arXiv:2303.17010v1 [cs.LG] 29 Mar 2023

面向语义感知模仿学习的规范引导数据聚合

Ameesh Shah*、Jon DeCastro†、John Gideon†、Beyazit Yalcinkaya*、Guy Rosman†、Sanjit A. Seshia*
*加州大学伯克利分校 †丰田研究院 联系方
式：ameesh@berkeley.edu

摘要——**仿真与形式化方法引导的环境采样 技术进步，使得人们能够在自动驾驶等多种安全关键场景中严格评估机器学习模型。将这些环境采样技术用于改进学习模型本身的潜力，尚未得到充分挖掘。本文提出一种新方法，通过利用规范引导采样技术，在新环境中聚合专家数据，从而以语义感知的方式改进模仿学习模型。具体而言，我们构建一组形式化规范，将所有可能环境的空间划分为语义相似的区域，并找出学习得到的模仿策略与专家行为差异最大的区域。随后，我们在这些区域对应的环境中聚合专家数据，从而更准确地模仿专家行为的语义。我们在CARLA驾驶模拟器中开展了一系列实验，验证了该方法得到的模型比采用其他环境采样方法训练的模型更加准确。**

I. 引言

大量研究已证明，模仿学习 （IL）在各种端到端控制和运动预测任务中都行之有效，尤其是在自动驾驶领域 [7, 27, 12, 10, 24]。模仿学习方法通过学习策略来模仿收集到的专家行为样本，通常具有很强的通用性，也能轻松应用于各种领域。尽管取得了这些成果，模仿学习方法的性能仍常因协变量偏移问题而受影响：在一个控制回合早期预测动作中的偏差，可能会在后续访问原训练分布之外的状态时累积成误差 [38]。在这项工作中，我们提出一种新方法来应对协变量偏移：利用可配置的环境领域（如仿真）在多样化环境中汇集专家数据，从而学习准确的模仿策略。

现有用于减少协变量偏移所导致误差的方法，即DAgger风格方法 [36, 33, 15, 25, 43]，提出在模仿学习智能体与专家策略之间切换控制，以便收集原始训练分布之外状态下的专家数据。这类算法通常专注于利用额外目标或学习模型的不确定性来确定何时切换控制并从专家处收集数据。我们设计了一种方法，用于选择智能体将在其世界中何处行动，因为这些方法通常将世界视为固定不变的，而不是模仿学习循环中可控制的一部分。

作为一个激励性示例，考虑模仿学生驾驶员的任务，以更好地理解学生在驾驶过程中可能如何以及何时犯错。由于学生是新手，我们希望避免让学生驾驶真实车辆，而是选择让学生在仿真环境中操作。此外，由于我们的目标是尽可能准确地模仿学生，因此我们并不明确寻求训练出最安全或性能最佳的学习得到的模仿策略。如果学生在某种特定情况下犯了错误，我们也希望学习得到的模仿策略犯下同样的错误。

在可控世界中，关键问题是：我们如何找到其中的环境，以拓展对专家行为的理解，从而实现更好的模仿？如果学生驾驶员通常驾驶得非常谨慎，并避免接近其他车辆，那么当另一辆车突然加塞时，学生驾驶员会如何应对？在驾驶等领域，低概率事件（例如突然加塞）具有很高的语义价值 ，对学生和教师而言都是如此（突然加塞可能导致碰撞）。学习模仿策略时，优先考虑这些情况并找到能诱发此类结果的环境至关重要，这样我们才能收集有价值的专家数据。

为了找到有意义的环境，我们提出了一种方法，借助逻辑属性形式的领域相关信息，对现实世界中众多可能的环境进行语义描述和区分。在驾驶示例中，一个属性可以描述学生驾驶员是否发生碰撞，另一个属性则可以衡量是否超出了限速。基于这些属性，我们通过枚举这些属性所有可能的真值赋值，创建一组形式化规范 。每个规范对应于该赋值所能涵盖的所有 可能结果。继续沿用上述示例，这些规范会将环境区分开来：学生发生碰撞但未超速的环境、学生超速但未发生碰撞的环境、学生既发生碰撞又超速的环境，以及学生既未发生碰撞也未超速的环境。

这组规范有效地划分 了现实世界中所有可能的（环境、智能体轨迹）对所构成的空间：单个组合只能满足（即属于）我们集合中的一个规范，而不能属于其他规范。这种划分为环境空间提供了语义刻画，随后我们可以利用它来指导数据聚合过程。

我们提供了数据聚合算法的可视化-

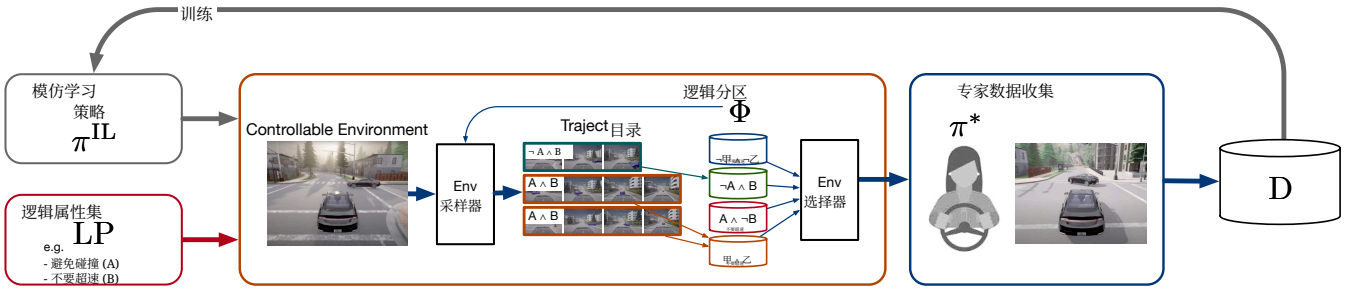


图1：我们用于数据聚合的环境采样方法。利用一组逻辑属性，我们对各种环境进行采样，并选择那些能够在我们的专家与模仿学习策略之间诱发语义差异行为的环境。

图1中展示了我们的数据聚合算法。我们将当前的模仿学习策略部署在各种环境中，利用上述划分从每种规范中公平抽样，收集一组语义多样的环境。从这组环境中，我们选出一个子集，在这些环境中，我们最预计专家策略与模仿学习策略在语义行为上会出现分歧，并在该子集上聚合专家数据。

我们将该方法具体实例化为一种称为 规范引导的数据聚合 （SGDA）的算法。我们在 CARLA 自动驾驶模拟器中，通过模仿两个不同的专家模型在十字路口执行不同驾驶操作的过程，展示了该算法的有效性。结果表明，使用算法所选环境进行数据聚合，相较于使用基线方法所选环境，能够获得更高的模仿学习准确率，尤其是在不太可能但具有语义意义的情况下。

总而言之，本文的贡献有三点：首先，我们提出了一种新颖的形式化方法——根据轨迹所满足的形式规范，对可能的（环境、轨迹）对进行语义划分——以系统化环境采样。其次，我们提出了一种算法，在数据聚合循环中利用上述划分来改进模仿学习模型。第三，在自动驾驶背景下，我们展示了该算法在人类驾驶员建模方面的良好实验结果，以及相较于环境采样基线方法的性能提升。

II. 相关工作

提升模仿学习的鲁棒性。多项研究通过修改训练准则，或将研究扩展到涵盖稀有事件等相关属性的环境，来解决模仿学习中的鲁棒性问题。[14] 承认并指出，数据集偏差会妨碍模型在边界案例中实现鲁棒性，并提出了算法改进和辅助训练任务。当任务已知时， [41, 45, 8] 均采用改进的生成对抗模仿学习（GAIL）结构来学习语义策略嵌入，并确保任务相关性。 [29] 通过引入强化学习（RL）智能体来增强纯粹的模仿学习，以应对分布偏移。其他研究探索了使用置信度度量来表征轨迹，并从中选择轨迹用于训练。 [44, 23] 利用置信度概念来对齐模仿与下游任务，同时 [34, 3]利用在多样化数据集上训练的新颖性检测概念。

我们同样研究鲁棒性问题，但从聚合专家数据的角度出发，以确保能够达到预期的模仿学习性能。与模仿学习方法解耦，可以消除对学习流程的限制，使用户能够采用任意模仿学习方法。

证伪与重要性采样。一些研究采用形式化方法和基于搜索的方法，来发现安全关键系统中的稀有事件和性质违例。其中一些研究处理证伪 问题，包括 [1], 提出的一种方法：该方法评估给定模型与参考模型的符合性，后者具有某些保证；而 [2, 28] 使用强化学习来减少寻找反例所需的仿真次数。 [11] 将证伪问题表述为控制综合问题，并以不变集为基础来集中采样。一些成熟的证伪方法已在 s-Taliro [4] 和 VerifAI [18] 等软件工具中实现。

