/TemplateVersion (2026.1)

STRIDE-QA: 用于城市驾驶场景时空推理的视觉问答数据集

Keishi Ishihara*[1] Kento Sasaki*[1, 2] Tsubasa Takahashi[1] Daiki Shiono[1, 3] Yu Yamaguchi[1]

[1]Turing Inc. [2]University of Tsukuba [3]Tohoku University
{ keishi.ishihara, kento.sasaki } @turing-motors.com

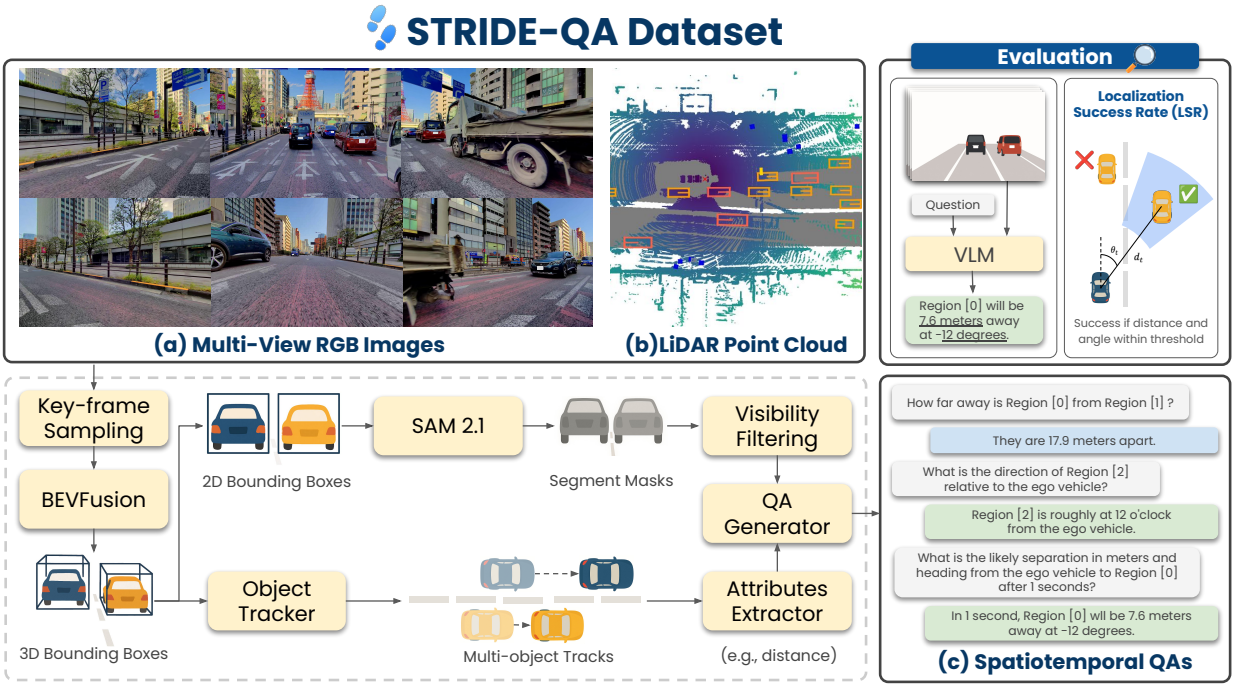

Figure 1: STRIDE-QA is a large-scale VQA dataset for spatiotemporal reasoning in autonomous driving, comprising 285 K frames and 16 M QA pairs from over 100 hours of urban driving in Tokyo. (a) It includes multi-view RGB images and (b) LiDAR point clouds, processed via a modular pipeline with 3D object detection, segmentation, tracking, and visibility filtering to produce spatially and temporally grounded annotations. (c) The annotations enable object-centric, ego-centric spatial, and spatiotemporal QA tasks, allowing structured evaluation of physically grounded reasoning over time.

Abstract

视觉-语言模型（VLMs）已被应用于自动驾驶，以支持复杂现实场景中的决策。然而，它们在静态、来源于网络的图像-文本对上的训练从根本上限制了理解和预测动态交通场景所需的精确时空推理。我们通过 STRIDE-QA 解决这一关键差距，这是一套用于从以自我为中心的视角进行物理推理的大规模视觉问答（VQA）数据集。该数据集由东京 100 小时的多传感器驾驶数据构建，捕捉了多样且具有挑战性的条件，STRIDE-QA 是城市驾驶时空推理最大规模的 VQA 数据集，在 285K 帧上提供 1600 万对问答。数据集由密集的自动生成注释，包括 3D 边界框、分割掩码和多物体轨迹奠基，独特地支持通过三项新颖问答任务实现的物体中心和自我中心的推理，这些任务要求进行空间定位和时间预测。我们的基准测试表明，现有的 VLMs 显著挣扎，在预测一致性上几乎得分为零。相反，经过 STRIDE-QA 微调的 VLMs 表现出显著的性能提升，在空间定位上达到 55 % 成功率，在未来运动预测的一致性上达到 28 %，与通用 VLMs 接近零的得分形成对比。因此，STRIDE-QA 为开发更可靠的用于安全关键自动驾驶系统的 VLMs 建立了一个全面的基础。

介绍

近期在视觉-语言模型（VLMs）方面的进展在跨模态任务中取得了显著的进步，如图像描述和视觉问答。这些

* Equal contribution.

在大规模图像-文本数据集上训练的模型表现出强大的语义理解和泛化能力。受这一成功的启发，VLMs 已被应用于物理人工智能领域，比如机器人学和自动驾驶。

这些方法强调了 VLMs 在实现整体场景理解和高层次驾驶决策方面的潜力。然而，这种潜力从根本上受到其训练数据性质的限制：大多数 VLMs 是在静态的、来源于网络的图像-文本对上训练的，因此缺乏真实世界应用所必需的空间推理能力（Fu et al. 2025）。这一限制在自动驾驶中尤为关键，因为缺乏适当的训练数据仍然是一个显著的挑战。

为了解决这个局限性，我们引入了 STRIDE-QA（Ego 视觉问答中的驾驶场景的时空推理），这是一个大规模 VQA 数据集，旨在真实世界驾驶场景中进行细粒度的空间和时空推理。数据集的概览及其自动化注释管道如图 1 所示。该数据集由在东京收集的超过 100 小时的多传感器驾驶数据构建，捕捉了包括交通拥堵、施工区和行人密集的交叉路口在内的多样且具有挑战性的场景。它包含了超过 16M 的问答对，这些问答对是通过一个结合 3D 物体检测、多物体追踪和实例分割的完全模块化和可扩展的注释管道生成的。

STRIDE-QA 专门构建用于支持对三个核心推理任务的监督训练和评估：

- 基于物体的空间问答：评估非自我代理（车辆、行人等）之间的空间关系。
- 自我中心空间问答：描述相对于自我车辆的主体距离、方向和大小。
- 自我中心时空问答：预测这些代理关系如何随时间演变。

这些任务旨在系统性地衡量空间和预测推理能力，这对于在安全关键的城市环境中的下游规划和决策制定至关重要。通过将每对问答对在物理和时间上保持一致的注释中结合起来，STRIDE-QA 为在现实世界自动驾驶中训练和基准测试 VLMs 提供了全面的基础。

贡献本文的主要贡献总结如下：

- 我们定义了三个新颖的以自我为中心的视觉问答任务，这些任务共同需要空间定位和短期预测推理，以应对自动驾驶系统在复杂交通场景中面临的核心挑战。
- 我们介绍了 STRIDE-QA，这是一项大规模数据集，包含 1600 万对问答对，密集标注于 28.5 万帧城市驾驶视频帧上，使视觉语言模型能够在真实世界交通动态的基础上进行细粒度的空间和短期时间推理的监督训练。
- 我们证明了现有的通用视觉语言模型在时空推理方面表现不佳，而在 STRIDE-QA 上进行微调的模型则显著超越了基线。我们的最佳模型 STRIDE-Qwen2.5-VL-7B 达到了最先进的性能，突显了我们数据集在驾驶场景中的时空理解效果的有效性。

这些贡献代表了将视觉语言模型整合到现实世界自主系统中的关键一步。通过从通用的视觉语言理解转向物理扎根的自我中心推理，STRIDE-QA 弥合了大规模多模态预训练与物理人工智能需求之间的差距。它不仅为时空视觉问答提供了一个基准，还推进了动态场景中的空间理解，促进了用于自动驾驶的更值得信赖的视觉语言模型。

相关工作

VLM 通过在 2D 视觉语言数据集（如 VQA v2（Goyal et al. 2017）和 GQA（Ainslie et al. 2023））上进行训练，已经在多模态任务中展示了强大的性能。然而，这些数据集缺乏 3D 空间信息，这限制了它们在现实世界环境中基于物理的推理能力的应用。

为了解决这一缺口，空间感知 VLMs，如 SpatialVLM（Chen et al. 2024a）和 Spatial-RGPT（Cheng et al. 2024），结合了几何注释，可以在 3D 中进行度量推理。然而，这些模型仍然局限于静态的单帧输入，无法捕捉到对例如自动驾驶等应用至关重要的时间动态。

作为回应，引入了几个针对驾驶场景的视频问答（VQA）数据集。例如，nuScenes-QA（Park et al. 2025）采用多视角视频来构建空间查询。ToD3Cap（Jin et al. 2024）和 NuPrompt（Wu et al. 2025）侧重于以自我为中心的空间定位，而 Refer-KITTI（Wu et al. 2023b）和 TUMTraffic-VideoQA（Zhou et al. 2025）则探索基于视频的问答和指代表达。

