NVIDIA Halos 安全评估框架

作者：Sever Topan (US)、Anitha Anand (US)、Joerg Rossow (DE)、Jonathan Mcneely (US)、Viola Wu (US)、Apoorva Sharma (US)、Martijn Tideman、Matteo Oldoni (IT)、Jonas Nilsson (US)、Riccardo Mariani (IT)

引言

证明智能汽车的安全性，需要证明其性能达到大约每 10^7 至 10^8 小时驾驶才发生一次失效的量级——这远超任何量产前车队能够实际行驶的规模，也不可能由单一工具独立完成。因此，需要将道路实测、基于仿真的评估以及量产车队监测以数据驱动的方式结合起来。随着行业转向基于 AI 的架构，汽车标准也正在日益明确和具体化安全要求。

本文介绍 NVIDIA Halos Safety Evaluation Framework (SEF，安全评估框架)——由一套工具生态及配套指南构成，用于生成足够的证据，以支持从 L2 ADAS 到 L4 无人驾驶出租车等不同自动化等级下的智能汽车安全案例。NVIDIA Halos SEF 建立在 NVIDIA Halos OS 内部 330 多篇研究论文和 1000 多项专利的基础之上。

我们的使命有两方面：

- 提出一套具体、可执行的方法，明确在 L2 到 L4 各自动化等级下开展符合标准的安全评估所需的关键组成部分。
- 向市场提供其中部分精选组件，例如 NVIDIA NuRec 和 Cosmos Transfer。许多生态合作伙伴同样深耕于一些组件，我们鼓励通过伙伴关系与协作，将 Halos SEF 落地为完整解决方案。

安全评估并不存在单一的"万能"方案，而是由数据、软件和流程共同作用，最终形成统计学证据。本文将对 NVIDIA Halos SEF 进行高层次概览。

我们的架构

智能汽车安全评估背后的工程挑战，源于需要被证明的安全目标极其严格。大多数智能汽车安全性能目标以人类驾驶基准为基础，要求系统导致死亡事故的频率不高于大约每 10^7 至 10^8 小时驾驶一次[1]。若仅依靠车内实车测试来覆盖如此巨大的驾驶时数，显然不可行。因此，仿真技术成为确立这些性能边界的关键手段，但它也立即引出了一个问题：仿真的真实度究竟如何？

除了数据量本身，我们还面临大规模数据采集、标注和策展的运营挑战。我们必须部署数据中心来存储 PB 级驾驶日志，并在其上执行仿真。我们还必须从数千项性能指标中层层筛查，识别安全问题、真值错误和仿真真实度缺口，并排定处理优先级。同时还需要解决如何有效防止流水线过拟合，以及如何处理多个串联自动化工具带来的误差累积。

Halos SEF 提供具体且可执行的指南，以及用于解决这些问题的业界领先工具。它支持测量关键性能指标 (KPI)，并据此生成支持汽车安全案例的证据。它包含三个主要组成部分：

- 流程管控 **(Process Governance)**：安全生命周期管理流程，涵盖需求管理，以及基于 KPI 结果的数据驱动风险评估。
- 数据基础 **(Data Foundation)**：用于创建、标注、策展和增强数据集的一组工具。
- 仿真核心 **(Simulation Core)**：由多种仿真器、自动分析软件和结果可视化平台组成，用于生成安全证据。

整个安全框架采用循环和迭代式设计：既往发现和版本发布会催生新的数据、标注、仿真与分析需求。Halos SEF 就像一个飞轮，持续生成支持汽车安全案例的证据。下面我们将沿着 Halos SEF 的流程逐一介绍各个组成部分。

[1] 人类驾驶的死亡事故率因地区显著不同，大约在每 10^6 至 10^7 小时驾驶发生一次之间。智能汽车应达到何种具体目标仍存在争议，但通常会在人类死亡率基础上再施加约 10 倍的改进系数，以涵盖酒驾、疲劳驾驶等驾驶能力受损情形及其他不确定性。本文中的目标值主要用于说明：若要证明智能汽车能够安全部署，所需评估基础设施必须具备多大的规模。

