arXiv:2303.00638v3 [cs.LG] 2 May 2024

MEGA-DAgger：与多个不完美专家进行模仿学习

Xiatao Sun[1*]、Shuo Yang[2*]、Mingyan Zhou[2]、Kunpeng Liu[2]、Rahul Mangharam[2]

摘要——得益于交互式算法的最新发展，模仿学习已广泛应用于各种自主系统。这些算法能够解决传统方法（如行为克隆）所导致的协变量偏移和累积误差。然而，现有的交互式模仿学习方法假设能够获得一位完美专家，而现实中更可能拥有多个不完美专家。本文提出 MEGA-DAgger，这是一种适用于与多个不完美专家进行交互式学习的新型 DAgger变体。首先，在汇总训练数据时会过滤掉不安全示范，从而使不完美示范在训练新手策略时影响甚微。接下来，根据特定场景的指标评估并比较专家，以解决专家之间相互冲突的标签。通过在自主赛车场景中的实验，我们证明，使用 MEGA-DAgger 学得的策略能够胜过专家，以及使用最先进的交互式模仿学习算法（如 Human-Gated DAgger）学得的策略。补充视频见 https://youtu.be/wPCht31MHrw。

I. 引言

基于学习的控制方法已成功应用于复杂的机器人系统中 [1]。模仿学习 [2]–[5], 适用范围广，只需收集易于大规模获取的专家示范即可。在各种模仿学习技术中, DAgger [6]和 Human-Gated DAgger (HG-DAgger) [7]等交互式方法日益流行，因为它们可以解决朴素行为克隆 [8]所导致的协变量偏移和累积误差。

交互式模仿学习 (IL) [6, 7, 9]–[12] 本质上是在新手策略训练期间间歇性地获取专家反馈。例如，DAgger 使用专家策略和新手策略提供的混合标签来训练新手。为了从人类专家那里学习更有效的策略，HG-DAgger 在 DAgger 的基础上引入了人类门控函数，以决定何时由专家接管。SafeDAgger [9], 等机器人门控方法允许机器人主动询问人类专家，并仅在必要时请求干预。尽管如此，交互式 IL 方法
通常有两个关键假设：
1）专家示范是完美的；2）所有示范都来自单一专家。
第一个假设在现实中很少成立，因为人类专家通常会犯错。例如，在2021年，

[1] Xiatao Sun 就职于美国康涅狄格州纽黑文市耶鲁大学计算机科学系，邮编 06510（电子邮箱：xiatao.sun@yale.edu）。[2] Shuo Yang、Mingyan Zhou、Kunpeng Liu 和 Rahul Mangharam 就职于美国宾夕法尼亚州费城宾夕法尼亚大学电气与系统工程系，邮编 19104（电子邮箱：{yangs1、derekzmy、kunpengl、rahulm}@seas.upenn.edu）。
*作者贡献相同。

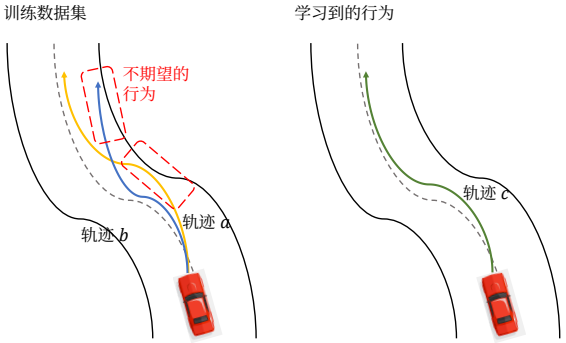


图 1：从多个不完美专家处学习的示意图。例如，左图中的两个回合 a 和 b （黄色和蓝色轨迹）分别来自两位不同的专家。每位专家都存在不理想的不安全行为（红框所示）；理想情况下，我们可以从两者那里学习互补的良好行为（右图中的绿色轨迹 c）。

宾夕法尼亚州每天发生323起人类驾驶车辆事故 [13]。驾驶员错误占所有交通事故的85%以上，这意味着人类专家有时作为驾驶员并不可靠。

当有多位专家试图教导新手时，第二个假设同样不成立。例如，不同驾驶员可能有不同的驾驶策略。有些人可能驾驶风格高效但激进，而另一些人可能选择安全但保守的风格。如果同时从不同专家那里学习策略，不同专家提供的示范标签可能会相互冲突。因此，存在多位专家的情形需要模仿学习中的新技术。如图 1 所示，来自两位专家的两组示范都不完美（左图），我们希望从中学习良好的行为（右图），并避免不期望的行为。

在本研究中，我们着力解决以下问题：如何从多个不完美专家中进行交互式学习？为此，我们提出了多专家门控 DAgger（MEGA-DAgger），这是一种专为从多个不完美专家中学习而设计的DAgger变体。具体而言，我们针对端到端自动驾驶竞速研究MEGA-DAgger，其中安全性和进度都至关重要。我们提出了一种基于控制屏障函数的安全评分器，并在聚合新数据时过滤掉不安全的专家示范。在运行时，为解决多个专家之间的动作冲突，我们根据安全分数和进度分数评估每位专家，并选择表现最佳的专家。实验结果表明，我们提出的解决方案优于现有的DAgger变体，并能够学习到超越专家的策略。需要注意的是，尽管我们的框架目前仅在赛车场景中进行了验证，但它可以轻松地

应用于采用经过修改的特定场景指标的一般自主系统。

A. 相关工作

自动驾驶的模仿学习：先前研究已成功将模仿学习应用于自动驾驶汽车，以实现安全驾驶，参见例如 [4], [14]–[16]。很少有研究采用模仿学习策略来应对自主赛车场景，而这类场景除安全性外，还要求维持性能。[17]提供了一个基于模仿学习的自主赛车基准。然而，这些方法都假定只有一位专家，无法处理多位专家的情况。此外，它们还假定专家示范是完美的，而本文放宽了这一假设。

基于不完美专家的模仿学习：过去几年也有研究探讨如何从不完美示范中学习，参见例如 [18]–[21]。这些研究侧重于使用逆强化学习，从单一专家的不完美示范中学习。相较之下，本文研究面向安全关键型自主系统、由多个不完美专家参与的交互式模仿学习。