与证伪不同，重要性采样 （IS）方法在寻找违反预期性质的反例方面展现出良好前景。 [37] 等研究采用交叉熵方法，在假设输入空间固定划分的情况下，调整对性质违例的搜索。 [5]通过训练具有效率保证的深度学习采样器来扩展标准重要性采样，而[32] 采用带有学习得到的搜索空间分布的自适应重要性采样程序，以进一步加速稀有事件的发现。由于许多此类方法会迅速聚焦于稀有事件，因此仍难以将其直接应用于补充用于学习的数据，因为它们可能会遗漏许多具有语义意义的非失败结果。

面向安全自主系统的场景生成。虽然上述研究侧重于一般性的采样，但也有大量文献专注于丰富用于学习安全自主系统的场景。 [16, 17] 训练基于流的生成模型，并利用层次结构的网络对环境进行采样，用于训练和验证。其他研究则将世界的动力学和仿真纳入考虑，以指导采样 [42, 20, 28]。 [32, 40, 39] 中的自适应采样研究在环境参数空间中搜索，以快速发现失效模式。一些研究（如 Scenic 语言 [19], ）探索了更具程序性且易于用户

阅读的模式指定方式。虽然这类研究也对环境进行采样，但我们的方法有所不同：我们在语义分区空间中平衡数据集，并将采样方法直接应用于训练模仿学习智能体。

数据聚合技术。 在线采样框架（如 DAgger [36]）已被用于解决训练模仿学习模型时的分布偏移问题。HG-DAgger [25] 会向人类专家请求补充的纠正动作，以提高学习者的鲁棒性。其他变体，如 Safe-DAgger [43]，（用于策略学习的数据聚合） [33] 和不确定性感知数据聚合 [15]，均提出了高效的查询采样方法。它们的目标与我们相似，都试图通过关注安全性和不确定性相关标准，减轻人类专家的数据负担。然而，与现有方法不同，我们的方法使用与任务相关的性质，并对学习者偏离预期属性集的不同环境进行采样。我们强调，我们的方法可以融入上述任意一种方法。

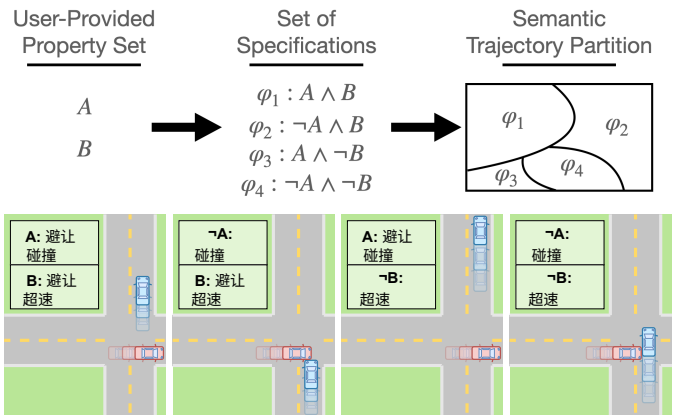


图 2：**上：** 展示如何将用户提供的属性视为命题变量，以构建我们的 STP。我们枚举每个属性的所有真值组合，生成一组规范，作为轨迹空间划分中的元素。**下：** 蓝色汽车智能体满足各项规范的轨迹示例。

III. 方法论

给定从专家策略收集的初始数据集、一个连续控制任务以及一个可控环境，我们旨在利用逻辑规范学习对专家的准确模仿。本节将阐述我们的方法：在数据聚合式循环中，对环境可能出现的条件进行语义采样。

我们首先概述在可控环境中给定策略下的回合控制。令 $\mathcal{E}$ 表示在模拟器中可表示的所有可能环境条件（环境的各种实例）。给定环境条件 $e \in \mathcal{E}$ 以及学习者可用策略类别中的一个策略 $\pi \in \Pi$，我们将一个控制回合定义为$\xi_\pi(e) = d(\pi, e)$。其中，$\xi_\pi(e)$是由专家生成的状态与动作对轨迹，而 d 表示一个函数，它使用固定策略 π 控制智能体，并使用 e 设置环境条件，从而展开一个控制回合的动力学过程。为简单起见，我们假设动力学是确定性的，且 d 返回单条轨迹。函数 d 会持续展开动力学，直至轨迹达到预定义的终止条件，例如抵达目标位置、违反安全条件，或达到最大时间步数。

本节首先介绍如何通过构建语义轨迹划分 （STP），对所有可能轨迹的空间进行语义划分；随后介绍如何在数据聚合式模仿学习循环中使用这一划分。

A. 创建语义轨迹划分

我们首先让用户提供一组（时序）逻辑属性，LP，这些属性具有特定领域的性质，并用于识别智能体轨迹中有意义的属性。例如，在自动驾驶场景中，这些属性可以表示避让碰撞或保持限速。此类属性的形式化已有充分研究，并且可以改编自诸如 [9]等研究。我们允许以灵活的方式表达这些属性：它们可以是使用度量时序逻辑（MTL） [30]和高阶逻辑（HOL） [35], 等逻辑表达的纯布尔属性（避让所有碰撞），也可以是伴随连续目标的布尔属性（与所有其他车辆保持超过阈值的距离），并使用信号时序逻辑（STL） [31, 21]等定量语义来表达。我们将LP的大小记为$|LP| = l$。

借助这个属性集 LP，我们将所有智能体轨迹构成的空间划分为若干部分。创建 STP 的方法如下。首先，将每个性质 $p \in LP$视为一个命题变量：当且仅当轨迹 $\xi_\pi(e)$ 在环境 e下满足 p 时，该变量为真；也就是说，$(\xi_\pi(e), e) \models p$。然后，我们枚举所有可能的布尔逻辑公式，这些公式由属性 p （属于 LP）或其否定项合取构成。这样，我们就创建了一组逻辑规范 Φ，涵盖 LP中每个属性的所有可能真值组合。我们将划分 $\varphi \in \Phi$ 中的元素称为规范。我们常会宽泛地使用这一说法：在聚合数据时，规范既指与某个划分元素相关联的逻辑公式，也 指满足该逻辑公式（即位于该划分元素内）的已收集轨迹集合。

作为一个简单示例，请看图 2 所示的场景。在这个示例中，我们有一个集合 LP，其中只有两个属性：(A) 避免碰撞，以及 (B) 遵守限速。我们的划分 Φ 将包含四个元素：$(A \wedge B)$（避免碰撞且不要超速）、$(\neg A \wedge B)$（发生碰撞且不要超速）、$(A \wedge \neg B)$（不发生碰撞且超速）以及 $(\neg A \wedge \neg B)$（发生碰撞且超速）。请注意，智能体能够生成的每条轨迹都只会满足我们划分中的一个规范。另外还要注意，为了探索完整的划分，智能体所处的某些环境会促使其在特定元素中采集

更多样本——例如，有利于较低速度的城市环境。

我们注意到，$|\Phi| = 2^l$ 的规模在实践中可能会限制用户能够提供的属性数量。不过，我们并不认为这是一个难以克服的问题，主要是因为正如我们的实验所证明的那样，用户只需指定少量属性，就能采样出一组多样化的数据。此外，我们的方法可以推广到这样一种设置：用户能够为每个属性分配重要性权重，从而降低特定划分元素的权重。我们将在附录中进一步阐述这一推广。

在训练过程中，我们可以使用 STP 从可能的语义结果空间中进行公平采样，并确保数据充分覆盖这些语义。接下来，我们将讨论如何使用 STP 及其元素，通过数据聚合来改进模仿学习器。

B. 规范引导的数据聚合

最终，我们的目标是通过聚合语义多样的数据集，使用 Φ 提高模仿学习智能体 π^{IL} 的准确率。为此，我们采用包含四个步骤的方法：(1) 训练、(2) 环境采样、(3) 环境选择和 (4) 数据聚合，如算法 1 所述。下面我们依次解释每个步骤：

算法 1： 规范引导的数据聚合（SGDA）

输入： 初始数据集 D，数据聚合轮数 n，STP Φ，专家策略 π^*，样本数 k，选择次数 m

1 学习 π^{IL} 相对于 D;
2 初始化 E_{prev}，D_{prev} 至 $\emptyset$;
3 **对每个迭代** $i \in 0 \dots n$ **执行**
4 $E_{\text{new}} := \text{EC-Sampling}(\pi^{\text{IL}}, k)$;
5 $E_{\text{new}} := \text{EC-Select}(E_{\text{new}}, E_{\text{prev}}, D_{\text{prev}}, m)$;
6 $D_{\text{new}} := \{d(\pi^*, e); \forall (e, \xi_{\pi^{\text{IL}}}(e)) \in E_{\text{new}}\}$;
7 $D = D \cup D_{\text{new}}$; // 聚合
8 学习 π^{IL} 关于 D;
9 将 E_{prev}，D_{prev} 设为 E_{new}，D_{new};
10 **返回** π^{IL}