虽然这些数据集在将空间和时间推理融入视觉语言模型方面代表了重要进展，但仍有几个挑战尚未解决。值得注意的是，大多数现有的基准测试旨在支持以物体为中心或以自我为中心的推理，而不是两者兼具。这一区分很重要，因为以物体为中心的推理对于捕捉代理之间的交互至关重要，而以自我为中心的推理则使得相对于自我车辆的空间理解成为可能。此外，许多数据集缺乏时间上对齐的 3D 注释，这是进行短期预测推理所必需的。可扩展性也仍然是一个挑战，因为现实世界的驾驶场景涉及高度的视觉多样性和复杂性。这些数据集的详细比较见表 1 。

为填补这一空白，我们引入了 STRIDE-QA。它包括超过 100 小时的同步 LiDAR 和多视图 RGB 数据，并定义了三类 QA 任务：以对象为中心的空间任务、以自我为中心的空间任务和以自我为中心的时空任务。这些任务共同支持在复杂交通条件下的细粒度预测性推理。

空间和时空问答定义

本节在介绍我们的数据集和基准测试之前，概述了自动驾驶中时空问答的关键要求。为了支持在城市交通场景中的细粒度推理，我们定义了三个针对不同场景理解方面的视觉问答类别：(1) 周围代理之间的关系，(2) 自我与代理之间的空间关系，以及 (3) 代理动态的短期预测。

面向对象的空间问答。 这一类别针对的是基于单个当前帧的两个周围代理之间的空间关系。它包括定性问题（例如，相对位置或是/否查询）和需要数值回答的定量问题（例如，“距离 1.53 米”或“5 度”），如图 2 (A) 所示。

自我中心空间问答。 从自我中心视角理解场景对自动驾驶至关重要。为支持这一点，我们设计了一类问题，专注于自车辆与单一周围代理之间的空间关系，基于单个当前帧。我们构建了涉及距离、相对方向和对象大小与自车辆相关的定性和定量问题，如图 2 (B) 所示。

以自我为中心的时空问答。 该类别通过结合短期时间预测扩展了以自我为中心的空间问答。该任务以以 2 Hz 采样的四个上下文帧作为输入，目标是在未来的时间范围内预测以下物理量 $t \in \{1, 2, 3\}$ s：

- 自车与目标代理之间的距离
- 从自车到目标代理的航向角

Dataset	Domain	Data Source	Modality	# Video	# QA	Viewpoint		Temporal Frames	Scope	
						Obj.	Ego		S	ST
Spatial-VQA (Chen et al. 2024a)	General	Web images	RGB	—	200 M	✓		1	✓	
Open Spatial Dataset (Cheng et al. 2024)	General	OpenImages	RGB	—	8.7 M	✓		1	✓	
Refer-KITTI (Wu et al. 2023a)	AD	KITTI	RGB + LiDAR	6 h	818		✓	1 – 400	✓	✓
ToD3Cap (Jin et al. 2024)	AD	nuScenes	RGB + LiDAR	5.5 h	468 K	✓	✓	1	✓	
nuScenes-QA (Qian et al. 2024)	AD	nuScenes	RGB + LiDAR	5.5 h	460 K	✓	✓	1	✓	
NuPrompt (Wu et al. 2025)	AD	nuScenes	RGB + LiDAR	5.5 h	87.3 K		✓	1 – 40	✓	✓
nuPlanQA (Park et al. 2025)	AD	nuPlan	RGB + LiDAR	119 h	1 M	✓	✓	3	✓	✓
TUMTraffic-VideoQA (Zhou et al. 2025)	AD	Self-collected	RGB	$\leq$ 33.3 h	87.3 K	✓	✓	1 – 101	✓	✓
STRIDE-QA (Ours)	AD	Self-collected	RGB + LiDAR	100 h	16M	✓	✓	4	✓	✓

Table 1: 从推理范围、模态、数据来源和时间结构方面比较 STRIDE-QA 和现有数据集。STRIDE-QA 是唯一一个提供大规模、自我收集的 RGB+LiDAR 数据的以自我为中心的时空监督的数据集，支持对自动驾驶（AD）至关重要的细粒度推理。（S：空间，ST：时空）

- 自车和目标体的速度

该任务的问答示例如图 2 (C) 所示。

STRIDE-QA 数据集

我们介绍了 STRIDE-QA（用于自我中心视觉问答的驾驶场景中时空推理的基准），这是一个大规模基准，旨在推进城市驾驶中的自我中心时空推理。STRIDE-QA 包括 285 K 帧，448 K 独特的对象对，以及 16 M 问答对。

第 节描述了数据收集设置，第 节详细介绍了自动化标注流程。

驾驶数据收集

行驶区域。 我们在东京收集了超过 100 小时的驾驶数据，东京以其交通拥堵、复杂的规定（如禁停区和单行道）以及行人和自行车密度高而闻名。该城市包括市中心区域、住宅社区和郊区。

传感器设置。 我们使用配备有传感器套组的多用途车辆，该套组包括一个 64 通道的 LiDAR 和六个摄像头，它们被同步并下采样到 2 Hz。每个摄像头拥有 60° 的视场（FOV），提供 360° 的视觉覆盖，分辨率各异（前/后：2880x1860 像素；侧面：1920x1240 像素）。为了精确估计车辆的状态和位置，我们还记录惯性测量单元（IMU）和实时动态全球导航卫星系统（RTK-GNSS）接收器的信号。在数据采集过程中，所有传感器和控制区域网络（CAN）信号都被记录为同步时间戳。依据 nuScenes（Caesar et al. 2020），记录的序列被分割成 20 秒的视频片段。有关传感器设置的详细信息见附录 。

注释流程

我们提出了一种自动化注释流程，该流程处理同步多视角 RGB 图像和 LiDAR 点云的序列，以生成时空问答对，每个对由一个问题、一个答案以及辅助的视觉证据（例如，边界框或分割掩码）组成。

我们的流程，图 1 中所示，由七个组件组成：关键帧采样、3D 目标检测、多目标跟踪、属性提取、语义分割、可见性过滤和问题生成。

关键帧采样。 在我们的流程中，关键帧作为每个时空问题的时间锚点。我们从每个剪辑的 2 到 17 秒之间以 1Hz 的频率进行采样，这为每个关键帧提供了一个从前 2 秒到后 3 秒的时间上下文窗口。

3D 目标检测。 我们采用了 BEVFusion（Liu et al. 2023），该方法融合了激光雷达和多摄像头图像，以进行高精度的 3D 目标检测。对于每个关键帧，我们估计所有可见物体的 3D 位置、方向和大小。物体类别集包括来自 nuScenes 的类别以及额外的自定义类别，总共 45 个类别。

对象跟踪。 为了捕捉时间动态，我们跟踪每个对象随时间的移动。选择了 PubTracker（Yin, Zhou, and Krahenbuhl 2021）因为它在基于点的 3D 对象跟踪中具有效率高和简单的特点，这使得可以在不依赖外观特征的情况下，在帧之间进行快速且一致的 ID 分配。我们在连续帧之间关联 3D 边界框，并为每个对象分配一致的实例 ID。这样，所有帧，包括与每个关键帧相关的上下文和未来帧，都标注了 3D 边界框，对象可以随着时间的一致性被跟踪。

属性提取器 为了支持时空问答生成，我们提取每一帧中每个对象的相关属性。基于三维边界框和自车的姿态，我们在以自车为中心的坐标系中计算出欧几里得距离和航向角。除了这些数值，我们还生成离散和定性的空间描述，比如从这些量导出的“左边”或“1 点钟”。

此外，通过利用在跟踪阶段获得的时间一致的目标轨迹，我们根据目标边界框中心位置随时间的变化来估算每个目标的速度。这些属性，包括距离、朝向和速度，在每一帧中都与每个目标相关联，并构成生成时空问答对的基础。

语义分割。为了获得精确的二维分割掩膜用于下游可见性过滤，并为视觉语言模型提供细粒度的像素级视觉锚定，我们应用了 SAM 2.1 Large（Ravi et al. 2025）。该模型在每个对象的投影三维边界框内生成一个无类别的掩膜。然后，将每个掩膜与跟踪阶段中对象的一致实例 ID 相关联。

可见性过滤。 由于边界框和分割掩码是自动生成的，因此过滤掉不可靠的对象以确保标注质量是很重要的。我们应用以下三条过滤规则：（1）边界框之间的交并比（IoU）必须大于 0.3，（2）覆盖率必须超过 0.8，以及（3）由 SAM 2.1 Large 预测的 IoU 评分也必须超过 0.8。通过应用这些规则，我们有效地消除了噪声检测，只保留高质量的标注。

