arXiv:2602.02864v1 [cs.RO] 2 Feb 2026

加速自主车辆中的结构化思维链

Yi Gu[*1] **Yan Wang**[1] **Yuxiao Chen**[1] **Yurong You**[1] **WenjieLuo**[1] **Yue Wang**[1,2]
Wenhao Ding[1] **Boyi Li**[1] **Heng Yang**[1,3] **Boris Ivanovic**[1] **Marco Pavone**[1,4]1NVIDIA [2]南加州大学 [3]哈佛大学 [4]斯坦福大学

摘要

思维链（CoT）推理增强了视觉-语言-动作模型在自动驾驶中的决策能力，但其自回归特性会导致显著的推理延迟，使其难以应用于实时场景。为解决这一问题，我们提出了 FastDriveCoT，这是一种加速模板结构化思维链的新型并行解码方法。我们的方法将推理过程分解为由不同子任务构成的依赖图，例如识别关键物体和总结交通规则，其中部分子任务可以并行生成。通过在单次前向传播中同时生成多个相互独立的推理步骤，我们大幅减少了顺序计算次数。实验表明，该方法可使思维链生成速度提升 3-4× 倍，并显著降低各种模型架构的端到端延迟，同时保留融入思维链推理所带来的原有下游任务性能提升。

1 引言

近年来，语言已成为机器人系统中的关键模态，推动了操作（Driess et al., 2023; Zitkovich et al., 2023）和自动驾驶（Huang et al., 2024; Tian et al., 2024b）等领域的发展。随着大语言模型（LLM）和视觉-语言模型（VLM）的快速进步，语言日益融入感知与决策流程，推动视觉-动作（VA）模型转变为视觉-语言-动作（VLA）模型（Black et al., 2024; Kim et al., 2024）。与传统的 VA 模型相比，VLA 得益于语言落地，增强了理解用户意图、分解任务并运用常识推理的能力，从而表现得更像人类。

这一方向的代表性技术是思维链（CoT）提示（Wei 等人，2022年），它以牺牲部分推理效率为代价，换取更高的推理准确率。通过鼓励 VLA 将复杂问题拆解为一系列更简单的子问题，思维链利用推理时扩展来提升策略性能。除提示之外，思维链也被纳入训练流程，例如通过监督微调（SFT）中的精选思维链数据集（Cui et al., 2025; Tian et al., 2024a）。自 DeepSeek v3（Liu et al., 2024）发布以来，借助扩展的结构化 思维链记录进行推理已在机器人学界获得广泛关注，并自然延伸至自主车辆（AV）（Zhao et al., 2025）。

自动驾驶对推理速度有着严格要求，这一点不同于纯语言任务或操作任务。在典型的自动驾驶策略中，必须以较高频率（通常为 10 Hz 或更高）更新决策，才能安全地应对快速变化的环境；这对每个规划周期内可生成的词元数量施加了严格限制。然而，标准的思维链轨迹通常包含多个阶段（例如环境描述、关键物体识别和元动作预测），累计会额外产生数百个词元，并带来显著的推理开销，因此难以应用于自动驾驶。

尽管自回归解码的序列特性使思维链推理成为推理过程中的瓶颈，但自主车辆中的推理包含多个在很大程度上相互独立的组成部分，因此

[*]在 NVIDIA 实习期间完成的工作。

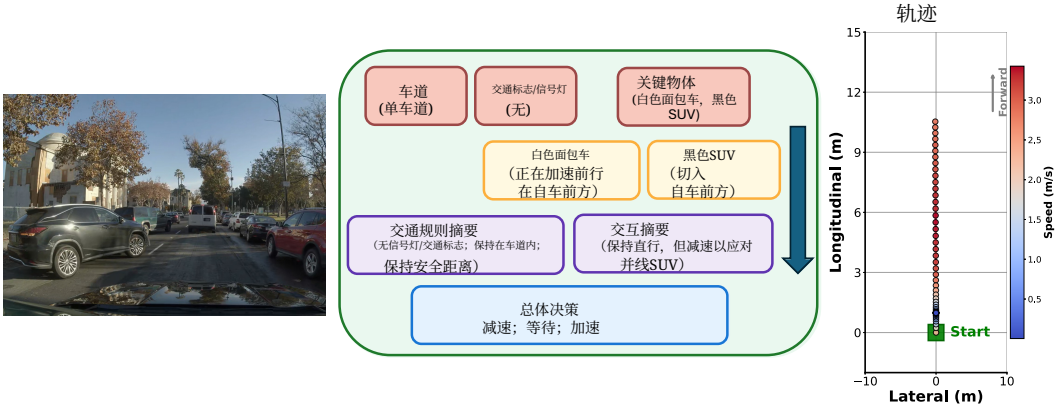


图 1：引入思维链模板有助于提升自主车辆中元动作预测和轨迹生成的准确性。通过包含特定主题依赖关系的模板进一步结构化思维链，可使思维链解码实现并行化。

适合进行并行化。与人类驾驶员类似，自动驾驶智能体可以并行评估道路状况、交通标志和关键物体等环境因素。自动驾驶场景的这一特性使其尤其适合高度并行化；相比之下，近期研究（Jin等人，2025年；Yang 等人，2025年b；Pan等人，2025年）聚焦于通用推理任务（例如数学），其中可分解的独立子任务数量通常有限，且因问题而异。此外，自动驾驶智能体通常遵循基于一组固定观测因素的结构化推理模式，以确保安全驾驶。这种规律性消除了其他领域所需的临时任务分解，并使得采用更精密的算法来协调思维链记录的并行生成成为可能。

贡献。 基于这一思路，我们提出了 **FastDriveCoT**，一种用于加速自动驾驶任务中思维链推断的方法。我们提出了一种将思维链记录格式化为结构化模板的系统方法。为了最大限度地提高并行性，我们采用一种最优算法，基于通用依赖图动态识别哪些字段可以并发生成。此外，我们通过并行解码策略优化大语言模型推理过程，以实现最高效率，尤其关注字段的排列与合并方式。实验表明，FastDriveCoT 能显著加速思维链生成；与自回归解码相比，其加速比达到 3.1 至 4.1×，同时保留了融入思维链推理所带来的下游任务改进。我们进一步验证，这些发现适用于多种设置，涵盖自回归和 Transfusion（Zhou 等人，2025a）框架下的元动作预测与轨迹预测任务。