训练。 首先，我们假设存在一个由专家收集的初始数据集 D。我们不对收集 D 时所依据的环境条件分布作任何假设。然后，我们使用任意一种模仿学习算法，从 D 中学习模仿策略 π^{IL}，例如行为克隆（BC）或 GAIL [22]。

环境采样（EC-Sampling）。 在此步骤中，我们通过对一组环境条件 $e \in E$ 进行采样并执行$\xi_{\pi^{\text{IL}}} = d(\pi^{\text{IL}}, e)$，尝试收集语义范围广泛的轨迹。

在采样和收集不同环境条件时，我们会权衡两个因素：首先，我们希望尽可能公平地为划分 Φ 中的每个元素收集轨迹。其次，我们不希望尝试为划分中那些非常难以满足的元素进行采样（即，我们尝试采样能够在给定划分元素 φ中生成轨迹的环境条件，但找到正确的环境条件颇具挑战）。为了找到给定 φ 的正确环境条件，我们利用基于优化的采样方法，例如贝叶斯优化或交叉熵。这些方法从已经采样的（环境条件、轨迹）对中学习分布，从而最大限度地提高后续样本满足目标规范 φ的可能性。

为了平衡上述两个因素，我们采用上置信界算法 [6] 来选择下一个要采样的目标规范，符号表示沿用文献 [26]。UCB算法在利用 （对代表不足的划分元素进行采样）和探索（不再对已进行大量采样尝试的划分元素采样）之间取得平衡。

我们在算法 2 中概述了我们的方法。我们记录针对带有N的选择进行采样的总次数，以及满足每个规范的（ξ_π（e） e）对的数量（第 2 行）。在每次迭代中，我们选择一个目标规范 φ_c，并尝试为其采样一个满足条件的环境条件。在第 5–6 行中，我们递增 φ_c 的计数，并使用我们的基于优化的采样器对其进行采样，我们将其记为 f。f 接收目标规范和收集到的样本作为参数，并生成一个新的样本环境条件，优化器认为该条件很可能满足 φ_c。在第 7–9 行中，我们展开动力学过程，收集与采样环境对应的轨迹，并通过检查集合中的每个规范的满足性，将该轨迹添加到其实际所属的元素划分中。在第 10–11 行中，我们通过优先选择相对于其他规范采样不足的规范，计算下一次迭代选择某个规范的价值。我们将这些价值以及针对给定规范尝试采样的次数（在我们的算法中为N ）用于第 13 行的 UCB 计算，以选择下一轮要进行采样的规范。

环境选择（EC-Select）。利用上一步收集的环境条件和轨迹集合，我们希望为下一轮数据聚合选出一组条件，以有意义的语义方式改进模型。为此，我们借鉴以下见解：在学习策略的轨迹与专家策略的轨迹满足不同规范的环境条件 e，在选择时应被赋予更高优先级。形式化地说，我们希望优先选择这样的环境：分别使用专家策略和模仿策略进行 rollout 时，会导致 $(\xi_{\pi^*}(e)\ e) \models \varphi_i$ 和 $(\xi_{\pi^{\text{IL}}}(e)\ e) \models \varphi_j$，其中 $\varphi_i \neq \varphi_j$。

在数据聚合的初始轮次中，我们没有一组同时包含学习模型和专家模型轨迹的环境条件。在这种情况下，我们只需从集合中均匀随机采样，以选择用于聚合的数据集。然而，在后续迭代中，我们拥有上一迭代的数据集，其中同时包含收集到的专家轨迹（$\xi_{\pi^*}(e)\ e$）以及学习模型的轨迹

算法 2： 环境条件采样（EC-采样器）

输入： 环境划分 Φ、优化引导采样器 f、学习策略 π^{IL}、迭代次数 k

从 Φ 中选择初始 φ_c；
将 N_{φ_j} 初始化为 $0\ \forall\varphi_j \in \Phi$;
将 S 初始化为 $\emptyset$;
对每次迭代 $t \in 0 \ldots k$ **执行**
  $e_t := f(\varphi_c, S)$; // 获取新样本
  递增 N_{φ_c};
  $\xi_{\pi^{\text{IL}}}(e_t) = d(\pi^{\text{IL}}, e_t)$; // 展开动力学
  将 $\xi_{\pi^{\text{IL}}}(e_t)$添加到满足它的 φ 中;
  对于每个 $\varphi_j \in \Phi$，**执行**
    将 Q_j 设为$max_{\varphi\in\Phi}(|\varphi|) - |\varphi_j|$;
  将$(e_t,\xi_{\pi^{\text{IL}}}(e_t))$添加到 S;
  设置 $\varphi_c := \text{UCB}(Q, N, t)$ // 选择下一个规范
回报 S

表 I：环境选择权重的计算示例。

专家结果	模仿学习结果 $A \wedge B$	$\neg A \wedge B$	$A \wedge \neg B$	$\neg A \wedge \neg B$
$A \wedge B$: 10	8	2	0	0
$\neg A \wedge \neg B$: 10	0	0	5	5
选择权重	0.2	1	1	0.5

$(\xi_{\pi^{\text{IL}}}(e)\ e)$，涵盖许多 e。随后，我们可以比较相同的环境条件是否使专家轨迹和学习模型轨迹满足了相同的规范。

例如，考虑第三节 A 中介绍的双属性划分。假设在上一轮数据聚合中，我们选取了 20 个环境条件。在这个假设情境中，专家在 10 个条件下避让了碰撞并遵守限速（$A \wedge B$），而在另外 10 个条件下发生碰撞并违反限速（$\neg A \wedge \neg B$）。在导致专家出现（$A \wedge B$）的同样 10 个条件下，我们的学习模型在其中 8 个条件下避让了碰撞并遵守限速，而在另外 2 个条件下，模型虽遵守限速，却发生了碰撞（$\neg A \wedge B$）。在导致专家出现（$\neg A \wedge \neg B$）的 10 个条件下，我们的学习模型也只在其中 5 个条件下发生碰撞并违反限速，而在另外 5 个条件下，模型没有发生碰撞，但违反了限速（$A \wedge \neg B$）。表 I 概述了这一情境。我们将如下设置选择权重：

$$1 - (M_\varphi / N_\varphi) \quad \text{if} \quad N_\varphi > 0, \quad \text{else} \quad 1$$

这里，N_φ 表示专家或模仿者满足某项规范的实例总数（不对两者都满足的情况重复计数），M_φ 表示模仿学习器匹配到的专家结果数量。这些权重被归一化为一个分类分布，然后我们从抽取的样本池中进行采样环境条件（来自前一步），并根据该分布进行加权，从而得出用于后续数据聚合迭代的选定批次。我们在算法 1 中将此选择步骤记为 EC-Select(E_{new}，E_{prev}，D_{prev}，m)，利用先前的专家数据和模仿数据计算选择权重，并从新集合 E_{new}中选出 m 个样本。

数据聚合。确定最终选定的环境条件集后，我们可以在这些条件下从专家策略收集轨迹，并将收集到的数据汇总到整体数据集中。随后可在下一次迭代中使用该数据集训练模仿学习模型；也就是说，我们在这个聚合数据集上重复训练步骤。

算法 1 总结了上述流程；它首先训练一个初始模仿策略（第 1 行），然后使用前述环境采样和选择方法选择新的环境条件（第 4–5 行），并根据这些条件向专家查询（第 6 行），以生成可供聚合的数据集（第 7 行）。这一流程可按需重复。

SGDA 的开销。在为每个 e 执行动力学操作并评估每个 $(\xi_\pi(e)\ e) \models \varphi$ 的满足性时，与不带任何指导地采样环境的方法相比，SGDA 会产生额外的计算开销。我们发现，该开销随 Φ 的规模线性增长；同时需要指出的是，在实践中，这种权衡的代价通常远低于专家参与的数据收集所需的高昂成本，而 SGDA 的目标正是降低这一成本。我们在附录中对这一开销进行了更详细的分析。

IV. 实验设置

我们在自动驾驶任务中模仿专家的任务上评估我们的方法。下面，我们将详细介绍我们的模拟器设置、基线方法和评估指标。

A. 模拟器与实验设计

我们使用 CARLA 模拟器 [13]，这是一款高保真的自动驾驶模拟器，能够调整多种环境因素，例如环境中其他智能体的位置和行为、道路布局以及障碍物的存在情况。在我们的实验领域中，我们考虑模仿专家在四向交叉路口中的行为，其中专家作为自车通过交叉路口到达目的地。