多位专家的模仿学习：当存在多位专家时，一个自然的问题是：我们应该选择哪位专家？在假设预先提供带标签数据集的分类场景中，已有研究探讨了这一问题，参见例如 [22, 23]。然而，在我们的工作中，针对复杂自主系统的训练过程是与无标签专家交互进行的，因此现有文献中的方法并不适用。

本文的贡献总结如下：

1）我们提出了多专家门控DAgger（MEGA-DAgger），这是一种新的DAgger变体，用于从多个不完美专家中进行交互式模仿学习；2）我们开发了一种数据过滤器，可以有策略地截断不期望的示范。与现有的DAgger变体相比，该方法显著提高了自主赛车场景中的安全性；3）我们提供了用于评估每位专家的指标，并通过实验证明，MEGA-DAgger能够学习出性能优于专家以及使用HG-DAgger（最先进的交互式模仿学习算法）所学习策略的策略。

II. 背景

A. DAgger与HG-DAgger

DAgger 是一种交互式模仿学习算法，通过同时运行专家策略和新手策略来聚合新数据 [6]。具体而言，在每次整合i时，新的训练数据 $\mathcal{D}_i$ 由以下方式生成：

$$\pi_i(x_t) = \phi_i \pi_{exp}(x_t) + (1-\phi_i)\pi_{N_{i-1}}(x_t), \qquad (1)$$

其中，x_t 是时间步 t的状态，$\phi_i \in [0,1]$ 是概率，π_{exp} 是专家策略，$\pi_{N_{i-1}}$ 是第 $i-1$次迭代训练得到的新手策略。接下来，可以通过 $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{D}_i$聚合数据集，并在 $\mathcal{D}$ 上训练新策略 π_{N_i}。通过允许新手策略影响采样状态分布，DAgger 可以缓解行为克隆 [8]导致的协变量失配和累积误差。

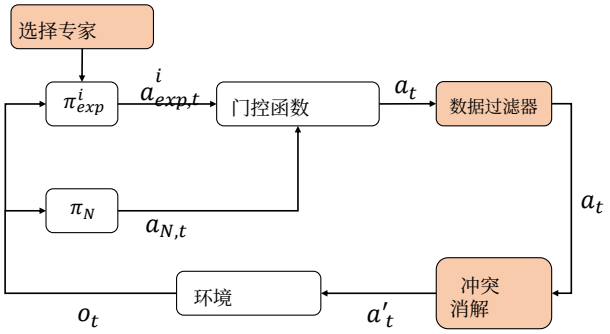


图 2：MEGA-DAgger 的控制循环。每次迭代都应选出一位专家作为主导专家。数据过滤器用于移除不安全示范，冲突消解用于消除专家之间的动作冲突。

人类门控 *DAgger*（*HG-DAgger*）是一种 DAgger 变体，旨在成为一种更适合从人类专家 [7]处学习的交互式算法。它的主要区别在于提出了新的轨迹生成方法：

$$\pi_i(x_t) = g(x_t)\pi_{exp}(x_t) + (1-g(x_t))\pi_{N_{i-1}}(x_t), \qquad (2)$$

其中，$g(x_t)$ 是一个门控函数：若状态 x_t 安全，其值为 0；若状态 x_t 不安全，其值为 1。HG-DAgger 假设专家是最优的，并且拥有关于当前状态安全性的特权信息。因此，只有观察到新手策略执行不安全行为时，专家才会接管控制。与 DAgger 相比，HG-DAgger 能够提高样本效率和训练稳定性。

与 DAgger 和 HG-DAgger 不同，MEGA-DAgger 面向多位非最优专家的情形。如图 2 所示，其整体循环与 HG-DAgger 类似，但增加了三个模块（以阴影颜色标示）：每次迭代中，首先选出一位专家作为主导专家；然后，数据过滤器会移除不安全示范，因为专家并非最优，其行为也可能不安全；最后，需要消解多位专家之间的冲突。

B. 使用自主赛车基准进行评估

本文专门研究一对一竞赛中自主赛车的智能体策略 [24]。自主赛车提供了一组定义明确的指标，用于平衡安全性与性能，因此是适合开展研究的场景。赛车竞赛为智能体设定了显而易见的目标——领先其他智能体；如果智能体发生碰撞，也会受到明确的惩罚。不同规模的赛车广泛用于研究和竞赛，例如全尺寸的 Indy Autonomous Challenge [25], Roborace [26], 、中等尺寸的 Formula Student Driverless [27],，以及 1/10 尺寸的 F1TENTH [28]。

基于学习的控制方法在自主赛车领域越来越受关注，例如 [29, 30]。模仿学习框架已成功应用于 [17] ，因为模仿学习只需要人类示范，而这些示范很容易收集。在 [17], 中，只考虑一位专家，并假设其表现完美。然而，赛车手的竞赛风格各不相同，而且都不完美。这

促使我们向多个不完美专家学习，并最终获得优于专家的策略。

III. 方法

算法1概述了MEGA-DAgger算法。该算法从一个空的全局数据集D 和一个从所有可能策略类 Π中随机初始化的新手策略 π_{N_0} 开始。与HG-DAgger类似，MEGA-DAgger在每次迭代中逐步将每次轨迹展开 j 产生的专家示范 D_j 纳入 D ，其中$j=\{1,...,M\}$， M是专家的数量。在轨迹展开过程中，新手策略执行推理并控制自车，直到专家发现新手进入不期望区域。此时，专家介入并接管控制，同时为当前观测 o提供动作标签 a。引导自车返回期望区域后，车辆控制权再次交还给新手策略。只有在专家介入并接管控制时，才会记录和收集示范中的观测-动作对。