2 相关研究

2.1 面向自动驾驶车辆的视觉-语言-动作模型

大语言模型（LLMs）具备广泛的世界知识和强大的上下文学习能力，因此有望应对自动驾驶中的长尾场景，而传统的感知–规划流程往往难以处理这些场景。近期研究通过视觉-语言-动作模型将这些能力引入规划环节。DriveVLM（Tian 等人，2024年b）提出了一种以视觉语言模型为中心的架构，用于场景描述、分析和分层规划。为降低计算延迟并弥补空间推理方面的不足，该方法采用双系统设计：由运行较慢的视觉语言模型执行高层规划，同时由快速专家策略负责实时控制，并在 nuScenes 基准和车载测试中取得了改进。与此同时，端到端视觉-语言-动作方法将多模态输入直接映射为动作。（Xu 等人，2024年；Renz 等人，2024年）采用接收二维视频特征的视觉-语言模型，并通过联合指令微调，使其能够生成以场景为中心的语言输出并进行动作预测。（Zhou 等人，2025年b）在输入中加入三维线索，例如鸟瞰视图（BEV）特征和结构化感知向量，从而提供更丰富的空间上下文，提升三维理解和动作预测能力。然而，这些方法尚未充分利用大语言模型的语言能力进行推理时推理，以提升动作预测性能。

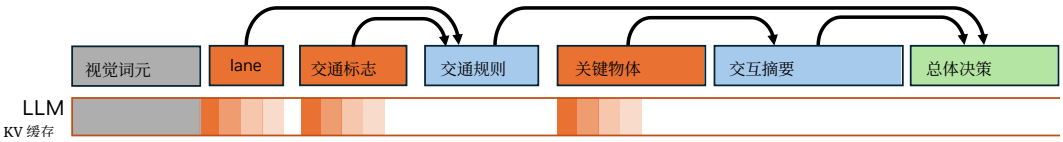


图 2：视觉-语言-动作（VLA）中的结构化思维链（思维链）并行解码。思维链被分解为一个包含若干字段的模板，这些字段之间具有特定的依赖关系，并以边表示。所有依赖关系均已满足的独立字段会同时进行解码，例如图中所示的车道、交通标志和关键物体。大语言模型中的背景阴影表示 KV 缓存的更新。

2.2 自动驾驶车辆中的思维链 CLES

思维链提示（Wei 等人，2022年）通过引出中间推理过程来提升多步推理能力；近期的推理优先系统（OpenAI，2024年；Comanici 等人，2025年；Guo 等人，2025年）表明，允许模型花费更多时间思考能够稳定地提高准确率。在自动驾驶领域，思维链已被整合进视觉-语言-动作流程中。（Wang 等人，2024年；Li 等人，2024年）发布了包含明确推理轨迹的端到端数据集。（Renz 等人，2025年）在预测动作之前，先生成关于“该做什么以及为什么”的评论，并在 CARLA 排行榜 2.0（CARLA 模拟器团队，2023年）和 Bench2Drive（Jia 等人，2024年）上取得了最先进的性能。AutoVLA（Zhou 等人，2025年c）通过 GRPO 微调引入快慢推理机制，仅在必要时调用长链推理。AutoDrive-R^2（Yuan 等人，2025年）将推理与轨迹对齐，进一步提升动作预测性能。尽管取得了这些进展，推理延迟仍然是一个瓶颈。我们的方法利用并行解码来加速思维链推理，而不牺牲动作预测性能。

2.3 大语言模型中的任务分解与并行解码

大语言模型推理领域的多项研究与我们的工作相关。其中一个重要方向是任务分解，即通过由少到多提示（Zhou 等，2022）、思维树（ToT）（Yao 等，2023）和经由规划的推理（RAP）（Hao 等，2023）等方法，将复杂问题分解为更简单的子任务。类似地，自一致性（Wang 等人，2022）和 Best-of-N（Brown 等人，2024）等并行解码技术会同时生成多条推理轨迹，并通过投票提高最终答案的质量。然而，这两个研究方向主要关注提升任务性能，通常会以增加延迟为代价。

相比之下，另一些研究则优先考虑效率。其中一类研究通过推测解码（Leviathan等人，2023年）和前瞻解码（Fu等人，2024年）等方法，在词元层面加速解码，不过其加速比从根本上受限于推测预测的准确性。与我们的方法更为密切相关的是近期关于独立子任务并行的研究，包括Hogwild（Rodionov等人，2025年）、Pasta（Jin等人，2025年）、APR（Pan等人，2025年）和 Multiverse（Yang 等人，2025年b）等方法。尽管这些方法旨在提升效率，但其加速比往往受限于通用推理任务中可并行化分支数量有限，以及必须先分析并分解问题所带来的计算开销。

3 方法

在现代GPU上，大语言模型的标准自回归生成受内存带宽限制（这是由KV缓存操作导致的），因此每次前向传播只处理一个词元，无法充分利用硬件的计算能力（Kwon等人，2023年；Dao等人，2022年；Dao，2023）。GPU利用率不足为并行解码创造了机会；这种技术可以同时生成多个词元，从而提高效率。

尽管先前的研究已探索通过子任务分解实现并行解码（Yang 等人，2025年b；Jin等人，2025年；Rodionov 等人，2025年），但其关注重点大多是数学和编程等通用推理领域。这些领域通常需要复杂且因任务而异的分解策略，且其并行性本身有限，因此加速比也有限。相比之下，自动驾驶任务通常遵循标准化的思维链模式，从

