arXiv:2302.00673v1 [cs.CV] 1 Feb 2023

ADAPT：动作感知驾驶字幕变换器

卜进[1,2], 刘新宇[2,5], 郑宇鹏[1,2],李鹏飞[3],赵昊[4],张桐[6],郑宇航[2,7],周谷越[2] 和刘晶晶[2]

摘要——端到端自动驾驶在交通运输行业具有巨大潜力。**然而，自动决策过程缺乏透明度和可解释性，阻碍了其在实践中的产业应用。已有一些初步尝试，使用注意力图或代价体来提升模型的可解释性，但普通乘客难以理解这些方法。为弥合这一差距，我们提出一种基于 Transformer 的端到端架构 ADAPT（动作感知驾驶字幕变换器），为自动驾驶车辆的控制与动作的每个决策步骤提供**用户友好的**自然语言叙述和推理。ADAPT通过共享视频表征，联合训练驾驶描述任务和车辆控制预测任务。在 BDD-X（Berkeley DeepDrive eXplanation）数据集上的实验表明，ADAPT框架在自动指标和人工评估中均达到了最先进的性能。为展示所提框架在现实应用中的可行性，我们构建了一个新颖的可部署系统，以原始汽车视频为输入，实时输出动作叙述和推理。代码、模型和数据可在 https://github.com/jxbbb/ADAPT 获取。**

一、引言

自主系统的目标是精准感知环境，实时做出安全决策，在无人干预的情况下可靠地执行动作，并为乘客提供安全舒适的乘坐体验。自动驾驶控制器的设计通常有两种范式：基于中间表示的方法 [1], [2] 和端到端学习方法 [3]–[23]。基于中间表示的方法依赖于识别人工指定的特征，例如车辆、车道标线等，这需要严格调优参数才能达到令人满意的性能。相比之下，端到端方法直接将传感器采集的原始数据作为输入，生成规划路线或控制信号。

将这类自主控制系统部署到真实车辆上，面临的一项关键挑战是：自动驾驶汽车中的智能决策策略往往过于复杂，普通乘客难以理解，而车辆的安全性和可控性对他们而言至关重要。

[1]中国科学院自动化研究所， 中国，{jinbu2022, zhengyupeng 2022}@ia.ac.cn.[2]清华大学人工智能产业研究院（AIR），中国，{ zhouguyue, JJLiu}@air.tsinghua.edu.cn.[3]清华大学计算机科学与技术系，中国，li-pf22@mails.tsinghua.edu.cn.[4]北京大学英特尔中国研究院，中国，zhao-hao@pku.edu.cn, hao.zhao@intel.com.[5]西安电子科技大学，中国，liuxinyu@stu.xidian.edu.cn.[6]南方科技大学，中国，11911611@mail.sustech.edu.cn.[7]北京航空航天大学机械工程及自动化学院，中国，zyh 021@buaa.edu.cn.

一些先前的研究探讨了自主导航的解释方式 [13], [14], [24]–[30]。例如， [13] 采用代价地图，通过可视化地图不同区域的通行难度，来解释自动驾驶系统的动作。 [24]利用视觉注意力筛除非显著图像区域，而 [31]则构建鸟瞰图（BEV），以可视化车辆的运动信息。然而，如果用户不熟悉系统，这些界面很容易导致误解。

理想的解决方案是，在自主控制模块的决策和动作执行过程中加入自然语言叙述，为用户提供指引，使其内容通俗易懂、便于使用。此外，为每项控制/动作决策提供推理解释，能够帮助用户了解车辆当前状态及周围环境，并为自动驾驶汽车采取的动作提供依据。例如，如图 1 所示，“[动作叙述:] 汽车靠向道路右侧停车， [推理:] 因为汽车正在停车”。通过自然语言叙述和推理解释车辆行为，整个自主系统将更加透明，也更容易理解。

为此，我们提出 ADAPT，这是首个具备动作感知能力、基于 Transformer 的驾驶动作描述架构，可为乘客提供易于理解的自然语言叙述和自动驾驶汽车的推理解释。为消除描述生成任务与车辆控制信号预测任务之间的差异，我们使用共享视频表征对这两项任务进行联合训练。通过集成文本生成模块，这一多任务框架可以构建于各种端到端自动驾驶系统之上。

我们在一个大规模数据集上展示了 ADAPT 方法的有效性，该数据集包含控制信号、视频以及动作叙述和推理解释。基于 ADAPT，我们构建了一种新颖的可部署系统，该系统以原始车辆导航视频为输入，并实时生成动作叙述和推理解释。

