。
纳入转向后续定位点，是为了在避让其他航空器时，通过提供

表 3 横向控制场景的状态变量

特征	定义	范围	单位
θ_f	从获准航向到下一个定位点的相对转向	[-180, 180]	度
θ_{sf}	从获准航向到后续定位点（下一个定位点加一）的相对转向	[-180, 180]	度
d_c	距航线中心线的截断距离	[-100, 100]	海里
Δ_t	距上次动作的截断时间	[0, 60]	秒

表4 横向避让场景的额外状态变量

特征	定义	范围	单位
$\theta_{1,2}$	从航空器1到航空器2的相对转向	[-180, 180]	度
$d_{1,2}$	航空器1与航空器2之间的距离	[0, 150]	海里

表5 垂直控制场景的状态变量

特征	定义	范围	单位
d_{xfl}	至出口定位点的距离与所需距离之间的截断差值 以每分钟 2000 英尺的固定速率从当前高度层爬升/下降至出口高度层 英尺/分钟	[-100, 100]	海里
FL_{sx}	所选飞行高度层与**s**退出飞行高度层之间的差值（截断）**x**	[-200, 200]	飞行高度层
FL_{nx}	入**n**场（初始）飞行高度层与**x**退出飞行高度层之间的差值（截断）	[-200, 200]	飞行高度层
Δ_t	距上次动作的截断时间	[0, 10]	秒

让航空器更清楚地了解退出目标。“获准航向”是指指示航空器遵循的航向。由于状态向量中的所有值都需要有明确的边界，因此将其与航线中心线的距离截断在 100 海里的范围内；对于该场景而言，100 海里足以涵盖“关注区域”。超出这些边界的行为与本研究无关，因为它们代表航空器大幅偏离空域和航线。自上次动作以来的时间截断为最多 60 秒，因为如后文公式7所示，我们对再次采取动作的惩罚范围与此一致。

运行双航空器场景时，该观测空间会加倍，每架航空器各对应一组数值。运行带避让的双航空器场景时，还会加入两个额外变量，如表4所示。两个航空器进行横向避让时的完整状态向量如下：

$$\mathcal{S}_t = (\theta_{f1}, \theta_{fs1}, d_{c1}, \Delta_{t1}, \theta_{f2}, \theta_{fs2}, d_{c2}, \Delta_{t2}, \theta_{1,2}, d_{1,2}) \tag{1}$$

最后，在训练简单的单航空器垂直控制场景时，我们提供的状态向量包含表5所示的元素。

状　态值会使用上述表格中定义的范围归一化至 [0, 1]范围内。

2. 动作

已知缩减动作空间也能带来更快、更可靠的策略训练 [36]。实际运行中的空中交通管制动作空间极其庞大。仅横向引导航向指令，就可以按 360° 度为增量 5° 下达指令（共 72 条），还可以选择不同范围的相对转向，至少包括左右转向，按 30° 度为增量 5°（共 12 条）。再加上航路导航指令、速度和垂直指令（航路飞行时，飞行高度层约为 70 至 450，按十个高度层为增量），每架航空器的动作空间轻易就会超过 100 个动作，很可能接近 200 个。在我们的场景中，我们先使用横向和垂直平面上的增量指令进行训练，然后采用我们创新的在线动作堆叠技术，将其组合成符合实际运行要求的指令。

单架航空器的动作空间如下：

$$\mathcal{A} = \{\phi, h_-, h_+\} \tag{2}$$

其中，ϕ 表示不执行动作，h_- 表示左转*10*度，h_+ 表示右转*10*度。选择10度的转向幅度，是为了在保证控制保真度合理的同时，确保完成一次转向所需的时间超过模拟器的一个步长。

对于包含两架航空器的场景，每架航空器都有一套复制的动作空间：

$$\mathcal{A} = \{\phi, h_-^1, h_+^1, h_-^2, h_+^2\} \tag{3}$$

当我们训练时　以较大的横向动作空间作为参考实现，单架航空器 i 的动作空间　　s:

一个 $=\{\phi,\ h_{-10},\ h_{-20}, h_{-30}, h_{-40}, h_{-50}, h_{-60}, h_{-70}, h_{-80}, h_{-90}, h_{+10}, h_{+20}, h_{+30}, h_{+40}, h_{+50}, h_{+60}, h_{+70}, h_{+80}, h_{+90}\}$ (4)

其中，ϕ 表示无动作，h_{-10} 表示左转*10*度，h_{-20} 表示左转*20*度.....总计可在 90° 的范围内向左或向右转动，每次增量为 10°。在双航空器场景中，每架航空器分别具有一组航向动作，总计为37个动作。

对于垂直控制场景，单架航空器的动作空间为：

$$\mathcal{A} = \{\phi, FL_-, FL_+\} \tag{5}$$

其中，ϕ 表示无动作，FL_- 表示下降*10*个飞行高度层，以及FL_+ 表示上升*10*个飞行高度层。

3. 奖励

在场景的每一步中，我们都会使用多个奖励函数，旨在对不理想的行为给予负奖励；这些奖励函数加权求和后构成需要最大化的总奖励。通过中心线距离奖励来抑制航空器偏离指定航线：

$$r_c = \exp\left(-(d_c/\lambda_c)^2\right) - 1 \tag{6}$$

其中，d_c 表示航空器与航线中心线之间的距离，λ_c 表示缩放因子。使用 $\lambda_c = 6$ 会在中心线附近产生逐渐增加的惩罚，并在航空器接近扇区边界（距中心线 10 海里）时急剧增大。这一设计反映了运行优先事项：轻微偏离航线可以接受，而可能引发安全风险的偏离空域行为则会受到严厉抑制。

动作阻尼奖励旨在限制发出的管制许可数量：

$$r_a = \begin{cases} \frac{n_s}{n_{max}} - 1 & \text{if } n_s < n_{max} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