在这种设定中，可能会有另一辆车（其他交通参与者）出现，并无视交通法规，疾驰穿过交叉路口。自车的轨迹会受到路口环境条件的影响，我们将这些条件定义为以下参数化变量：首先，其他车辆可能在交叉路口的任意一侧生成（位于自车对面或自车任一侧），与交叉路口之间的距离也各不相同。其他交通参与者还可能在交叉路口执行三种机动动作之一：右转、左转或直行。自车与交叉路口之间的初始距离，以及其他车辆的最高和最低速度，也都是可变的环境条件参数。其他交通参与者的行为由一个简单的程序化控制器控制

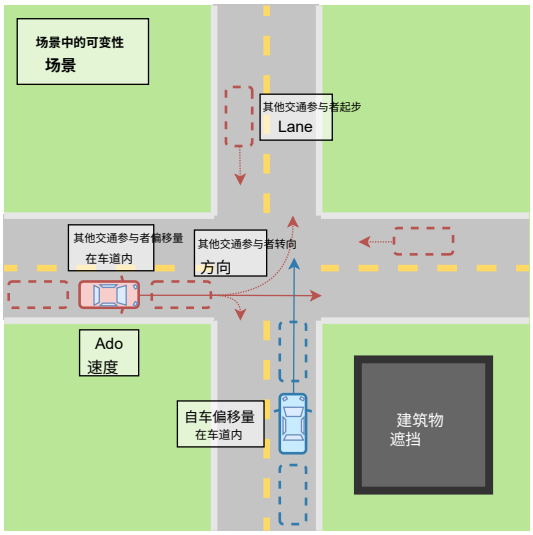


图 3：我们在 CARLA 仿真器中的实验设置。我们概述了可用于改变场景的环境参数。

该控制器会在其他车辆穿过交叉路口时将其加速至最高速度。我们改变所有参数，以便在数据聚合过程中采样新的环境。我们的实验设置如图 3 所示。

我们用信号时序逻辑（STL）定义了一组性质，并为每个性质定义了各自的定量语义，用以衡量其距离满足的程度。我们的实验所用的属性集 LP 如下：

1) **避免碰撞**。 如果检测到自车与环境中的任何物体发生碰撞，则此性质不成立。我们将此性质的定量语义定义为自车与其他车辆在一条轨迹中的最小距离。我们将所有碰撞一视同仁。

2) **不要停止**。 如果自车在轨迹过程中完全停下，则此性质不成立。此性质的定量语义由自车在一条轨迹中达到的最低速度来衡量。

3) **不要突然制动**。 如果自车的制动强度超过预先设定的阈值（设为自车最大制动能力的 40%），则此性质不成立。此性质的定量语义由自车在一条轨迹中施加的最大制动值来衡量。

我们将 SGDA 与以下环境采样方法作为基线方法进行比较：

1) **均匀先验采样**：为了从专家处汇集更多数据，我们从环境的先验分布中采样（将所有参数设为均匀分布），并在每一轮中利用这些样本从专家处收集更多轨迹用于训练。2) **单一规范证伪**：我们不使用 STP 对环境进行采样，而是定义一个单一规范，用来表示我们感兴趣的结果并针对能够证伪该规范的环境条件进行采样。我们将该规范设为上述所有原始（未取反）属性的合取，以描述自车的“典型”驾驶行为。

3) **单项属性重要性采样**：不采用 SGDA 方法中创建的划分时，另一种利用属性集合中所有属性的方法是，分别对每项属性进行采样以实现证伪（彼此独立）。我们将待聚合的环境数量按属性数量进行划分，并对每项属性进行重要性采样；对于每项属性，我们都采用与 [32] 中类似的方法，并使用固定数量的样本。

我们通过一组定量指标评估所提方法的效果，这些指标既衡量模仿结果在语义上与专家的相似程度，也衡量专家轨迹与模仿轨迹在物理上的相似程度：

1) **结果匹配：** 与第 III-B 节所述的环境选择方法类似，我们衡量在相同 e 下，专家轨迹满足与学习模型轨迹相同规范（即结果）的频率。2) **轨迹长度距离：** 为衡量轨迹层面的差异，我们使用动态时间规整（DTW）距离，比较专家与学习模型在相同 e 下的轨迹。3) **测试集损失：** 我们使用从专家收集的状态和动作测试集上的 L1 损失，衡量学习模型在相同状态下所采取的动作与专家动作的差异。

我们通过模仿两种不同的专家来评估该方法：CARLA 内置自动驾驶车辆控制器（AP），以及一个神经网络控制器（NN）。该神经网络专家控制器使用从人类驾驶员操作我们环境中车辆所收集的数据进行训练。训练数据共计两小时，采集方式是让人类反复操纵方向盘和踏板，在仿真中驾驶自车穿过交叉路口。我们指示驾驶员尽可能快速地通过交叉路口，同时尽可能避免碰撞。对于自动驾驶仪，我们还针对另外两种机动动作进行了两项实验：在交叉路口右转或左转。在自动驾驶仪的右转和左转实验中，我们只学习机动动作中油门分量的模仿，并让专家策略控制转向，以避免在数据受限的情况下同时学习具有挑战性的横向控制技能，从而掩盖模仿的结果。

我们使用行为克隆（BC）算法来学习专家的模仿策略，其中状态空间由自车信息（位置、速度、加速度）以及对其他交通参与者位置和速度的带噪估计组成；当其他交通参与者处于自车的视线范围内时，才包含这些估计。

我们收集了一个包含 200 条轨迹的初始专家数据集（平均约对应 90,000 个状态-动作对），

其中的环境条件从模拟器对所有环境参数设定的均匀先验中采样。后续每轮数据聚合包含 100 个回合。对于 SGDA、单规范基线和单项属性采样基线，采样步骤会生成 200 个样本，并选取其中 100 个用于聚合。在这 200 个样本中，前 100 个通过均匀随机采样收集，用于为每个规范初始化我们的优化引导的证伪器。在 SGDA 中，后续 100 个样本使用算法 2 收集。在我们的基线方法中，这 100 个样本要么用于证伪该规范（在单规范基线中），要么在各项性质之间均分，并通过重要性采样选取（在单项属性采样基线中）。我们的优化引导的证伪器使用贝叶斯优化来学习环境参数上的分布，其中提供的目标值是优化器试图证伪的 STL 规范的定量评估值。除该目标值外，我们还加入一项用于鼓励探索的项，以防止分布将过多概率质量集中在参数空间的狭小区域（详情见附录）。我们报告了每种方法经过两轮数据聚合后的结果。

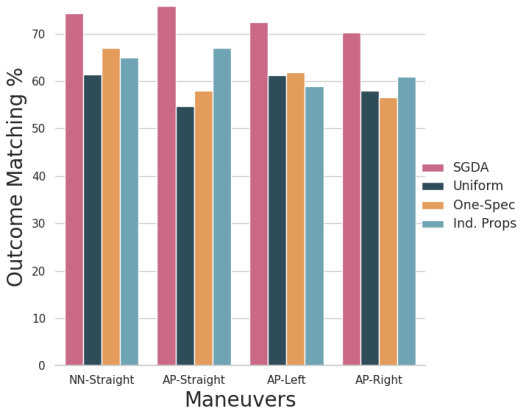


图 4：模仿专家所执行机动动作的结果匹配率。AP 指由自动驾驶仪控制的机动动作，NN 指由我们的神经网络专家控制的机动动作。在学习语义上更准确的模仿方面，SGDA方法优于所有基线方法。

V. 结果

在本节中，我们将性能与基线方法进行比较，以考察 SGDA 是否能生成与专家在语义上更相似的模仿结果（图 4 和表 II），SGDA 生成的轨迹是否更接近专家的轨迹（图 5），以及 SGDA 是否能取得具有竞争力的测试集损失（表 III）。

我们在从每位专家针对每种机动收集的测试数据集上评估各方法的性能。收集测试集时，我们力求避免数据集中不太可能出现的结果占比过高，同时确保每种结果都至少得到一定程度的体现。为避免测试数据集偏向 SGDA（它会明确地对代表不足的规范进行采样），我们在每项实验中都从专家处收集了一个包含 500 条轨迹的数据集，并从所有环境参数的均匀分布中采样。为确保每种规范都得到最低限度的体现，我们首先针对每种规范进行拒绝采样，直到满足该规范的轨迹达到至少 20 条（占整个数据集的 4%）。