问题生成器。 最后，我们基于先前提取的对象属性

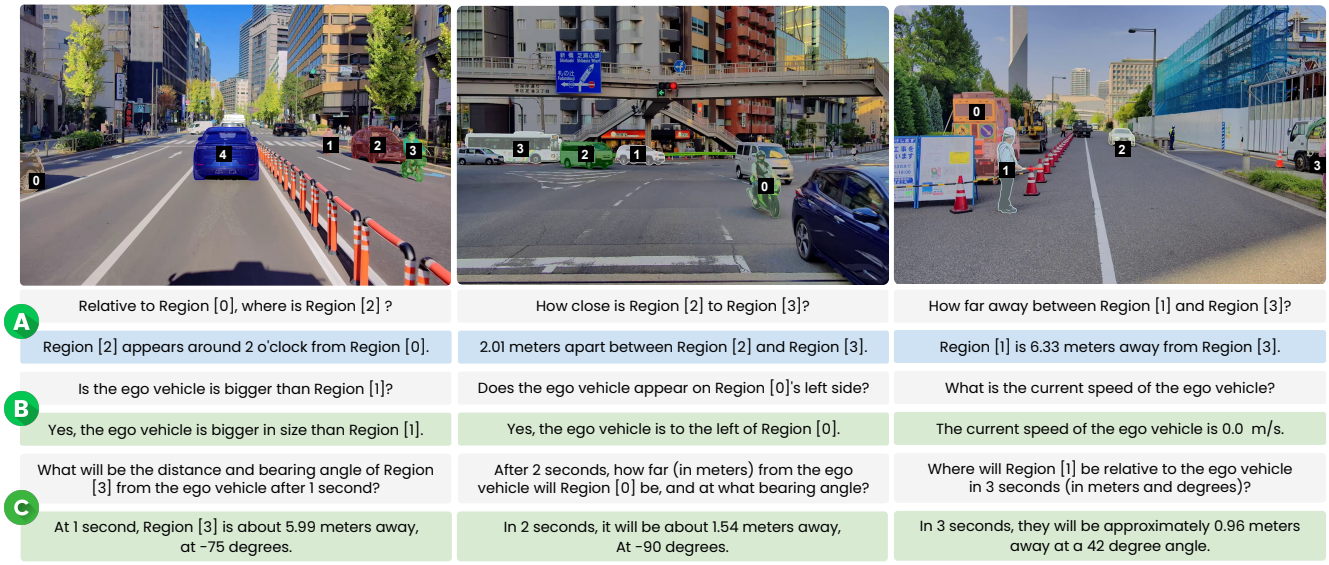

Figure 2: 来自我们 STRIDE-QA 数据集的示例数据。从上到下，每个问答对分别对应于（A）以物体为中心的空间问答，（B）以自我为中心的空间问答，以及（C）以自我为中心的时空问答。

（如距离、方向和速度）为每个关键帧生成基于模板的问答对。每个对象在问题文本中被称为 Region [X] ，其中数字 X 对应于图像中显示的视觉标识符。这些问答对与当前、背景和未来帧的 2D 边界框和分割掩膜对齐。每个问答对 (q, a, s) 都与关键帧 $t_k \in \mathcal{K}$ 相关联，并在时间窗口 $[t_k - 2, t_k + 3]$ 内引用目标对象或对象对 (o_i, o_j) 。空间定位 s 由相关帧 t 中的 2D 边界框 $\mathcal{B}_t^{(i)}$ 和分割掩膜 $\mathcal{M}_t^{(i)}$ 组成，通过一致的实例 ID 对齐。如图 2 所示，这产生了一个由成对的图像、分割和问答示例组成的数据集。

这些阶段自动从同步的多摄像头 RGB 和 LiDAR 数据生成时空 QA 对，确保用于 VLM 训练和评估的时序对齐、实例一致的标注；标注质量细节在附录 中。

实验

我们进行实验以评估 STRIDE-QA 数据集在增强 VLMs 在交通场景中的时空推理能力方面的有效性。具体来说，我们在空间和时空问答任务上对多个 VLMs 进行基准测试，以揭示它们当前的优势和局限性，并评估 STRIDE-QA 是否有助于改善这些能力。

实验设置

微调模型。 我们微调了两个开源的视觉语言模型（VLMs），Qwen2.5-VL-7B-Instruct 和 Cosmos-Reason1-7B，在 STRIDE-QA 的训练集上，只使用前置摄像头的 RGB 序列（$t \in \{-1.5, -1.0, -0.5, 0\}$ s）。为了匹配下一小节详细说明的评估设置，侧面和后置摄像头的图像以及 LiDAR 被省略。我们将得到的模型称为 STRIDE-Qwen2.5-VL-7B 和 STRIDE-Cosmos-Reason1-7B。

基准模型。我们将微调后的模型与一系列视觉语言模型（VLMs）进行评估。为了估计通用视觉语言模型的上限，我们包括了专有的 GPT-4o 和 GPT-4o mini。我们进一步与开源模型 InternVL2.5-8B 和 Qwen2.5-VL-7B-Instruct、空间增强的 SpatialRGPT-8B、以及自动驾驶特定的视觉语言模型 Senna-VLM 和 Cosmos-Reason1-7B 进行了比较。训练过程的详细信息在附录中提供。

空间基准。 我们在 SpatialRGPT-Bench（Cheng et al. 2024）的户外拆分上评估空间推理。关于 SpatialRGPT-Bench 的细节在附录 中提供。

时空基准。 除了空间推理评估外，我们还采用了三种评估指标。简要描述在下面的小节（ ）中提供，详细信息请参见附录 。

时空问答基准

超越空间问答基准，我们评估在第 节中定义的时空问答任务的性能。我们报告了四个指标：定位成功率（LSR）、平均定位成功率（MLSR）、时间定位一致性（TLC）和诊断每个维度的成功率（SR）。

模型接收到从前置车载摄像头以 60° 视场角在 $t \in \{-1.5, -1.0, -0.5, 0\}$ s 处捕获的四帧 RGB 序列。在此序列中，模型观察到来自六类之一的单个目标主体（汽车、大型车辆、巴士、行人、摩托车或自行车），通过其在所有四帧中的分割掩码进行识别，以提供足够的背景信息。模型的任务是在 $t \in \{0, 1, 2, 3\}$ s 处预测以下数量：目标主体的距离、速度和航向角，以及自车的速度。主体的航向角是在自车的参考框架中定义的，其中 0° 为前方向，正角在 $(-180°, 180°]$ 的范围内逆时针。这一定义在提示中明确提供，以便进行公平评估。值得注意的是，任务包括预测可能在 $t > 0$ 处移出摄像头视场的主体。

评估数据集。 评价数据集是从我们 STRIDE-QA 语料库中保留的录音日期构建，以确保训练和测试的分离。为了专注于有挑战性的时空推理，我们过滤数据以排除静态和重复的场景，仅保留那些具有动态交互的场景。在对所有语义掩码注释进行人工验证后，该基准由 409 个独特的场景组组成。这些组被归类为六种动态情境：迎面通过、维持状态、超车、路径分歧、远离自车以及

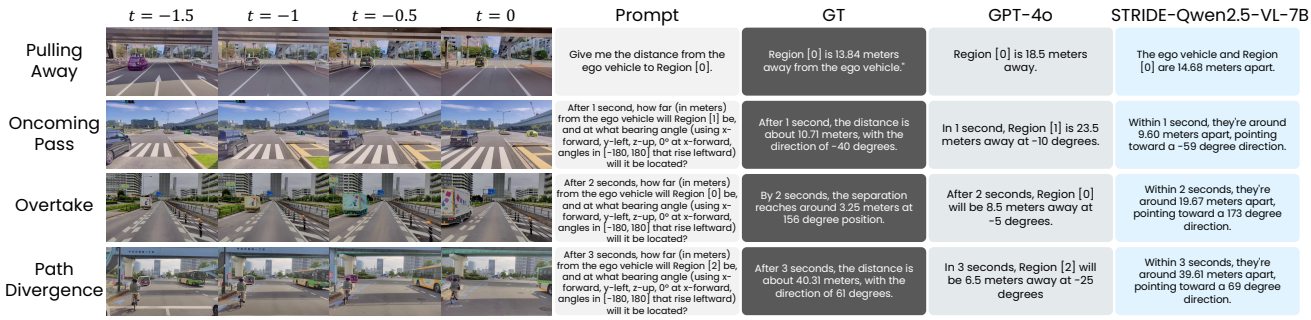

Figure 3: 关于时空 QA 基准的定性结果。在四种驾驶场景和 $t = 0$–$3\,s$ 的预测未来帧中，我们精调的 STRIDE-Qwen2.5-VL-7B 始终比 GPT-4o 提供更准确的距离和角度估计，这突显了其卓越的时空推理能力。

次要关系。我们定义离视（OOV）事件为目标物体在未来的任何帧（$t \in \{1, 2, 3\}$ 秒）中完全离开前摄像头视野的情况。这些情境在其 OOV 可能性上表现出明显的差异；例如，一些情境如迎面通过和超车通常涉及 OOV 事件，而其他情境如维持状态则很少涉及。所有情境的详细统计信息见附录 。每个场景组包含一组关于自车和目标代理在多个时间步（$t \in \{0, 1, 2, 3\}$s））的状态的问题，总计 5,317 个问答对。

定位成功率 (LSR)。 LSR 是一个二进制指标，用于评估模型的空间预测在距离和方向上是否同时准确。在这里，$\hat{d}_t$ 和 $\hat{\theta}_t$ 表示模型预测，而 d_t^* 和 θ_t^* 则是相应的真实值。仅当以极坐标表示的估计距离和方位 $\hat{\boldsymbol{p}}_t = (\hat{d}_t, \hat{\theta}_t)$ 满足条件

$$|\hat{d}_t - d_t^*| < 0.25\, d_t^* \quad \text{and} \quad |\hat{\theta}_t - \theta_t^*| < 10^\circ.$$

时，预测才被视为成功。我们采用了先前工作中的 ±25% 距离边界 (Cheng et al. 2024)，并设置 ±10° 航向边界，以便在 10 米处，横向偏移等于标准的 3.5 米车道宽度。

平均定位成功率 (MLSR)。 MLSR 对序列中的每一帧 LSR 取平均，提供了一个更柔和的时间稳定性度量：

$$\mathrm{MLSR} = \frac{1}{G} \sum_{g} \frac{1}{T+1} \sum_{t=0}^{T} s_{g,t},$$

其中 $\boldsymbol{s}_g = [s_{g,0}, \ldots, s_{g,T}]$ 表示场景 g 的 LSR 成功位。

时间定位一致性 (TLC)。 TLC 是一种严格的度量标准，衡量代理在所有四个时间步内成功定位的序列比例，即目标从未丢失：

$$\mathrm{TLC} = \frac{1}{G} \sum_{g} \mathbf{1}\big[s_{g,0} \wedge s_{g,1} \wedge s_{g,2} \wedge s_{g,3}\big].$$