其中，n_s 表示自上次动作以来经过的时间步数，n_{max} 定义了负奖励停止生效的阈值。在本实验中，$n_{max} = 10$，相当于两次管制许可之间间隔一分钟。该奖励会强烈抑制过于频繁地发出指令，符合实际运行惯例。结合导航奖励和安全奖励，它会促使策略避免细小的修正动作，并在需要进行较大机动时，改为成组发出一系列指令。动作发出后，负奖励缓慢衰减，这会直接激励动作突发序列；连续发出动作而不留间隔，可以最大限度地减少累积的负奖励。这些自然形成的原始动作突发序列非常适合在线动作堆叠。

航空器之间的安全间隔通过安全奖励来实现：

$$r_s = \begin{cases} \left(\frac{d_{1,2}}{d_{max}} - 1\right) e^{-\left(\frac{d_s}{\lambda_s}\right)^2} & \text{if } d_{1,2} < d_{max} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

其中，$d_{1,2}$ 表示两架航空器之间的当前距离，d_s 表示两架航空器沿路径长度为d_{max}的路径向前预测时，它们之间的预计最小距离，λ_s 为缩放因子。该公式会对两架航空器之间的预计最小间隔变小或违反5海里标准的状态施加惩罚。当预计间隔远高于5海里时，指数项会产生较小的惩罚；而当预计间隔接近或低于规定的最小值时，惩罚则会迅速增大。该惩罚会根据航空器之间的当前距离进行线性缩放，最大距离为d_{max}，以体现航空器越接近，安全问题就越紧迫。我们将$d_{max} = 150$ 设为与X-Plus扇区的大小大致相当，并将 $\lambda_s = 5$ 设为使奖励曲线在5海里的最小间隔标准附近呈现预期的变化趋势，从而强力抑制安全违规。

对于垂直场景，我们采用简单垂直奖励：

$$r_v = \exp\left(\frac{-|\Delta_{FL}|}{\lambda_v}\right) - 1, \tag{9}$$

其中，Δ_{FL} 是航空器选定飞行高度层与其出口飞行高度层之差，λ_v 是设为40个飞行高度层的缩放因子，用于在预期的 ±200 个飞行高度层范围附近提供有效梯度。

表 6 奖励配置权重

配置	中心线距离	动作阻尼	安全性	垂直方向	终止集
无阻尼横向导航	1.0	0.0	0.0	0.0	1.0*
横向导航	1.0	0.25	0.0	0.0	1.0
垂直导航	0.0	0.5	0.0	1.0	1.0
横向导航与避障	1.0	0.3	2.0	0.0	1.0

*在此配置中，终止集未包含动作激励

在场景结束时，还会评估一组终止奖励，并将其加到累积奖励中，从而得到该场景的最终奖励。这些终止奖励同时适用于横向场景和垂直场景。终止奖励集由一项固定奖金奖励组成，其取决于智能体是否满足以下3项标准:

1) 在正确的飞行高度层导航至出口 2) 始终处于扇区的横向和垂直边界内 3) 每架航空器发出的动作少于 30 次

每满足一项标准即可获得 5 点奖励，满足所有标准时还可额外获得 20 点奖励。除了塑形奖励函数外，这些奖励还提供了额外的粗粒度成功信号，有助于进一步促进策略向期望行为收敛。

这些奖励根据场景配置按加权和的方式组合。表 6 展示了所使用的配置。这些奖励权重通过反复测试多种配置而确定，同时依据终端奖励集中所述的标准监测成功情况。

总之，对于横向导航场景，我们使用中心线距离奖励、动作阻尼奖励和终端奖励集进行训练。对于横向导航与避障场景，我们使用中心线距离奖励、动作阻尼奖励、安全奖励和终端奖励集进行训练。对于垂直导航场景，我们使用简单垂直奖励、动作阻尼奖励和终端奖励集进行训练。

C. 智能体训练配置

我们的策略采用全连接神经网络实现，包含2个隐藏层，每层有64个神经元，并使用ReLU激活函数。我们的PPO实现部分参考了Brittain and Wei的参数 [12]，其中学习率为$l_r = 0$。0001，折扣因子为 $\gamma = 0$。99，熵系数为 $\beta = 0$。01。场景回合在经过300个时间步（30分钟模拟时间）后终止。每个场景总计训练200万时间步。在训练横向避让策略时，我们采用一种简单的基于课程的方法：首先对策略进行200万步预训练，暂不纳入安全奖励，以学习简单的导航至出口。随后纳入安全奖励，并继续训练3000万时间步，以适应任务难度的增加。

D. 在线动作堆叠

在线动作堆叠利用了公式7中所述的动作阻尼惩罚。该惩罚缓慢衰减，促使智能体以突发 方式连续执行动作，也就是说，在单个决策点通过一系列连续指令实现预期的航向变化，从而将累积惩罚降至最低。

为实现动作堆叠，当策略持续对同一架航空器发出指令时，我们会反复查询策略；将相同的 10 度增量（向左或向右）累积为一条宏指令，并在策略输出“无动作”或对另一架航空器采取动作时停止堆叠。我们保持底层训练马尔可夫决策过程不变。在推理时，我们会在单个物理时间步内多次查询策略，同时更新指令航向，而环境状态（位置、速度）保持不变。每次中间查询都以完全可观测的状态为条件，该状态包含更新后的指令航向，因此底层过程的马尔可夫结构得以保留。

不同于依赖预定义宏动作或结构化分解的分层式或时间抽象强化学习方法， [30, 35], 在线动作堆叠会在运行时根据策略的输出动态编排原始动作。该方法在训练期间将每架航空器的动作空间维持在仅包含三个离散动作的精简规模，而非使用原本所需的数十种航向指令，因此具有显著的计算优势 [36]。这种方法尤其适用于空中交通管制应用，因为这些应用在实际操作中要求使用特定的航向值（而不是