环境描述逐步推进到关键物体识别，最后进行元动作预测。这种结构化推理过程天然适合进行细致的子任务分解，从而实现高度并行的思维链生成。

为提升自动驾驶任务的推理效率，我们提出 FastDriveCoT，这是一种利用上述结构化推理过程的并行解码方法。我们首先介绍一个针对自动驾驶任务上述标准化思维链模式设计的思维链模板（第 3.1 节）。为了管理并最大限度地提升并行性，我们构建了一个依赖图和动态规划算法，使其能够适应不同字段长度（第 3.2 节）。最后，我们详细介绍对注意力掩码、位置 ID、填充和 KV缓存的处理设计。该实现充分考虑了大语言模型推理中的关键瓶颈，以确保最佳性能（第 3.3 节）。图 2 展示了 FastDriveCoT 的概览。

3.1 思维链模板

我们的思维链模板将自动驾驶推理任务分解为一系列特定字段。初始字段用于捕捉驾驶环境及其中的关键实体，包括光照、道路状况、天气、路口类型、道路类型、车道、关键物体、交通信号灯、交通标志以及其他交通规则。其中，关键物体 字段尤其用于涵盖车辆、行人、骑行者、障碍物及其他任何与驾驶安全相关的物体。对于每个关键物体，我们还会纳入其相对位置、对象类型和依据。接下来的一组字段对场景进行结构化概述，涵盖交通规则以及非交互式和交互式元素。最后，模板以对预期自车行为的整体总结收尾。每个字段均可用长度不一的自由形式自然语言填写。这里展示的思维链模板仅为示例；实际应用中，可根据具体应用和系统配置对这类模板进行调整。

某些字段（例如车道和关键物体）包含的实例数量可能不固定。例如，车辆行驶过程中车道配置会发生变化，而关键物体的数量也因情况而异。若采用简单的方法，在单个自由格式字段中描述所有实例，就会导致生成过程缓慢且按顺序进行。为实现并行性，我们针对这些多实例字段引入了一个两阶段流程。

- 阶段 1（枚举）：模型首先生成一个高层概述。对于车道，这涉及识别待分析的不同时间范围。对于关键物体，则是枚举需要描述的每个独立物体。
- 阶段 2（细化）：随后，模型会详细描述第一阶段中识别出的每个实例。

这种结构使多个车道时间范围或多个关键物体的详细描述能够并行生成，从而显著提高推理效率。为简化实现，我们为这些多实例字段定义了固定数量的槽位。我们的模板为车道分配了 3 个时间范围的空间，为关键物体分配了 4 个槽位，分别对应训练数据中观察到的各类别最大数量。如果实际实例数量少于分配的数量，则剩余槽位将填入“N/A”占位符。更具适应性的方法是根据第一阶段的枚举结果动态调整模板，我们将此留作未来的研究方向。

字段采用每行一个条目的格式，结构为“字段名称：字段内容”。随后将这些行串联起来，形成完整的思维链文本，通常总长度为 300 到 500 个词元。

3.2 依赖图

我们的思维链模板中的字段存在多种依赖关系。有些字段彼此独立，例如天气 和道路状况，因此可以并行解码。然而，其他字段之间存在依赖关系，要求遵循特定的生成顺序。例如，只有在交通标志 和交通信号灯 字段完成后，才能生成交通规则摘要 。同样，对于车道和关键物体等多实例字段，必须先进行枚举 ，再进行

详细阐述 。在推理过程中，必须严格遵循这一生成顺序。

管理这些依赖关系需要一种能够适应各种关系结构的通用方法。长度可变性使这一挑战更加复杂：模板中的各个字段长度彼此不同，且同一字段在不同情形下的长度也可能不同。这种不可预测性使得固定的解码调度变得不切实际。为应对这些挑战并优化性能，我们引入

算法 1使用依赖图进行并行思维链解码

要求： 依赖图 $\mathcal{G}$

$d_v \leftarrow$ 入边的数量 v，$\forall v \in V(\mathcal{G})$
$S \leftarrow \{v | d_v = 0\}$
当 $S \neq \varnothing$ **时**
  并行地在每个 $v \in S$ 中解码一个新词元
  对 $v \in S$ **执行**
    如果对应于 v 的字段已完成**则**
      从 S中移除 v
      对于 u：$(v, u) \in E(\mathcal{G})$ **执行**
        $d_u \leftarrow d_u - 1$
        如果 $d_u = 0$**则**
          将 u 添加到 S
        结束如果
      结束循环
    结束如果
  结束循环
结束当

依赖**图**。这种数据结构使我们的系统能够动态跟踪生成过程，并在任意步骤识别哪些字段已满足其先决条件，可以进行并行解码。

依赖图是一个有向无环图（DAG），其中每个节点代表模板中的一个字段。从字段 A 对应的节点指向字段 B 对应节点的有向边，表示字段 B 的生成直接依赖于字段 A先前完成。图 3 展示了一个简化依赖图的示例。依赖图的设计直观且灵活。它只需基于简单的直接依赖关系构建，并且可以迁移到任意可分解推理任务。

我们使用动态规划（DP）算法，根据依赖图全自动地 安排并行解码。首先，执行初始化步骤：将所有源节点（即没有任何入边的节点）加入“就绪”字段集合。随后，算法迭代执行，直到所有字段都完全生成。在每次迭代中：

- 对于就绪集合中的每个字段，在大语言模型的一次前向传播中并行生成一个词元。
- 当一个字段生成完成后，其节点会向所有依赖节点发出信号。
- 节点一旦收到其所有前置节点的信号，就会被加入就绪集合。 te 节点。

该算法的伪代码见算法 1。

我们的调度算法在前向传播次数方面是最优的，因为它使用可能的最少次数完成生成。该最小次数等于关键路径的长度，关键路径定义为图中任意依赖链上的词元累积数量的最大值。正如我们将在实验中进一步展示的那样，这种对前向传播次数的最小化所实现的最优性，会直接转化为调度算法加速比的最大化。

3.3 语言模型推理

并行生成多个字段，需要一种高效的方法来管理和合并这些字段的输出。许多以往将推理分解为子任务的研究，只是独立生成各个部分，然后将生成的自然语言文本拼接起来（Zhou 等，2022；Yao 等，2023；Hao 等，2023）。然而，这种方法的计算效率较低，因为在并行生成不同字段时，KV 缓存无法共享或重复利用，从而导致冗余的内存操作和计算。