算法1MEGA-DAgger

过程MEGA-DAGGER($\pi^{1:M}_{exp}$)
  初始化 $\mathcal{D} \leftarrow \emptyset$
  初始化 π_{N_0} 初始化为 Π 中的任意策略
  对于迭代 $i=1:K$ **执行**
    对于轨迹展开 $j=1:M$，使用专家 π^j_{exp} **do**
      对于轨迹展开 j 的时间步 $t \in T$ **执行**
        如果 π^j_{exp} 接管**则**
          $o \leftarrow$轨迹展开$^t_{i,\ j}$
          $a \leftarrow \pi^j_{exp}(o)$
          $D_j \leftarrow o,\ a$
          $D_j,\ \sigma_t \leftarrow$ 数据 过滤(D_j)
        如果结束
      结束循环
      $D_j \leftarrow$ 冲突消解$(D_j,\ D,\ \sigma_t)$
      $D \leftarrow D \cup D_j$
    结束循环
    训练 π_{N_i} 于 $\mathcal{D}$
  结束循环
结束过程

MEGA-DAgger考虑从多个不完美专家 $\pi^{1:M}_{exp}$处学习，这与其他DAgger变体不同，后者假设只能访问一位最优或近似最优专家 [6, 7]。在每次迭代 i中，我们让每位专家 π^j_{exp} 轮流进行观测，并在必要时介入轨迹展开$_{i,\ j}$。

挑战：多个不完美专家的场景带来了两个主要挑战。首先，专家示范可能并不安全。在 HG-DAgger 中， [7], 安全性通过最优专家的干预得到保证。由于在我们的场景中缺少这样的最优专家，不安全示范可能会被纳入训练数据集，这会损害新手策略，并可能导致自主赛车进行推理时发生碰撞。此外，

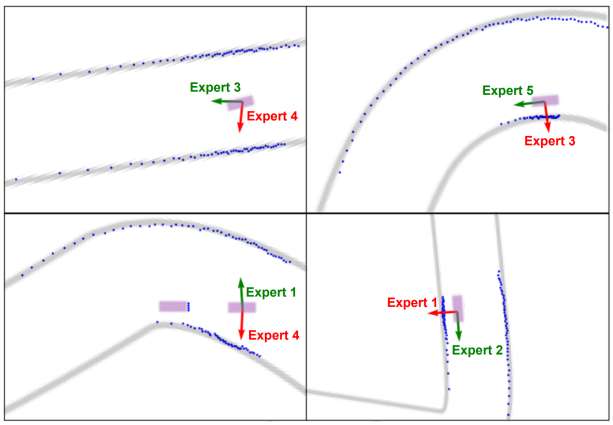


图3：不同专家给出的冲突标签示意图。蓝点表示激光雷达扫描的命中点。红色和绿色箭头表示标签中的转向角单位向量。粉色矩形表示自车和对手车辆。灰线表示赛道边界。

如图3所示，不同专家可能会为来自相邻状态的相似观测提供截然不同的标签，从而在推理过程中干扰使用所学习的新手策略进行的插值。

针对不期望示范的解决方案：为缓解不安全数据带来的挑战，我们设计了一种基于控制屏障函数（*CBF*）[31]的数据过滤器。数据过滤器以D_j 作为输入。它首先检查当前激光雷达观测并获取当前自车位置（x^e_t， y^e_t），然后输出CBF值 h（x^e_t， y^e_t），其定义为

$$h(x^e_t, y^e_t) = (x^e_t - x^p_t)^2 - (y^e_t - y^p_t)^2 - \alpha^2, \quad (3)$$

其中，(x^p_t， y^p_t) 是当前障碍物（例如对手智能体、最近边界点）的位置， α 是相应的最小安全距离。利用离散时间控制屏障函数条件 [32], 的结果，我们将安全分数定义为：

$$\sigma_t = h(x^e_{t+1}, y^e_{t+1}) - (1-\gamma)h(x^e_t, y^e_t), 0 < \gamma \leq 1. \quad (4)$$

请注意， σ_t 值越高表示安全鲁棒性越高[1]。因此，一旦安全分数 σ_t 变为负值，数据过滤器就会立即终止当前轨迹展开。由于自车很可能在进入不安全区域前的几个时间步就已经偏离了期望的超车行为，因此，一旦车辆进入不安全区域，也会截去之前的 β 个时间步，以尽可能移除不期望的示范。

冲突示范的解决方案：为了解决不同专家给出的冲突标签，每次轨迹展开后都会执行一个冲突消解函数 j ，在将其纳入 D之前 D_j 。冲突消解以D_j、 D和 σ_t 作为输入。由于余弦相似度广泛应用于相似度检测，适用于各种传感器（机器人学中常见的传感器包括）余弦相似度可用于识别并选取相似的观测

[1]构造有效的控制屏障函数有时成本很高。在本文中，我们不要求控制屏障函数有效，因为这里的控制屏障函数条件仅用于提供启发式安全分数。

，例如激光雷达扫描 [33, 34] 和 RGBD 相机 [35, 36]。为了有效利用并行处理，通过将 O 和 O_j 的点积除以 O 和 O_j的欧几里得范数逐元素相乘的结果，计算 D_j中的所有观测 O_j 与 D 中的所有观测 O 之间的余弦相似度 Θ：

$$\Theta = \frac{O \cdot O_j}{\|O\| \odot \|O_j\|}.$$

选取 Θ 中高于相似度阈值 ϵ 的元素索引，以从 D 和 D_j 中检索相似示范，用于计算评估分数 ω_t[2]。在自主赛车场景中，每个相似示范的 ω_t 可计算为归一化安全分数（安全性指示）与自车归一化速度（进度指示）之和[2]：

$$\omega_t = \frac{\|\sigma_t\| - \min_t\|\sigma_t\|}{\max_t\|\sigma_t\| - \min_t\|\sigma_t\|} + \frac{\|v_t\| - \min_t\|v_t\|}{\max_t\|v_t\| - \min_t\|v_t\|}.$$

在一组相似示范中，随后使用 ω_t 最高的示范的动作标签，替换所有其他相似示范的动作标签。完成冲突消解后，将 D_j 与 D 合并。最后，在迭代 i 结束时，使用 D 训练策略 $\pi_{N_{i+1}}$。

IV. 实验

本节将进行实验评估，并证明与专家和 HG-DAgger 相比，我们提出的 MEGA-DAgger 在安全性和性能方面均有显著提升。实验代码、rosbag 数据、视频链接等可在 https://github.com/derekhanbaliq/f1tenth-MEGA-DAgger 获取。