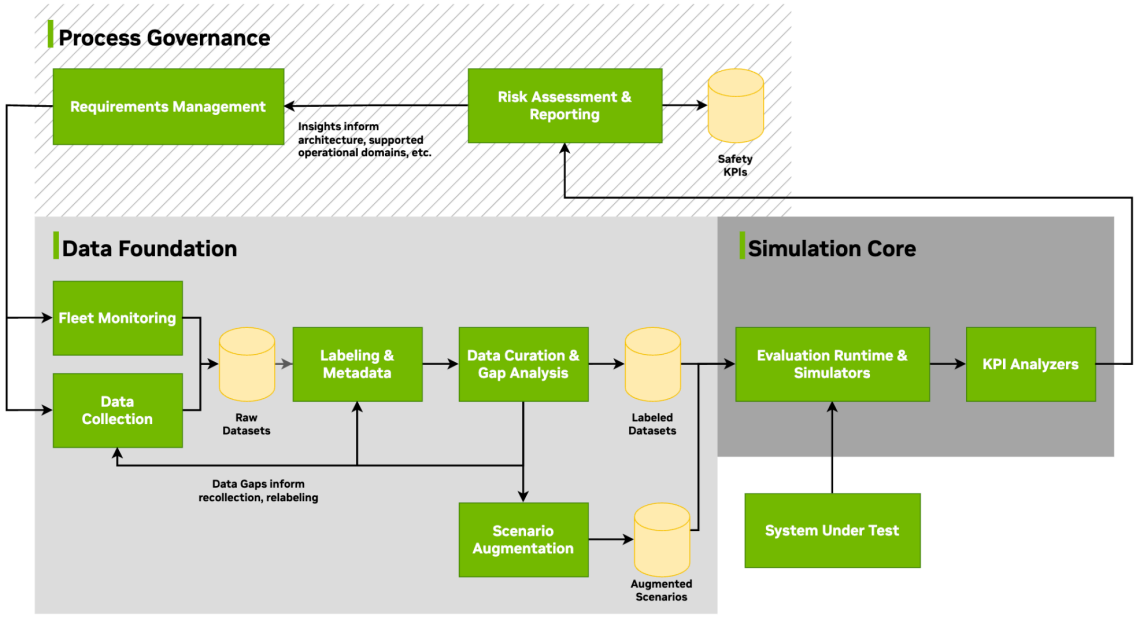

图 *1*：*Halos Safety Evaluation Framework* 的主要组成部分。

流程管控：需求管理

Halos SEF 的核心目标，是提供统计学保证，以确认系统符合其安全目标。简而言之，它遵循这样一个原则：安全要求不仅必须规定系统所需具备的属性，还必须定义验证该属性时所需达到的客观置信水平。因此，对这些安全概念和要求进行定义与管理，是 Halos SEF 的起点。

图 *2*：*Halos SEF* 中的流程治理，重点展示安全需求管理流程。

安全要求通过基于第一性原理、整体性的安全案例驱动方法推导得出。我们以应用的 ODD (Operational Design Domain，运行设计域) 为分析锚点，识别系统故障以及由环境触发因素、已知性能局限和运行假设所产生的风险。随后，将这些分析转化为安全概念，在系统、软件、硬件、AI/数据生命周期以及运营控制层面分配安全要求与验收准则。这些分配后的要求将定义技术措施、接口、监控机制、冗余、容错行为，以及将风险维持在可接受范围内所需的 V&V (Verification & Validation，验证与确认) 证据。关于如何推导可量化安全 KPI，以及如何在 AI 和机器学习生命周期中管理触发条件，可参见 Effective and reflective assurance for AI-based autonomy 一文。

所有内容最终汇总为结构化的安全保证论证，也就是所谓的安全案例，用于证明残余风险已经降至可接受水平。它是一份持续更新的文档，不仅在产品首次发布时提供安全保证，还贯穿整个产品生命周期。整个安全案例框架建立在行业标准的基础要求之上。ISO 26262、ISO 21448/SOTIF、ISO PAS 8800 和 UL 4600 等标准，为安全案例中的保证论证提供了结构化的组织框架和论证依据。TÜV Rheinland 已成功完成对 NVIDIA NDAS SAE Level 4 自动驾驶系统安全管理规划的独立安全评估。该系统基于 NVIDIA DRIVE AGX Hyperion 10 平台构建，其功能安全、SOTIF 和 AI 安全管理规划，以及 V&V 策略，均按照 ISO 26262、ISO 21448 (SOTIF) 和 ISO/PAS 8800 标准接受评估。在整个评估过程中，NVIDIA 展示了开展下一代自动驾驶出行所需的安全活动与安全驱动流程的相关能力。