我们的贡献可概括如下：

- 我们提出 ADAPT，这是一种面向自动驾驶车辆的新型端到端 Transformer 动作叙述与推理框架。
- 我们提出一种多任务联合训练框架，使驾驶动作描述任务与控制信号预测任务相协调。
- 我们为 ADAPT 在模拟器环境和现实世界中的应用开发了一套可部署流程。

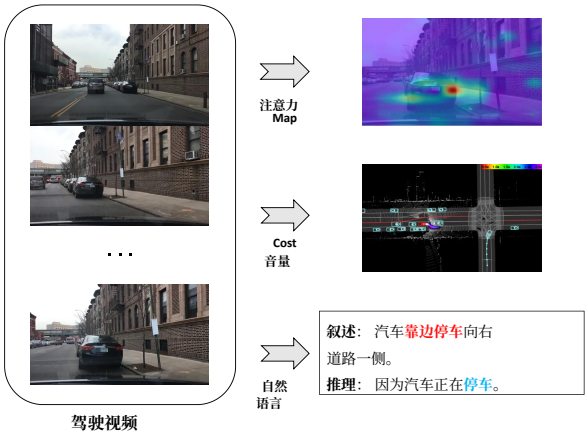

图 1. 自动驾驶车辆的不同解释方法，包括注意力图 [24], 代价体 [13] 和自然语言。尽管注意力图或代价体效果显著，但基于语言的解释对于普通乘客而言更加友好。

II. 相关工作

A. 视频描述

视频描述任务的主要目标是用自然语言描述给定视频中的物体及其相互关系。早期研究 [32]–[35] 通过将识别出的元素填入固定模板来生成具有特定句法结构的句子，但这种方法缺乏灵活性，表达也不够丰富。[36]–[45] 采用序列学习方法生成句法结构灵活、自然流畅的句子。具体而言，这些方法利用视频编码器提取帧特征，并通过语言解码器学习视觉文本对齐，以实现描述生成。为了在字幕中融入细粒度物体和动作， [46]–[48] 采用物体级表示，以捕捉视频中细致的物体感知交互特征。 [49] 进一步开发了一种新颖的双分支卷积编码器，用于联合学习视频的内容和语义信息。此外， [50] 将单模态变换器应用于视频描述，并采用稀疏边界感知池化来减少视频帧中的冗余。场景理解的发展[51]–[62] 也为描述生成任务作出了重要贡献。最近， [63] 提出了一种端到端的 Transformer 模型 SWINBERT，利用稀疏注意力掩码减少连续视频帧中的冗余和无关信息。

虽然现有架构在通用视频描述方面取得了令人瞩目的成果，但不能直接将其应用于动作表示，因为简单地将视频字幕转换为自动驾驶动作表示会遗漏一些关键信息，例如车辆的速度，而这对自主系统至关重要。如何有效利用这些多模态信息生成句子仍是一个谜，这也是我们工作的重点。

B. 端到端自动驾驶

基于学习的自动驾驶是一个活跃的研究领域 [64], [65]。一些基于学习的驾驶方法，例如可供性 [3], [4] 和强化学习 [5]–[7]，已展现出令人期待的性能。模仿方法 [8]–[13]也被用于从人类示范中回归控制指令。例如， [14]–[16]对车辆、骑行者或行人等驾驶主体的未来行为进行建模，以预测车辆航点；而 [17]–[23] 则根据传感器输入直接预测车辆控制信号，这与我们的控制信号预测子任务类似。

C. 自动驾驶的可解释性

可解释性，即提供全面解释的能力，在人工智能的社会接受度方面发挥着重要作用， [66], [67] 自动驾驶也不例外。大多数可解释的自动驾驶汽车方法都基于视觉 [24]–[27], [31]或激光雷达[13], [14], [28]–[30]。 [24] 首先利用注意力图的可视化来过滤非显著图像区域，使自动驾驶汽车的行为合理且可解释。然而，注意力图可能很容易包含一些不太重要的区域，从而让乘客产生误解。 [25]–[27], [31]基于车载摄像头构建鸟瞰图（BEV），以可视化车辆的运动信息和环境状态。 [13] 以激光雷达和高精地图为输入，预测驾驶主体的边界框，并利用代价体解释规划器做出决策的原因。此外， [14] 根据分割结果以及驾驶主体状态构建在线地图，以避免过度依赖高精地图。

尽管基于视觉或基于激光雷达的方法取得了令人期待的成果，但由于缺乏语言解释，这些方法过于复杂，像老年人这样的乘客难以理解。 [68] 首次探索了为自动驾驶车辆提供文本解释的可能性：先离线提取控制信号预测任务中的视频特征，再进行视频描述。不幸的是，这两项任务之间的差异导致离线提取的特征并非下游描述生成任务的最佳特征，而这正是我们工作的重点。

