基于自动驾驶场景的多模态视觉语言助手

一、项目概述

本项目旨在利用多模态大模型 Qwen2.5-VL-7B-Instruct 构建一个自动驾驶场景助手。模型通过对交通场景图像的分析理解，识别车辆、行人、信号灯等关键交通要素，并解释其对驾驶行为的影响。当用户提出交通状况相关问题，模型可以实时给出交通状况分析以及开车决策。本项目包含数据处理、LoRA 微调以及模型效果评估，并基于 Gradio 实现了可视化问答交互系统。完整代码保存在 GitHub 仓库：https://github.com/Dq-sss/Autonomous-Driving-Assistant。

二、项目具体流程

2.1 数据处理

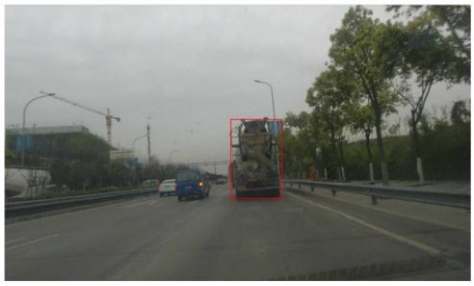
（1）数据集介绍

本实验选用专为自动驾驶场景设计的视觉-语言数据集 CODA-LM，该数据集包含 9,768 个真实驾驶场景，其中影响自动驾驶决策的关键道路实体被分为七大类：车辆（轿车、卡车、公交车等）、易受伤害的道路使用者（行人、自行车骑行者、摩托车骑行者）、交通标志（禁止停车、警告、指示等）、交通信号灯（红灯、绿灯、黄灯）、交通锥、障碍物，以及杂物（碎片、垃圾桶、动物等）。

CODA-LM 数据集共包含三个任务，分别是通用感知、区域感知和驾驶建议。通用感知任务的基础在于全面理解驾驶场景中的关键道路实体，包括其外观、位置及其影响本车驾驶行为的原因，如图 1 所示。区域感知任务用于训练模型在给定特定边界框时理解特定案例对象的能力，包括描述框内对象并解释其如何影响自动驾驶行为，如图 2 所示。驾驶建议任务旨在训练模型制定驾驶策略的能力，要求模型在准确感知当前驾驶环境的通用特征与区域特征后，为本车提供最优驾驶建议，如图 3 所示。


```
{
    "from": "human",
    "value": "<img>\n这是一张从自车视角捕捉的交通图像。重点关注影响自车驾驶行为的物体：车辆（轿车、卡车、公交车等）、易受伤害的道路使用者（行人、自行车骑行者、摩托车骑行者）、交通标志（禁止停车、警告、指示等）、交通信号灯（红灯、绿灯、黄灯）、交通锥、障碍物，以及杂物（碎片、垃圾桶、动物等）。请不要讨论以上七类以外的任何物体。请描述每个物体的外观、位置、方向，并解释为什么它会影响自车的行为。"
  },
  {
    "from": "assistant",
    "value": "在所提供的交通图像中，场景包括了几个影响驾驶行为的关键元素。\n\n首先，正前方有一辆白色的SUV，它位于右转车道上，表明它可能会右转。这个位置暗示着主车可能需要根据SUV的行动调整速度或准备停车。此外，在左车道上有一辆银色的汽车，位于主车的稍前方，看起来将直行。虽然这辆车的轨迹不太可能直接影响主车即时的驾驶决策，但应该密切注意以防万一它突然变道或刹车。\n\n路面上还有一个方向箭头指示右车道是用来右转的。这个标志告诉主车关于车道的使用情况，意味着如果主车在右车道上，它应该准备右转。此外，主车所在车道前方有一个红色交通灯，要求它停在停止线后或前车后，直到交通灯变绿。\n\n主车右侧有一个交通锥，位于路面的一个损坏区域旁边。这标示着道路上的一个障碍需要避免，可能暗示如果车道空间允许，轻微向左移动。"
  }
```


图 1 通用感知案例




```
{
  "from": "human",
  "value": "<img>\n请描述图像中红色矩形内的物体，并解释为什么它会影响自车的驾驶。"
},
{
  "from": "assistant",
  "value": "该物体是一种施工车辆，很可能是一辆水泥搅拌车，通过其用于混合水泥成分的旋转
鼓进行识别。这类车辆通常用于城市和道路建设场景。道路上这种车辆的出现表明附近可能有建设活动，可
能涉及车道封闭或交通模式改变。驾驶员应保持警惕，减速行驶，并准备应对潜在危险，如道路附近的碎片
或建筑工人。"
}
```