下方图 3 展示了这一背景下具有代表性的安全案例论证结构。

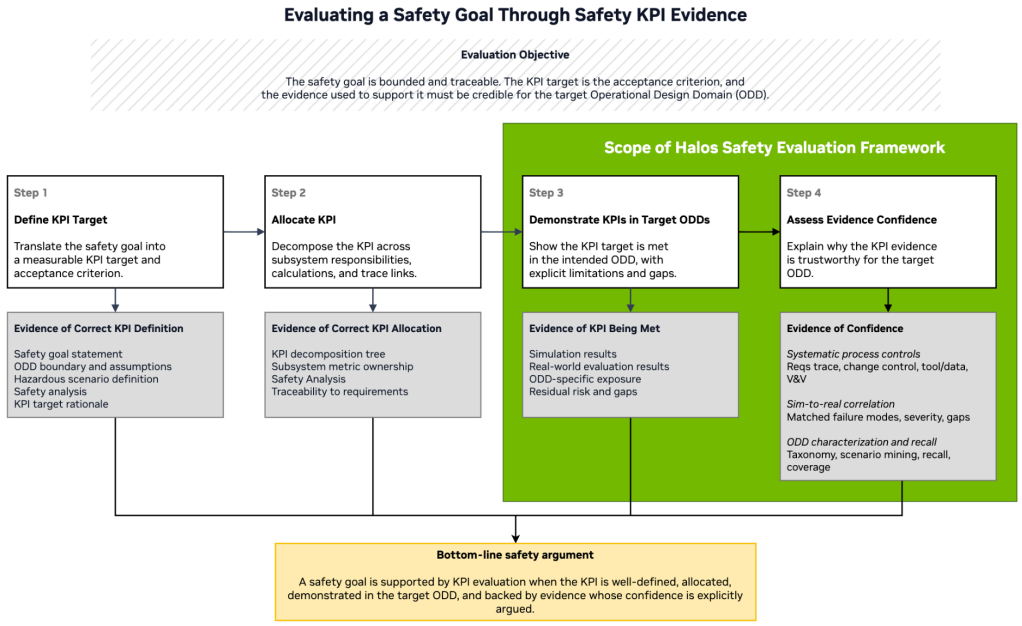

图 *3*：具有代表性的安全案例论证结构。

每一项安全要求，都可以被视作沿着安全评估框架中的某一条特定路径展开，只经过证明该要求所必需的元素。聚焦不同自动驾驶级别的生态伙伴，可自主选择与其业务战略契合度最高的演进路径。下面在介绍 Halos SEF 各组成部分时，我们将使用这些安全要求作为示例，说明不同组件如何串联成完整的安全论述。

示例安全要求	自动驾驶级别	示例安全要求目标
车辆的 *AEB* 系统应在 *CCFTap NCAP* 场景中触发	L2	通过所有 CCFTap NCAP 场景变体，并进行额外模糊测试 (fuzzing)
车辆应避免高速行驶时的 *AEB* 幽灵制动 *(Ghost Brakes)*	L2	每 10,000 小时最多 1 次失效
车辆在无保护交叉口通行过程中，应避免可能导致严重伤害的碰撞	L2++	每 20,000 小时 L2++ 运行中，最多 1 次非通常可控 (non generally controllable) 的交通冲突
在变道相关场景中，车辆不应出现摄像头障碍物感知错误	L4	每 10,000 小时最多 1 次失效