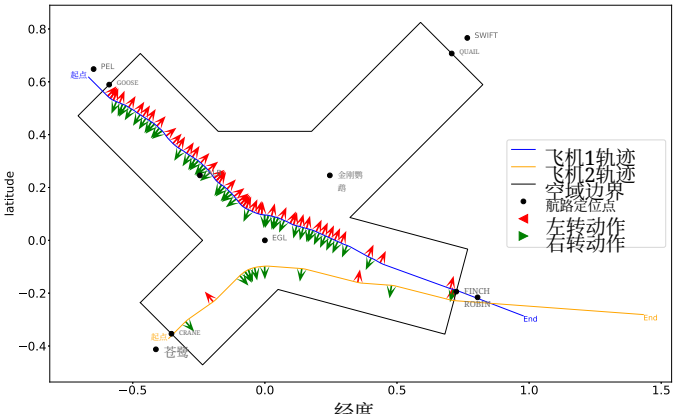

图 4 未施加动作阻尼惩罚且未采用动作堆叠时，策略控制 2 架航空器通过 X-Plus扇区的情况。总共执行了 92 个独立动作。

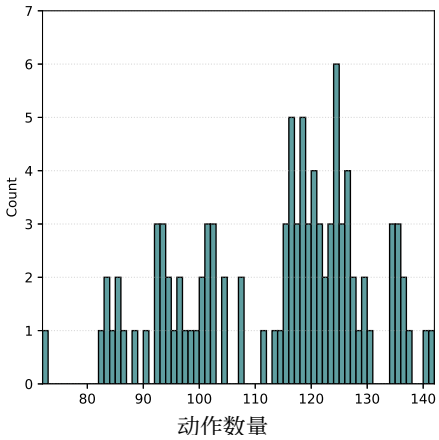

图 5 使用无阻尼横向导航策略进行 2 架航空器导航时，样本中 100 个回合内每回合采取的动作数量。观测到动作均值为 113.0，标准差为 15.9。

仅使用相对变化） [17], , 并且多架飞机之间的协调需要精确安排指令的发送时机 [12]。

此外，通过在不推进仿真时间的情况下组合动作，我们避免了宏动作在多个未观测时间步中执行所引入的额外部分可观测性。在传统的时间抽象方法中，宏动作跨多个时间步执行时，环境会持续演变，但智能体在宏动作完成前无法观测中间状态，也无法调整决策。这会造成显著的部分可观测性，因为环境状态在执行过程中可能发生不可预测的变化。我们的方法则通过在环境推进之前的单个决策点组合多个原始动作，避免了这一问题。

III. 结果与讨论

为有效测试并展示我们的贡献，我们训练了一系列广泛的实验策略。以下小节将介绍并考察这些策略。

A. 无动作阻尼导航

图4展示了一种策略的行为，该策略使用公式6中所述的横向中心线距离奖励以及为成功导航至航线出口且不偏离航线而设定的终端奖励，训练了200万时间步。由于没有对执行多个动作施加惩罚，该策略几乎每个时间步都会发出一个动作，因此每回合的动作总数非常高。图5展示了100个样本回合中每回合执行的动作数量，该策略的动作均值为113。除了动作频率很高之外，如图4所示，该策略在尝试与航线中心线保持最佳距离时，还频繁地在左右控制动作之间来回振荡。这种行为阻碍了在线动作堆叠的应用，因为我们要求同一动作能够连续重复执行。

B. 阻尼导航与动作堆叠

为了解决未采用阻尼的策略所存在的不理想特征，我们训练了一个经过阻尼的策略，并应用了在线动作堆叠技术。公式7中所述动作阻尼惩罚的引入，大幅减少了策略发出的动作数量，如图6所示。现在，策略发出稀疏指令，仅在航线上的关键拐点处调整航空器的横向轨迹。在抽取的100次运行中，该策略每回合平均发出14.5个动作，与未采用阻尼的策略相比减少了87%。然而可以看出，要完成较大的转弯，仍需要多条指令；而在实际空中交通管制中，通常会用一条复合指令取而代之。

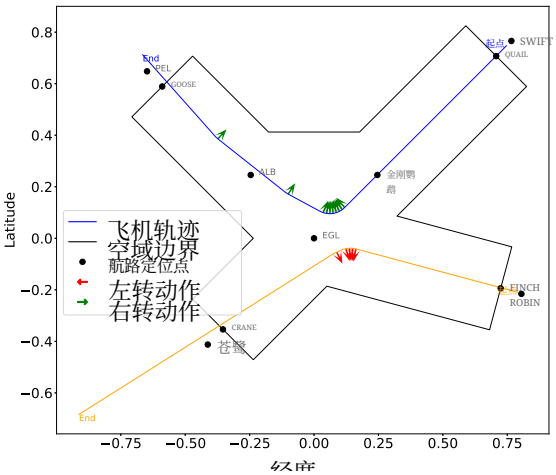

图 6 使用动作阻尼惩罚且不采用动作堆叠时，策略控制 2 架航空器穿越 X-Plus 扇区。该策略总共发出了 15 个独立动作。

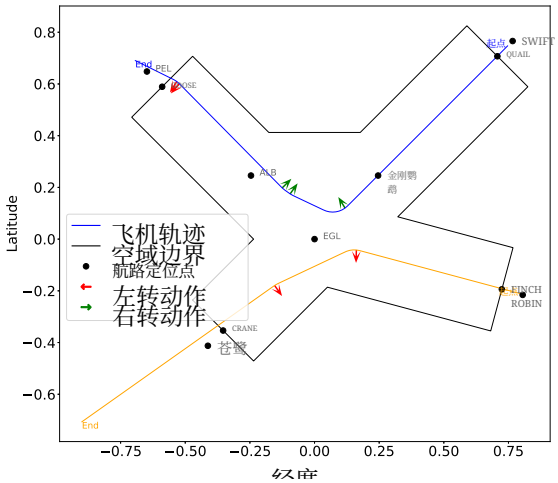

图 7 使用动作阻尼惩罚和动作堆叠时，策略控制 2 架航空器穿越 X-Plus 扇区。该策略总共发出了 7 个独立动作。

在采用经过阻尼的策略时，最北侧航空器在EGL附近进行的中央转弯由七条重复的“右转10度”动作组成。图7所示的在线动作堆叠技术有效解决了运行真实性问题。在此，中央航线转弯被合并为一条“右转70度”指令。图8和图9展示了动作堆叠显著降低动作频率的效果。在抽取的100个回合中，每回合动作数量平均减少 ~50 %。动作阻尼与动作堆叠相结合，将发出的动作数量降至符合实际运行需求的水平。