图 2　区域感知案例


```
{
  "from": "human",
  "value": "<img>\n这是一张从自车视角捕捉的交通图像。重点关注影响自车驾驶行为的物体：车辆（轿车、卡车、
公交车等）、易受伤害的道路使用者（行人、自行车骑行者、摩托车骑行者）、交通标志（禁止停车、警告、指示等）、交
通信号灯（红灯、绿灯、黄灯）、交通锥、障碍物，以及杂物（碎片、垃圾桶、动物等）。请不要讨论以上七类以外的任何
物体。请根据当前场景为自车提供驾驶建议。"
},
{
  "from": "assistant",
  "value": "本车应保持与前方载有起重机的卡车安全距离，随时准备减速或变道，以防运输车辆侵占车道导致不
安全。此外，司机应为可能的变道做准备，以依据即将出现的交通标志沿正确出口或直行。关注周围车辆，尤其是SUV和小型
巴士，对于在交通流中安全操控至关重要。随着车辆接近路口，监控任何可能出现的交通灯或标志也是必要的。"
}
```


图 3　驾驶建议案例

由于显存限制，本实验从三种任务中共选取 7044 条数据用于训练模型微调，选取 1592 条数据用于模型测试和评估，两组数据都涵盖三种任务。

（2）数据格式处理

CODA-LM 数据集以 Parquet 格式存储，因此使用 process_data.py 读取 Parquet 文件，将图片提取到 converted_images 目录，并生成对应的 json 图文问答数据。

为适配 Qwen2.5-VL-7B 模型的输入格式要求，将 json 问答数据转换为模型可解析的对话格式。首先从原始嵌套结构中提取纯文本对话内容；然后在指令文本起始处插入特定的视觉标记 <img>，以建立文本指令与关联图像之间的显式联系；最后将样本重组为模型所需的固定 JSON 格式，包含样本 ID、图像路径以及一个由‘human’和‘assistant’组成的对话列表，确保模型能够正确关联图像与文本对话，为后续的模型微调与评测提供了格式合规的输入数据。