A. 实验设置

为了更好地理解 MEGA-DAgger 中各组成部分对利用多个不完美专家进行学习的影响，我们将该方法用于双车竞争式自主赛车场景中的超车行为学习。实验使用二维赛车模拟器 f1tenth gym[3] [28]。模拟器中的每辆车以转向角和速度作为输入，并配备一个二维平面激光雷达，可输出长度为 1080 的激光扫描数据数组。除了激光雷达扫描数据外，模拟器还会在每个时间步提供每辆车的位姿。

作为对比，我们选择 HG-DAgger 作为基线，因为它是最先进的交互式模仿学习算法，而 MEGA-DAgger 正是在此基础上提出的，旨在迈向从不完美专家中学习。在训练和评估期间，每次轨迹展开都会在自车成功超越对手车辆，或自车与对手车辆或其他障碍物发生碰撞时终止。轨迹展开终止后，环境会重置。超车比例和碰撞比例

[2]也可以选择更复杂的分数计算方法，例如根据偏好为安全分数和进度分数分配不同的权重。[3]https://github.com/f1tenth/f1tenth_gym

表 I：超参数取值。新手策略采用一个含 256 个隐藏单元的两层多层感知机（MLP）。

超参数	**值**
神经网络结构	2×256
输入维度[4]	108
评估回合数	100
最小安全距离 α	0.42
截断步长 β[5]	70
余弦相似度阈值 ϵ	0.95

是我们在整个实验中选用的主要评估指标。在评估期间，每个学习策略都会接受 100 次轨迹展开测试。我们分别记录超车或发生碰撞的回合数，并将其除以回合总数来计算相应的比例。

我们采用 F1TENTH ICRA 2022竞赛 [37]中的获胜策略——车道切换器，既将其用作对手车辆，也以其为基础构建专家规划器。车道切换器将赛道划分为两条车道。它会使用纯追踪持续跟踪其中一条车道，直到遇到对手车辆。在这种情况下，它会切换到另一条未被占用的车道，并使用纯追踪跟踪该车道。虽然车道切换器被直接用作对手车辆，但在专家规划器中，车道切换器会以概率 $P(U)$输出反向转向角，以生成不期望的行为，其中U 表示发生不期望行为的事件。通过这种方式，我们可以为 $P(U)$设置不同的随机种子，从而生成多个不同的不完美专家。实验参数列于表 I。

B. 数据过滤器

为了公平评估数据过滤器的效果，我们首先禁用冲突解决函数，仅使用 HG-DAgger 和我们提出的数据过滤器，从一位专家处学习，并设置不同的 $P(U)$，取值范围为 0.1 至 1.0。在每次试验中，新手策略以固定的 $P(U)$ 从一位专家处学习 1000 个回合。作为对比，我们还以相同方式仅使用 HG-DAgger 训练新手策略。为确保对比公平、示范数量相同，在仅使用 HG-DAgger 时，我们会在每个轨迹展开后随机截断示范，使其数量与使用 HG-DAgger 和所提出的数据过滤器时的示范数量相同。不同 $P(U)$ 对应的截断比例见表 II。

如图 4 所示，与使用随机截断的示范进行 HG-DAgger 相比，使用所提出的数据过滤器进行 HG-DAgger 能显著提升性能。然而，随着 $P(U)$ 增大，采用两种方法分别学到的新手策略的性能都会逐渐下降。需要注意的是，由于 $P(U)$ 仅与反转转向角有关，并不能保证发生碰撞，因此专家规划器仍可能通过对手车辆的另一侧，以反转后的转向角成功超越对手车辆，并提供良好的示范。因此，即使 $P(U)$ 等于 1.0

[4]网络并未直接使用长度为 1080 的原始激光雷达扫描数据作为输入，而是采用长度为 108 的均匀下采样激光雷达数组作为输入，以加快训练速度。5

有关选择的更多详情 β 可在附录中找到 .

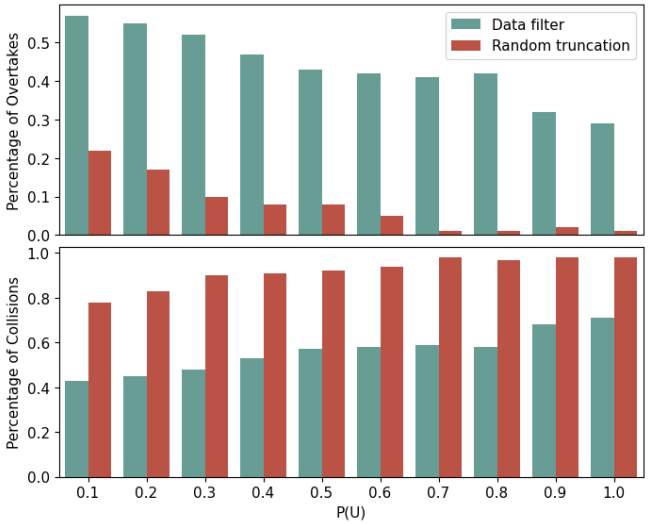


图 4：数据过滤器对超车率（上）和碰撞率（下）的影响。图中展示了不同不良行为概率 $P(U)$ 下的结果。

表 II：不同$P(U)$ 下被移除示范的比例，其中 r_β表示从所有收集到的示范中移除的示范所占比例。

$P(U)$	0.1	0.2	0.3	0.4	0.5
r_β	0.22	0.25	0.28	0.32	0.34
$P(U)$	0.6	0.7	0.8	0.9	1.0
r_β	0.38	0.49	0.57	0.69	0.82

并且数据过滤器会截断较差的示范，学到的新手策略仍可能实现约 30% 的成功超车。这也体现在通过数据过滤器的示范中：其中 18%（如表 II 所示）。

C. 冲突消解

为评估所提出的冲突消解方法的效果，我们基于修改后的车道切换器生成了 5 个不同的专家，并将 $P(U)$ 固定为 0.5。作为比较，我们使用 MEGA-DAgger、带数据过滤器的 HG-DAgger，以及仅使用 HG-DAgger，在两张不同的地图上训练新手策略。使用带数据过滤器和不带数据过滤器的 HG-DAgger 时，新手策略仅从一位专家处学习。每种方法均用于训练一个随机初始化的新手策略，总计进行 1000 次轨迹展开；实验中每 100 次训练轨迹展开保存并评估一次网络。我们进行了 5 次试验，以计算 95% 置信区间。