评估指标分析。在图 4 中，我们展示了使用 SGDA 和环境采样基线方法训练的模仿学习模型在结果匹配方面的结果。在每种机动中，对于每个模型，使用 SGDA 训练的模型，其结果匹配率都高于所有基线方法。直观地说，我们可以将这一结果解读为：与基线方法相比，SGDA 训练的模型总体上在语义准确性方面更接近专家。此外，在图 5 中，我们展示了 SGDA 与基线方法的对比结果，比较在相同环境条件下采集的一对学习模型轨迹和专家轨迹之间的平均轨迹长度距离。我们发现，SGDA 训练的模型始终具有更低的平均轨迹长度距离，这支持了以下观点：SGDA 能使模型不仅在语义上更贴近地模仿专家，而且在（模拟的）物理意义上也能实现更接近的模仿。作为最后一项比较，我们计算了每个模型的测试集损失，结果见 III。尽管这一指标无法捕捉语义相似性，也无法反映轨迹长度执行过程中可能累积的误差，但我们注意到，SGDA 训练的模型具有与基线方法相当（且略低）的测试损失。

鉴于 SGDA 学习模型展现出的模仿能力提升，我们希望了解哪些语义结果对这些提升的贡献最大。为此，我们计算测试数据集中每种可能规范（结果）的结果匹配率，并将 SGDA 与基线方法进行比较。

结果专项分析。表 II 展示了在人类学习神经网络模仿实验中，专家在测试数据集上实现的各项规范对应的结果匹配率。结果表明，SGDA 学习模型相较于基线方法的提升，主要归功于对较少见结果的匹配。例如，基线方法完全无法匹配专家发生碰撞并急刹车但未停止的结果（$\neg 1 \wedge 2 \wedge \neg 3$），而我们的 SGDA 方法能够匹配近一半此类结果。专家发生碰撞并在急刹车时停止的结果（$\neg 1 \wedge \neg 2 \wedge \neg 3$）也有类似情况。要正确匹配这些与其他结果高度相关的结果（即必须刹车才能停止），就需要能够模仿专家所展现的细腻行为。

相比之下，SGDA 学习得到的模型在更常见的结果中准确率相当，但往往略低，例如专家未发生碰撞、未停止或未急刹车的结果（$1 \wedge 2 \wedge 3$），或专家未发生碰撞但停止并急刹车的情况（$1 \wedge \neg 2 \wedge \neg 3$）。这种差异

表 II：专家测试数据集中每项规范的样本数，以及神经网络专家直行操控实验中各基线方法的结果匹配率。SGDA 在大多数结果特定准确率指标上优于基线方法，并且在所有结果上的准确率都相当。

(1) 避让碰撞	(2) 不要 Halt	(3) 不要急刹车	# 结果	神经网络直行匹配率 SGDA	独立属性	单一规格	均匀
✓	✓	✓	197	94%	96%	**97%**	95%
✓	✓	✗	20	**50%**	40%	0%	0%
✓	✗	✓	44	**52%**	41%	14%	23%
✓	✗	✗	65	78%	62%	**86%**	57%
✗	✓	✓	20	**95%**	85%	70%	50%
✗	✓	✗	20	**45%**	0%	0%	0%
✗	✗	✓	94	**80%**	61%	79%	66%
✗	✗	✗	40	**23%**	5%	0%	0%

表 III：各实验中每个基线在我们的专家测试数据集上的 L1 测试损失值。SGDA 始终取得较低的 L1 误差。

测试损失	NN-Straight	AP-Straight	AP-Left	AP-Right
SGDA	**.059**	**.044**	**.066**	**.052**
均匀	.072	.049	.068	.054
单一规格	.067	.067	.071	.061
独立属性	.063	.057	.083	.056

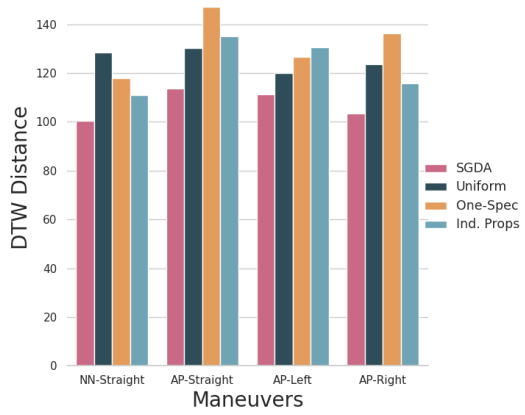


图 5：采用动态时间规整（DTW）距离衡量专家所执行机动动作的模仿效果，其轨迹长度距离结果。SGDA方法在生成与专家最相似的轨迹方面优于所有基线方法。

在性能上的差异可直接归因于各基线的训练数据分布不均衡。以我们的单一规格基线采样为例，该方法旨在仅证伪一项规范，并找到了大量以非常相似方式实现这一点的环境条件（具体而言，是找到满足（$1 \wedge \neg 2 \wedge \neg 3$）的条件）。事实上，在单一规格基线的专家训练数据中，38% 的数据满足（$1 \wedge 2 \wedge 3$），53% 满足（$1 \wedge \neg 2 \wedge \neg 3$）。相比之下，对于 SGDA 的训练数据，这些规范分别为 35% 和 24%。总体而言，我们的结果表明，SGDA 将极大地帮助改善在罕见但语义上有意义的结果中的模仿行为，而不会显著改变常见结果中的行为准确率。

属性鲁棒性分析。最后，为了深入了解 SGDA 在集合 LP 中属性变化时的鲁棒性，我们针对“不要突然制动”这一性质中的制动值采用不同阈值，重复进行 NN-Expert 实验。我们注意到，该性质出现的罕见程度确实会影响 SGDA 模仿专家在该性质上的行为的能力。不过，我们注意到，即使该性质变得罕见，SGDA 仍优于与属性无关的基线方法——均匀采样；而当该性质出现得更频繁时，SGDA 通常能够稳健应对变化。有关该实验的完整细节，请参见附录。

VI. 结论

本文提出了一种方法，利用与任务相关的信息，在模仿学习场景中系统地采样新的环境，以进行数据聚合。在该方法中，我们引入了语义轨迹划分这一概念，将可能的环境空间划分为若干元素，每个元素代表对智能体而言语义不同的结果。我们的实验表明，使用该方法聚合数据的效果优于现有的环境采样方法，尤其是在那些可能性较低、但对用户而言具有语义意义的情形中。

未来研究有许多可能的方向。可以将 SGDA 与其他 DAgger 式方法结合，通过控制何时以及何处收集专家数据，进一步改进学习得到的模仿策略。另一个方向是将 SGDA 用于人类建模场景中的模仿学习，尤其是针对正在进行仿真训练的个体，以便更好地了解他们有待提升之处。