表 *1*：若干示例性安全要求。

数据基础

Halos SEF 的数据基础负责构建安全评估所需的数据集，覆盖数据采集、标注、策展以及增强等环节。

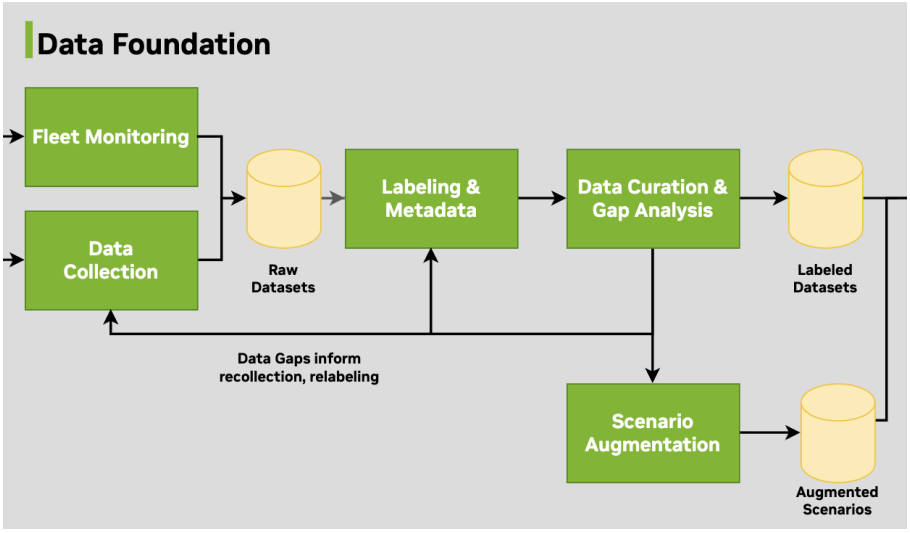

图 *4*：*Halos Safety Evaluation Framework* 的数据基础。

数据采集

指导数据采集的首要因素是目标数据集的规模及其分布。若要以 95% 的置信度证明一个“20,000 小时”的 KPI 目标，大致需要约三倍规模、且期间无错误的驾驶数据。此外，为使基于这些数据的统计论证成立，样本必须按照预期的客户使用分布进行采样。例如，一个过度代表高速公路里程的数据集，不能用于证明城市交叉口中的性能。因此，我们的车队需要专用工具来规划采集作业，并分析采集产生的数据，以确保在数据进入下游流水线前，已经满足数据量和分布目标。

标注与元数据

为了开展数据分布分析，同时支持许多下游 SEF 组件，摄取的数据必须得到恰当标注。不同 KPI 对标注的要求各不相同，从障碍物长方体框、车道折线到天气元数据不等。NVIDIA 在 Hyperion 传感器平台之上构建了多种自动化及人工监督的数据标注工具，使我们能够大规模生成相关标签。

需要特别指出的是，对于任何此类工具，标签本身的精确率 (precision) 与召回率 (recall) 都必须被量化刻画，并在下游安全论证中加以考虑。例如，车道图自动标注中的系统性问题，可能沿整个安全论证链条传播，并最终造成结果偏差。

作为 NVIDIA 已构建的数据采集和标注工具的一个示例，我们在今年早些时候开源了包含 1,700 小时已标注数据的 NVIDIA 物理 AI 数据集 (Physical AI Autonomous Vehicles Dataset)。

策展与缺口分析

60,000 小时的已标注驾驶日志，其存储与仿真的数据量相当可观。不过，我们的示例 L2++ 安全要求关注的是交叉口表现。因此，可以从这 60,000 小时数据中策展出一个显著更高效的测试子集，重点覆盖与当前安全要求相关的场景。将数据集针对安全要求进行定制，往往可以优化数据结构，并将仿真成本降低若干数量级。

在这一阶段，我们还可能发现数据集存在系统性缺口，需要重新采集。NVIDIA 将传统统计方法与前沿 AI 工具相结合用于该任务。以 NVIDIA Cosmos Dataset Search 为代表的向量嵌入搜索技术，为语义化策展提供了独特机会，使工程师不仅可以根据元数据寻找匹配日志，还可以寻找底层驾驶情景相匹配的日志。此阶段发现的缺口会反馈到数据采集计划中，从而在安全要求所需内容与车队实际采集内容之间形成闭环。