其中 $\boldsymbol{s}_g = [s_{g,0}, \ldots, s_{g,T}]$ 是场景组 g 的 LSR 成功位序列，具有 $T = 3$ 。

每维成功率。 单轴成功率针对距离、航向角和速度进行了计算以便诊断分析。完整的指标定义和详细结果见附录 。

SpatialRGPT-Bench 的结果

表 2 总结了结果。STRIDE-Qwen2.5-VL-7B 和 STRIDE-Cosmos-Reason1-7B 在所有拆分中均显著

Model	Original ↑		Obj. Spatial ↑		Ego Spatial ↑	
	Qual.	Quant.	Qual.	Quant.	Qual.	Quant.
GPT-4o	80.5	32.5	68.1	39.4	55.7	27.7
GPT-4o mini	54.7	30.6	56.3	32.1	57.1	44.4
Intern-VL 2.5 8B	64.1	18.8	48.4	26.6	36.3	29.1
Qwen2.5-VL 7B-Instruct	67.2	24.4	64.0	12.8	47.1	29.3
SpatialRGPT-VILA-1.5-8B	75.0	46.9	58.4	42.2	25.3	16.9
Senna-VLM	18.0	0.63	9.14	4.59	5.88	2.03
Cosmos-Reason1-7B	53.9	30.0	55.2	21.1	33.2	20.3
STRIDE-Qwen2.5-VL-7B	69.5	37.5	61.1	61.5	77.9	70.3
STRIDE-Cosmos-Reason1-7B	71.1	30.0	62.2	58.7	79.9	68.9

Table 2: SpatialRGPT-Bench 结果。在原始 Spatial-RGPT 数据集上进行定性（Qual.）和定量（Quant.）空间推理性能评估，以及我们的对象中心和自我中心扩展。

优于其基础模型。在定量对象中心的空间问答中，STRIDE-Qwen2.5-VL-7B 从 12.8 上升到 61.5（+48.7 pt，×4.8），而 STRIDE-Cosmos-Reason1-7B 从 21.1 上升到 58.7（+37.6 pt，×2.8）。在定量自我中心拆分中也出现了类似的提升，两个模型的分数分别从 29.3 提升到 70.3（+41.0 pt）和从 20.3 提升到 68.9（+48.6 pt）。这些结果表明，我们的 STRIDE 对齐策略不仅限于 STRIDE-QA，并在外部基准上提高了细粒度的空间推理能力。

虽然 GPT-4o 和 SpatialRGPT-8B 在原始划分上取得了最高的准确率，但它们在 Object-centric 子集上的性能显著下降（定性准确率 −12.4 pt 和 −16.6 pt）。这一差距表明，通用的大模型并不会自动转移到以相机为中心的设置中，可能是由于视角变化和注释噪声导致的。详细的消融试验见附录 。

时空问答基准结果

如表 3 所示，所有的基线模型，包括强大的通用 VLMs，在第 节定义的时空基准测试中表现不佳。它们的定位成功率（LSR）得分很低，并且在多帧（MLSR）和时间一致性（TLC）指标上几乎接近于零的结果。这表明现有模型在复杂驾驶场景中进行可靠的时空推理方面存在根本上的不足。

相比之下，在 STRIDE-QA 上微调的模型表现出显著的性能提升。我们表现最好的模型，STRIDE-Qwen2.5-VL-7B，在 $t = 0$ s 上达到 96.3% 的 LSR，比其基线提高了 96× ，证明了其获取精确的当前时刻空间理解的

Model	LSR ↑				MLSR ↑	TLC ↑
	0s	1s	2s	3s		
GPT-4o	18.1	6.6	6.1	7.6	9.6	0.7
GPT-4o mini	4.6	2.0	0.7	0.7	2.0	0.0
InternVL2.5-8B	2.4	1.0	1.7	0.7	1.5	0.0
Qwen2.5-VL-7B-Instruct	1.0	3.4	4.4	1.0	2.4	0.0
SpatialRGPT-VILA-1.5-8B	0.5	0.2	0.2	0.0	0.2	0.0
Senna-VLM	1.0	0.0	0.2	0.0	0.3	0.0
Cosmos-Reason1-7B	1.5	3.2	2.0	1.5	2.0	0.0
STRIDE-Qwen2.5-VL-7B	96.3	46.2	38.4	38.9	55.0	28.4
STRIDE-Cosmos-Reason1-7B	96.8	43.5	37.4	36.2	53.5	25.4

Table 3: 时空 QA 基准结果。我们微调的 STRIDE-Qwen2.5-VL-7B 在联合 (LSR@t)、多帧 (MSLR) 和时间一致性 (TLC) 指标方面实现了最佳性能。

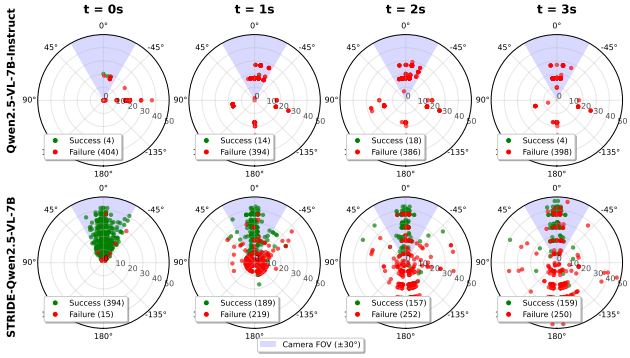

Figure 4: 在极坐标系上比较定位性能（角度 = 航向方向，半径 = 自我的距离）。第一行（Qwen2.5-VL-7B-Instruct）：基线的稀疏、有偏预测表明一种简单的、记忆化的行为。第二行（STRIDE-Qwen2.5-VL-7B）：我们微调的模型生成了密集的、合理的预测，在 $t = 0\,s$ 处实现了几乎完美的定位。绿色点表示成功（LSR），红色表示失败，蓝色楔形标记相机的 ±30° 视野。

能力。这种能力还扩展到未来的预测，在 $t = 3$ s 上维持 38.9% 的 LSR（提高了 39× ）。此外，它获得了 55.0 的 MLSR 和 28.4 的 TLC ，确认我们的数据集有效地教导了跨视角和时间的一致推理（定性示例见图 3 ）。

然而，结果也突出了这一任务的巨大难度。尽管 TLC 得分达到 28.4 较 0 有了巨大改进，但这也表明实现长期、完全一致的推理仍然是一个艰巨的挑战。这表明虽然对我们的数据集进行微调提供了一个坚实的基础，但可能不足以完全掌握现实世界动态预测的复杂性。

总之，我们的工作有两个关键贡献。首先，我们表明在 STRIDE-QA 上进行微调可以弥补视觉语言模型在时空推理能力方面的一个关键差距。其次，通过量化长期一致性方面的剩余挑战，我们的基准测试引导未来研究朝着构建鲁棒且安全关键的自主系统方向发展。

除了性能指标之外，我们还进行了深入分析，以更好地理解使用 STRIDE-QA 训练的 VLMs 的优点、局限性和泛化行为。

对图 4 中预测模式的定性分析揭示了模型行为的根本区别。基线 VLM（顶部行）表现出一种一致的失败模式：其预测不仅稀疏而且系统性地偏向于无论视觉输入如何都反复猜测类似的错误答案。这表明模型默认采用一种简化的、死记硬背的行为，而不是基于视觉背景

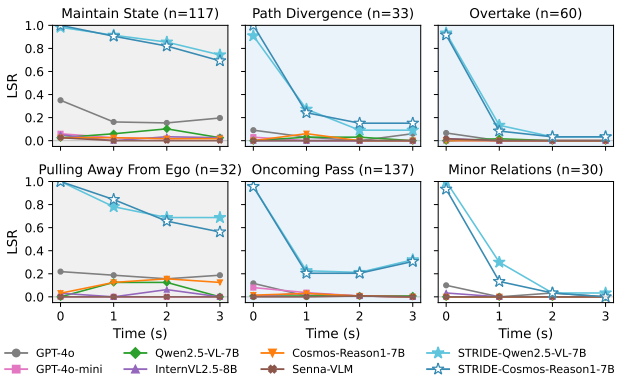

Figure 5: 在动态场景中，按视野外（OOV）可能性分组的 LSR 趋势。相比视野内场景（灰色背景），在高 OOV 可能性场景（蓝色背景）中性能的急剧下降突显出 OOV 预测是主要挑战。

的推理。相反，我们的微调模型（底层）具有响应能力，能够在 360 度空间中生成密集的预测，并准确定位当前帧中的物体（$t = 0s$ ）。基线模型未能提供上下文感知的连续预测，表明缺乏时间一致性，而这对运动规划来说是一个基本前提。这个前提在以前将 VLMs 与规划模块连接的工作中常常被忽视（Sima et al. 2024; Tian et al. 2024），而我们的框架旨在解决这一差距，提供一个显式训练和评估此技能的框架。

视野外预测的误差分析

为了理解上一节中识别出的时间退化的根本原因，我们分析了我们的模型在基准测试中定义的不同动态场景下的性能（图 5 ）。一个明显的模式浮现出来：在目标体可能保持在摄像机视场内的场景中，例如保持状态，LSR 会比较平缓地下降。相反，在目标体倾向于离开视场的场景中，例如超车、迎面而来的通过和路径偏离，LSR 的退化要显著得多。这种明显的对比强烈地表明，长期预测的主要失效模式是模型无法推理出视场外物体轨迹的能力。这一发现揭示了仅依赖单摄像机输入进行预测的一个关键局限性，并强调整合多摄像机信息是未来工作的一个关键方向，以构建真正强健的自动驾驶系统。

结论

我们引入了 STRIDE-QA，这是一个大规模的视觉问答数据集，旨在解决自动驾驶视觉语言模型中的时空推理差距。在该数据集上进行微调提升了性能：我们最好的模型达到了 55.0 % MLSR 和 28.4 TLC，显著超越了接近于零的基线。尽管这一结果令人鼓舞，但其在下游规划任务中的泛化能力仍不明确。STRIDE-QA 为自动系统中稳健且物理基础深厚的视觉语言理解奠定了基础。

本文基于从项目 JPNP20017 获得的结果，该项目由新能源和产业技术开发组织 (NEDO) 资助。

References

Ainslie, J.; Lee-Thorp, J.; De Jong, M.; Zemlyanskiy, Y.; Lebrón, F.; and Sanghai, S. 2023. GQA: Train-

ing Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. arXiv:2305.13245.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Baruch, G.; Chen, Z.; Dehghan, A.; Dimry, T.; Feigin, Y.; Fu, P.; Gebauer, T.; Joffe, B.; Kurz, D.; Schwartz, A.; et al. 2021. Arkitscenes: A Diverse Real-World Dataset for 3D Indoor Scene Understanding Using Mobile RGB-D Data. https://arxiv.org/abs/2111.08897. arXiv:2111.08897.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11621–11631.