致谢

丰田研究院为本研究提供了资金支持。A.Shah 获得美国海军研究办公室提供的 NDSEG奖学金资助。

参考文献

[1] Houssam Abbas、Bardh Hoxha、Georgios Fainekos、Jy- otirmoy V Deshmukh、James Kapinski 和 Koichi Ueda。将符合性测试作为信息物理系统的证伪方法。*arXiv*预印本 *arXiv:1401.5200*，2014年。[2] Takumi Akazaki、Shuang Liu、Yoriyuki Yamagata、Yihai Duan 和 Jianye Hao。利用深度强化学习对信息物理系统进行证伪。载于形式化方法：第*22*届国际研讨会 *FM 2018*，作为联合逻辑会议 *FloC 2018* 的一部分举行，英国牛津，*2018*年*7*月*15*日至*17*日，论文集第*22*卷，第456–465页。施普林格，2018年。[3] Alexander Amini、Wilko Schwarting、Guy Rosman、Brandon Araki、Sertac Karaman 和 Daniela Rus。用于自动驾驶端到端控制并结合新颖性检测和训练去偏的变分自编码器。载于*2018*年 *IEEE/RSJ* 机器人与智能系统国际会议，*IROS 2018*，西班牙马德里，*2018*年*10*月*1*日至*5*日，第568–575页。IEEE，2018年。数字对象标识符：10.1109/IROS.2018.8594386。网址：https://doi.org/10.1109/IROS.2018.8594386。[4]Yashwanth Annpureddy、Che Liu、Georgios Fainekos 和 Sriram Sankaranarayanan。S-taliro：用于混合系统时序逻辑证伪的工具。载于系统构建与分析工具和算法：第*17*届国际会议 *TACAS 2011*，作为软件理论与实践联合欧洲会议 *ETAPS 2011* 的一部分举行，萨尔布¨吕肯，德国，*2011*年*3*月*26*日至*4*月*3*日。论文集第*17*卷，第254–257页。施普林格，2011年。[5]Mansur Arief、Zhiyuan Huang、Guru Koushik Senthil Kumar、Yuanlu Bai、Shengyi He、Wenhao Ding、Henry Lam 和 Ding Zhao。深度概率加速评估：一种适用于黑盒安全关键系统的稳健、可认证稀有事件仿真方法。载于*AISTATS*，机器学习研究论文集第*130*卷，第595–603页。PMLR，2021年。[6] Peter Auer。利用置信界权衡利用与探索。机器学习研究杂志，3：397–422，2002年。网址：http://jmlr.org/papers/v3/auer02a.html。[7]Mayank Bansal、Alex Krizhevsky 和 Abhijit Ogale。Chauffeurnet：通过模仿最佳驾驶者并综合最差驾驶行为来学习驾驶。*arXiv*预印本 *arXiv:1812.03079*，2018年。[8] Raunak P. Bhattacharyya、Blake Wulfe、Derek J. Phillips、Alex Kuefler、Jeremy Morton、Ransalu Senanayake 和 Mykel J. Kochenderfer。通过生成对抗模仿学习对人类驾驶行为建模。*CoRR*，abs/2006.06412，2020年。网址：https://arxiv.org/abs/ 2006.06412。[9]Andrea Censi、Konstantin Slutsky、Tichakorn Wong- piromsarn、Dmitry S. Yershov、Scott Pendleton、James Guo Ming Fu 和 Emilio Frazzoli。利用规则手册规范责任、伦理与文化感知行为。载于机器人与自动化国际会议，*ICRA 2019*，蒙特利尔，魁北克省，加拿大，*2019*年*5*月*20*日至*24*日，第8536–8542页。IEEE，2019年。数字对象标识符：10.1109/ICRA. 2019.8794364。网址：https://doi.org/10.1109/ICRA.2019. 8794364.[10] Jianyu Chen、Bodi Yuan 和 Masayoshi Tomizuka。通过增强安全性的深度模仿学习，在通用城市场景中实现自动驾驶。载于*2019*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第2884–2890页。IEEE，2019年。[11]Glen Chou、Yunus Emre Sahin、Liren Yang、Kwesi J Rutledge、Petter Nilsson 和 Necmiye Ozay。利用控制综合生成边界案例：自动驾驶案例研究。*IEEE*集成电路与系统计算机辅助设计汇刊，37(11)：2906–2917，2018年。[12]Felipe Codevilla、Matthias M¨uller、Antonio L´opez、Vladlen Koltun 和 Alexey Dosovitskiy。通过条件模仿学习实现端到端驾驶。载于*2018*年*IEEE*机器人与自动化国际会议（*ICRA*），第4693–4700页。IEEE，2018年。[13] Felipe Codevilla、Matthias M¨uller、Antonio M. L´opez、Vladlen Koltun 和 Alexey Dosovitskiy。通过条件模仿学习实现端到端驾驶。载于*2018*年*IEEE*机器人与自动化国际会议，*ICRA 2018*，澳大利亚布里斯班，*2018*年*5*月*21*日至*25*日，第1–9页。IEEE，2018年。数字对象标识符：10.1109/ICRA.2018.8460487。网址：https://doi.org/10.1109/ICRA.2018.8460487.[14]Felipe Codevilla、Eder Santana、Antonio Lopez 和 Adrien Gaidon。探索行为克隆在自动驾驶中的局限性。载于*2019*年*IEEE/CVF*计算机视觉国际会议（*ICCV*），第9328–9337页，2019年。数字对象标识符：10.1109/ICCV.2019.00942。[15] Yuchen Cui、David Isele、Scott Niekum 和 Kikuo Fu-jimura。用于深度模仿学习的不确定性感知数据聚合。载于机器人与自动化国际会议，*ICRA 2019*，蒙特利尔，魁北克省，加拿大，*2019*年*5*月*20*日至*24*日，第761–767页。IEEE，2019年。数字对象标识符：10.1109/ICRA.2019.8794025。网址：https://doi.org/10. 1109/ICRA.2019.8794025.[16] Wenhao Ding、Baiming Chen、Bo Li、Kim Ji Eun 和 Ding Zhao。用于决策算法评估的多模态安全关键场景生成。*IEEE*机器人与自动化快报，6(2)：1551–1558，2021年。[17]Wenhao Ding、Bo Li、Kim Ji Eun 和 Ding Zhao。在显式知识指导下实现语义可控的场景生成。*CoRR*，abs/2106.04066，2021年。[18] Tommaso Dreossi、Daniel J Fremont、Shromona Ghosh、Edward Kim、Hadi Ravanbakhsh、Marcell Vazquez-Chanlatte 和 Sanjit A Seshia。Verifai：用于人工智能系统形式化设计与分析的工具包。载于计算机辅助验证：第*31*届国际会议，*CAV 2019*，美国纽约市，纽约州，*2019*年*7*月*15*日至*18*日，论文集，第一部分 *31*，第432–442页。施普林格，2019年。[19]Daniel J. Fremont、Edward Kim、Tommaso Dreossi、Shromona Ghosh、Xiangyu Yue、Alberto L. Sangiovanni-

Vincentelli 和 Sanjit A. Seshia。Scenic：用于场景规范和数据生成的语言。*CoRR*，abs/2010.06580，2020年。[20] Shangding Gu、Long Yang、Yali Du、Guang Chen、Florian Walter、Jun Wang、Yaodong Yang 和 Alois Knoll。安全强化学习综述：方法、理论与应用。*arXiv*预印本 *arXiv:2205.10330*，2022年。[21]Mohammad Hekmatnejad、Shakiba Yaghoubi、Adel Dokhanchi、Heni Ben Amor、Aviral Shrivastava、Lina Karam 和 Georgios Fainekos。使用信号时序逻辑对自动驾驶车辆的责任敏感安全规则进行编码与监控。载于第*17*届 *ACM-IEEE*系统设计形式化方法与模型国际会议论文集，MEMOCODE '19，美国纽约州纽约，2019年。计算机协会。国际标准书号：9781450369978。数字对象标识符：10.1145/3359986.3361203。网址：https://doi.org/10.1145/ 3359986.3361203.[22] Jonathan Ho 和 Stefano Ermon。生成对抗模仿学习。*CoRR*，abs/1606.03476，2016年。网址：http://arxiv.org/abs/1606.03476。[23] Xin Huang、Stephen McGill、Brian C Williams、Luke Fletcher 和 Guy Rosman。城市交叉路口的不确定性感知驾驶员轨迹预测。2019年1月。[24]Parham M Kebria、Abbas Khosravi、Syed Moshfeq Salaken 和 Saeid Nahavandi。基于卷积神经网络的自主车辆深度模仿学习。*IEEE/CAA*自动化学报，7(1)：82–95，2019年。[25] Michael Kelly、Chelsea Sidrane、Katherine Rose Driggs-Campbell 和 Mykel J. Kochenderfer。HG-DAgger：与人类专家进行交互式模仿学习。载于机器人与自动化国际会议，*ICRA 2019*，蒙特利尔，魁北克省，加拿大，*2019*年*5*月*20*日至*24*日，第8077–8083页。IEEE，2019年。数字对象标识符：10.1109/ICRA. 2019.8793698。网址：https://doi.org/10.1109/ICRA.2019. 8793698.[26] Piyush Khandelwal、Elad Liebman、Scott Niekum 和 Peter Stone。蒙特卡罗树搜索中复杂备份策略的分析。载于 Maria-Florina Balcan和Kilian Q. Weinberger编，《第*33*届国际机器学习会议论文集，*ICML 2016*，美国纽约市，*2016*年*6*月*19*日至*24*日》，第48卷*JMLR*研讨会与会议论文集，第1319–1328页。JMLR.org，2016年。网址：http://proceedings.mlr.press/v48/khandelwal16.html.[27] Alex Kuefler、Jeremy Morton、Tim Wheeler 和 Mykel Kochenderfer。利用生成对抗网络模仿驾驶员行为。载于*2017*年*IEEE*智能车辆研讨会（*IV*），第204–211页。IEEE，2017年。[28] Ritchie Lee、Ole Mengshoel、Anshu Saksena、Ryan Gardner、Daniel Genin、Joshua Silbermann、Michael Owen 和 Mykel Kochenderfer。自适应压力测试：利用强化学习寻找可能的失效事件。人工智能研究杂志，69：1165–1201，2020年12月。数字对象标识符：10.1613/jair.1.12190。[29] Yiren Lu、Justin Fu、George Tucker、Xinlei Pan、Eli Bron- stein、Becca Roelofs、Benjamin Sapp、Brandyn White、Aleksandra Faust、Shimon Whiteson 等。模仿并不足够：利用强化学习增强模仿，以应对具有挑战性的驾驶场景。*arXiv*预印本 *arXiv:2212.11419*，2022年。[30]Sebastian Maierhofer、Anna-Katharina Rettinger、Eva Charlotte Mayer 和 Matthias Althoff。时序逻辑中的州际公路交通规则形式化。在*2020*年*IEEE*智能车辆研讨会（*IV*），第752–759页，2020年。数字对象标识符：10.1109/IV47402.2020.9304549。[31] Oded Maler 和 Dejan Nickovic。连续信号的时序性质监测。在 *FORMATS 2004* 和 *FTRTFT 2004*，法国格勒诺布尔，*2004*年*9*月*22*日至*24*日，第152–166页。施普林格，2004年。[32] Matthew O'Kelly、Aman Sinha、Hongseok Namkoong、Russ Tedrake 和 John C. Duchi。通过稀有事件仿真实现可扩展的端到端自动驾驶汽车测试。在 *NeurIPS*，第9849–9860页，2018年。[33]Aditya Prakash、Aseem Behl、Eshed Ohn-Bar、Kashyap Chitta 和 Andreas Geiger。探索基于视觉的城市自动驾驶策略学习中的数据聚合。在*2020*年*IEEE/CVF*计算机视觉与模式识别会议，*CVPR 2020*，美国华盛顿州西雅图，*2020*年*6*月*13*日至*19*日，第11760–11770页。计算机视觉基金会 / IEEE，2020年。数字对象标识符：10.1109/CVPR42600.2020.01178。URL https://openaccess.thecvf.com/content CVPR 2020/html/Prakash Exploring Data Aggregation in Policy Learning for Vision-Based Urban Autonomous CVPR 2020 paper.html。[34] Charles Richter 和 Nicholas Roy。通过深度学习和新颖性检测实现安全的视觉导航。在机器人学：科学与系统第十三届会议。机器人学：科学与系统基金会，2017年7月。[35] Albert Rizaldi 和 Matthias Althoff。为自主车辆的问责实现交通规则形式化。在*2015*年第*18*届 *IEEE*智能交通系统国际会议，第1658–1665页，2015年。数字对象标识符：10.1109/ITSC.2015.269。[36] St´ephane Ross、Geoffrey Gordon 和 Drew Bagnell。将模仿学习和结构化预测简化为无悔在线学习。在第十四届国际人工智能与统计学会议论文集，第627–635页，2011年。[37] Sriram Sankaranarayanan 和 Georgios Fainekos。使用交叉熵方法对混合系统的时序性质进行证伪。在第*15*届 *ACM*混合系统：计算与控制国际会议论文集，第125–134页，2012年。[38] Jonathan C. Spencer、Sanjiban Choudhury、Arun Venka- traman、Brian D. Ziebart 和 J. Andrew Bagnell。模仿学习中的反馈：协变量偏移的三种状态。*CoRR*，abs/2102.02872，2021年。URL https://arxiv. org/abs/2102.02872。[39] J. Uesato，A. Kumar，Cs. Szepesv´ari，T. Erez，A. Ruder-