数据增强

构建一个代表 60,000 小时交叉口通行的数据集，虽足以证明统计性能边界，但有些场景尽管属于合理可预见的情形，却极少能被真实采样到，属于辅助驾驶领域著名的“长尾”场景。

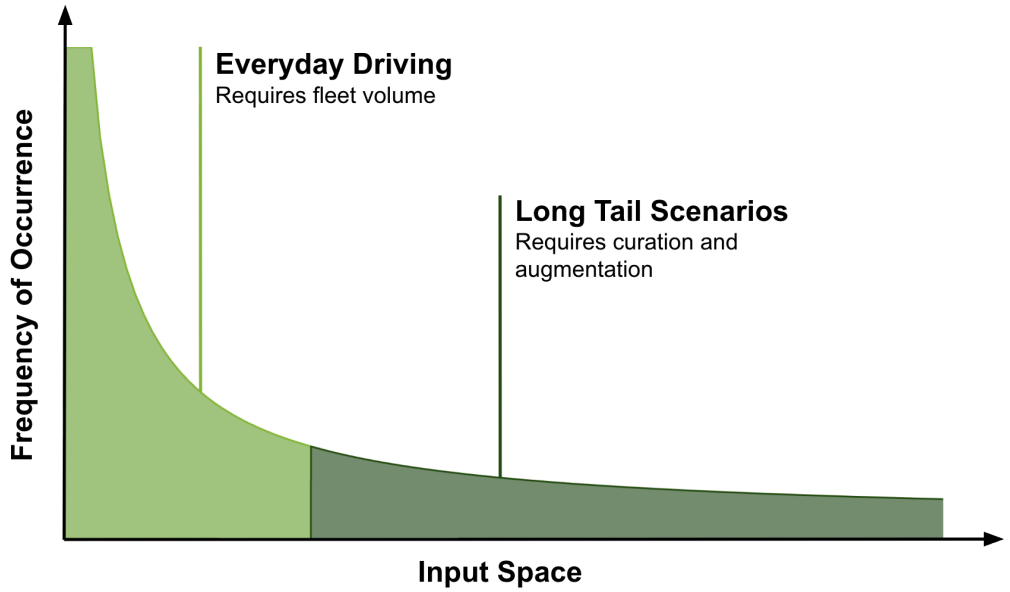

图 5：*Halos SEF* 必须同时覆盖日常驾驶和长尾场景；仅依靠暴力式数据采集，往往难以发现这些罕见但危险的场景。

长尾场景的规范可以从文献和国家级碰撞事故数据库中挖掘，但这类场景本身就极为罕见，而且实地采集有时具有危险性。因此，数据增强技术是实现长尾场景覆盖的理想方案。

增强可以采用两种形式：合成场景生成，以及场景扰动。两种方法都有效，但 NVIDIA 更关注后者。在场景扰动过程中，我们修改已有日志，以生成原始遭遇的变体，例如调整自车接近速度、调整横穿方向的交通参与者出现的时机，或在关键时刻加入一辆全新的遮挡车辆。尤其是 NVIDIA NuRec，它允许我们利用神经重建技术，以高真实度渲染扰动后的场景。Cosmos Transfer 等技术还可以通过建模雨、雪或野火烟雾等环境效应，扩展测试的运行域。

图 *6*：*Cosmos Transfer* 用于覆盖多样化的天气和光照条件的数据。

真实驾驶日志让我们的统计结论锚定在真实客户分布之上，而增强数据则确保罕见但可预见的场景具备足够的样本密度。与自动标注类似，增强工具本身的真实度特性，也必须在下游安全论证中予以考虑。增强数据生成后，会与真实日志一道进入流水线，形成可供仿真核心使用的数据集。

车辆监测

在新功能推向市场之前，必须证明其满足安全要求。在最初的产品导入阶段，此类评估通常由专用且受到严密监控的开发车辆完成。一旦功能发布到市场，客户日志就会成为极具价值的数据来源[2]。客户车辆数据集可用于计算发布后的 KPI，并回答如下问题：我们在开发阶段对客户运行分布作出的假设是否准确？测得的失效率是否与实际部署车队经历的情况一致？是否存在某些场景需要针对性地重新采集或增强？