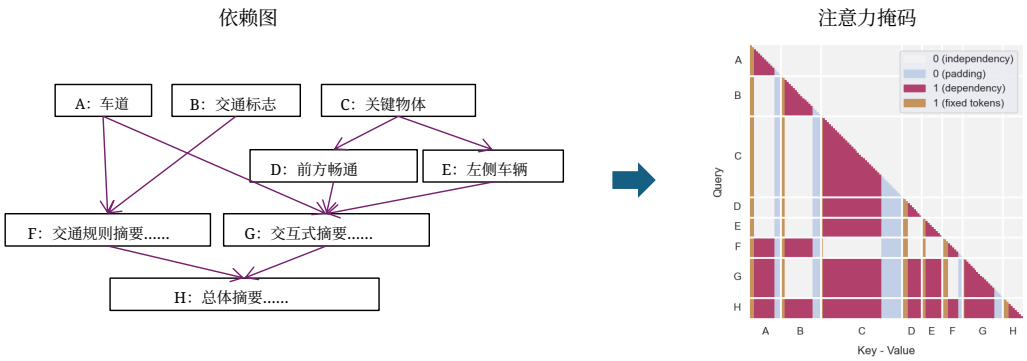


图 3：依赖图的一部分示例及其对应的注意力掩码。由于 彼此独立，因此可以并行处理，例[illegible]和 E 之间不能相互注意。其他存在 依赖关系的字段对，可以像往常一样让后者注意前者。具体而言，固定词元会预先填充，因此无论依赖关系如何，所有后续词元都可以注意到它们。 其[illegible]无法注意到填充词元。

为克服这一低效问题，我们提出将整个模板 CoT 格式化为一个连续序列。在并行解码多个字段的词元时，我们沿序列长度维度将它们全部打包在一起。这种方法使模型能够在保持原始批量大小的同时，通过一次前向传播同步计算所有并行字段的下一令牌逻辑值。因此，KV 缓存可以在所有字段之间充分共享和重复利用，从而最大限度地提高计算效率。

为了在单个序列中实现这种并行解码策略，我们设计了一种自定义注意力掩码，以强制遵循正确的因果依赖。该掩码会阻止为某个字段生成的词元注意到其他字段中尚未保证完整的词元，并且可以根据依赖图自动 计算得出。形式上，只有当 A是 B的祖先时，字段 B 中的词元才可以注意到字段 A 中的词元，即存在一条路径

$$X_1 \to X_2 \to \cdots \to X_n; \quad X_1 = A \wedge X_n = B \wedge \forall i : (X_i, X_{i+1}) \in E(\mathcal{G}) \tag{1}$$

在依赖图 $\mathcal{G}$ 中。具体而言，初始模板中的固定令牌对序列中后续的所有词元都可见，因为这些令牌从一开始就已知。图 3 展示了此类注意力掩码的示例。由于生成的字段长度可变，我们会将其输出填充或截断至预先设定的固定长度。这样可以确保正确确定序列中所有后续词元的位置编码。每个字段的最大长度都经过选择，以容纳至少 99% 训练数据中的输出。

需要注意的是，朴素的实现也会处理填充令牌，这会造成计算资源的浪费。为避免这种额外开销，我们修改注意力掩码，使所有填充令牌对序列中的其他词元不可见，包括为后续元动作和轨迹生成的词元。这种策略使填充能够预留空间并确定词元位置，同时无需在推理前向传播期间处理填充令牌本身。总体而言，FastDriveCoT 在效率和实现方面具有以下几个主要优势：

- **零计算开销：**与标准自回归解码相比，我们的方法不会增加任何浮点运算次数。通过在单个序列中管理所有生成过程，自定义注意力掩码可确保在所有前向传播中完全复用 KV 缓存，无需重新计算已有的键值对。
- **降低内存瓶颈：**沿序列长度维度打包词元，使多个查询词元（每个并行字段对应一个）能够在单个 CUDA注意力内核中访问共享的 KV 缓存。该策略显著减少了 GPU 内存访问总量，而这正是现代 GPU 上大语言模型推理的主要瓶颈（Zhong等人（2024））。
- **实现简洁性：**该方法易于实现。它通过对标准 Transformer 组件——即注意力掩码和位置 ID——进行细致而精确的设计来获得效率提升，而无需复杂的架构修改。

表1：主要结果总结

基础模型	CoT	元动作（交并比）↑[1]	轨迹（平均位移误差 @ 3秒）↓	轨迹（6.4 秒时的平均位移误差）↓	思维链时间（秒）↓	总时间（秒）↓
Qwen2 0.5B + VLA-AR	None	0.784	0.798	3.761	-	1.724
	自回归	**0.811**	0.753	2.948	9.191	11.305
	FastDriveCoT	0.804	**0.666**	**2.911**	2.239（4.1 倍）[2]	4.723（2.4 倍）
Qwen3 1.7B + VLA-AR	None	0.799	0.717	3.470	-	2.569
	自回归	0.801	0.686	2.863	13.186	16.033
	FastDriveCoT	**0.815**	**0.639**	**2.686**	4.189（3.1倍）	8.246（1.9倍）
Qwen2.5-VL 3B + VLA-AR	None	**0.865**	0.617	1.970	-	5.107
	自回归	0.850	0.511	2.045	14.866	18.244
	FastDriveCoT	0.856	**0.482**	**1.908**	4.608 (3.2x)	8.415 (2.2x)
Qwen2 0.5B + Transfusion	None	-[3]	0.839	4.023	-	0.330
	自回归	-	**0.564**	**2.225**	7.977	8.256
	FastDriveCoT	-	0.720	2.783	2.428 (3.6x)	2.688 (3.1x)

[1] ↑ 表示数值越高越好；↓ 表示数值越低越好。[2] 括号中的数字表示相较于自回归思维链的相对加速比。[3] 使用 Transfusion 时，我们不会将元动作纳入序列。