优于 HG-DAgger：图5展示了地图以及学习所得策略的实验结果。就超车和避碰指标而言，DAgger 和行为克隆的表现最差。与原版HG-DAgger相比，MEGA-DAgger 在超车和避碰方面的平均表现均提升约 45%；与使用数据过滤器的 HG-DAgger 相比，平均提升约15%。仅使用 HG-DAgger 时，新手几乎无法从多位不完美专家的示范中学习——

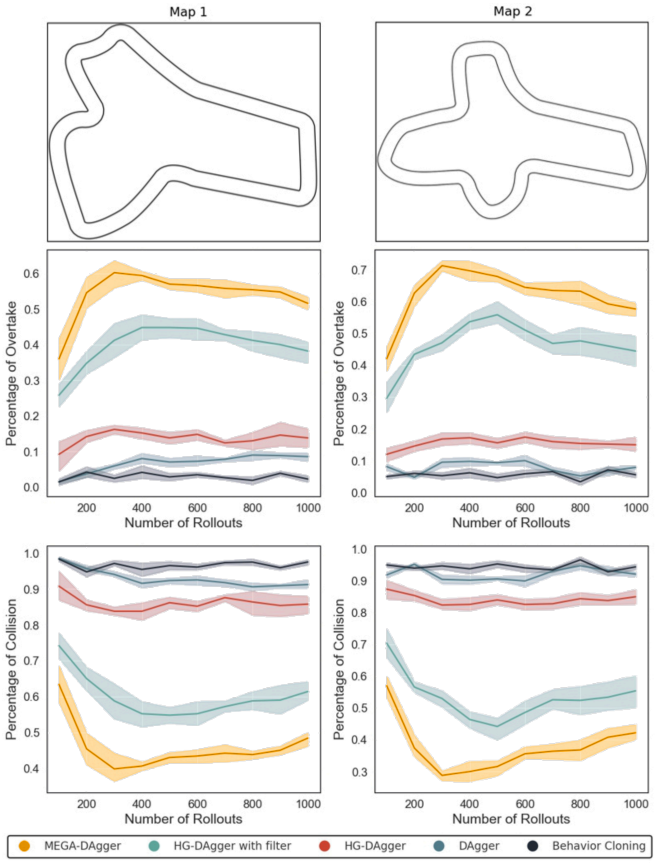


图5：MEGA-DAgger、带数据过滤器的 HG-DAgger、HG-DAgger、DAgger 和行为克隆在两张不同地图上的比较。左列和右列分别显示地图1和地图2上的结果。每完成100次训练滚动轨迹，便保存一次策略网络以供评估。每张图均为5次实验的平均结果，阴影区域表示95%置信区间。

的示范中学习。与原版HG-DAgger相比，数据过滤器和冲突解决功能在超车和避碰方面均带来了显著提升，但这两项指标在约300个训练滚动轨迹后达到峰值，随后逐渐下降。这表明，MEGA-DAgger 能够减少不安全且存在冲突的示范数量，而非将其彻底消除。随着更多示范被纳入全局数据集，MEGA-DAgger 的效果逐渐减弱；我们推测，原因是在训练过程中纳入了更多不理想的数据。

优于专家：我们通过实证将 MEGA-DAgger 相较于采用数据过滤器的 HG-DAgger 所提升的性能，归因于其能够从不同专家提供的互补性良好示范中进行学习。通过可视化 MEGA-DAgger 学习策略与所有专家在 200 次评估轨迹展开中的碰撞点（如图6所示），我们可以看到，每个专家经常会在地图的不同区域发生碰撞，而学习策略能够从这些专家那里学习互补性的良好行为，从而减少碰撞。表III显示，MEGA-DAgger 优于所有专家，并且在避碰和超车方面的平均性能分别提升了 13.6% 和 13.2%。此外，我们发现

表III：MEGA-DAgger策略与各位专家的碰撞和超车百分比。均值和标准差根据5次试验计算得出。

指标	MEGA-DAgger	专家1	专家2	专家3	专家 4	专家 5	专家累计值
碰撞百分比	$\mathbf{0.212 \pm 0.019}$	0.340 ± 0.025	0.401 ± 0.033	0.291 ± 0.025	0.392 ± 0.028	0.317 ± 0.032	0.348 ± 0.051
超车百分比	$\mathbf{0.781 \pm 0.016}$	0.657 ± 0.027	0.594 ± 0.036	0.706 ± 0.022	0.605 ± 0.024	0.681 ± 0.030	0.649 ± 0.051

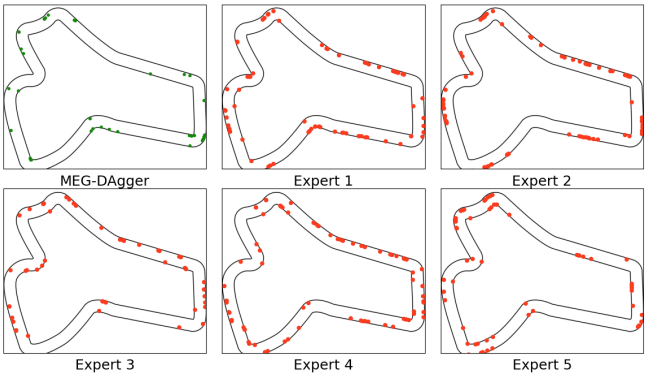


图 6：在地图1上进行的 200 次评估轨迹展开中，MEGA-DAgger 策略和不同专家造成的碰撞点可视化。MEGA-DAgger 策略以及专家1至5的碰撞次数分别为 40、71、79、51、79 和 63。

MEGA-DAgger 训练得到的策略比专家更加稳定（标准差更小）。如图3所示，由于 MEGA-DAgger 能够在相似观测下通过选择最佳动作来解决冲突，因此它能够有效利用碰撞点的互补性，解决冲突，并最终学习出优于专家的策略。

