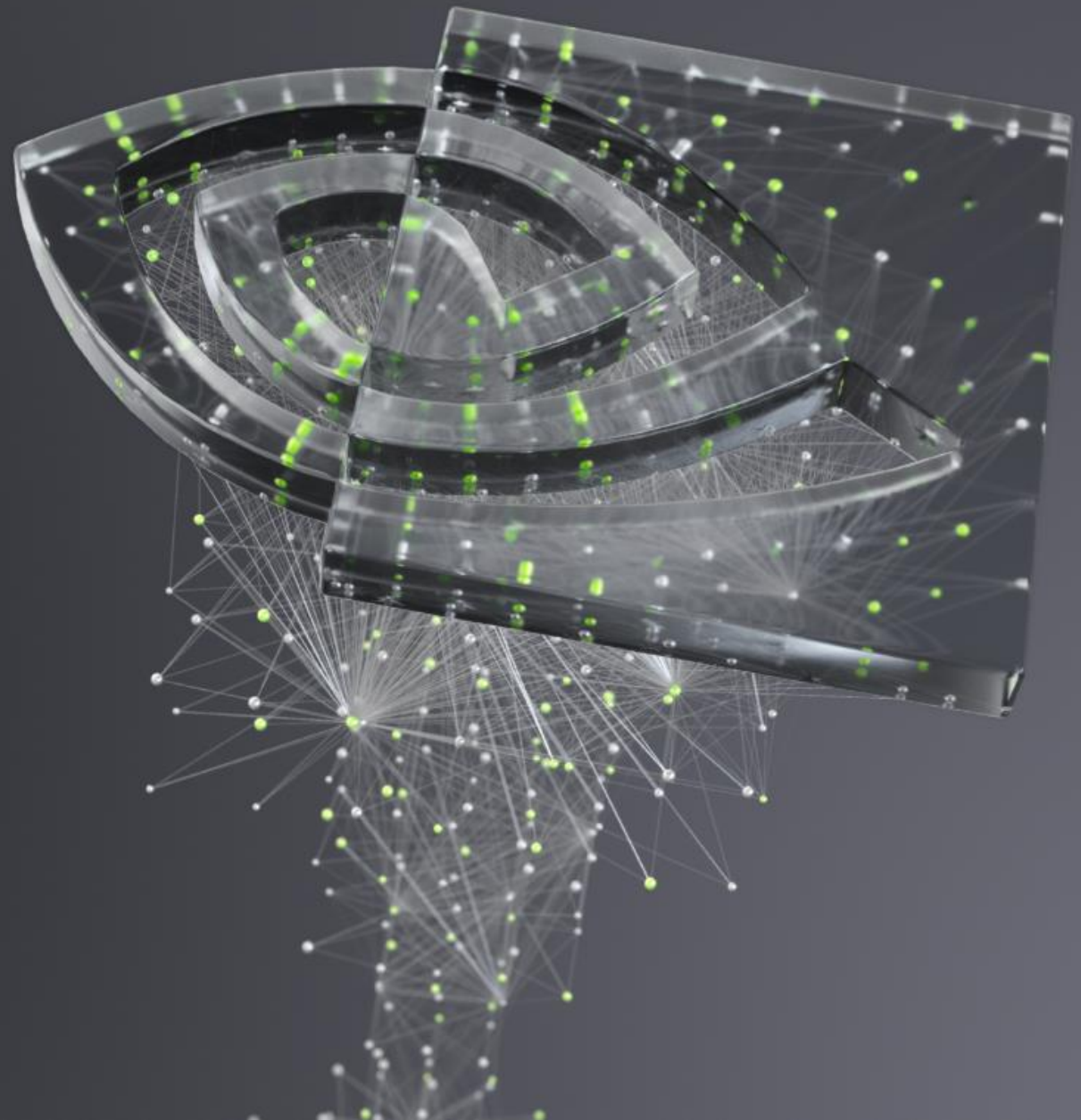


加速 AI 迈入新纪元

NVIDIA 助力的推理案例

AI 部署达到全新速度

推理与 AI 并存。推理可以帮助在线助手作出即时而相关的响应，帮助医生更快了解疾病并作出更准确的诊断，还可优化我们的购物体验，助力开展强大的新产品设计。推理还有助提升农作物的健康水平，保护野生动植物，甚至为科学家提供外太空探索的全新视野。