C. 大动作空间下的横向导航

图 10 展示了一个策略示例，该策略采用 II.B.2 节所述的更大横向动作空间（37 个动作）进行训练。经过相同数量的时间步训练后，该策略学会了有效地导航至出口。然而，如更大的动作空间所示，这会导致转弯效率低下和过度修正。虽然进一步训练可能会解决这一问题，但这正是采用在线动作堆叠的一个关键动因：鉴于完整空中交通管制任务的复杂性极高，在实现相当性能的同时缩短训练时间，是一项非常理想的特征。图 11 表明，该策略每回合发出的动作平均数量较少，均值为 6.3 个。尽管大动作空间策略的平均动作数量更低（6.3 对 7.2），但这一差异在统计上并不显著（标准差 ≈ 3）。至关重要的是，堆叠方法只需提供少得多的可用动作，就能实现相当的运行保真度，从而降低复杂空中交通管制建模中常见的“维数灾难”：5 维动作空间的表现与使用 37 维动作空间训练的策略相近。

图 12 和图 13 展示了阻尼横向导航策略和采用较大动作空间的导航策略在训练期间的奖励。可以看出，较小动作空间的收敛速度确实更快，不过对于这个较简单的问题，两种策略训练后都达到了相近且合理的收敛水平和奖励。成功率反映了智能体满足第 II.B.3 节所述全部 3 项终止奖励标准的频率：成功导航至出口、没有偏离空域，以及发出的动作少于 30 个。可以看出，大动作空间策略要晚得多才能稳定取得成功。检查各项指标的具体数值后发现，尽管理论上完成该场景所需的动作数量更少，但导致大动作空间策略成功率下降的是动作数量这一标准。鉴于这种空中交通管制任务建模方式相对简单，我们预计，在尝试采用广阔动作空间实现更复杂的方案时，将面临显著的扩展挑战。

这些结果表明，通过应用在线动作堆叠，有可能以较小的动作空间近似达到解决方案的保真度和效能。因此，可以将训练复杂度重新分配到处理更高的交通密度或更稳健的安全定义上，而不会产生高昂的计算成本。