4 实验

4.1 实验设置

模型。 我们在三种架构与规模各异的基础模型上评估 FastDriveCoT，分别为 Qwen2-0.5B（Qwen 团队，2024年）、Qwen3-1.7B（Yang 等人，2025a）和 Qwen2.5-VL-3B（Bai 等，2025）。对于本身不接受视觉输入的 Qwen2-0.5B 和 Qwen3-1.7B，我们使用 DINOv2（Oquab 等人，2023年）从输入帧中提取特征，并将其作为连续输入提供给语言模型。我们还将 1.6 秒的轨迹历史编码为一个嵌入（通过一个小型多层感知机），作为大语言模型的额外输入。在思维链输出之后，我们还会生成元动作和未来轨迹，以评估模型的驾驶性能。我们探索了两种架构：（1）纯自回归变换器（VLA-AR），将思维链、元动作和未来轨迹词元化为离散序列，使用下一词元预测损失进行训练，并以自回归方式解码；（2）Transfusion（Zhou 等人，2025a），其中思维链和元动作的处理方式与（1）相同，但未来轨迹通过流匹配建模（Lipman 等人，2022），同时共享同一变换器骨干网络（如 NVIDIA 等人（2025）所述）。

为表示（1）中的轨迹，我们采用自编码预训练分词器，将每条 6.4 秒的未来轨迹压缩为 6 个离散词元。对于（2）中的轨迹表示，我们首先将未来轨迹转换为 64 个（Δx，Δy，Δ偏航角）10 Hz 动作。然后使用正弦位置编码对这些动作进行嵌入，再经由多层感知机处理，生成 64 个连续嵌入。最后，使用轻量级多层感知机将嵌入解码回（Δx，Δy，Δ偏航角）动作空间。

为评估我们方法的有效性，我们将其与两种基线方法进行比较：

1. **无思维链：** 一种以端到端方式训练、不进行中间思维链生成的模型。此基线用于衡量引入思维链对整体性能的贡献。2. **自回归思维链：** 一种采用完整思维链模板、但使用标准自回归解码按顺序生成各字段的模型。此基线用于单独评估并量化并行解码方法带来的效率提升。

数据。 为了评估 FastDriveCoT 的效率和任务性能，我们利用了一个大型内部数据集，其中包含来自 2500+ 个城市和 25 个国家[1]的多辆自车的 20,000 小时驾驶数据。该数据集涵盖各种道路和天气条件、昼夜时段以及不同程度的交通流量。我们使用轨迹数据，以及来自前视和后视摄像头的同步记录，并将其下采样至 320×512 的分辨率。我们采用了一个自动标注流水线

[1]此数据集的一个子集已发布在 https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles

其中包含 Qwen2.5-VL-72B（Bai 等，2025），用于为我们的实验生成结构化思维链数据。对于每个数据点，我们随机采样一个时间戳，并将相应的 2Hz 前视视频和轨迹历史作为输入提供给模型。随后，我们提示模型生成思维链数据，并使用受限解码来确保输出严格遵循预定义的模板。该流水线生成了 717,344 个高质量训练样本和 950 个测试样本。

评估指标。为验证 FastDriveCoT 在保持高任务性能的同时提升计算效率的能力，我们采用以下评估指标：

- **思维链时间：**从模型接收输入到思维链生成完毕所需的延迟。该指标直接衡量 FastDriveCoT 所针对的核心推理组件的加速比。
- **总时间：**从模型输入到最终轨迹输出的总延迟。该指标衡量实际应用中的端到端效率提升。
- **元动作交并比：**模型预测 6.4 秒时间范围内的一系列元动作。我们通过计算每个 0.1 秒时间间隔内预测动作与真实动作的交并比（IOU），评估这些预测。该指标评估模型的场景理解和高层决策质量。
- **轨迹平均位移误差：**我们使用预测轨迹与真实轨迹之间的平均位移误差（ADE），衡量最终运动规划的准确性。

训练与推理配置。 我们从预训练检查点初始化模型，并根据需要添加随机初始化的嵌入和层。随后，模型在包含 717,344 个样本的数据集上训练 50,000 步。我们使用批量大小为 64 的配置，采用 AdamW 优化器（Loshchilov 与 Hutter，2017），学习率为 3×10^{-5}，并使用余弦衰减调度。

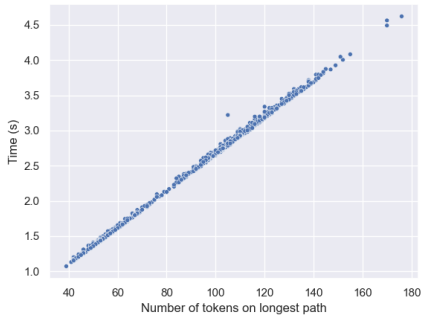
所有推理实验均在单台 NVIDIA A100 80GB SXM GPU 上进行。大多数计算采用 BFloat16 精度，但处理输入轨迹历史和最终输出 logits 时使用 Float32。我们的实现使用 PyTorch 的缩放点积注意力（SPDA）；对于标准因果掩码（基线方法中使用），它会调用 FlashAttention-2（Dao，2023）；对于并行解码方法所需的自定义掩码，则会调用 xFormers（Lefaudeux 等，2022）。为准确测量延迟，我们使用 CUDA 事件计时器，并舍弃每次运行的前 10 次迭代，以消除内核预热效应的影响。

4.2 主要结果

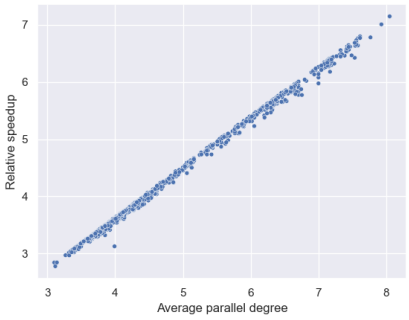
效率结果。 如表1所示，与标准自回归基线相比，FastDriveCoT 在思维链生成时间上实现了 $3.1 - 4.1\times$ 加速比。根据后续元动作和轨迹生成所占用的时间比例，整体推理时间的端到端加速比可达 1.9 –$3.1\times$ 。