D. 相似度阈值的影响 ϵ

为研究余弦相似度阈值ϵ 对 MEGA-DAgger 的影响，我们通过使用不同的 ϵ 值训练策略开展实验，并将 ϵ、训练滚动轨迹数量与性能指标（碰撞和超车的百分比）之间的关系可视化。附录中的图 10 和图 11 展示了实验结果，表明 ϵ 的选择会显著影响策略的性能。在我们的场景中，将 ϵ 设为 0.95 时效果最佳。当 ϵ 设为 1 时，冲突解决机制会被禁用。相反，过度降低 ϵ 会导致策略无法学会期望行为，因为冲突解决机制会错误地将不相似的状态识别为相似状态。

这些发现可以结合我们从图6中的观察来理解，图6展示了不同专家之间碰撞点的互补性。我们推测，经过谨慎选择的 ϵ 值能够让MEGA-DAgger通过有选择地解决冲突来利用不同专家的优势，从而形成更稳健的策略。然而，过于激进的 ϵ 值可能导致MEGA-DAgger高估状态之间的相似性，进而产生次优策略。因此，ϵ 的选择体现了在利用专家示范的多样性与维持有效学习所必需的状态区分水平之间的权衡。

E. 真实世界实验

我们在F1TENTH自主赛车平台上实现了MEGA-DAgger。该算法部署在自车上，使其能够安全地进行巡航和超车。对手车辆采用纯追踪，以较低速度进行路径跟踪。我们在ROS 2环境下，分别于仿真和真实场景中评估MEGA-DAgger的性能。对于这两种场景，我们都记录了自车和对手车辆的轨迹。如图8所示，即使存在仿真到现实间隙，MEGA-DAgger仍能执行可靠的策略。如图7所示，自车成功完成了驾驶任务，并通过有效超过对手车辆展现出安全的超车操作。

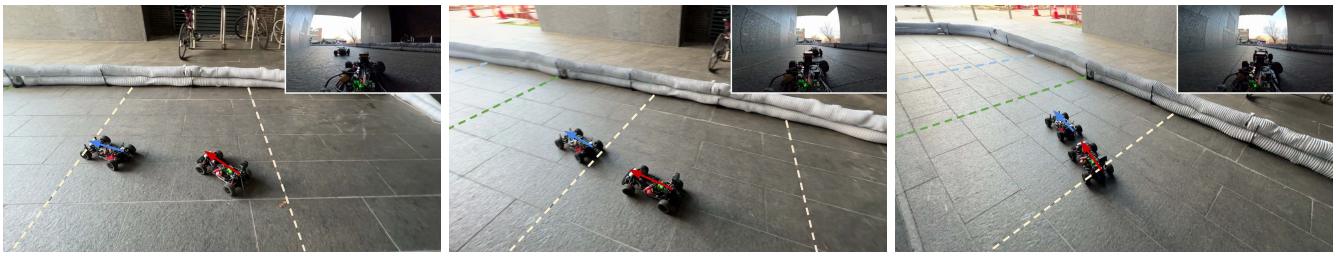
V. 结论与讨论

尽管 DAgger 及其变体等交互式模仿学习方法已成功应用于许多自主系统，但它们都假设能够获得一位最优专家。然而，更常见的情况是只能获得多位非最优专家。本文研究了如何通过交互式模仿学习有效利用这些专家。具体而言，我们提出了一种新的 DAgger 变体 MEGA-DAgger，用于筛除不安全示范并解决专家冲突。通过在端到端自动驾驶竞速任务上进行全面实验，我们证明，与 HG-DAgger 相比，MEGA-DAgger 提升了安全性和性能。我们还展示了 MEGA-DAgger 能够学得优于专家的策略。

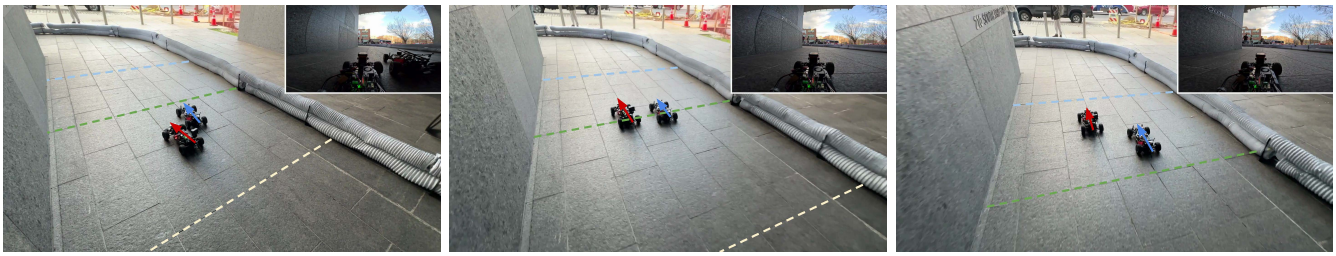
值得注意的是，我们使用进度分数和安全性分数来启发式地评估示范，但它们不会用作训练反馈。这与强化学习语境中的奖励函数 不同，后者用于指导训练过程，通常需要经过精心设计。未来一个有趣的研究方向是自动学习置信度分数，以评估和比较专家的动作。此外，我们计划在真实世界的自动驾驶汽车上开展实验，并尝试缩小仿真到现实间隙。

附录

我们使用带数据过滤器的 HG-DAgger，针对从 0 到 100 步的不同 β 值训练了 10 个初步策略，并发现将 β 设为 70 最适合我们的场景（超车率和碰撞率均表现最佳），如图9所示。

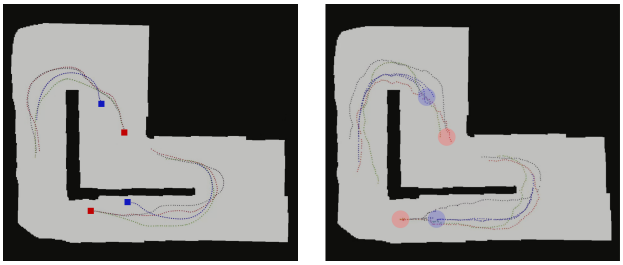


(a) (b) (c)