客户数据集覆盖的驾驶里程，可能比开发车队高出若干数量级。在这一规模下，我们必须依赖 Halos SEF 其他环节同样使用的数据基础工具，对进入系统的日志进行摄取、存储、标注和策展，把海量数据流筛选为与各项安全主张相关的子集。车队监测使安全案例不再只是一次性的产品导入活动，而成为已部署系统需要持续维护的属性。

[2] 当然，任何时候采集和存储客户数据，隐私都是关键问题。此类流水线从一开始就必须围绕用户同意、匿名化以及不同地区的监管要求进行架构设计。

仿真核心

现在我们已经构建了用于评估系统的数据集，下一步就是对系统进行仿真，并生成我们所需的安全性能证据。

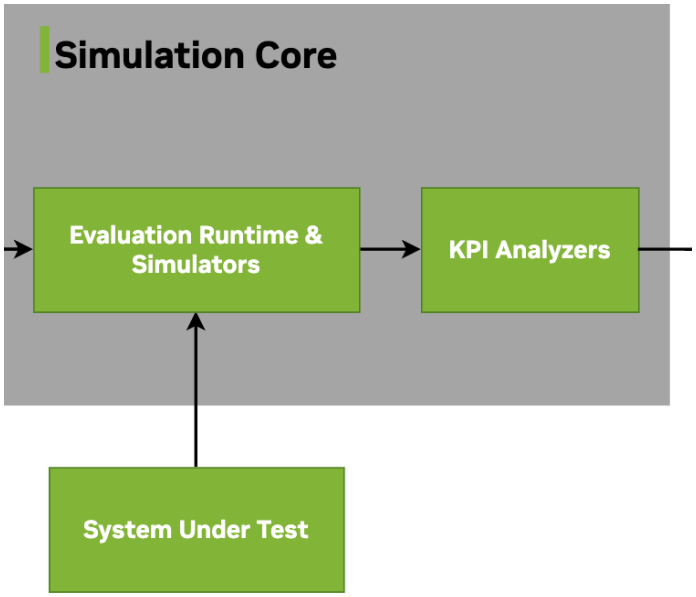

图 7：*Halos Safety Evaluation Framework* 的仿真核心。

评估运行态与仿真器

不同的安全主张，对所需仿真的规模、真实度和成本提出不同要求。Halos SEF 并不规定必须使用某一种仿真器，而是提供多种仿真技术，以及针对当前安全主张选择适当工具的指南。以贯穿全文的 L2++ 交叉口示例为例，为评估无保护转弯中的碰撞规避能力，需要采用具有高传感器真实度的闭环仿真：系统的执行器输出必须能够影响周围交通参与者的轨迹，同时渲染出的传感器流还必须足够逼真，使感知堆栈的行为与真实世界中保持一致。与之相比，对于 AEB 误触发 (false-positive) 论证，使用已记录传感器数据进行开环回放不仅已经足够，而且成本低几个数量级——只要检测到一次幽灵制动触发，就构成相关证据，无需进一步建模后续车辆动力学。当我们进一步考虑 HiL (Hardware-in-the-Loop，硬件在环) 和 PiL (Processor-in-the-Loop，处理器在环) 配置时，仿真技术的选择空间会更加复杂，因为要正确建模系统行为，可能需要将量产车辆中的实际硬件组件纳入测试。

在 NVIDIA，我们最先进的仿真器采用 NuRec——一种神经重建与渲染技术，可生成照片级真实且传感器层面准确的真实驾驶场景重建。NuRec 允许我们读取已有驾驶日志、重建三维场景，并通过神经渲染的新视角合成，在闭环条件下重新进行仿真。如前面“数据增强”一节所述，NuRec 允许我们扰动自车行为，并向场景中注入新的交通参与者，同时保持 AI 感知堆栈所要求的传感器真实度。正是这一能力，使得类似 L2++ 交叉口目标这