这一显著加速降低了自回归推理原本难以接受的延迟，使思维链更适用于现实应用。此外，在所有测试的模型架构和规模上，这些性能提升都保持一致，表明 FastDriveCoT 广泛适用于各类视觉-语言模型和大语言模型。

任务性能。 表1还显示，引入思维链可显著改善元动作生成和轨迹生成。在 Qwen2.5-VL 3B 模型的 3 秒轨迹预测中，提升最为显著：平均位移误差从 0.617（无思维链基线）降至自回归思维链下的 0.511，并在采用我们的并行解码思维链后进一步降至 0.482。其他模型在较长期（6.4 秒）预测上的平均位移误差也有明显改善，例如 Qwen2 0.5B 和 Qwen3 1.7B，这证实了思维链在自动驾驶任务中的普遍有效性。
与自回归思维链基线相比，我们的并行解码方法展现出极具竞争力的性能。在视觉-语言-动作-自回归实验中，我们的并行方法表现略胜一筹；我们推测，这得益于结构化解码过程本身能够促使模型更好地遵循模板格式。在 Transfusion 实验中，尽管与自回归思维链相比，轨迹平均位移误差略有下降，FastDriveCoT 仍显著优于不使用思维链的基线。总而言之，FastDriveCoT 带来了



(a) 思维链生成时间和词元数量在最长路径上。



(b) 思维链生成的相对加速比和平均并行度。

图4：对测试集中每个数据样本生成时间的补充分析。

sign计算效率的显著提升，同时始终保持了实质性的 1
task 所有实验中思维链的准确率提升。

4.3 推理分析

为更好地理解 FastDriveCoT 加速比的来源，我们使用 Qwen2 0.5B +VLA-AR 配置，对推理过程进行了详细分析。

我们的分析聚焦于依赖图中的关键路径：必须按顺序生成的最长依赖词元链。如图4a所示，思维链生成时间与关键路径上的词元数量呈现显著的线性关系。这一结果证实，延迟的主要决定因素是顺序前向传播的次数，而非每次前向传播生成的词元数量，也非所有字段生成的词元总数。这一发现为未来研究带来了两个有希望的方向：

- 通过专门设计思维链模板来缩短关键依赖路径，可以进一步提升推理速度。
- 由于并行生成更多词元并不会增加延迟，FastDriveCoT 可以使智能体同时生成“快速响应”轨迹（几乎不含或完全不含思维链）和“全面思考”思维链，而无需额外耗时。双响应系统的应用将留待未来研究。

从图 4b 所示的已实现加速比与并行程度之间的关系中，我们还得到了另一项关键发现。此处，相对加速比是自回归基线生成思维链所需时间与 FastDriveCoT 所需时间的比值，而平均并行度是每次前向传播处理的词元平均数量。图中显示二者呈现显著的线性关系，证实加速比与并行程度成正比。我们还观察到，加速比始终仅略低于平均并行度。这表明，在这一规模下，FastDriveCoT 的加速比大致等于其平均并行度。我们推测，这一细微差异是由于效率较低的注意力内核无法充分支持 FastDriveCoT 所用的自定义注意力掩码。我们将这一进一步优化留待未来研究。

5 结论

本文提出 FastDriveCoT，这是一种利用并行解码加速自动驾驶任务中模板结构化思维链的方法。FastDriveCoT 采用直观的 且具有泛化能力的 依赖图，以及全自动的 最优动态规划算法，动态识别可并行生成的字段。实验表明，FastDriveCoT 在不同的视觉语言模型/大语言模型架构和规模下，思维链推断速度显著提升 3-4× ，同时

在提升元动作准确率和生成轨迹质量的同时，保持了下游任务性能。

参考文献

Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。Qwen2.5-VL 技术报告。*arXiv* 预印本 *arXiv:2502.13923*，2025年。

Kevin Black、Noah Brown、Danny Driess、Adnan Esmail、Michael Equi、Chelsea Finn、Niccolo Fusai、Lachy Groom、Karol Hausman、Brian Ichter 等。 π_0：一种用于通用机器人控制的视觉-语言-动作流模型。*arXiv* 预印本 *arXiv:2410.24164*，2024年。

Bradley Brown、Jordan Juravsky、Ryan Ehrlich、Ronald Clark、Quoc V Le、Christopher R´e 和 AzaliaMirhoseini。大型语言猴：通过重复采样扩展推理计算。*arXiv* 预印本 *arXiv:2407.21787*，2024年。

CARLA 模拟器团队。CARLA 自动驾驶排行榜 2.0。https://leaderboard. carla.org/，2023年。排行榜网站。

Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein、Ori Ram、Dan Zhang、Evan Rosen 等。Gemini 2.5：通过高级推理、多模态、长上下文和新一代智能体能力拓展前沿。*arXiv* 预印本 *arXiv:2507.06261*，2025年。

Yixin Cui、Haotian Lin、Shuo Yang、Yixiao Wang、Yanjun Huang 和 Hong Chen。面向自动驾驶的思维链：全面综述与未来展望。*arXiv* 预印本 *arXiv:2505.20223*，2025年。

Tri Dao。FlashAttention-2：通过更优的并行性和工作划分实现更快的注意力计算。*arXiv* 预印本 *arXiv:2307.08691*，2023年。

TriDao、DanFu、StefanoErmon、Atri Rudra 和 Christopher R´e。FlashAttention：具备 IO 感知的快速、节省内存的精确注意力机制。神经信息处理系统进展，35：16344–16359，2022年。