2.2 模型微调

（1）整体架构图

模型的整体架构如图 4 所示，包含输入层、多模态对齐的编码器层、微调的解码器层以及文本输出层。

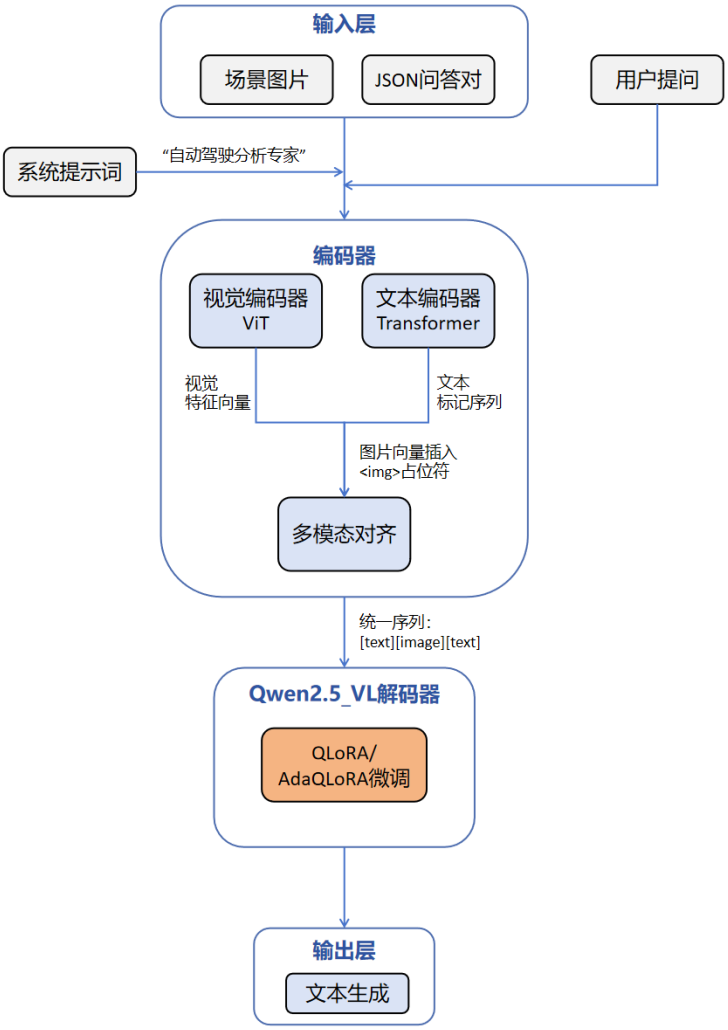


图 4 整体架构图

（2）多模态指令对齐

为实现基于 Qwen2.5-VL-7B 的自动驾驶场景分析微调，本项目设计了一个标准化的视觉语言对齐流程，具体操作流程如下：

首先引入一条系统提示词，赋予模型“自动驾驶决策与风险分析专家”角色，并约束其推理仅基于给定的图像与文本内容，如图 4 所示。这一先验语言信息能够使模型聚焦于当前场景，实现自动驾驶特定领域的对齐强化。随后，基于模型原生的多模态处理器对图像与文本进行联合编码。其中图像由基于 Vision Transformer 的视觉编码器进行处理，该编码器将图像分割为块序列，并通过多层自注意力机制提取出一系列高级的视觉语义特征向量；文本内容由基于 Transformer 的分词器转化为对应的文本标记序列。然后将视觉特征向量整体插入文本标记序列的占位符位置，生成视觉与文本特征交错排列的统一序列，为后续微调提供数据支撑。

```
# 定义系统提示词
self.system_prompt= (
    "你是一名专业的自动驾驶决策与风险分析专家。\n"
    "请严格基于给定的交通场景文字描述回答问题。\n"
    "约束：\n"
    "1. 只能基于描述中明确提到的交通要素进行分析。\n"
    "2. 不得引入描述中未出现的车辆、行人或信号设施。\n"
    "3. 所有决策必须有明确理由，且逻辑清晰。\n"
    "4. 语言专业、客观、简洁，避免泛泛而谈。"
)
```

图 4 系统提示词

（3）模型微调方法

1. 量化低秩微调（QLoRA）

为在有限的计算资源下对大规模视觉语言模型 Qwen2.5-VL-7B 进行领域适配，本研究采用量化低秩适配（Quantized Low-Rank Adaptation, QLoRA）方法进行参数高效微调，以在显存受限的硬件环境下实现高效训练。

QLoRA 的核心思想是在冻结原始大模型参数的前提下，仅对少量可训练的低秩矩阵进行优化，从而显著降低训练所需的参数存储和梯度计算量。同时，QLoRA 在此基础上结合 4 位量化技术，将原模型的权重以高保真方式压缩至低精度表示，使得大模型能够在单张 24GB 显存的 GPU 上进行训练，而不影响推理性能和最终任务表现。相比标准 LoRA，QLoRA 通过量化和低秩参数化的双重策略，使得微调过程的显存占用大幅下降，从而在有限资源条件下仍能保持模型的表达能力与收敛效率。

在具体实现中，本研究将模型的主体参数完全冻结，仅对关键投影矩阵 "q_proj"、"v_proj" 和 "down_proj" 应用低秩适配，同时结合梯度检查点与梯度累积策略，进一步降低单步计算对显存的压力。这些低秩矩阵在训练过程中能够有效捕捉领域特定的参数调整，使模型在保持原有多模态理解能力的同时，快速适应自动驾驶场景下的视觉与语言指令任务。通过这种设计，QLoRA 不仅实现了高效的训练资源利用，也确保了模型在复杂多模态输入条件下的稳定性和专业性输出，是在有限硬件条件下进行大规模视觉语言模型领域微调的一种可行且高效的方法。

2. 自适应量化低秩适配（AdaQLoRA）

为了在资源受限条件下进一步提升微调性能，本研究在 QLoRA 框架基础上引入自适应量化低秩适配方法（Adaptive Quantized Low-Rank Adaptation, AdaQLoRA）。该方法针对 QLoRA 中低秩参数分配策略静态化的问题进行改进，能够更有效地应对自动驾驶视觉问答等高复杂度多模态任务。

标准 QLoRA 通过量化技术极大降低了显存占用，使大模型微调变得可行，但其采用的 LoRA 模块为所有目标网络层分配了固定的秩，而忽略了不同层对下游任务的贡献度存在显著差异。未能充分考虑不同层在下游任务中的功能差异与语义贡献。在自动驾驶场景理解任务中，模型各层的作用并不均衡，静态的参数分配方式在一定程度上限制了参数利用效率。