样的安全主张具备可操作性：我们可以从真实交叉口遭遇出发，在该场景中模拟新的自车轨迹，并以纯合成仿真很难实现的规模测量碰撞结果。

无论最终选择哪一种仿真器，其自身的真实度都必须先经过刻画，才能把仿真输出作为可信的安全证据。Halos SEF 中常用的一种方法是 log-to-sim 一致性检查：选取一条真实道路驾驶日志，保持所有变量不变——相同车辆配置、相同软件版本、相同传感器输入——然后在仿真中重新运行。道路真实结果与仿真器报告结果之间的任何差异，都可以用来衡量仿真的真实度缺口。随着仿真器和车辆软件不断演进，仿真真实度应在整个项目生命周期内持续测量。不同安全主张对应的真实度缺口将分别跟踪，并纳入下游安全论证。

KPI 分析器

仿真完成后，我们需要分析产生的日志，以理解与安全要求相关的哪些行为实际发生过。Halos SEF 将仿真与一套自动化 KPI 分析器配合使用：这些分析器同时读取仿真器输出和真值数据，并验证系统是否遵守相关安全属性。对于 L2++ 交叉口目标，分析器会检查每一次模拟的交叉口通行，判断是否发生严重度二级及以上的碰撞事件。对于 AEB 幽灵制动论证，它会检测正常驾驶工况下的非必要触发事件。部分更先进的分析器内部还编码了可形式化验证的安全属性，例如 NVIDIA 首创的基于 HJ-Reachability 的安全区域 (Safety Zones)，用于理解在特定驾驶任务中哪些障碍物实际上与安全相关。

Halos SEF 的一个关键架构特性，是位于仿真器、被测系统 (SUT) 和分析器之间的抽象层。这三个组件通过标准化接口相互通信，因此其中任一组件都可以被替换，而无需改动流水线的其他部分。例如，我们可以把开环回放仿真器换成基于 NuRec 的闭环运行，将被测系统从一个版本迁移到另一个版本，或将人工标注真值替换为自动标注真值，而下游分析器仍可保持不变。

值得注意的是，安全分析器通常会被调校为偏保守，即以牺牲精确率为代价来提高召回率。漏掉一次真实失效的代价远高于一次误报，因此，系统会对被标记事件加入人工复核，用于区分真实失效与无害触发。最终经过复核的结果，才会形成上报给流程治理环节的 KPI 数值。

流程治理：KPI 分析与报告

当支持安全案例的证据生成完成后，我们再次回到流程治理环节，在这里对结果进行分析并开展风险评估。

图 *8*：再次回到 *Halos Safety Evaluation Framework* 的流程治理环节。

一个原始故障本身并不能构成安全论证。必须结合置信区间、数据分布，以及产出这些结果的工具已知的精确率和召回率特性，对结果进行解读。同时还必须证明不存在 Risk Transfer (风险转移)[3]——即总体失效率虽然满足规范，但失效分布却以非预期方式集中在某些群体或场景中。

我们的结果可视化工具与底层数据科学平台专门为这种分析而构建，使安全工程师能够从汇总 KPI 一路下钻到构成该 KPI 的单条驾驶日志、标签以及仿真器输出。NVIDIA 首创的 Sim2Val 等统计工具也在这里发挥作用，使我们能够高效量化最终安全保证中的仿真真实度。

Halos SEF 生成的每一份安全证据，都被设计为能够通过 KPI 管理工具和结果可视化平台追溯到某一项具体、可审计的要求。这使整个安全框架首尾闭环：KPI 结果分析和风险评估最终重新连接回需求管理。

智能汽车安全飞轮

随着对评估结果的分析反馈进入需求管理流程，一个反馈闭环由此形成，使我们能够迭代，并基于数据驱动决策逐步弥合性能缺口。

观察结果	下一轮评估中的含义
覆盖雨天场景的数据不足	需要收集、挖掘或增强更多雨天数据。
仿真真实度不足	需要改进仿真器，或改用其他替代方案。
小路-主路交叉口中的安全 KPI 低于阈值	需要改进产品在这些区域的表现，或限制产品的运行范围，例如在这些情况下请求驾驶员接管。
当前 ODD 中所有 KPI 均达标	有机会扩展到新的地理区域。更新数据分布并评估就绪度。