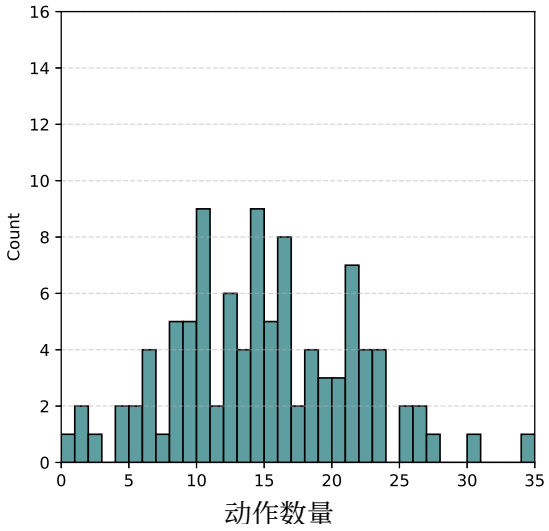

图 8 使用阻尼横向导航策略控制 2 架航空器时，100 个回合样本中每回合采取的动作数量。观察到动作数量均值为 14.5，标准差为 6.6。

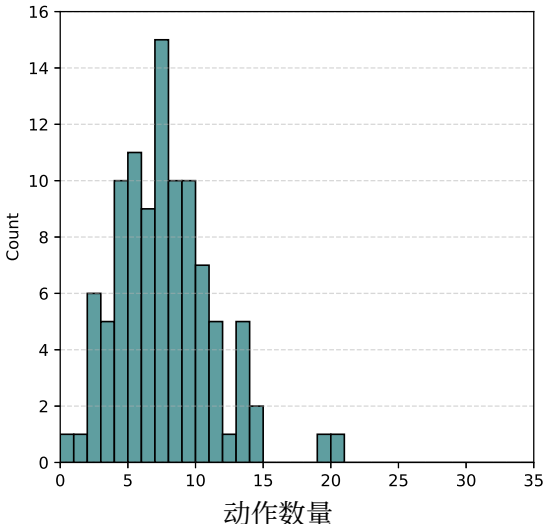

图 9 使用阻尼且堆叠的横向导航策略控制 2 架航空器时，100 个回合样本中每回合采取的动作数量。观察到动作数量均值为 7.2，标准差为 3.5。

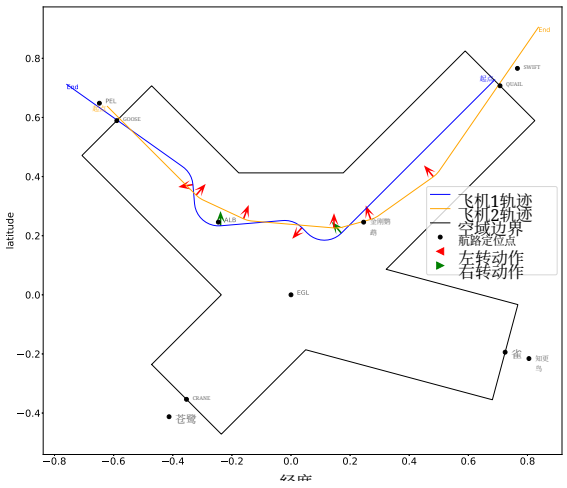

图 10 使用广阔动作空间的策略控制 2 架航空器通过 X-Plus 扇区。

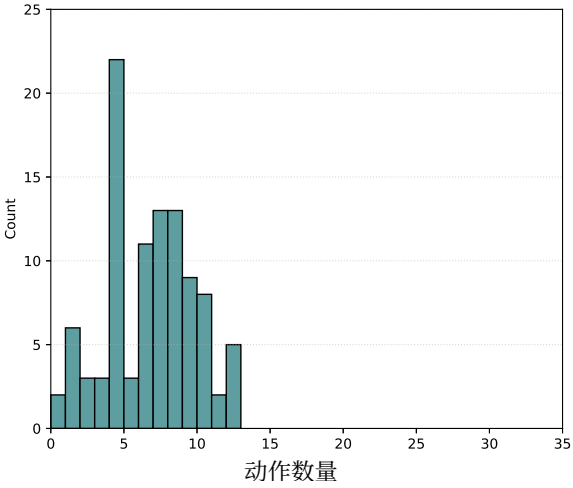

图 11 使用大动作空间的横向导航策略控制 2 架航空器时，100 个回合样本中每回合采取的动作数量。观察到动作数量均值为 6.3，标准差为 3.0。

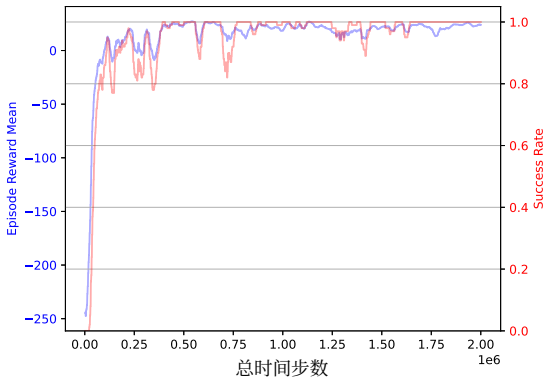

图 12：2 架航空器采用阻尼横向导航策略时的奖励收敛情况。根据第 II.B.3 节终止奖励中定义的成功标准监测任务是否成功。

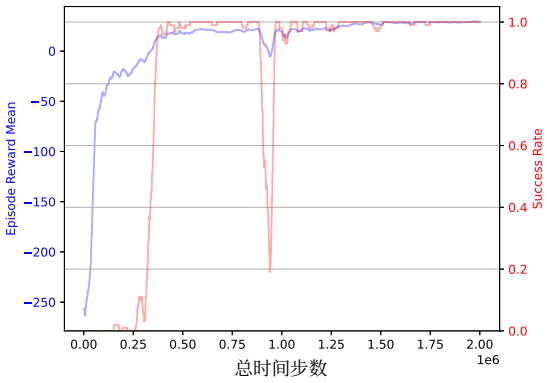

图 13：2 架航空器采用大动作空间横向导航策略时的奖励收敛情况。根据第 II.B.3 节终止奖励中定义的成功标准监测任务是否成功。

D. 垂直控制的动作堆叠

垂直控制是动作堆叠的一种自然延伸，因为在空中交通管制运行中，航空器飞行高度层经常发生大幅变化。为研究这一点，我们训练了一个简单策略，使其引导航空器飞往目标飞行高度层；训练采用方法部分讨论的奖励和状态配置，持续 200 万个时间步。图 14 和图 15 展示了在垂直领域应用动作堆叠的效果：原本需要连续发出 15 条指令才能实现的垂直指令，被整合为一条符合实际运行要求的爬升指令。将在线动作堆叠 扩展到垂直领域尤其值得关注，因为它为解决空中交通管制动作空间中一个原本极具挑战性的方面提供了方案。航空器性能限制意味着，不同航空器在垂直方向上的可行动作各不相同，同时还存在高维性问题。高性能喷气式航空器的运行高度层范围可能高于飞行高度层 400；由于有效飞行高度层以 10 为间隔，仅一架航空器就会产生 40 维动作空间。应用在线动作堆叠 有望将其缩减至仅 2 维。