Chen, B.; Xu, Z.; Kirmani, S.; Ichter, B.; Sadigh, D.; Guibas, L.; and Xia, F. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14455–14465.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 24185–24198.

Cheng, A.-C.; Yin, H.; Fu, Y.; Guo, Q.; Yang, R.; Kautz, J.; Wang, X.; and Liu, S. 2024. SpatialRGPT: Grounded Spatial Reasoning in Vision-Language Models. In NeurIPS.

Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2025. BLINK: Multimodal Large Language Models Can See but Not Perceive. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., Computer Vision – ECCV 2024, 148–166. Cham: Springer Nature Switzerland. ISBN 978-3-031-73337-6.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, 3354–3361.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, 6904–6913.

Gupta, V. 2023. Dashcam Anonymizer.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. ICLR, 1(2): 3.

Jiang, B.; Chen, S.; Liao, B.; Zhang, X.; Yin, W.; Zhang, Q.; Huang, C.; Liu, W.; and Wang, X. 2024. Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving. arXiv:2410.22313.

Jin, B.; Zheng, Y.; Li, P.; Li, W.; Zheng, Y.; Hu, S.; Liu, X.; Zhu, J.; Yan, Z.; Sun, H.; Zhan, K.; Jia, P.; Long, X.; Chen, Y.; and Zhao, H. 2024. TOD3Cap: Towards 3D Dense Captioning in Outdoor Scenes. In Computer Vision –ECCV 2024: 18th European Conference, Milan, Italy, September 29 –October 4, 2024, Proceedings, Part XVIII, 367–384. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-72648-4.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.; and Han, S. 2023. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In IEEE International Conference on Robotics and Automation (ICRA).

NVIDIA; :; Azzolini, A.; Brandon, H.; Chattopadhyay, P.; Chen, H.; Chu, J.; Cui, Y.; Diamond, J.; Ding, Y.; Ferroni, F.; Govindaraju, R.; Gu, J.; Gururani, S.; Hanafi, I. E.; Hao, Z.; Huffman, J.; Jin, J.; Johnson, B.; Khan, R.; Kurian, G.; Lantz, E.; Lee, N.; Li, Z.; Li, X.; Lin, T.-Y.; Lin, Y.-C.; Liu, M.-Y.; Luo, A.; Mathau, A.; Ni, Y.; Pavao, L.; Ping, W.; Romero, D. W.; Smelyanskiy, M.; Song, S.; Tchapmi, L.; Wang, A. Z.; Wang, B.; Wang, H.; Wei, F.; Xu, J.; Xu, Y.; Yang, X.; Yang, Z.; Zeng, X.; and Zhang, Z. 2025. Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning. arXiv:2503.15558.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Łukasz Kaiser; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Łukasz Kondraciuk; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Le-

ung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov, M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; and Zoph, B. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Park, S.-Y.; Cui, C.; Ma, Y.; Moradipari, A.; Gupta, R.; Han, K.; and Wang, Z. 2025. NuPlanQA: A Large-Scale Dataset and Benchmark for Multi-View Driving Scene Understanding in Multi-Modal Large Language Models. arXiv:2503.12772.

Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; and Jiang, Y.-G. 2024. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 4542–4550.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollar, P.; and Feichtenhofer, C. 2025. SAM 2: Segment Anything in Images and Videos. In The Thirteenth International Conference on Learning Representations (ICLR).

Roberts, M.; Ramapuram, J.; Ranjan, A.; Kumar, A.; Bautista, M. A.; Paczan, N.; Webb, R.; and Susskind, J. M. 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In Proceedings of the IEEE/CVF international conference on computer vision, 10912–10922.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; and Li, H. 2024. DriveLM: Driving with Graph Visual Question Answering. In Proceedings of the European Conference on Computer Vision (ECCV), 256–274.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 567–576. Los Alamitos, CA, USA: IEEE Computer Society.

Tian, X.; Gu, J.; Li, B.; Liu, Y.; Wang, Y.; Zhao, Z.; Zhan, K.; Jia, P.; Lang, X.; and Zhao, H. 2024. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. In 8th Annual Conference on Robot Learning (CoRL).

Wu, D.; Han, W.; Liu, Y.; Wang, T.; Xu, C.-Z.; Zhang, X.; and Shen, J. 2025. Language Prompt for Autonomous Driving. Proceedings of the AAAI Conference on Artificial Intelligence, 39(8): 8359–8367.

Wu, D.; Han, W.; Wang, T.; Dong, X.; Zhang, X.; and Shen, J. 2023a. Referring Multi-Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14633–14642.

Wu, D.; Han, W.; Wang, T.; Dong, X.; Zhang, X.; and Shen, J. 2023b. Referring Multi-Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14633–14642.

Yang, J.; Zhang, H.; Li, F.; Zou, X.; Li, C.; and Gao, J. 2023. Set-of-mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. https://arxiv.org/abs/2310.11441. arXiv:2310.11441.

Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. Center-Based 3D Object Detection and Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11784–11793.

Zhou, X.; Larintzakis, K.; Guo, H.; Zimmer, W.; Liu, M.; Cao, H.; Zhang, J.; Lakshminarasimhan, V.; Strand, L.; and Knoll, A. 2025. TUMTraf VideoQA: Dataset and Benchmark for Unified Spatio-Temporal Video Understanding in Traffic Scenes. In Forty-second International Conference on Machine Learning.

STRIDE-QA 数据集详情

传感器设置和校准

传感器设置　这六个摄像头和 LiDAR 传感器按照图 6 所示进行安装。此摄像头配置提供了 360 度的视觉覆盖。

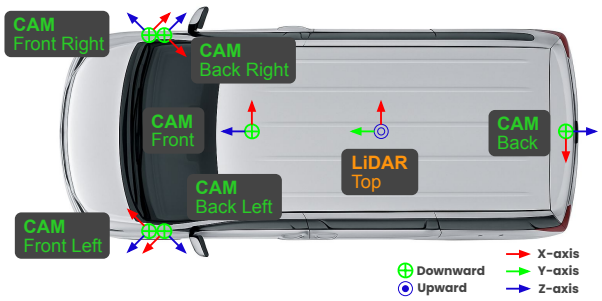

Figure 6: 我们数据收集平台的传感器设置。

传感器校准　为了获得高质量的多传感器数据集，需要对传感器的内部和外部参数进行精确校准。我们执行了一项全面的校准过程，涉及 LiDAR、摄像头和 VRTK（视觉-RTK）。关键步骤如下所述：

- 高精度映射和轨迹估计
  我们采用 LiDAR SLAM（同步定位与地图构建）来估计 3D 点云地图和车辆的轨迹。初始位姿由 VRTK 系统提供，该系统集成了 RTK-GNSS、IMU 和相机数据。生成的地图随后使用迭代最近点（ICP）与全球参考地图对齐。
- 相对位姿估计：VRTK、LiDAR 和地面
  我们通过刚性变换拟合对齐它们的轨迹来估计 VRTK-LiDAR 的位姿，固定平移并优化旋转。LiDAR-地面的位姿是通过从点云中检测地面平面，并计算跨帧的中值高度、俯仰和滚转来确定的。
- 相机标定和尺度对齐：
  从运动中重建 (SfM) 用于恢复相机内部参数和相机间的外部参数。为了解决尺度歧义问题，我们使用 7 自由度的具有尺度感知的 ICP，将 SfM 重建与度量 LiDAR 地图对齐，该过程从轨迹比较开始初始化。

此过程产生所有传感器之间的最终外部参数以及每个摄像机的内部参数。

统计学

我们提供了 STRIDE-QA 数据集的统计信息，以突出其规模和内容多样性。表 4 显示了定性和定量问答对的分布。此外，表 5 展示了不同问答类别中的对象类别分布。大量问题针对常见的交通参与者，如车辆、摩托车和行人。这个对象分布紧密反映了东京的城市交通模式，该数据集就是在那里收集的。

管道验证和质量评估

鉴于 STRIDE-QA 数据集的庞大规模（1600 万对问答对），直接对所有注释进行人工评估是不可行的。因此，我们采用了一种间接但严格的方法：评估决定流程质量的核心组件的性能，即 3D 目标检测和多目标跟踪。

QA Category	Qualitative	Quantitative
Object-centric Spatial QA	2.20 M	1.10 M
Ego-centric Spatial QA	3,10 M	4.33 M
Ego-centric Spatiotemporal QA	—	2.18 M