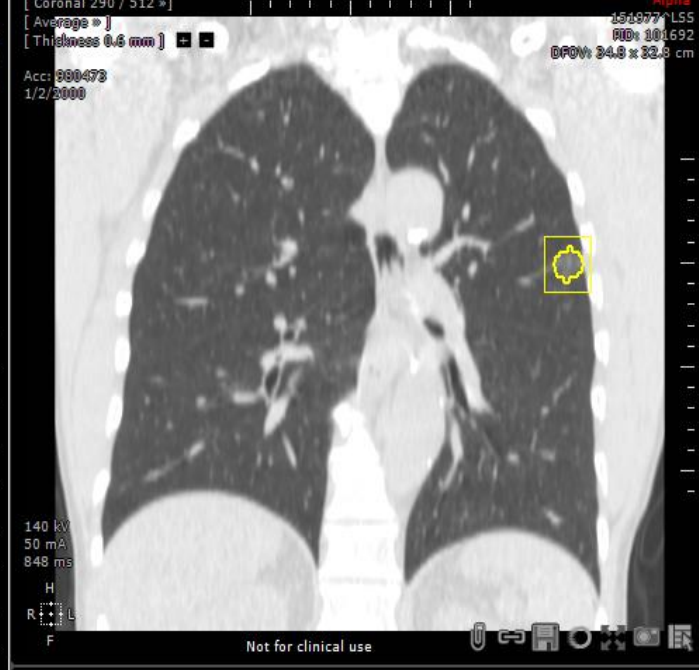
放眼各行各业，推理正在转变、加速和改进我们的工作内容和工作方式，最终惠及我们的生活。

NVIDIA 技术正在让这一切成为可能。从数据中心到边缘节点乃至物联网 (IoT) 设备，NVIDIA GPU 加速解决方案为全球跨学科用例提供了领先的推理能力。

下面将介绍一些相关案例。

利用 AI 及早检测出肺癌

在 NVIDIA Clara 平台和 NVIDIA GPU 技术的助力下，12 Sigma Technologies（图玛深维）公司的 σ-Discover/Lung 系统能够自动检测图像中小至 0.01% 的肺结节，对恶性肿瘤的分析准确率超过 90%，并可为放射科医生提供决策支持工具。在利用 NVIDIA T4 集群进行优化后，该系统的运行速度提升 18 倍。