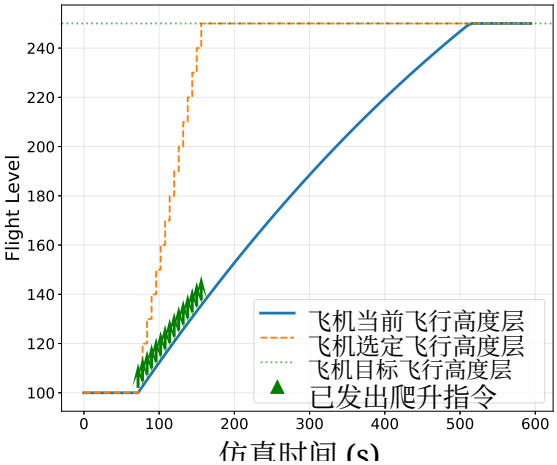

图 14：未采用堆叠的垂直控制策略执行爬升。可以看到，控制指令序列在每个仿真步都提高了航空器所选的飞行高度层。

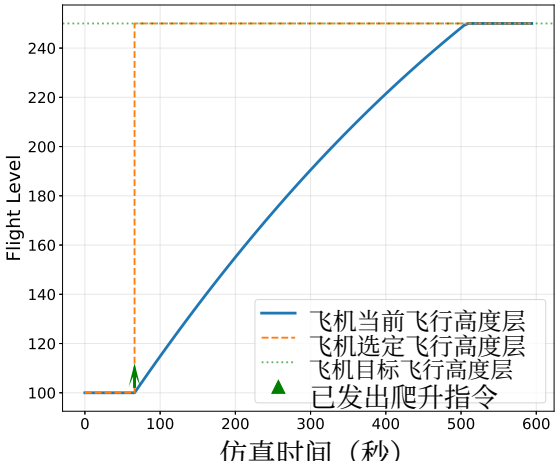

图 15：采用动作堆叠的垂直控制策略，在动作启动时立即获得飞往目标飞行高度层的许可。本例中，15 条“提高飞行高度层”指令被合并为一条命令。

E. 横向导航与避障的动作堆叠

加入横向避让行为会显著增加控制任务的复杂性。方程 8 中描述的安全性设定鼓励采用构成空中交通管制基础的“故障安全”程序。图 16 和图 17 展示了经过训练、用于实现防撞的阻尼策略和堆叠策略，它们控制两架航空器沿相向航线飞行。可以看出，和之前一样，堆叠策略有效地整合了中间的转弯动作。在此示例中，阻尼策略发出了 15 个动作，而堆叠策略仅发出了 7 个，其中包括两条指令：“左转 40 度”和“右转 70 度”。此外，通过引导航向使一架航空器向北飞行，同时让另一架进一步向南飞行，从而实现了有效间隔。在这两种场景中，最小间隔均达到 18 海里，远超 5 海里的最低要求。

我们提出在线动作堆叠的核心依据之一是：使用大动作空间时，横向导航与避让等更复杂的任务会更难训练。因此，除了使用方程 3 所述的标准增量动作空间训练横向导航与避让策略外，我们还使用由方程 4 推导出的大动作空间训练了第二个横向导航与避让策略（每架航空器有 19 个动作，两架航空器共有 37 个动作）。图 18 和图 19 展示了这两种策略在各自抽取的 100 个回合中实现的最小间隔。堆叠策略仅发生了一次间隔丧失（小于 5 海里），而大动作空间策略则发生了九次间隔丧失。此外，我们报告的堆叠策略每回合动作均值为 81.2，标准差为 13.2；大动作空间策略每回合发出的动作均值为 126.3，标准差为 12.3。这直接支持了我们的观点：在固定训练预算下，较小的动作空间能让复杂的空中交通管制任务得到更有效的训练，因为阻尼／堆叠策略在动作稀疏性和间隔性能方面均更优。两种策略的训练步数都超过了 3000 万步（在现代服务器上约需 60 小时），因此尽管总可以继续训练，但如果能在不损失保真度的情况下使用较小的动作空间，我们就能获得明显优势。

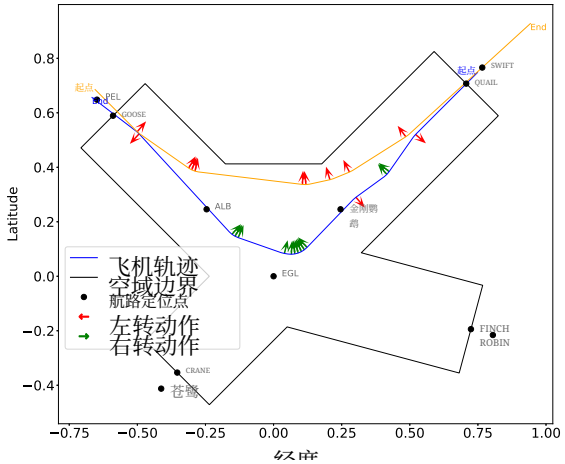

图 16 阻尼式横向导航与避让策略控制 2 架航空器。

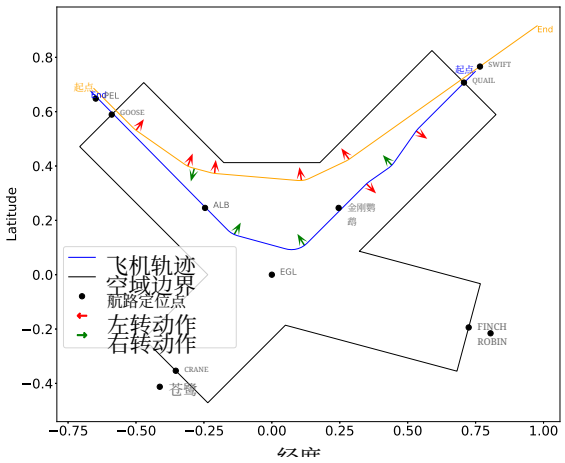

图 17 堆叠式横向导航与避让策略控制 2 架航空器。

IV. 结论

我们提出了一种新方法，将在线动作堆叠与突发动作激励相结合，以实现动作稀疏性、复现大得多的动作空间中的复合指令，并直接应对空中交通管制强化学习训练所面临的挑战。该技术能够实现逼真的管制员式指令，同时保持训练效率。对于横向导航，我们已证明，使用 5 维动作空间进行在线动作堆叠，可以达到与使用 37 维动作空间相当的效果。

我们目前在双航空器避让和导航方面的工作取得了令人鼓舞的成果。未来，我们将扩展应用，将横向和垂直控制与冲突消解相结合，从而全面处理空中交通管制的基本任务，并填补现有文献中的一个重大空白。