Table 4: STRIDE-QA 数据集中三个类别中的问答对的分布。

Class	Obj. (S)	Ego. (S)	Ego. (ST)
vehicle	3,780,995	3,844,934	1,448,028
pedestrian	1,187,035	1,213,703	195,642
large vehicle	965,995	1,029,173	348,618
bicycle	337,045	357,987	64,476
bus	179,395	201,948	80,790
motorcycle	150,450	158,710	45,582
kick scooter	3995	4,160	894
Total	6,604,910	6,810,615	2,184,030

Table 5: STRIDE-QA 数据集中，问题解答类别中目标类的分布。Obj. (S)：以物体为中心的空间问题解答，Ego. (S)：以自我为中心的空间问题解答，Ego. (ST)：以自我为中心的时空问题解答。请注意，一个单一的目标实例可以与多个类别的问题相关联；因此，这些列并不是相互排斥的。

为了进行此评估，我们首先准备了一个与主要的 STRIDE-QA 驾驶日志分开收集的基准数据集，但具有相同的传感器配置。该数据集的创建外包给专家人工标注员，使用 Appen® MatrixGo 工具（如图 7 所示）并根据我们的详细指南进行。我们团队审核并修正了所有提交的标注以确保质量。标注遵循 nuScenes 格式，每个三维边界框分配一个一致的实例 ID，并在每个 10 秒的场景中进行跟踪。该数据集被分为 1688 个场景（大约 281 分钟）的训练集和 100 个场景（约 17 分钟）的验证集。

使用这个 100 个场景验证集作为真实数据，我们通过五个标准指标（mAP, mATE, mASE, mAOE, mAVE）评估检测性能，并通过四个指标（AMOTA, AMOTP, Recall, IDSW）评估跟踪性能。结果如表 6 所示。

Task	Metric	Score	Range
Detection	mAP ↑	0.701	$[0, 1]$
	mATE ↓ (m)	0.136	$[0, \infty)$
	mASE ↓	0.168	$[0, 1]$
	mAOE ↓ (rad)	0.146	$[0, \pi]$
	mAVE ↓ (m/s)	1.280	$[0, \infty)$
Tracking	AMOTA ↑	0.676	$[0, 1]$
	AMOTP ↓ (m)	0.556	$[0, \infty)$
	Recall ↑	0.700	$[0, 1]$
	IDSW ↓	687	$[0, \infty)$

Table 6: 我们的自动注释流水线实现的类别平均检测和跟踪指标。

检测质量。该检测器达到了高精度，其 mAP

(IoU@0.5:0.95) 为 0.701，平移误差（mATE）仅有 0.136 米（13.6 厘米）。形状（mASE=0.168）和航向（mAOE=0.146 弧度 8.4 度）同样准确，速度误差保持在足够低的水平（mAVE=1.28 米/秒）。这些数值表明，嵌入 STRIDE-QA 问题中的值是来源于厘米级的几何估计。± 25 % 的距离容差基于以前的工作，而 ± 10 ° 的角度容差是一个实用的阈值设置，使得在 10 米距离处的横向误差对应于一个标准车道宽度（大约 3.5 米）。
跟踪质量。 整体多目标跟踪（MOT）性能达到了一个良好的分数，AMOTA = 0.676（在包含 10,461 条独特目标轨迹的 100 个场景中进行评估）。该指标对跟踪失败（ID 切换）、漏检（召回率）和虚警进行了全面评估，这个分数表明我们的管道在这些错误之间保持了良好的平衡。对于我们从稳定轨迹生成问答对的任务，我们强调两个指标：跟踪可靠性和位置精度。为此，我们的系统实现了高可靠性，在总共 10,461 条轨迹中只有 687 次 ID 切换（大约每 10 秒场景 6.9 次）。对于位置精度，平均多目标跟踪精度（AMOTP）保持在 0.556 m。这种高位置精度对于生成空间精确的问题和答案至关重要，因为它们的内容依赖于准确的目标位置。
总结。 总之，所提出的流程在位置精度上取得了高水平，平均检测误差（mATE）低于 14 cm，平均跟踪误差（AMOTP）低于 0.6 m。这次验证的结果强有力地支持了这样的论断：STRIDE-QA 数据集将大规模 QA 对与几何精度相结合，为在 VLMs 中的时空推理进行严格评估提供了必要条件。

隐私与匿名化

我们通过使用 Dashcam Anonymizer（Gupta 2023）自动模糊面部和车辆牌照来匿名化整个数据集，并移除所有可识别的时间戳元数据，同时保留原始帧顺序。

实现细节

我们微调了两个开源的 VLMs，Qwen2.5-VL-7B（Bai et al. 2025）和 Cosmos-Reason1-7B（NVIDIA et al. 2025），在 STRIDE-QA 数据集的训练拆分中进行微调。每个训练样本由四个上下文帧组成，调整到 532×336 像素，使用 Set-of-Mark (SoM) 方法生成的区域 ID-标注的遮罩（Yang et al. 2023）。我们采用 LoRA（Hu et al. 2022）方法进行参数高效的微调。所有模型均在 16 个 NVIDIA H100 GPU 上使用 DeepSpeed (ZeRO Stage 2) 进行分布式训练优化，所有超参数设置如表 7 中所总结的一样。在所有训练运行中使用固定的随机种子 42 以确保可重复性。在评估期间，输入帧采用与训练期间相同的方式进行预处理。我们使用 0 的解码温度以确保确定性推理。

Parameter	Value
Training Configuration	
Epochs	2
Global Batch Size	64
Precision	bfloat16
Optimizer (AdamW)	
Learning Rate (LLM)	5e-5
Learning Rate (Vision Encoder)	2e-6
Learning Rate (Projection Head)	1e-5
Betas (β_1, β_2)	(0.9, 0.999)
Epsilon (ϵ)	1e-8
Weight Decay	0.1
Scheduler	
Type	Cosine
Warmup Ratio	0.05
LoRA Configuration	
Rank	16
Alpha	32
Dropout	0.05
Bias	none

Table 7: 用于微调的超参数。

时空问答基准

本节提供了有关时空问答基准的更多详细信息。

评估设置和提示

评估任务涉及将四个连续的上下文帧（$t \in \{-1.5, -1.0, -0.5, 0\}$ s ）作为输入，以预测目标代理在未来时间步的距离、航向角和速度（$t \in \{0, 1, 2, 3\}$ s ）。输入包括由标记集方法（SoM）生成的指定区域的掩码（Yang et al. 2023）。然而，其中一个基线模型 SpatialRGPT-VILA-1.5-8B 包含一个内置区域提取器，并且不支持多帧输入。因此，为了确保公平比较，我们采用了与其他模型相同的基于 SoM 的方法。
评估期间用于所有模型的完整指令提示在图 8 (b) 中提供。

评估数据集的统计及定义

我们的评估数据集由 409 个独特的场景组组成。这些场景在六种动态场景中的分布及其各自的视野外（OOV）率详见表 ?? 。此外，图 9 展示了整个基准测试中关键物理量的分布，证明了我们的数据集涵盖了自车和目标代理多种距离、航向角和速度的多样性。评估中包含的目标类别的详细信息见表 8 。

Class	Count
vehicle	303
large vehicle	61
pedestrian	17
bicycle	13
bus	8
motorcycle	7
Total	409

Table 8: 空间时间问答基准中的对象类别分布。

每个动态场景描述如下：
正面交会通行：目标代理从相反方向接近并通过自车。在这种情境中经常涉及 OOV 事件。
保持状态：目标代理和自车大致在相同方向上行驶，并且速度相似，保持相对恒定的空间关系。
超车：在未来 3 秒的时间范围内，自车超过目标代理。这也常常涉及到 OOV 事件。