D. 自动驾驶中的多任务学习

我们的端到端框架采用多任务学习，在联合目标下训练模型，同时进行文本生成和控制信号预测。多任务学习利用不同任务之间的归纳偏置，有助于提取更多有用信息 [69] ，并已在自动驾驶领域展现出良好的应用前景。 [70], [71] 表明，检测与跟踪可以联合训练。 [72] 进一步将联合检测器和轨迹预测器整合到单一模型中，并取得了良好效果。 [73] 将这一思路拓展至同时预测参与者意图。最近， [13] 又在联合模型中加入了基于代价地图的控制信号规划器。这些工作表明，联合训练不同任务能够通过更充分的数据利用和共享特征提升各个任务的性能，这也启发我们采用控制信号预测任务与文本生成任务的联合训练策略。

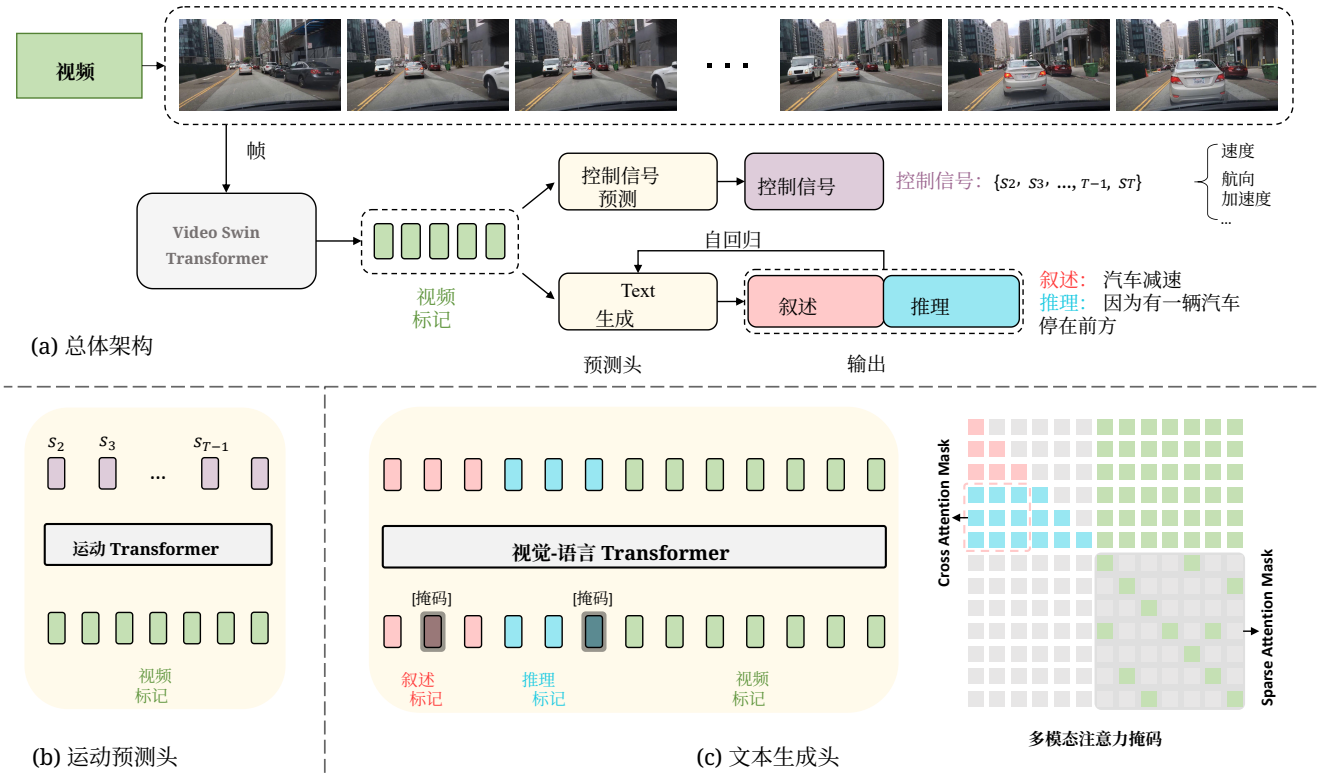

图 2. ADAPT框架概览。(a) 输入为车辆前视视频，输出为预测的车辆控制信号，以及对当前动作的叙述和推理。我们首先从视频中密集且均匀地采样 T 帧，并将其送入可学习的 Video Swin Transformer，将其标记化为视频令牌。不同的预测头生成最终的运动结果和文本结果。(b)(c) 分别展示了预测头。

III. 方法

A. 概述

ADAPT 架构如图 2 所示，旨在处理两项任务：驾驶描述生成（DCG）和控制信号预测（CSP）。DCG以一系列原始视频帧作为输入，并输出两个自然语言句子：一个描述车辆的动作（例如，“汽车正在加速”），另一个解释采取该动作的原因（例如，“因为交通信号灯变绿了”）。CSP以相同的视频帧作为输入，并输出一系列控制信号，例如速度、航向或加速度。