AdaQLoRA 的核心改进在于引入了动态的参数重要性评估与分配机制。它将传统的低秩适配矩阵重新参数化为由奇异值分解定义的结构，从而可以独立评估和调整每个奇异值方向的重要性。在训练过程中，算法持续监控这些参数对整体训练损失的影响，并根据评分动态地执行预算再分配：在总参数量的约束下，它会周期性地裁剪掉重要性低的奇异值方向，同时将节约出的参数预算重新分配给那些被识别为更重要的方向。

也就是说，该方法在完全继承 QLoRA 显存优势的基础上，引入了面向任务需求的自适应参数配置能力，使模型能够将有限且关键的可训练参数集中用于对自动驾驶场景分析

最具贡献的特征表达方向。预期在相同硬件资源与参数规模约束下，AdaQLoRA 能够获得更快的收敛速度与更优的下游任务性能。

2.3 模型评估与打分

（1）模型评估指标

为客观评估微调后的 Qwen2.5-VL-7B 模型在自动驾驶视觉问答任务中的文本生成质量与语义一致性，本项目构建了完整评测流程，从实体覆盖、语义一致性以及整体文本合理性等多个层面进行综合分析。

在评测前，对模型生成的预测文本与人工标注的参考答案进行统一的文本预处理，包括无关符号清洗、格式规范化以及中文分词处理，以适配基于 n-gram 统计和语义相似度计算的自动评测指标。针对中文自然语言生成任务中分词粒度不一致、标点干扰等问题，上述预处理步骤能够有效降低格式差异对评测结果的影响，从而保证评估的公平性与稳定性。

在评价指标选择方面，综合考虑自动驾驶场景下文本生成任务对安全相关信息完整性、语义准确性以及决策一致性的要求，本项目采用 Entity Recall、BERTScore、BLEURT 和 ROUGE-L 四种互补指标进行评估。

1. BLEU：用于衡量模型生成文本中对关键交通实体的覆盖程度，重点评估模型是否正确识别并描述了影响本车驾驶行为的核心对象；
2. BERTScore：借助预训练语言模型对上下文语义进行建模，从语义相似度角度评估模型输出与参考答案之间的深层语义一致性；
3. BLEURT：通过学习人类主观评分信号，对生成文本的整体合理性、连贯性与语义匹配程度进行综合评估，用于衡量模型输出在整体质量层面与人工参考之间的一致程度；
4. ROUGE-L：基于最长公共子序列计算生成文本与参考答案在整体结构上的一致性，关注生成内容的连贯性与信息覆盖能力；

通过上述指标联合评估，评测结果不仅能够反映模型在词汇层面和句法层面的生成质量，还能够全面衡量其在语义理解、信息表达及领域适应能力方面的整体表现，为分析不同微调策略对自动驾驶视觉问答任务性能的影响提供了可靠依据。

（2）模型评测打分

为进一步评估模型在自动驾驶视觉问答任务中生成文本的语义合理性与决策相关性，本项目基于大语言模型的自动化文本主观评分机制，对模型生成的描述文本进行细粒度评分。具体流程如下：

首先，构建统一的评判提示词，通过系统指令明确设定评估模型的“评判者”角色，并规定三大评分标准：准确性、幻觉抑制和相关性。为提升评估模型对评分标准的理解与一致性，流程在正式评测前引入少量高质量与低质量示例作为上下文示范（few-shot prompting）。这些示例以“参考文本—预测文本—评分”的形式嵌入对话历史中，使模型能够在推理过程中隐式学习评分尺度与判别边界。

在实际评分阶段，采用智谱 AI 的 GLM-4-Flash 模型作为评估模型，每一对参考文本与预测文本被成对输入评估模型。模型在综合理解两段文本后，首先生成简要的判断依据说明，然后严格按照预定义格式输出 1 到 10 的整数评分，量化预测文本相对于参考答案的整体质量。