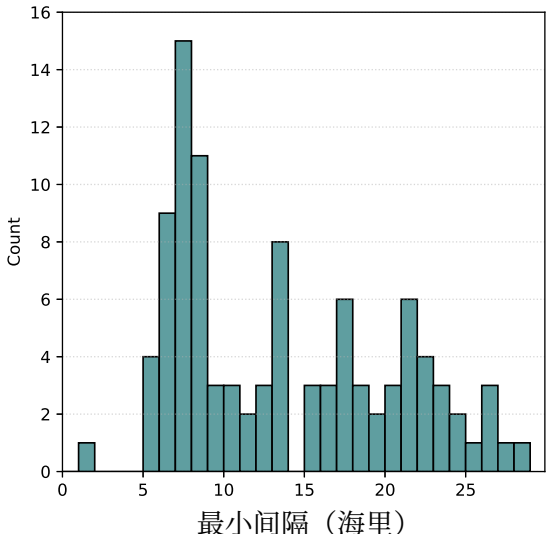

图 18 堆叠式横向导航与避碰策略每回合实现的最小间隔。

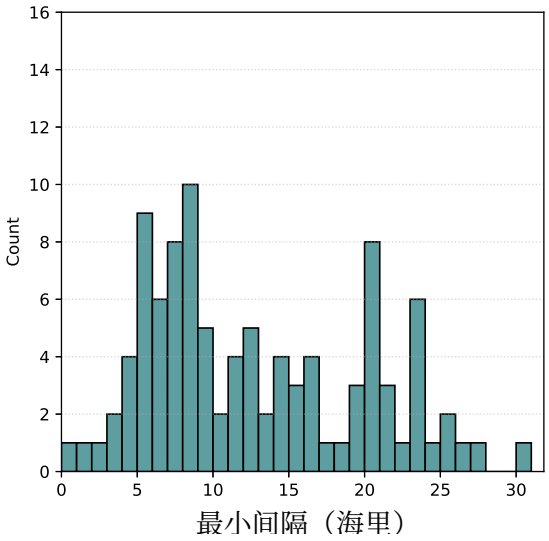

图 19 大动作空间横向导航与避碰策略每回合实现的最小间隔。

致谢

本研究得到 EPSRC（EP/V056522/1）和 NATS 的支持。

V. 参考文献

[1] IATA，《行业统计》，2025年。网址：https://www.iata.org/en/iata-repository/pressroom/fact-sheets/industry-statistics/，访问日期：19-05-2025。[2] ICAO，《*ICAO 2026—2050*年战略计划》，2024年。网址：https://www.icao.int/sites/default/files/about-icao/ StrategicPlan/ICAO-Strategic-Plan-2026-2050-WEB-F5.pdf。[3] 交通部，《“喷气式飞机零排放”战略：到*2050*年实现航空业净零排放》，2022年。网址：https://assets.publishing. service.gov.uk/media/62e931d48fa8f5033896888a/jet-zero-strategy.pdf。[4] NATS，《英国空中交通管制中的*iFACTS*：飞行员须知》，2014年。网址：https://www.ukfsc.co.uk/files/External%20Meetings/ NATS%20SPA%202012/NATS%20SPA%20iFacts%20May%202014.pdf，访问日期：19-05-2025。[5] Indra，《*NATS*与*Indra*将在普雷斯特威克中心上层空域部署*iTEC*系统》，2014年。网址：https://www.indracompany. com/pt-br/node/48013，访问日期：19-05-2025。[6] Nuic, A.、Poles, D. 和 Mouillet, V.，“BADA：面向当前及未来空中交通管理系统的先进航空器性能模型”，国际自适应控制与信号处理期刊，第24卷，第10期，2010年，第850–866页。https: //doi.org/10.1002/acs.1176。[7] Wesson, R. B.，“空中交通管制员世界中的规划”，第五届人工智能国际联合会议论文集——第*1*卷，Morgan Kaufmann Publishers Inc.，美国加利福尼亚州旧金山，1977年，第473–479页。[8] Eurocontrol，《*ARGOS*情况说明书》，2023年。网址：https://www.eurocontrol.int/publication/argos-factsheet，访问日期：19-05-2025。[9] SESAR，《*SESAR*联合事业 | *AGENT*项目最终成果报告》，2018年。网址：https://www.sesarju.eu/node/3183，访问日期：26-06-2024。[10] SESAR，《*SESAR*联合事业 | *HYPERSOLVER*——能够在单一框架内管理空中交通管制（*ATC*）和空中交通流量管理（*ATFM*）的人工智能控制器》，2023年。网址：https://www.sesarju.eu/projects/hypersolver，访问日期：26-06-2024。[11] Brittain, M. 和 Wei, P.，“基于分层深度强化学习的自主航空器排序与间隔”，国际航空运输研究会议，西班牙卡斯特尔德费尔斯，2018年。