man、K. Anderson、K. Dvijotham、N. Heess 和 P. Kohli。《严格的智能体评估：一种揭示灾难性故障的对抗方法》。载于*ICLR*，第3692–3702页，2019年6月。[40] Jingkang Wang、Ava Pun、James Tu、Sivabalan Mani- vasagam、Abbas Sadat、Sergio Casas、Mengye Ren 和 Raquel Urtasun。Advsim：为自动驾驶车辆生成安全关键场景。载于*CVPR*，第9909–9918页。计算机视觉基金会 / IEEE，2021年。[41] Ziyu Wang、Josh Merel、Scott Reed、Greg Wayne、Nando de Freitas 和 Nicolas Heess。鲁棒模仿多样化行为。NIPS'17，第5326–5335页，美国纽约州雷德胡克，2017年。Curran Associates Inc. 国际标准书号9781510860964。.. .[42] Bowen Weng、Linda Capito、Umit Ozg¨uner 和 Keith A. Redmill。通过场景采样对黑箱系统安全性能进行形式化刻画。*IEEERoboticsAutom. Lett.*，7(1)：199–206，2022年。[43] Jiakai Zhang 和 Kyunghyun Cho。面向端到端模拟驾驶的高效查询模仿学习。载于 Satinder Singh 和 Shaul Markovitch（编辑），第三十一届*AAAI*人工智能大会论文集，*2017*年*2*月*4*日至*9*日，美国加利福尼亚州旧金山，第2891–2897页。AAAI出版社，2017年。URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14590。[44]Songyuan Zhang、Zhangjie Cao、Dorsa Sadigh 和 Yanan Sui。基于不同最优程度示范的置信度感知模仿学习。神经信息处理系统进展，34：12340–12350，2021年。[45] Konrad Zolna、Scott E. Reed、Alexander Novikov、Ser- gio Gomez Colmenarejo、David Budden、Serkan Cabi、Misha Denil、Nando de Freitas 和 Ziyun Wang。与任务相关的对抗模仿学习。载于机器人学习会议，2019年。

附录

更多实验细节与结果

实验实现细节 正如正文中所述，我们采用贝叶斯优化与高斯过程，将其作为优化引导采样器。当使用这些优化器之一进行采样时，我们会最大化一个目标值，该目标值等价于所采样 STL 规范的定量语义。对于单项属性基线，其规范简单表示为$G(p)$，其中 $p \in LP$ 是各个属性本身，而 G 是时序逻辑语义中的全局算子。为鼓励探索，我们在优化器的目标值中加入一项，仅当当且仅当该样本及其对应轨迹满足该规范时，才会增加以下量：

$$a \cdot \min_{\xi_\pi \in \varphi} (\text{distance}(\xi_\pi, \xi_\pi^c))$$

这里， ξ_π^c 是当前正在评估的采样环境条件所对应的轨迹，而 a 是我们的探索超参数。直观而言，我们计算当前轨迹与所有满足相同轨迹的其他轨迹之间的最小距离，并将其用作探索奖励。尽管这无法保证最低程度的探索，但确实会激励我们在环境条件空间中寻找同样满足我们规范的新区域。对于所有采用基于优化的采样的基线方法，我们都使用相同的探索奖励，并采用相同的超参数乘数 a 。

对于我们的行为克隆模型，我们采用一个由 3 层组成、每层 256 个单元的多层感知机网络作为模型类别。我们在其训练数据集上训练模型 100 个轮次，学习率为 0.0001，批量大小为 500。所有训练均在 NVIDIA RTX A4500 上进行；CARLA 环境的采样和仿真也使用了同一硬件。我们的训练数据是一个特征向量，包含轨迹和环境的相关信息，即自车的位置、速度、航向和加速度；如果其他交通参与者位于视线范围内（即未被环境中的遮挡挡住），还包括该参与者的位置和速度。我们通过注入高斯噪声对这些数据进行扰动（位置坐标和速度所加噪声的均值均为零，方差分别为二和一。）

其他结果。 在表 IV、V 和 VI 中，我们展示了 CARLA 自动驾驶仪各项机动动作的结果特定匹配率。与神经网络专家的结果类似，我们可以看到，SGDA 在获得更高的结果特定匹配率方面总体上优于基线方法，尤其是在结果极不可能出现的情况下。对于更可能出现的结果，SGDA 的表现能够接近基线方法的水平；由于基线方法在训练中见过更多此类结果，因此表现更好，正文中提到的示例便属于这种情况。这些差异部分可归因于训练期间提供给每个模型的数据量有限。我们预计，随着数据量增加，相较于基线方法，SGDA 的表现还会进一步提升，因为 SGDA 会聚合其匹配效果较差的结果上的数据（包括不太可能出现的结果和

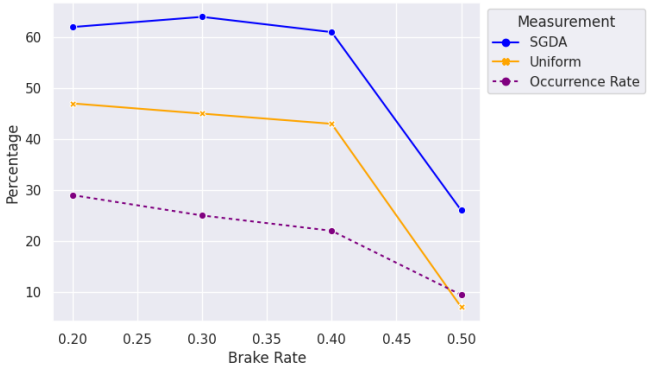


图 6：改变制动率阈值时，SGDA 和均匀基线在违反制动属性情况下的匹配率（橙色和蓝色）。我们还展示了阈值设为不同值时该性质被违反的频率（紫色）。

较可能出现的结果），而基线方法没有一种有原则的方法来采样许多不同的不太可能出现的结果，并在此类情境下改进其模仿效果。

鲁棒性实验。如正文所述，我们进行了一项简单实验，观察当我们改变集合 LP 中某个性质时，SGDA 的表现如何变化。具体而言，我们重复 NN-Expert 实验，并改变“不要突然制动”这一性质中的制动阈值，同时保持其他性质不变。改变该阈值后，该性质在不同环境条件下被专家违反的频率也会随之变化。我们按照正文所述的实验设置，使用相同的附加性质（避免碰撞和不要停止）运行两轮 SGDA。然后，我们测量 SGDA 在仅我们所改变的性质上的匹配百分比。具体来说，我们观察：如果专家在阈值设为某个值时违反了制动性质，那么我们学习得到的模仿策略在该阈值下是否也会违反这一性质？我们单独研究该性质上的表现，不考虑其他性质。