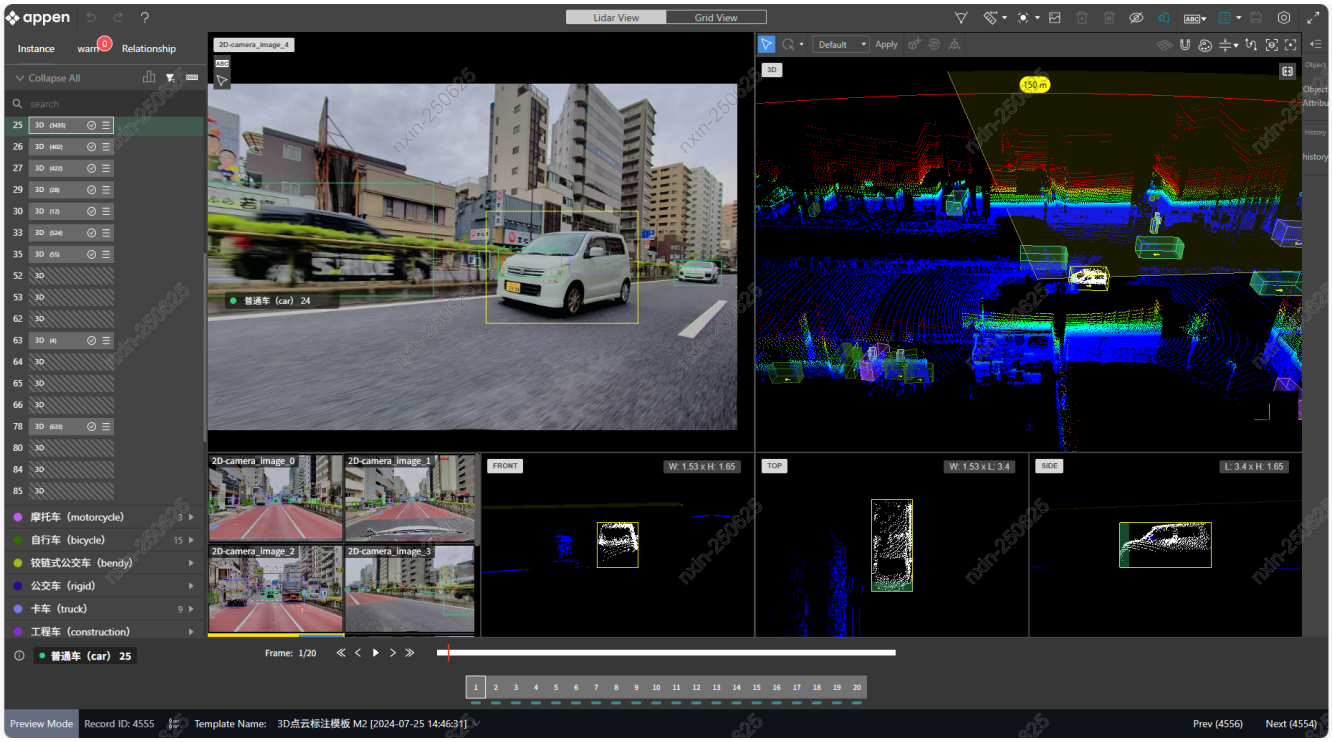

Figure 7: 使用 Appen® MatrixGo 在我们的驾驶数据集上制作的代表性 2D 和 3D 标注示例，显示了带有类别特定边界框的同步相机图像和 LiDAR 点云。

```
You are a spatial reasoning assistant that analyzes images with marked regions.
Always refer to regions as "Region [X]" where X is the region number.

For quantitative questions:
- Always provide exact numerical values with appropriate units (feet, meters).
- Never respond with "cannot determine", "not enough information", or
uncertainty expressions.
- Use visual references like typical building heights, street widths, vehicle sizes
for scale estimation.
- Format: "Region [X] is Y.YY [unit] [measurement_type]."
- Example: "Region [0] is 6.91 feet in height."

For qualitative questions:
- Give definitive comparative answers based on visual analysis.
- Use clear spatial language (left, right, taller, larger, bigger).
- Never express uncertainty, make definitive judgments.
- Format: "Region [X] [comparison_result]"
- Example: "Region [0] appears more on the left side.
- If a value cannot be obtained exactly, make your best estimate and answer..
```

(a) Instruction prompt used in SpatialRGPT-Bench

```
You are a precise assistant interpreting 4 images (t = -1.5 s, -1.0 s, -0.5 s, 0 s).
All images have marked objects (SoM). Base every reply solely on the
images.

Every question requires numeric answers. Respond with ONE short sentence
that lists all requested numbers and their units—nothing more.

Examples:
- distance: "Region [0] is 20.12 meters away."
- distance + bearing: "In 1 second they are 23.54 meters at -24 degrees."
- velocity: "Ego speed after 2 seconds is 11.77 m/s."

If a value cannot be obtained exactly, make your best estimate and answer.
No extra text, symbols, or formatting.
```

(b) Instruction prompt used in Spatiotemporal QA Benchmark

Figure 8: 定义评估任务和输出格式的指令提示：(a) SpatialRGPT-Bench 和 (b) 我们的时空问答基准。

路径分歧：目标代理和自车的轨迹在交叉路口处出现分歧，例如当一辆车转弯而另一辆车直行时。这通常涉及 OOV 事件。

远离自车：目标车辆在自车前方行驶，加速并增加分离距离。

次要关系：一个类别，汇集那些无法清晰归类到上述类别中的不常见或模糊场景，例如不完整超车、避让行人或远距离穿越。

指标定义

尽管正文中报告了整合指标如 LSR、MLSR 和 TLC，但本节提供了用于诊断分析的每个维度成功率 (SR) 的补充定义。SR 评估预测的物理量的误差是否在预定义的容差范围内。每个物理量的容差设置如下：

- 距离：预测误差必须在真实值的 $\pm$ 25 % 之内。
- 航向角：预测误差必须在真实值的 $\pm$ 10 ° 范围内。

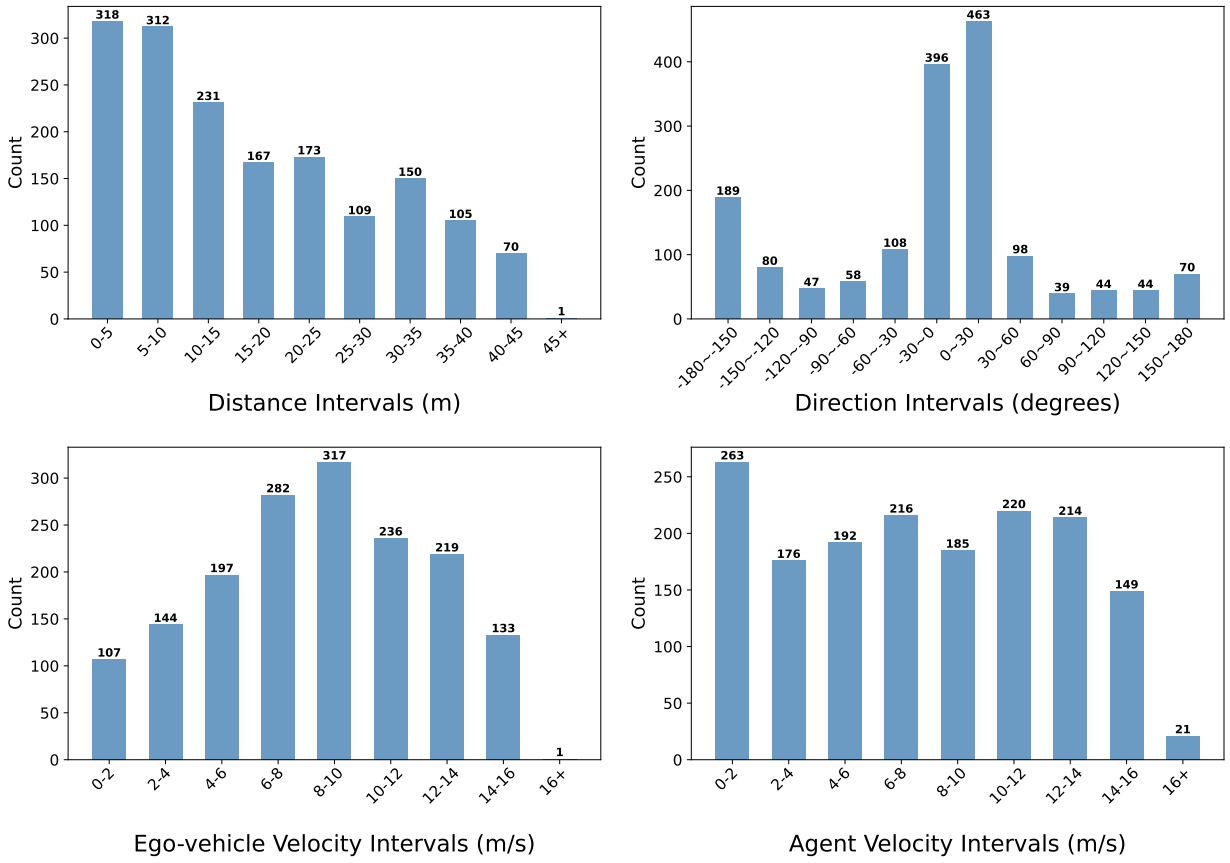

Figure 9: 时空问答基准样本分布。

- 速度：使用混合容差；预测误差必须在地面真实值的 ± 20 % 以内，若地面真实速度低于 1.0 m/s，则切换为绝对容差 ± 0.5 m/s。

逐维评估结果

表 9 报告了每个模型在预测范围内的每维成功率（SR）。

SpatialRGPT 基准测试

空间 RGPT-Bench 包含来自户外场景（nuScenes（Caesar et al. 2020），KITTI（Geiger, Lenz, and Urtasun 2012））、室内场景（SUNRGBD（Song, Lichtenberg, and Xiao 2015），ARKitScenes（Baruch et al. 2021））和模拟场景（Hypersim（Roberts et al. 2021））的数据。在本研究中，我们仅使用户外部分来评估 VLMs 在城市环境中的空间理解力。我们排除了与交通场景不太相关的问答类别，例如 Below/Above，并略去了已被时空问答基准覆盖的评估项目。用于此基准的具体指令提示如图 8 (a) 所示。

表 10 报告了 SpatialRGPT-Bench 上每个 QA 类别的结果。SpatialRGPT-Bench 的户外分集和我们的以对象为中心的 QA 分集之间的平均得分差异巨大。由于评估的 QA 类别并不完全一致，直接比较是困难的。然而，有两个主要因素似乎导致了 SpatialRGPT-Bench 户外分集和以对象为中心的 QA 分集之间明显的差距。第一个因素是标注错误。我们的数据集是通过 LiDAR 和多视角传感器融合流水线标注的（见第 节），具有很高的几何精度，而原始的 SpatialRGPT-Bench 分集通过单个前视摄像头估计深度，这可能产生更大的误差。第二个因素涉及摄像头视角范围。以对象为中心的分集使用的视角范围比户外分集更窄，去除了许多周围空间上下文，使相对位置推理更加困难。这些观察进一步强调了在评估空间理解能力时进行公平跨基准比较的难度。

局限性和未来工作

限制。 本研究有三个主要限制。首先，由于 STRIDE-QA 是自动驾驶时空推理的首个基准测试，并且不存在可比较的公共数据集，我们目前无法量化跨数据集的泛化能力。其次，资源限制使我们只能进行轻量级的 LoRA 适应；因此，用全参数监督微调可以达到的上限性能尚未被探索。第三，尽管 STRIDE-QA 的设计是为了增强视觉-语言模型的时空推理核心，但其对如运动规划和行为预测等安全关键下游任务的影响尚未经过严格评估。

未来工作。 从自车的角度来看，当目标物体离开前置摄像头的视野时，VLM 在预测周围代理的距离、方向和相对速度方面的性能下降。这一发现强调了自动驾驶 VLM 需要整合多摄像头输入的必要性，并激励未来的工作关注于高效结合多视图信息。此外，我们将通过