(d) (e) (f)

图 7：从 (a) 到 (f) 的连续帧以相等的时间间隔展示了 MEGA-DAgger 在 F1TENTH 平台直道上的安全超车过程。红色和蓝色箭头分别表示估计的自车和对手车辆位姿。等间距的虚线表示赛道的纵向进程。每一帧中的画中画显示了自车视角。



(a) 仿真 (b) 真实场景

图 8：在 F1TENTH 仿真和 RViz2 中真实场景下，使用 MEGA-DAgger 的自车（红色、绿色、紫色）超车轨迹，以及对手车辆（蓝色）的轨迹。红色和蓝色标记分别表示自车和对手车辆的初始位置。

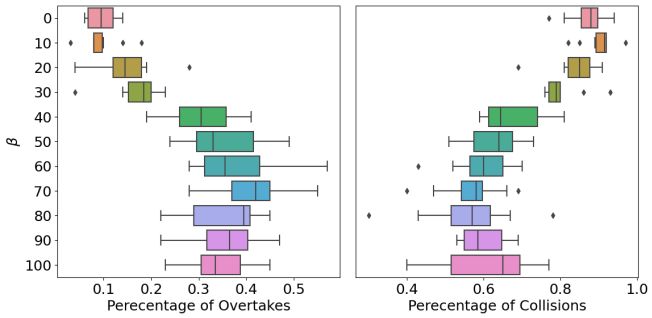


图 9：对预训练策略进行评估，以确定最适合自主赛车超车任务的 β（不安全示范的截断步数）。当 $\beta = 70$ 时，超车率和碰撞率均为最佳。

参考文献

[1] H. A. Pierson 和 M. S. Gashler。机器人学中的深度学习：近期研究综述。《*Advanced Robotics*》，31(16):821–835，2017。[2] B. D. Argall、S. Chernova、M. Veloso 和 B. Browning。机器人示范学习综述。《*Robotics and Autonomous Systems*》，57(5):469–483，2009。[3] B. Fang、S. Jia、D. Guo、M. Xu、S. Wen 和 F. Sun。机器人操作模仿学习综述。《*International Journal of Intelligent Robotics and Applications*》，3:362–369，2019。[4] L. Le Mero、D. Yi、M. Dianati 和 A. Mouzakitis。端到端自动驾驶汽车模仿学习技术综述。*IEEE*智能交通系统汇刊，2022。[5] M. Bojarski、D. Del Testa、D. Dworakowski、Be. Firner、B. Flepp、P. Goyal、L. D. Jackel、M. Monfort、U. Muller、J. Zhang 等。自动驾驶汽车的端到端学习。*arXiv* 预印本 *arXiv:1604 07316*，2016。。[6] S. Ross、G. Gordon 和 D. Bagnell。将模仿学习和结构化预测归约为无悔在线学习。载于第十四届人工智能与统计国际会议论文集，第627–635页。《机器学习研究杂志》研讨会与会议论文集，2011。[7] M. Kelly、C. Sidrane、K. Driggs-Campbell 和 M. J. Kochenderfer。HG-DAgger：与人类专家进行交互式模仿学习。载于*2019* 国际机器人与自动化会议 *(ICRA)*，第 8077–8083 页。IEEE，2019年。[8] S. Ross 和 D. Bagnell。用于模仿学习的高效归约。载于第十三届人工智能与统计国际会议论文集，第 661–668 页。《机器学习研究杂志》研讨会与会议论文集，2010年。[9] J. Zhang 和 K. Cho。面向端到端自动驾驶的高查询效率模仿学习。*arXiv* 预印本 *arXiv:1605.06450*，2016年。[10] K. Menda、K. Driggs-Campbell 和 M. J. Kochenderfer。Ensembledagger：一种安全模仿学习的贝叶斯方法。载于*2019*年 *IEEE/RSJ*智能机器人与系统国际会议 *(IROS)*，第 5041–5048 页。IEEE，2019年。[11] R. Hoque、A. Balakrishna、C. Putterman、M. Luo、D. S. Brown、D. Seita、B. Thananjeyan、E. Novoseller 和 K. Goldberg。Lazydagger：减少交互式模仿学习中的上下文切换。载于*2021*年 *IEEE*第*17*届自动化科学与工程国际会议 *(CASE)*，第 502–509 页。IEEE，2021年。[12] R. Hoque、A. Balakrishna、E. Novoseller、A. Wilcox、D. S. Brown 和 K. Goldberg。Thriftydagger：面向交互式模仿学习的预算感知新颖性与风险门控。*arXiv* 预印本 *arXiv:2109.08273*，2021年。[13] 宾夕法尼亚州交通部。2021年宾夕法尼亚州交通事故事实与统计。https://www.penndot.pa.gov/TravelInPA/Safety/Documents/2021_CFB_linked.pdf。