我们在图 6 中展示了结果。我们收集了一个包含 500 个回合的测试数据集，从环境条件的均匀先验分布中采样（与主实验不同，这里没有进行拒绝采样）。紫色（虚线）表示在给定制动阈值时，测试数据集中专家轨迹违反该性质的频率。可以看到，随着制动阈值（x 轴）增大，观察到违规的频率略有下降，随后在阈值从 0.4 增至 0.5 时显著降低。蓝线表示 SGDA 对这一单一性质的**违规**匹配百分比。为与基线进行比较，我们还展示了均匀基线的相同结果，以橙色表示。可以看到，当阈值从 0.2 变至 0.4 时，即使违规频率从略低于 30% 降至略高于 20%，SGDA 的匹配率变化也很小。然而，当阈值为 0.5、违规频率降至 10% 以下时

，SGDA 和均匀基线的表现都急剧下降。这表明，当某一性质出现的频率低于特定阈值时，该频率会对 SGDA 的表现产生显著影响。

我们认为，在实践中，扩大可供采样和选择的环境池，以及增加数据聚合轮次，都有助于缓解此类性能下降。测量 LP 中所含属性的相对频率，有助于确定使用 SGDA 时的超参数选择，例如需要聚合的数据量、需要采样和选择的环境数量，以及在环境采样期间如何在 UCB 算法中平衡利用与探索。属性的具体性质同样重要——我们观察到，在神经网络专家以较高制动率急刹车的情况下，这一高制动率只会应用于最初的少数几个时间步，随后制动率下降，专家的减速过程也变得更加渐进。我们的模仿策略在学习过程中难以复现这一精确行为，这进一步解释了为什么当阈值增大到 0.5 时性能会下降。

算法补充细节

属性权重。 在第三节 A 中，我们指出，STP方法的一个潜在问题是，其划分大小随属性集 $p \in LP$ 的大小呈指数增长。如果我们在一个极大的划分上运行原始算法，就很可能会为一些相较于其他规范而言价值较低的规范进行采样，从而导致环境采样和选择效率低下。为解决这一问题，我们将 SGDA 和 STP构造推广到这样一种情形：用户可以为其属性集中的每个属性提供权重 ，以量化某一属性的存在相对于其他属性在语义上的重要程度。

形式上，我们将加权属性集定义为一个元组 $(w, p) \in LP^w$，其中与某个性质 w相关联的每个权重都是范围为 (0,1) 的实数值。然后，我们可以将这些权重相乘（对于否定属性，则将其补数相乘），从而为整个规范赋予权重。为了更具体地说明这一点，回想一下第三节 A 中介绍的简单示例，其中只有两个属性：**A** （避免碰撞）和**B** （避免超速）。假设用户非常关注发生碰撞的结果，而对看到汽车没有超速的结果只稍有兴趣。基于这一考虑，用户为以下属性赋予权重：**A** 的权重为 0.1，**B** 的权重为 0.6。这意味着**A** 的否定（存在碰撞时）被赋予与**A**互补的权重，即 (1 - 0.1) = 0.9，而**B**的否定则被赋予 (1 - 0.6) = 0.4。随后将这些权重相乘，以形成 STP 中每个规范的权重。在此示例中，$(A \wedge B)$ 的权重为 $(0.1 * 0.6) = 0.06$，$(\neg A \wedge B)$ 的权重为 $(0.9 * 0.6) = 0.54$，$(A \wedge \neg B)$ 的权重为 $(0.1 * 0.4) = 0.04$，而 $(\neg A \wedge \neg B)$ 的权重为 $(0.9 * 0.4) = 0.36$。

一旦计算出 STP 中每个规范的权重，在 SGDA 算法中利用这些权重在概念上就很简单。我们可以将每个规范的权重用作乘数，应用于算法 2 中为每个规范定义的 Q值。随后，UCB算法将使用这些加权 Q值来选择后续要采样的规范。这意味着权重较低的规范会具有较低的 Q值，在 SGDA 的采样步骤中被采样的可能性也较低。

SGDA的计算开销。与其他环境采样方法不同，SGDA算法相对于算法的超参数会产生渐近计算开销。我们在此对这一开销作非正式讨论。

SGDA所承担的计算开销主要在采样步骤（EC-Samp）中。算法 2 概述了该步骤，其中以下计算会带来不可忽略的开销：(1) 调用基于优化的采样器以获取新的环境条件；(2) 执行动力学以收集轨迹 $\xi_{\pi^{\mathrm{IL}}}(e) = d(\pi^{\mathrm{IL}}, e)$，并检查其落入划分中的哪项规范（$e$，$\xi_{\pi^{\mathrm{IL}}}(e)$）；(3) 计算每项规范的 Q 值，并调用 UCB 计算，利用这些 Q 值选择下一项要采样的规范。

关于 (1)，优化引导采样器学习到的分布通常是通过迭代方式在线学习的，因此这项计算只会带来较小的开销。关于 (3)，这些操作的计算复杂度会随划分规模线性增长，但计算每项规范的 Q 值和 UCB 值只需要简单的算术运算，不会给计算带来太大负担。无论从算法层面还是实践中，我们都发现 (2) 带来的开销最大：针对给定环境条件调用动力学可能代价高昂，而且每次采样都要调用。监测每项规范，以确定给定的 $(e, \xi_{\pi^{\mathrm{IL}}}(e))$ 对是否满足 $\varphi \in \Phi$，最坏情况下还需要评估每个 φ 的语义。

我们指出，这些操作的组合在实际应用中会带来不小但尚可承受的开销，而且该开销会随 Φ 的规模线性增长，并随 LP 的规模呈指数增长。不过，这些开销中的许多部分也为其他环境采样技术（例如我们的两种基线方法）所共有，具体包括调用优化引导采样器和驱动动力学。更重要的是，正如实验所表明的，我们以计算量的增加换取模型性能的提升，从而减少对专家模型的需求。当专家模型的查询成本高昂时，例如在人类研究中，这类改进至关重要；我们认为这是 SGDA 的一项主要贡献。

表 IV：自动驾驶直行机动实验中专家测试数据集中每项规范的样本数量，以及各基线的结果匹配率。

(1) 避让碰撞	(2) 不要 Halt	(3) 不要急刹车	# 结果	AP 直行匹配率 SGDA	独立属性	单一规格	均匀
✓	✓	✓	226	91%	85%	**92%**	**92%**
✓	✓	✗	20	**15%**	0%	0%	0%
✓	✗	✓	67	**92%**	76%	21%	34%
✓	✗	✗	65	88%	**91%**	75%	44%
✗	✓	✓	20	**80%**	**80%**	55%	50%
✗	✓	✗	26	**31%**	23%	15%	0%
✗	✗	✓	20	**70%**	50%	20%	0%
✗	✗	✗	56	**20%**	0%	0%	7%

表 V：专家测试数据集中每种规范的样本数量，以及自动驾驶仪左转机动实验中各基线的结果匹配率。

(1) 避让碰撞	(2) 不要 Halt	(3) 不要急刹车	# 结果	AP-左匹配率 SGDA	独立属性	单一规格	均匀
✓	✓	✓	191	82%	**89%**	**89%**	88%
✓	✓	✗	20	**20%**	0%	0%	0%
✓	✗	✓	115	92%	**96%**	95%	87%
✓	✗	✗	20	**65%**	0%	**65%**	10%
✗	✓	✓	20	**35%**	**35%**	10%	0%
✗	✓	✗	68	**37%**	0%	13%	**37%**
✗	✗	✓	20	**90%**	40%	5%	0%
✗	✗	✗	46	**70%**	0%	0%	20%

表六：专家测试数据集中每种规范的样本数量，以及自动驾驶仪右转机动实验中各基线的结果匹配率。

(1) 避让碰撞	(2) 不要 Halt	(3) 不要急刹车	# 结果	AP-右匹配率 SGDA	独立属性	单一规格	均匀
✓	✓	✓	195	88%	**93%**	91%	**93%**
✓	✓	✗	20	**20%**	5%	0%	0%
✓	✗	✓	97	**73%**	67%	34%	60%
✓	✗	✗	20	**30%**	0%	0%	0%
✗	✓	✓	20	**80%**	50%	55%	70%
✗	✓	✗	55	**62%**	25%	18%	16%
✗	✗	✓	37	**84%**	41%	79%	38%
✗	✗	✗	56	28%	34%	**39%**	23%