Danny Driess、Fei Xia、Mehdi S. M. Sajjadi、Corey Lynch、Aakanksha Chowdhery、Brian Ichter、Ayzaan Wahid、Jonathan Tompson、Quan Vuong、Tianhe Yu、Wenlong Huang、Yevgen Chebotar、Pierre Sermanet、Daniel Duckworth、Sergey Levine、Vincent Vanhoucke、Karol Hausman、Marc Toussaint、Klaus Greff 和 Andy Zeng。PaLM-E：一种具身多模态语言模型。*arXiv* 预印本，arXiv:2303.03378，2023年。URL https://arxiv.org/abs/2303.03378。预印本。

Yichao Fu、Peter Bailis、Ion Stoica 和 Hao Zhang。利用前瞻解码打破大语言模型推理的顺序依赖。*arXiv*预印本 *arXiv:2402.02057*，2024年。

Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等。Deepseek-r1：通过强化学习激励大语言模型的推理能力。*arXiv*预印本 *arXiv:2501.12948*，2025年。

Shibo Hao、Yi Gu、Haodi Ma、Joshua Jiahua Hong、Zhen Wang、Daisy Zhe Wang 和 Zhiting Hu。借助语言模型进行推理，就是利用世界模型进行规划。*arXiv*预印本 *arXiv:2305.14992*，2023年。

X. Huang 等。扩展用于驾驶的自回归行为模型（DriveGPT）。*arXiv*预印本，arXiv:2412.14415，2024年。网址：https://arxiv.org/abs/2412.14415。预印本。

Xiaosong Jia、Zhenjie Yang、Qifeng Li、Zhiyuan Zhang 和 Junchi Yan。《Bench2Drive：面向闭环端到端自动驾驶多能力基准测试》。神经信息处理系统进展，37：819–844，2024年。

Tian Jin、Ellie Y Cheng、Zack Ankner、Nikunj Saunshi、Blake M Elias、Amir Yazdanbakhsh、Jonathan Ragan-Kelley、Suvinay Subramanian 和 Michael Carbin。《学会信守承诺：利用学习式异步解码扩展语言模型解码并行化》。*arXiv* 预印本 *arXiv:2502.11517*，2025年。

Moo Jin Kim、Karl Pertsch、Siddharth Karamcheti、Ted Xiao、Ashwin Balakrishna、Suraj Nair、Rafael Rafailov、Ethan Foster、Grace Lam、Pannag Sanketi 等。OpenVLA：一种开源视觉-语言-动作模型。*arXiv* 预印本 *arXiv:2406.09246*，2024年。

Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph Gonzalez、Hao Zhang 和 Ion Stoica。使用 PagedAttention 高效管理大语言模型服务的内存。发表于第 *29* 届操作系统原理研讨会论文集，第 611–626 页，2023年。

Benjamin Lefaudeux、Francisco Massa、Diana Liskovich、Wenhan Xiong、Vittorio Caggiano、Sean Naren、Min Xu、Jieru Hu、Marta Tintore、Susan Zhang 等。xFormers：一个模块化且易于修改的 Transformer 建模库，2022年。

Yaniv Leviathan、Matan Kalman 和 Yossi Matias。通过推测解码实现 Transformer 的快速推理。载于国际机器学习会议，第 19274–19286 页。PMLR，2023年。

Yiheng Li、Cunxin Fan、Chongjian Ge、Zhihao Zhao、Chenran Li、Chenfeng Xu、Huaxiu Yao、Masayoshi Tomizuka、Bolei Zhou、Chen Tang 等。WOMD-Reasoning：用于驾驶交互推理的大规模数据集。*arXiv* 预印本 *arXiv:2407.04281*，2024年。

Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matt Le。用于生成建模的流匹配。*arXiv* 预印本 *arXiv:2210.02747*，2022年。

Aixin Liu、Bei Feng、Bing Xue、Bingxuan Wang、Bochao Wu、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan 等。DeepSeek-V3 技术报告。*arXiv* 预印本 *arXiv:2412.19437*，2024年。

Ilya Loshchilov和 FrankHutter。《解耦权重衰减正则化》。*arXiv* 预印本 *arXiv:1711.05101*，2017年。

NVIDIA、Yan Wang、Wenjie Luo、Junjie Bai、Yulong Cao、Tong Che、Ke Chen、Yuxiao Chen、Jenna Diamond、Yifan Ding、Wenhao Ding、Liang Feng、Greg Heinrich、Jack Huang、Peter Karkus、Boyi Li、Pinyi Li、Tsung-Yi Lin、Dongran Liu、Ming-Yu Liu、Langechuan Liu、Zhijian Liu、Jason Lu、Yunxiang Mao、Pavlo Molchanov、Lindsey Pavao、Zhenghao Peng、Mike Ranzinger、Ed Schmer- ling、Shida Shen、Yunfei Shi、Sarah Tariq、Ran Tian、Tilman Wekel、Xinshuo Weng、Tianjun Xiao、Eric Yang、Xiaodong Yang、Yurong You、Xiaohui Zeng、Wenyuan Zhang、Boris Ivanovic 和 Marco Pavone。《Alpamayo-R1：弥合推理与动作预测，助力长尾场景下的泛化自动驾驶》，2025年。网址：https://arxiv.org/abs/2511.00088。

OpenAI。使用大语言模型进行推理。OpenAI o1 模型系列的 OpenAI 系统卡，2024年。通过强化学习训练以执行复杂推理；引入了用于推理的“测试时计算”。

Maxime Oquab、Timoth´ee Darcet、Th´eo Moutakanni、Huy Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby 等。Dinov2：无需监督即可学习鲁棒视觉特征。*arXiv* 预印本 *arXiv:2304.07193*，2023年。

Jiayi Pan、Xiuyu Li、Long Lian、Charlie Snell、YifeiZhou、Adam Yala、TrevorDarrell、Kurt Keutzer 和 Alane Suhr。使用语言模型学习自适应并行推理。*arXiv* 预印本 *arXiv:2504.15466*，2025年。

Qwen 团队。Qwen2 技术报告。*arXiv* 预印本 *arXiv:2407.10671*，2024年。

Katrin Renz、Long Chen、Ana-Maria Marcu、Jan H¨unermann、Benoit Hanotte、Alice Karnsund、Jamie Shotton、Elahe Arani 和 Oleg Sinavski。CarLLava：用于仅依靠摄像头进行闭环驾驶的视觉语言模型。*arXiv* 预印本 *arXiv:2406.10165*，2024年。