表 *2*：*Halos SEF* 通过安全飞轮的连续迭代所支持的数据驱动决策示例。

随着产品生命周期推进，安全飞轮将不断转动，直到我们建立起坚实的证据基础，足以支持汽车安全案例。

[3] 仅仅让智能汽车的总体死亡率低于人类驾驶员还不够。我们还必须证明，系统不会让任何特定人群承受不成比例的更高风险。Philip Koopman 在《How Safe Is Safe Enough》第 5.2 章中对此有较充分的讨论，其中举例说明了这样一个系统：总体安全水平高于人类驾驶，但大多数事故都发生在穿着高可视背心的行人身上。这样的系统仍可能被认为不适合部署。

让 Halos Safety Evaluation Framework 具备可信性与可审计性

为了真正支持汽车安全案例，框架中的所有要素——从数据集、仿真技术到 KPI 分析器——都必须针对其所支持的具体安全主张具备可信性。UN-R ADS 第 7.2.1 条以及 UN-R171 DCAS 附件 5 为这种可信性论证提供了一个有用的结构。

对于研究人员而言，这种可信性论证直接对应“可复现评估方法”的概念：其核心原则与计算机视觉社区进行基准测试设计时相同，即让评估流水线透明、可审计并可复制。

在 Halos SEF 中，这种可信性论证围绕五类证据组织：

- 基础与范围 **(Foundation and scope)**：定义虚拟测试的预期用途、相关安全要求、测试目标、ODD 与系统边界、假设、局限、真实度期望、验证策略、验收准则，以及潜在工具链错误对安全论证的关键程度。
- 流程与治理 **(Process and governance)**：管理证据背后的数据、人员与版本发布。包括输入/输出数据谱系、数据质量与覆盖范围、从 KPI 结果回溯到场景、参数以及工具链配置与版本的可追溯性、人员能力、第三方输入以及生命周期支持。
- 技术验证（**Technical verification**）：证明工具链实现正确，并且对于有效输入能够稳健运行。证据包括代码验证、参数空间探索、合理性与一致性检查，以及有界的数值误差，例如离散化、舍入和收敛效应。
- 确认与不确定性分析 **(Validation and uncertainty analysis)**：证明针对预期用途，仿真输出与相关物理或真实世界参考数据具有相关性。包括确认指标、拟合优度准则、校准、敏感性分析、不确定性量化，以及在不确定性影响结果解释时采用的安全裕度。
- 最终适用性评估 **(Final suitability assessment)**：综合所有证据，并明确工具链和 SEF 要素是否适用于所定义的安全主张、范围、版本、假设和局限。

Halos SEF 旨在通过总体 SEF Credibility Manual以及针对各个 SEF 元素的配套手册，把上述证据结构明确化。这些手册将可信性预期映射到依据既有标准与实践生成的证据，包括 ISO 26262、ISO 21448、ISO/PAS 8800、ISO 34502、ISO 34503 和 ISO/IEC 17025。

结论与展望

向 AI 架构转型，正在改变“如何证明智能汽车安全”这一问题本身的含义。传统功能安全方法建立在确定性软件和演绎式失效分析之上，这些方法仍然必要，但已经不再充分。AI 驱动系统要求对安全采用数据驱动、统计学的方法；而这种方法又进一步要求评估框架具备过去行业从未需要面对的证据规模和真实度。

Halos SEF 是 NVIDIA 对这一挑战的回答。通过把数据基础、仿真核心和流程治理组合成一套统一、连贯的方法，我们提供了一条从安全要求出发，走向可复现、符合标准的安全案例的具体且可执行路径。我们邀请合作伙伴和客户与我们共同建设这一未来，让 AI 驱动的自动化真正兑现其价值，同时满足公众道路交通安全所要求的统计严谨性。

对于研究人员而言，Halos SEF 还代表着一种超越智能汽车的更普遍原则：当物理 AI 系统——从机器人、手术辅助设备到智能基础设施——从研究实验室走向真实世界时，同样严谨、可复现的评估方法将成为必需。为智能汽车安全评估开发的神经渲染、生成式仿真和统计分析工具，正是 NVIDIA 与社区共同推进整个物理 AI 研究议程的基础构件。