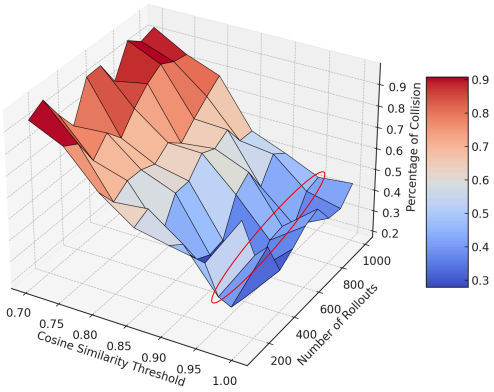


图 10：展示余弦相似度阈值、训练回合数与碰撞率之间关系的曲面图。在我们的场景中，阈值为 0.95 时效果最佳（红色椭圆标出）。

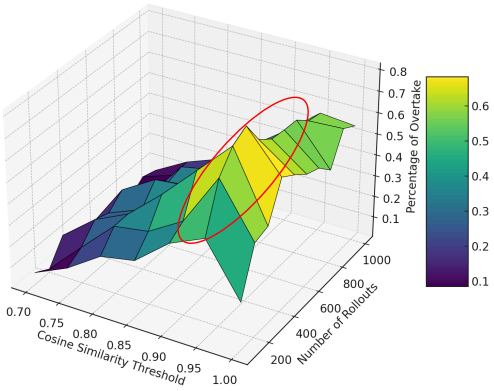


图 11：展示余弦相似度阈值、训练回合数与超车率之间关系的曲面图。该图显示的趋势补充了图 10 所示的结果。

[14] P. Abbeel 和 A. Y. Ng。通过逆强化学习进行学徒学习。载于第二十一届国际机器学习会议论文集，第 1 页，2004年。[15] Y. Pan、C.-A. Cheng、K. Saigol、K. Lee、X. Yan、E. Theodorou 和 B. Boots。使用端到端深度模仿学习实现敏捷自动驾驶。*arXiv* 预印本 *arXiv:1709.07174*，2017年。[16] J. Zhou、R. Wang、X. Liu、Y. Jiang、S. Jiang、J. Tao、J. Miao 和 S. Song。探索使用反馈合成器和可微光栅化的自动驾驶模仿学习。载于*2021*年 *IEEE/RSJ*智能机器人与系统国际会议 *(IROS)*，第 1450–1457 页。IEEE，2021年。[17] Xiatao Sun、Mingyan Zhou、Zhijun Zhuang、Shuo Yang、Johannes Betz 和 Rahul Mangharam。自主赛车基于模仿学习的控制策略基准比较。载于*2023*年 *IEEE*智能车辆研讨会 *(IV)*，第 1–5 页，2023年。[18] Y.-H. Wu、N. Charoenphakdee、H. Bao、V. Tangkaratt 和 M. Sugiyama。来自不完美示范的模仿学习。载于国际机器学习会议，第 6818–6827 页。PMLR，2019年。[19] D. S. Brown、W. Goo 和 S. Niekum。通过自动排序的示范实现优于示范者的模仿学习。载于机器人学习会议，第 330–359 页。PMLR，2020年。[20] Y. Wang、C. Xu、B. Du 和 H. Lee。学习为不完美示范赋权。载于国际机器学习会议，第 10961–10970 页。PMLR，2021年。[21] D. Brown、W. Goo、P. Nagarajan 和 S. Niekum。通过基于观测的逆强化学习，从次优示范中进行外推。载于国际机器学习会议，第 783–792 页。PMLR，2019。[22] V. C. Raykar、S. Yu、L. H. Zhao、A. Jerebko、C. Florin、G. H. Valadez、L. Bogoni 和 L. Moy。多位专家的监督学习：当所有人都有些不诚实时，该信任谁？载于第 *26* 届年度国际机器学习会议论文集，第 889–896 页，2009。[23] V. C. Raykar、S. Yu、L. H. Zhao、G. H. Valadez、C. Florin、L. Bogoni 和 L. Moy。从众包中学习。机器学习研究杂志，11(4)，2010。[24] J. Betz、H. Zheng、A. Liniger、U. Rosolia、P. Karle、M. Behl、V. Krovi 和 R. Mangharam。极限操控下的自动驾驶汽车：自动驾驶汽车竞赛综述。*IEEE*智能交通系统开放期刊，3:458–488，2022。[25] A. Wischnewski、M. Geisslinger、J. Betz、T. Betz、F. Fent、A. Heilmeier、L. Hermansdorfer、T. Herrmann、S. Huch、P. Karle 等。Indy Autonomous Challenge——极限操控下的自主赛车。载于*2021*年第*12*届慕尼黑国际底盘研讨会：*chassis.tech plus*，第 163–182 页。施普林格，2022。[26] J. M. Rieber、H. Wehlan 和 F. Allgower。Roborace赛事。*IEEE*控制系统杂志，24(5):57–60，2004。[27] M. Zeilinger、R. Hauk、M. Bader 和 A. Hofmann。Formula Student Driverless（FSD）自动赛车设计。载于*OAGM*与*ARW*联合研讨会，2017。[28] M. O'Kelly、H. Zheng、D. Karthik 和 R. Mangharam。F1TENTH：用于连续控制和强化学习的开源评估环境。机器学习研究会议录，123，2020。[29] J. Herman、J. Francis、S. Ganju、B. Chen、A. Koul、A. Gupta、A. Skabelkin、I. Zhukov、M. Kumskoy 和 E. Nyberg。Learn-to-race：用于自主赛车的多模态控制环境。载于*IEEE/CVF*计算机视觉国际会议论文集，第 9793–9802 页，2021。[30] J. Kabzan、L. Hewing、A. Liniger 和 M. N. Zeilinger。面向自主赛车的基于学习的模型预测控制。*IEEE*机器人与自动化快报，4(4):3363–3370，2019。[31] A. D. Ames、X. Xu、J. W. Grizzle 和 P. Tabuada。基于控制屏障函数的二次规划，用于安全关键系统。*IEEE*自动控制汇刊，62(8):3861–3876，2016。[32] J. Zeng、Z. Li 和 K. Sreenath。利用离散时间控制屏障函数提高非线性模型预测控制的可行性与安全性。载于*2021*年*IEEE*第*60*届决策与控制会议（*CDC*），第 6137–6144 页。IEEE，2021。[33] K. Liu、W. Wang 和 J. Wang。基于单模板匹配的激光雷达点云行人检测。《电子学》，8(7):780，2019。[34] C. Yang、Z. Zhou、H. Zhuang、C. Wang 和 M. Yang。基于网格化高斯分布和Wasserstein距离的全局位姿初始化。*IEEE*智能交通系统汇刊，第 1–11 页，2023。[35] J. Zhou、J. Yan、T. Wei、K. Wu、X. Chen 和 S. Hu。利用RGB-D相机系统识别室内环境中的尖锐角点/边缘。*IEEE*电路与系统汇刊第二辑：快报，62(10):987–991，2015。[36] Z. Hu、D. Yang、S. Cheng、L. Zhou、S. Wu 和 J. Liu。我们通过RGB-D相机了解他们的注视方向：在 3 14, 2022 71 1 d 中进行视线跟踪。*IEEE*仪器与测量汇刊，:–。[37] F1TENTH ICRA 2022：结果。https://icra2022-race. f1tenth.org/results.html。