最终，系统自动解析 GLM-4-Flash 模型的文本响应，统计各样本评分并计算平均值，以得到模型在自动驾驶视觉问答任务中的综合文本质量评价指标。

（3）消融实验

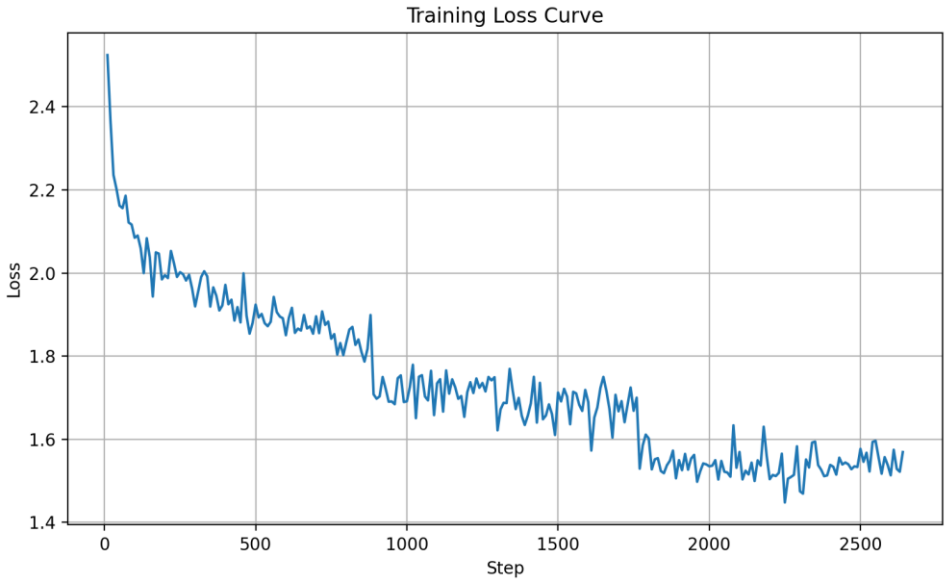


图 5　QloRA 微调 loss 曲线

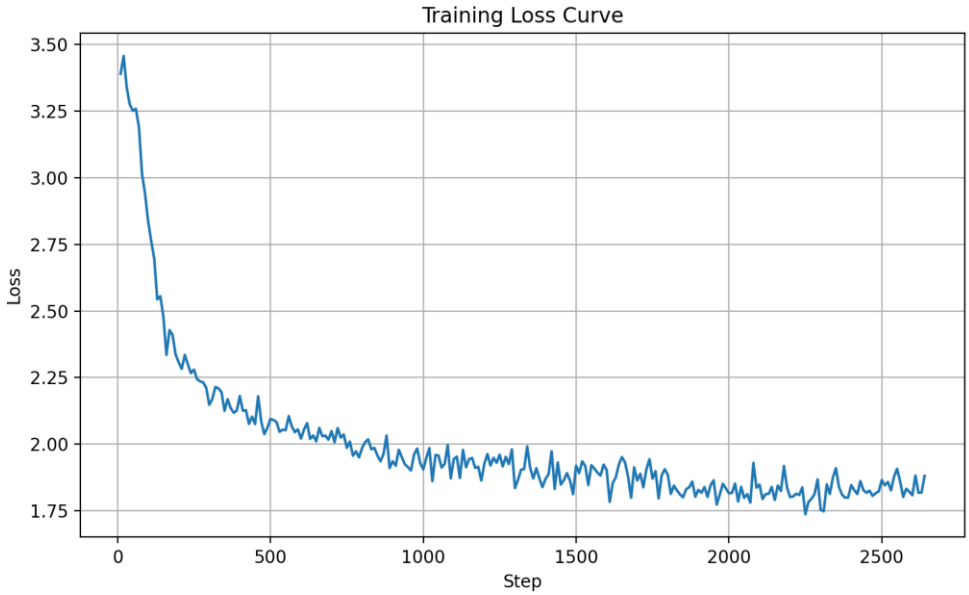


图 6　AdaQloRA 微调 loss 曲线

如图 5、6 所示，从训练损失曲线分析来看，QLoRA 的训练 loss 在初期快速下降后逐步趋于平稳，整体下降过程较为平滑，最终收敛于较低的 los 区间，表明固定低秩参数在训练过程中具有较强的稳定性和可预测性。相比之下，AdaQLoRA 的训练 loss 初期下降更为迅速，但在中后期呈现出更明显的波动性，最终收敛值略高于 QLoRA。这种现象反映了 AdaQLoRA 在训练过程中持续进行低秩参数重分配，使模型参数空间保持更高的动态调整能力，从而在一定程度上牺牲了收敛的平滑性。