[12] Brittain, M. 和 Wei, P.，“自主空中交通管制员：一种深度多智能体强化学习方法”，国际机器学习会议：*ICML 2019* 现实生活强化学习研讨会，2019年。网址：http://arxiv.org/abs/1905. 01303。[13] Brittain, M. 和 Wei, P.，“一对多：结合深度多智能体强化学习与长短期记忆的分布式冲突消解”，*AIAA Scitech 2021* 论坛，2021年。https://doi.org/10.2514/6.2021-1952。[14] Wang, Z.、Li, H.、Wang, J. 和 Shen, F.，“基于深度强化学习的空中交通管制冲突探测与解脱”，*IET*智能交通系统，第13卷，第6期，2019年，第1041–1047页。https://doi.org/10.1049/iet-its.2018.5357。[15] Sui, D.、Ma, C. 和 Dong, J.，“基于深度强化学习的空中交通管理冲突消解策略”，航空，第27卷，第3期，2023年，第177–186页。https://doi.org/10.3846/aviation.2023.19720。[16] Sui, D.、Ma, C. 和 Wei, C.，“利用深度强化学习辅助空中交通管制员的战术冲突求解器”，航空航天，第10卷，第2期，2023年，第182页。https://doi.org/10.3390/aerospace10020182。[17] Dalmau, R. 和 Allard, E.，“利用消息传递神经网络和多智能体强化学习进行空中交通管制”，第十届*SESAR*创新日（*SID*），2020年。[18] Vouros, G.、Papadopoulos, G.、Bastas, A.、Cordero, J. M. 和 Rodrigez, R. R.，“飞行冲突消解自动化：深度强化学习服务于空中交通管制员”，人工智能与应用前沿，A. Passerini 和 T. Schiex 编，IOS Press，2022年。https://doi.org/10.3233/FAIA220066。[19] Vinyals, O.、Babuschkin, I.、Czarnecki, W. M.、Mathieu, M.、Dudzik, A.、Chung, J.、Choi, D. H.、Powell, R.、Ewalds, T.、Georgiev, P.、Oh, J.、Horgan, D.、Kroiss, M.、Danihelka, I.、Huang, A.、Sifre, L.、Cai, T.、Agapiou, J. P.、Jaderberg, M.、Vezhnevets, A. S.、Leblond, R.、Pohlen, T.、Dalibard, V.、Budden, D.、Sulsky, Y.、Molloy, J.、Paine, T. L.、Gulcehre, C.、Wang, Z.、Pfaff, T.、Wu, Y.、Ring, R.、Yogatama, D.、Wünsch, D.、McKinney, K.、Smith, O.、Schaul, T.、Lillicrap, T.、Kavukcuoglu, K.、Hassabis, D.、Apps, C. 和 Silver, D.，“利用多智能体强化学习在星际争霸II中达到宗师级水平”，*Nature*，第575卷，第7782期，2019年，第350–354页。https://doi.org/10.1038/s41586-019-1724-z。[20] Silver, D.、Huang, A.、Maddison, C. J.、Guez, A.、Sifre, L.、van den Driessche, G.、Schrittwieser, J.、Antonoglou, I.、Panneershelvam, V.、Lanctot, M.、Dieleman, S.、Grewe, D.、Nham, J.、Kalchbrenner, N.、Sutskever, I.、Lillicrap, T.、Leach, M.、Kavukcuoglu, K.、Graepel, T. 和 Hassabis, D.，“用深度神经网络和树搜索掌握围棋”，*Nature*，第529卷，第7587期，2016年，第484–489页。https://doi.org/10.1038/nature16961。[21] Berner, C.、Brockman, G.、Chan, B.、Cheung, V.、Dębiak, P.、Dennison, C.、Farhi, D.、Fischer, Q.、Hashme, S.、Hesse, C. 等，“大规模深度强化学习玩Dota 2”，*arXiv*预印本 *arXiv:1912.06680*，2019年。[22] Smith, L.、Kostrikov, I. 和 Levine, S.，“展示公园漫步：利用无模型强化学习在20分钟内学会行走”，机器人学：科学与系统（*RSS*）演示，第2卷，第3期，2023年，第4页。[23] Sutton, R. S. 和 Barto, A. G.，《强化学习：导论，第*2*版》，强化学习：导论，第2版，麻省理工学院出版社，美国马萨诸塞州剑桥，2018年。[24] Bellman, R.，“马尔可夫决策过程”，数学与力学杂志，第6卷，第5期，1957年，第679–684页。网址：https://www.jstor.org/stable/24900506。[25] Schulman, J.、Wolski, F.、Dhariwal, P.、Radford, A. 和 Klimov, O.，“近端策略优化算法”，2017年。网址：http://arxiv.org/abs/1707.06347。[26] Ouyang, L.、Wu, J.、Jiang, X.、Almeida, D.、Wainwright, C. L.、Mishkin, P.、Zhang, C.、Agarwal, S.、Slama, K.、Ray, A.、Schulman, J.、Hilton, J.、Kelton, F.、Miller, L.、Simens, M.、Askell, A.、Welinder, P.、Christiano, P.、Leike, J. 和 Lowe, R.，“训练语言模型遵循指令并利用人类反馈”，第*36*届神经信息处理系统国际会议论文集，Curran Associates Inc.，美国纽约州雷德胡克，2022年。[27] Hoekstra, J. 和 Ellerbroek, J.，“BlueSky空管模拟器项目：一种开放数据与开源方法”，第七届航空运输研究国际会议，第7卷，费城，2016年，第132页。[28] Biedenkapp, A.、Rajan, R.、Hutter, F. 和 Lindauer, M.，“TempoRL：学习何时采取行动”，第*38*届国际机器学习会议论文集，PMLR，2021年，第914–924页。网址：https://proceedings.mlr.press/v139/biedenkapp21a.html。[29] Carvell, B.、Thomas, M.、De Ath, G.、Everson, R.、Pepper, N.、Keane, A. 和 Cannon, R.，《基于受监管评估框架的人在回路测试：面向空中交通管制的人工智能智能体》，AIAA，2026年。

[30] Sutton, R. S.、Precup, D. 和 Singh, S.，“MDP 与半马尔可夫决策过程之间：强化学习中的时间抽象框架”，人工智能，第 112 卷，第 1 期，1999 年，第 181–211 页。https://doi.org/10.1016/S0004-3702(99)00052-1。[31] Sharma, S.、Lakshminarayanan, A. S. 和 Ravindran, B.，“学会重复：深度强化学习中的细粒度动作重复”，*5 2017, 24 26,* 第十 *2017,* 2017 = 1 5 七届国际学习表征会议，*ICLR*，法国土伦，*2020* 年 *4* 月——会议论文集，OpenReview.net，。URL https://openreview.net/forum?id B GOWV eg。[32] CAA，*CAP 493*：《空中交通服务手册》（*MATS*）第 *1* 部分——民航局，2024年。URL https://www.caa.co.uk/our- work/publications/documents/content/cap-493/，访问日期：19-05-2025。[33] 艾伦·图灵研究所，*Project Bluebird*：用于空中交通管制的人工智能系统，2025年。URL https://www.turing.ac.uk/research/ research-projects/project-bluebird-ai-system-air-traffic-control，访问日期：19-05-2025。[34] Pepper, N.、Keane, A.、Hodgkin, A.、Dewi Gould, E. H.、Lauritsen, L.、Vlahos, C.、De Ath, G.、Everson, R.、Cannon, R.、Castro, A. S.、Korna, J.、Carvell, B. 和 Thomas, M.，用于训练和评估空中交通管制人工智能智能体的英国航路空域概率数字孪生，AIAA，2026年。[35] Asadi, M. 和 Huber, M.，“分层强化学习中的状态空间缩减”，第十七届国际佛罗里达人工智能研究学会会议论文集，美国佛罗里达州迈阿密海滩，2004 年。[36] Kanervisto, A.、Scheller, C. 和 Hautamäki, V.，“深度强化学习中的动作空间塑形”，*2020* 年 *IEEE* 游戏会议（*CoG*），2020年，第 479–486 页。https://doi.org/10.1109/CoG47356.2020.9231687。