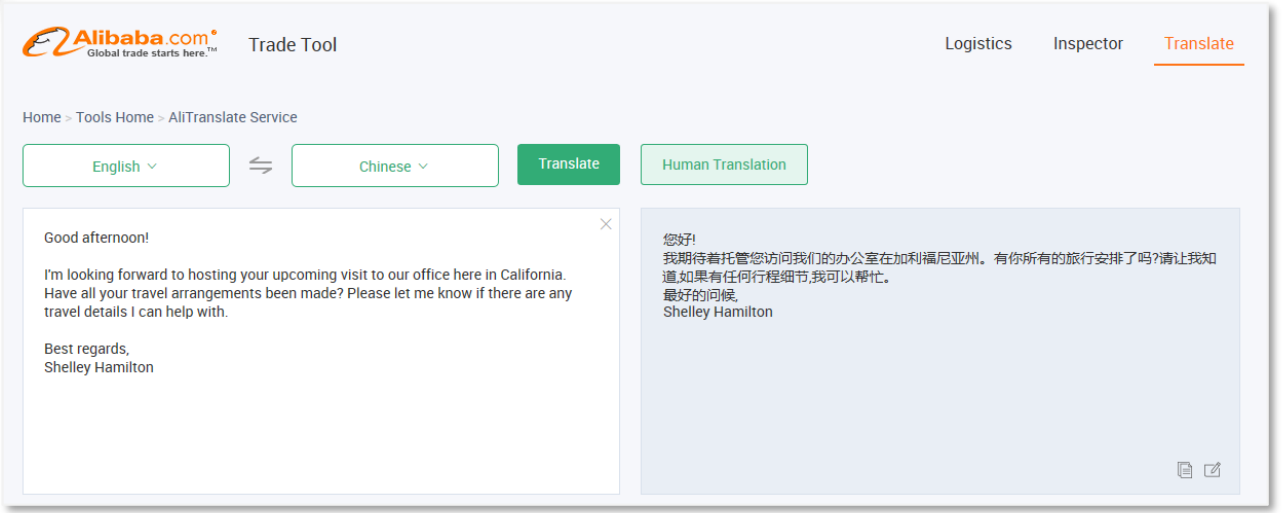
打破
商业壁垒

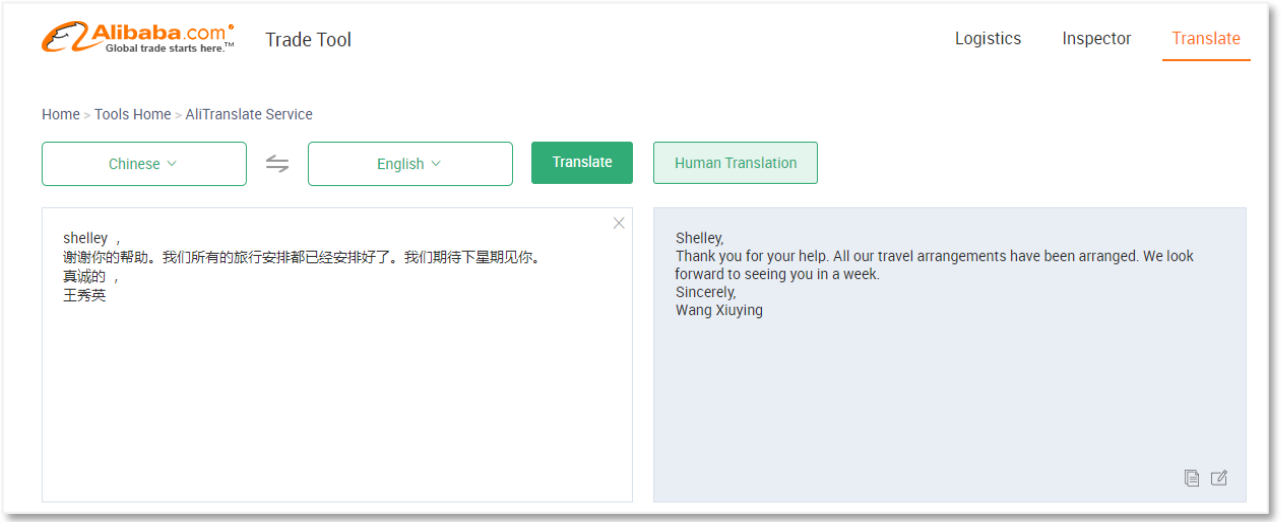
阿里巴巴集团每天要处理 83 亿个翻译请求，来支持国际商业贸易。

阿里巴巴使用神经网络机器翻译 (NMT) 显著改善了翻译质量，但增加了延迟和计算成本。

为补其不足并加速 NMT 在线服务，阿里巴巴部署了 NVIDIA Tesla GPU，处理的请求数量增加 3 倍的同时，响应速度缩短到原来的 1/3。

NVIDIA Alibaba Group





扩大服务规模，降低总体拥有成本 (TCO)

语音翻译帮助游客、企业、学生等群体克服了语言障碍。科大讯飞希望扩展其普通话的语音服务，支持多种口音和方言。

鉴于此，该公司将其推理运算迁移到 Tesla GPU 和 TensorRT，扩大了对 GPU 的采用。

科大讯飞现可处理的并发请求数量已增加为原来的 10 倍，准确率提高了 20%，而且 TCO 运营成本也已降低 20%。

NVIDIA 科大讯飞 iFLYTEK

http://www.nvidia.cn/deep-learning-ai/startups

提供
合家欢内容

网络视频流量不断增加，这就要求运营企业加大监控力度，以过滤不当内容。

京东在 Tesla P40 GPU 上使用 NVIDIA DeepStream SDK 和 TensorRT 来识别和过滤全高清直播视频的 1000 个频道。

该公司在使用推理来过滤视频内容时，吞吐量已增至 20 倍，而搭载 Tesla 的每个服务器则可同时处理 20 个视频。

NVIDIA

JD.COM

通过 AI 实现签章和配送

京东 X 事业部利用由 NVIDIA Jetson 超级计算机提供支持的智能机器，将 AI 引入物流和配送领域。

JDrone 在为偏远地区配送新鲜食品和药品时可将物流成本降低 70%。JDrover 机器人可在行人和交通网中轻松导航，为选定地点递送包裹。

京东 X 事业部利用分拣机器人设立了全球首个自主分拣中心，每小时可分拣多达 16000 个包裹，准确率高达 99.999%。

NVIDIA

JD.COM

自主配送货物的货运卡车

长途运输为物流公司带来了挑战：美国法规规定，卡车司机每天累计驾驶时间需少于 11 小时，因此司机缺口正变得越来越大。

自动驾驶技术可帮助物流公司提高效率，缓解司机数量日益紧张的压力，并能更快配送更多货物。

自动驾驶卡车公司 TuSimple（图森）开展了一项试验：通过 USPS 在凤凰城和达拉斯之间的逾 1000 英里路程中运送邮件。

TuSimple 的卡车利用 NVIDIA DRIVE 技术实时识别物体（可见距离长达 1000 米），直接在卡车上处理图像，以提升夜视功能，减少车头灯炫光。

NVIDIA. tusimple UNITED STATES POSTAL SERVICE

大规模提供实时语音服务

作为一家用户规模约达 10 亿的中国社交媒体领先平台，微信希望提升其语音转文本服务。

但在部署新的声学模型时，其 CPU-only 服务器却无法有效运行新版本。

于是微信部署了搭载 Tesla P4 GPU 推理加速器的服务器，由此将语音推理吞吐量提高 2.5 倍，模型内准确率提升 20%，同时仍将延迟估算保持在较低水平。