不过训练 loss 的绝对值并不能完全反映生成文本的实际质量，具体效果还要看下面的评估指标和模型打分。

表 1 指标对比表

Model	Entity Recall↑	BERTScore↑	BLEURT↑	ROUGE-L↑
Qwen2.5-VL-7B	0.4186	0.6677	0.3137	0.1768
+ QloRA 微调	0.5850	0.7166	0.3810	0.2441
+ AdaQLoRA 微调	0.5560	0.7135	0.3974	0.2392

从表 1 可以看出，相较于未进行微调的 Qwen2.5-VL-7B 基线模型，采用 QLoRA 和 AdaQLoRA 微调后，各项评测指标均获得了显著提升，说明参数高效微调策略能够有效增强模型在自动驾驶视觉问答任务中的文本生成质量与语义表达能力。

进一步比较两种微调策略可以发现，相较于 QLoRA 微调模型，AdaQLoRA 微调模型在 Entity Recall、BERTScore 和 ROUGE-L 这几个基于文本匹配与语义相似度的指标上略有下降，但在 BLEURT 指标上有一定提升。 这一现象主要源于两种微调方法在低秩参数分配机制上的差异。QLoRA 采用固定低秩参数分配方式，而 AdaQLoRA 能够在总参数预算受限的条件下对低秩参数进行自适应重分配，使模型更加关注对下游任务贡献度更高的语义方向，从而在一定程度上减少了对次要或冗余实体的描述，导致实体覆盖率和表层结构相似度指标略有下降。

另外，BLEURT 指标的提升表明，AdaQLoRA 所生成文本在整体语义合理性、逻辑连贯性以及与人工参考的主观一致性方面具有更好的表现。这说 AdaQLoRA 并非简单追求参考文本的复现，而是在生成过程中对信息进行更具选择性的组织与表达，更符合自动驾驶场景中高质量的自然语言解释的需求。

表 2 模型评分表

Model	Score↑
Qwen2.5-VL-7B	6.0880
+ QloRA 微调	7.1640
+ AdaQLoRA 微调	7.3760

如表 2 所示，与未进行微调的 Qwen2.5-VL-7B 基线模型相比，采用参数高效微调策略的模型在综合评分上均取得了显著提升，表明微调方法能够有效增强模型在自动驾驶视觉问答任务中生成文本的整体质量与实用价值。

尽管 AdaQLoRA 在部分评测指标上不占优势，其模型最终评分却明显高于 QLoRA 微调模型。这一结果反映了评测指标与主观语义评估在关注重点上的差异。大模型评分更加关注生成文本在自动驾驶场景中的决策相关性、语义合理性以及幻觉抑制能力，而非单纯的词汇或结构匹配程度。AdaQLoRA 通过动态低秩参数分配，使模型在生成过程中更聚焦于真正影响本车驾驶行为的关键交通要素，从而提升了整体决策解释的自然性、连贯性与可信度。这种以任务语义为导向的生成特性在主观评估中更容易被判定为高质量输出，因此使 AdaQLoRA 微调模型在综合评分中取得了更高表现。

2.4 可视化问答交互

本项目采用 Gradio 实现问答交互，用户上传当前从自车拍摄的图片，并提出交通决策相关问题，模型进行交通状况实时分析并回答驾驶决策。

通用感知、区域感知和驾驶建议这三个任务的问答分别如图 7、8、9 所示。

图 7 通用感知问答

图 8 区域感知问答

图 9 驾驶建议问答

三、总结

综上所述，本文围绕自动驾驶场景下多模态信息理解与决策解释任务，基于视觉语言大模型 Qwen2.5-VL-7B-Instruct，构建了一套完整的自动驾驶场景助手系统。通过引入专为自动驾驶设计的 CODA-LM 视觉语言数据集，从数据处理、多模态指令对齐、参数高效微调以及模型评估等多个方面，对模型在真实驾驶场景中的感知理解能力与文本生成质量进行了系统研究。

总体而言，本文的研究验证了在计算资源受限条件下对大规模多模态模型进行高效微

调并服务于自动驾驶场景理解与决策辅助的可行性与有效性。未来工作可引入更加多样化和复杂的驾驶场景数据，如极端天气、夜间场景以及高密度交通环境，以提升模型在复杂条件下的泛化能力与鲁棒性，从而推动多模态大模型在智能驾驶领域中的进一步落地与发展。