Katrin Renz、Long Chen、Elahe Arani 和 Oleg Sinavski。Simlingo：借助语言-动作对齐实现的纯视觉闭环自主驾驶。载于计算机视觉与模式识别会议论文集，第 11993–12003 页，2025年。

Gleb Rodionov、Roman Garipov、Alina Shutova、George Yakushev、Erik Schultheis、Vage Egiazar- ian、Anton Sinitsin、Denis Kuznedelev 和 Dan Alistarh。Hogwild! 推理：通过并发注意力实现并行大语言模型生成。*arXiv* 预印本 *arXiv:2504.06261*，2025年。

Ran Tian、Boyi Li、Xinshuo Weng、Yuxiao Chen、Edward Schmerling、Yue Wang、Boris Ivanovic 和 Marco Pavone。将世界拆解为对象级知识，以应对自动驾驶中的长尾事件。*arXiv* 预印本 *arXiv:2407.00959*，2024a。

Xiaoyu Tian、Junru Gu、Bailin Li、Yicheng Liu、Yang Wang、Zhiyong Zhao、Kun Zhan、Peng Jia、Xianpeng Lang 和 Hang Zhao。DriveVLM：自动驾驶与大型视觉语言模型的融合。发表于机器人学习会议 *(CoRL)*，2024b。网址：https://arxiv.org/abs/2402.12289。arXiv 预印本 arXiv:2402.12289；亦于 CoRL 2024 发表。

Tianqi Wang、Enze Xie、Ruihang Chu、Zhenguo Li 和 Ping Luo。Drivecot：将思维链推理与端到端驾驶相结合。*arXiv* 预印本 *arXiv:2403.16996*，2024年。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdh- ery 和 Denny Zhou。自一致性提升语言模型中的思维链推理。*arXiv* 预印本 *arXiv:2203.11171*，2022年。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。思维链提示能够激发大语言模型的推理能力。神经信息处理系统进展，35：24824–24837，2022年。

Zhenhua Xu、Yujia Zhang、Enze Xie、Zhen Zhao、Yong Guo、Kwan-Yee K Wong、Zhenguo Li 和 Hengshuang Zhao。Drivegpt4：通过大语言模型实现可解释端到端自动驾驶。*IEEE*机器人与自动化快报，2024年。

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, ChangGao, ChengenHuang, Chenxu Lv 等。Qwen3 技术报告。*arXiv* 预印本 *arXiv:2505.09388*，2025a。

XinyuYang、YuweiAn、Hongyi Liu、陈天奇和Beidi Chen。Multiverse：你的语言模型会暗中决定如何并行化并合并生成。*arXiv* 预印本 *arXiv:2506.09991*，2025年b。

Shunyu Yao、Dian Yu、Jeffrey Zhao、Izhak Shafran、Tom Griffiths、Yuan Cao 和 Karthik Narasimhan。思维树：借助大语言模型进行审慎的问题求解。神经信息处理系统进展，36:11809–11822，2023年。

Zhenlong Yuan、Jing Tang、Jinguo Luo、Rui Chen、Chengxuan Qian、Lei Sun、Xiangxiang Chu、Yu- jun Cai、Dapeng Zhang 和 Shuo Li。AutoDrive-R^2：提升自动驾驶中 VLA模型的推理与自我反思能力。*arXiv* 预印本 *arXiv:2509.01944*，2025年。

Qingqing Zhao、Yao Lu、Moo Jin Kim、Zipeng Fu、Zhuoyang Zhang、Yecheng Wu、Zhaoshuo Li、Qianli Ma、韩松、Chelsea Finn 等。CoT-VLA：面向视觉-语言-动作模型的视觉思维链推理。载于计算机视觉与模式识别会议论文集，第 1702–1713 页，2025年。

Yinmin Zhong、Shengyu Liu、Junda Chen、Jianbo Hu、Yibo Zhu、Xuanzhe Liu、Xin Jin 和 Hao Zhang。 {DistServe}：通过分离预填充与解码，实现有效吞吐量优化的大语言模型服务。载于第 *18* 届 *USENIX*操作系统设计与实现研讨会（*OSDI 24*），第 193–210 页，2024年。

Chunting Zhou、LILI YU、Arun Babu、Kushal Tirumala、Michihiro Yasunaga、Leonid Shamis、Jacob Kahn、Xuezhe Ma、Luke Zettlemoyer和Omer Levy。Transfusion：用一个多模态模型预测下一个词元并扩散图像。载于第十三届学习表征国际会议，2025a。URL https://openreview.net/forum?id=SI2hI0frk6。

Denny Zhou、Nathanael Sch¨arli、Le Hou、Jason Wei、Nathan Scales、Xuezhi Wang、Dale Schuurmans、Claire Cui、Olivier Bousquet、Quoc Le等。由少到多提示使大语言模型能够进行复杂推理。*arXiv*预印本 *arXiv:2205.10625*，2022年。

Xingcheng Zhou、Xuyuan Han、Feng Yang、Yunpu Ma和Alois C Knoll。Opendrivevla：迈向基于大型视觉-语言-动作模型的端到端自动驾驶。*arXiv*预印本 *arXiv:2503.23463*，2025年b。

Zewei Zhou、Tianhui Cai、Seth Z Zhao、Yun Zhang、Zhiyu Huang、Bolei Zhou和Jiaqi Ma。Autovla：一种面向端到端自动驾驶的视觉-语言-动作模型，采用自适应推理与强化微调。*arXiv*预印本 *arXiv:2506.13757*，2025年c。

Brianna Zitkovich、Tianhe Yu、Sichun Xu、Peng Xu、Ted Xiao、Fei Xia、Jialin Wu、Paul Wohlhart、Stefan Welker、Ayzaan Wahid等。RT-2：视觉-语言-动作模型将网络知识迁移至机器人控制。载于机器人学习会议，第2165–2183页。PMLR，2023年。