为直播内容保驾护航

YY 为 1 亿名并发直播参与者提供引人入胜的有趣体验。在直播期间审核内容以检测并过滤不当内容时，需进行实时推理。

通过在 GPU 上部署 NVIDIA TensorRT，YY 将推理工作负载吞吐量提高了 30%，并将内存需求降低了 40%。

借助实时推理，YY 成功将不当内容阻挡在直播画面之外。

NVIDIA

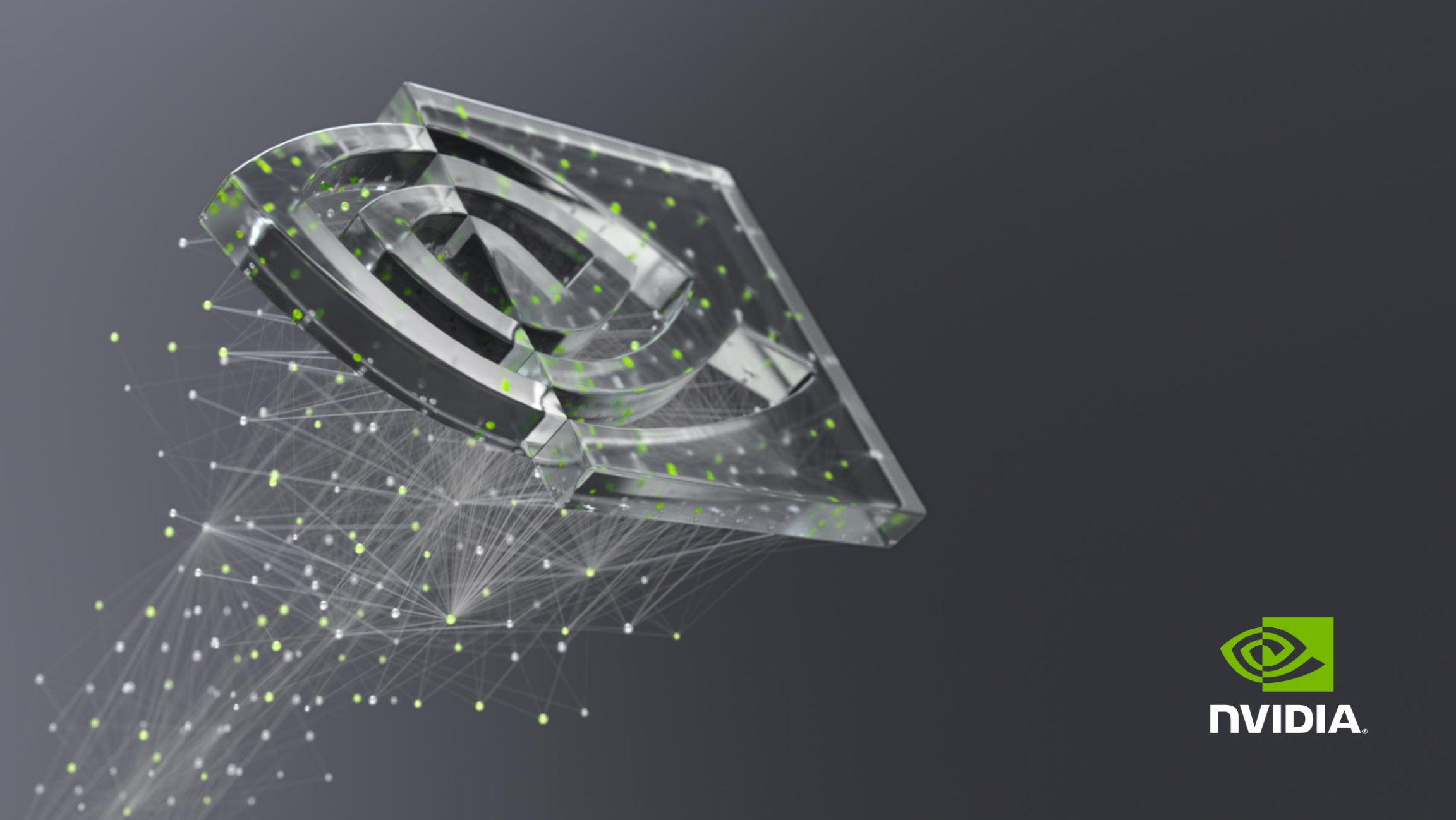
YY.COM 就是爱YY

加速推理工作
实现全新功能

从数据中心到边缘节点以及各个 AI 用例，NVIDIA 技术正在塑造 AI 部署的现在和未来。

NVIDIA 技术不仅能为新服务提供助力，而且还能推动创新、创造新品、增加收入、简化运营，进而改变我们的生活。

详情请访问 https://www.nvidia.cn/deep-learning-ai/solutions/inference-platform/


NVIDIA