Model	SR_d ↑				SR_θ ↑				SR_v^{ego} ↑				SR_v^{agent} ↑			
	0s	1s	2s	3s	0s	1s	2s	3s	0s	1s	2s	3s	0s	1s	2s	3s
GPT-4o	34.7	23.0	24.9	20.5	41.3	21.3	21.5	25.9	18.6	25.7	31.5	26.2	13.0	12.7	11.2	14.2
GPT-4o mini	17.4	15.6	21.0	20.8	35.7	17.4	10.0	10.5	27.6	29.6	27.6	29.8	20.5	24.4	26.7	24.2
InternVL2.5-8B	15.9	18.6	20.8	22.0	13.9	7.8	6.1	3.2	0.5	3.7	25.9	15.2	3.4	2.7	17.8	7.6
Qwen2.5-VL-7B-Instruct	21.8	12.0	21.0	12.5	1.0	21.5	17.8	20.0	8.6	23.5	33.3	10.8	10.3	12.2	21.5	10.5
SpatialRGPT-VILA-1.5-8B	7.3	4.2	6.8	3.2	24.0	11.0	9.0	6.1	0.0	0.2	9.3	2.0	2.7	1.2	8.3	3.4
Senna-VLM	11.2	2.7	6.4	4.2	3.7	0.7	1.0	0.0	4.6	5.1	7.8	2.7	0.7	3.4	6.4	4.9
Cosmos-Reason1-7B	13.0	11.2	17.8	13.4	5.6	12.7	13.2	12.2	28.9	31.1	33.5	28.1	16.4	26.9	31.5	31.8
STRIDE-Qwen2.5-VL-7B	96.3	66.0	51.3	48.9	100	57.9	56.2	61.1	68.0	65.3	63.6	58.9	59.4	60.4	53.3	52.8
STRIDE-Cosmos-Reason1-7B	96.8	69.9	49.6	47.2	100	53.8	55.7	61.1	62.8	59.7	58.4	54.8	59.4	58.2	53.5	46.7

Table 9: 关于时空问答基准的详细每维成功率（SR）结果。所有值均为成功率（%）。我们的微调模型（蓝色背景），特别是 STRIDE-Qwen2.5-VL-7B，在所有基线的每项指标和时间范围内表现出显著的性能优势，尤其是与其原始基础模型（灰色背景）相比时。

添加更精细的指标来改进评估协议，并与 STRIDE-QA 一同发布，以便它可以作为广泛采用的基准。

Original (outdoor)

Model	Below/Above	Left/Right	Big/Small	Tall/Short	Wide/Thin	Behind/Front	Avg.
GPT-4o (OpenAI et al. 2024)	—	92.11	73.33	65.38	85.71	78.57	80.47
GPT-4o mini (OpenAI et al. 2024)	—	68.42	66.67	30.77	52.38	53.57	54.69
InternVL2.5-8B (Chen et al. 2024b)	—	89.47	73.33	46.15	66.67	39.29	64.06
Qwen2.5-VL-7B-Instruct (Bai et al. 2025)	—	92.11	53.33	50.00	61.90	60.71	67.19
SpatialRGPT-VILA-1.5-8B (Cheng et al. 2024)	—	73.68	66.67	80.77	61.90	85.71	75.00
Senna-VLM (Jiang et al. 2024)	—	21.05	6.67	11.54	14.29	28.57	17.97
Cosmos-Reason1-7B (NVIDIA et al. 2025)	—	76.32	53.33	46.15	33.33	46.43	53.91
STRIDE-Qwen2.5-VL-7B	—	81.58	66.67	61.54	71.43	60.71	69.53
STRIDE-Cosmos-Reason1-7B	—	81.58	86.67	50.00	80.95	60.71	71.09

Model	Direct	Horizontal	Vertical	Width	Height	Direction
GPT-4o (OpenAI et al. 2024)	28.21	15.00	—	55.56	79.31	5.88
GPT-4o mini (OpenAI et al. 2024)	7.69	17.50	—	55.56	82.76	14.71
InternVL2.5-8B (Chen et al. 2024b)	10.26	15.00	—	44.44	37.93	2.94
Qwen2.5-VL-7B-Instruct (Bai et al. 2025)	33.33	32.50	—	11.11	34.48	2.94
SpatialRGPT-VILA-1.5-8B (Cheng et al. 2024)	20.51	27.50	—	44.44	82.76	70.59
Senna-VLM (Jiang et al. 2024)	2.56	0.00	—	0.00	0.00	0.00
Cosmos-Reason1-7B (NVIDIA et al. 2025)	12.82	17.50	—	33.33	72.41	26.47
STRIDE-Qwen2.5-VL-7B	20.51	12.50	—	66.67	82.76	32.35
STRIDE-Cosmos-Reason1-7B	12.82	5.00	—	66.67	79.31	17.65

Object-centric Spatial QA

Model	Below/Above	Left/Right	Big/Small	Tall/Short	Wide/Thin	Behind/Front	Avg.
GPT-4o (OpenAI et al. 2024)	—	44.55	36.36	37.21	31.48	42.22	39.23
GPT-4o mini (OpenAI et al. 2024)	—	50.91	52.27	60.47	38.89	57.78	52.51
InternVL2.5-8B (Chen et al. 2024b)	—	50.91	50.00	52.33	50.00	48.89	50.74
Qwen2.5-VL-7B-Instruct (Bai et al. 2025)	—	61.82	61.36	50.00	40.74	53.33	54.28
SpatialRGPT-VILA-1.5-8B (Cheng et al. 2024)	—	35.45	38.64	33.72	33.33	53.33	37.46
Senna-VLM (Jiang et al. 2024)	—	15.45	4.55	5.81	0.00	15.56	9.14
Cosmos-Reason1-7B (NVIDIA et al. 2025)	—	54.55	45.45	40.70	35.19	66.67	48.38
STRIDE-Qwen2.5-VL-7B	—	63.64	59.09	68.60	57.41	46.67	61.06
STRIDE-Cosmos-Reason1-7B	—	63.64	75.00	63.95	59.26	46.67	62.24

Model	Direct	Horizontal	Vertical	Width	Height	Direction
GPT-4o (OpenAI et al. 2024)	20.00	16.67	13.64	55.00	63.64	—
GPT-4o mini (OpenAI et al. 2024)	13.33	16.67	4.55	70.00	59.09	—
InternVL2.5-8B (Chen et al. 2024b)	13.33	3.33	4.55	40.00	36.36	—
Qwen2.5-VL-7B-Instruct (Bai et al. 2025)	6.67	3.33	0.00	5.00	18.18	—
SpatialRGPT-VILA-1.5-8B (Cheng et al. 2024)	26.67	30.00	0.00	70.00	77.27	—
Senna-VLM (Jiang et al. 2024)	0.00	10.00	9.09	0.00	0.00	—
Cosmos-Reason1-7B (NVIDIA et al. 2025)	20.00	3.33	4.55	15.00	31.82	—
STRIDE-Qwen2.5-VL-7B	40.00	53.33	40.91	75.00	95.45	—
STRIDE-Cosmos-Reason1-7B	33.33	56.67	36.36	75.00	86.36	—

Ego-centric Spatial QA

Model	Below/Above	Left/Right	Big/Small	Tall/Short	Wide/Thin	Behind/Front	Avg.
GPT-4o (OpenAI et al. 2024)	—	36.84	43.55	28.81	50.00	—	37.37
GPT-4o mini (OpenAI et al. 2024)	—	45.61	30.65	39.83	51.92	—	41.18
InternVL2.5-8B (Chen et al. 2024b)	—	21.05	41.94	32.20	46.15	—	34.60
Qwen2.5-VL-7B-Instruct (Bai et al. 2025)	—	49.12	30.65	45.76	44.23	—	42.91
SpatialRGPT-VILA-1.5-8B (Cheng et al. 2024)	—	45.61	1.61	18.64	32.69	—	22.84
Senna-VLM (Jiang et al. 2024)	—	12.28	3.23	5.08	3.85	—	5.88
Cosmos-Reason1-7B (NVIDIA et al. 2025)	—	29.82	30.65	29.66	34.62	—	30.80
STRIDE-Qwen2.5-VL-7B	—	80.70	82.26	74.58	76.92	—	77.85
STRIDE-Cosmos-Reason1-7B	—	77.19	80.65	82.20	76.92	—	79.93

Model	Direct	Horizontal	Vertical	Width	Height	Direction
GPT-4o (OpenAI et al. 2024)	25.71	3.33	13.33	68.97	33.33	—
GPT-4o mini (OpenAI et al. 2024)	5.71	33.33	3.33	86.21	37.50	—
InternVL2.5-8B (Chen et al. 2024b)	17.14	20.00	0.00	89.66	37.50	—
Qwen2.5-VL-7B-Instruct (Bai et al. 2025)	20.00	0.00	0.00	17.24	20.83	—
SpatialRGPT-VILA-1.5-8B (Cheng et al. 2024)	0.00	13.33	0.00	68.97	0.00	—
Senna-VLM (Jiang et al. 2024)	0.00	3.33	0.00	0.00	8.33	—
Cosmos-Reason1-7B (NVIDIA et al. 2025)	14.29	0.00	6.67	75.86	20.83	—
STRIDE-Qwen2.5-VL-7B	42.86	56.67	63.33	100	100	—
STRIDE-Cosmos-Reason1-7B	51.43	40.00	63.33	100	100	—

Table 10: SpatialRGPT-Bench 结果。从上到下，表格呈现了原始的、以对象为中心的问答和以自我为中心的问答。所有问答类别都报告成功率 (↑) 。