通常，DCG 和 CSP 任务共享同一个视频编码器，但采用不同的预测头来生成最终预测结果。对于 DCG 任务，我们使用视觉语言 Transformer 编码器，通过序列到序列生成来生成两个自然语言句子。对于 CSP 任务，我们使用运动 Transformer 编码器来预测控制信号序列。

B. 视频编码器

沿用 Swinbert [63], ，我们采用 Video Swin Transformer（video swin） [74] 作为视觉编码器，将视频帧编码为视频特征令牌。给定一段从第一视角拍摄的汽车视频，我们首先进行均匀采样，得到 T 帧，每帧大小为 $H \times W \times 3$。这些帧作为输入传入 video swin，生成大小为 F_V 的特征，其中 C 是 video swin 中定义的通道维度。然后，将视频特征输入不同的预测头，以执行各项任务。

C. 预测头

文本生成头 文本生成头旨在生成两个句子，分别描述车辆的动作及其背后的原因。如第 III-B 节所述，视频帧被编码为大小为 $\frac{T}{2} \times \frac{H}{32} \times \frac{W}{32} \times 8C$ 的视频特征F_V。随后，我们沿通道维度对视频特征进行标记化，得到 $\frac{T}{2} \times \frac{H}{32} \times \frac{W}{32}$ 个维度为 $8C$ 的标记。对于文本输入（动作、叙述和推理），我们首先对每个句子进行标记化，并将其填充至固定长度。然后，我们将这两个句子拼接起来，并通过嵌入层将其嵌入。为了区分动作叙述和推理，我们采用分段嵌入方法（常用于 Bert [75]）来标记二者。我们还使用一个可学习的多层感知机变换视频标记的维度，以确保视频标记与文本词元的维度一致。最后，将文本词元和视频令牌输入视觉语言 Transformer 编码器，由其生成一个同时包含动作叙述和推理的新序列。

控制信号预测头 CSP 头的目标是根据视频帧预测车辆的控制信号（例如加速度）。给定 T 帧的视频特征，以及相应的控制信号记录 $\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, .., \boldsymbol{s}_T\}$，CSP 头的输出是一系列

控制信号 $\hat{} \boldsymbol{S} = \{\hat{}\boldsymbol{s}_2$、.....$\hat{}\boldsymbol{s}_T\}$。每个控制信号 $\boldsymbol{s}_i$或 $\hat{}\boldsymbol{s}_i$ 都是一个 n 元组，其中 n 表示我们使用的传感器类型数量。我们首先对视频特征进行词元化，然后利用另一个 Transformer（运动 Transformer）来生成这些控制信号的预测。损失函数$\hat{}\mathcal{L}_{CSP}$ 定义为 $\boldsymbol{S}$ 与 $\boldsymbol{S}$之间的均方误差：

$$\mathcal{L}_{CSP} = \frac{1}{T-1}\sum_{i=2}^{T}(s_i - \hat{s}_i)^2 \tag{1}$$

请注意，我们不会预测与第一帧对应的控制信号，因为第一帧的动态信息有限，而其他信号可以很容易地从前面的帧中推断出来。

D. 联合训练

在我们的框架中，我们假设 CSP 和 DCG 任务在视频表示的语义层面上是对齐的。直观而言，动作叙述和控制信号数据是自动驾驶车辆动作的不同表达形式，而推理解释则集中于影响车辆动作的环境要素。我们认为，在单一网络中联合训练这些任务，可以利用不同任务之间的归纳偏置来提升性能。

在训练过程中，CSP 和 DCG 会同时执行。我们只需将 $\mathcal{L}_{CSP}$ 和 $\mathcal{L}_{DCG}$ 相加，即可得到最终的损失函数：

$$\mathcal{L} = \mathcal{L}_{CSP} + \mathcal{L}_{DCG} \tag{2}$$

尽管对两个任务进行了联合训练，但每个任务的推理仍可独立进行。对于 DCG 任务，ADAPT 以视频序列作为输入，并输出包含两个片段的驾驶描述。文本生成以自回归方式进行。具体而言，我们的模型以“[CLS]”标记开始，每次生成一个词令牌，并将此前生成的标记作为视觉语言 Transformer 编码器的输入。生成过程会持续进行，直到模型输出结束标记“[SEP]”，或达到单个句子的最大长度阈值。将第一个句子填充至最大长度后，我们会在输入中拼接另一个“[CLS]”，并重复上述过程。

IV. 实验

本节使用标准描述生成任务的指标评估 ADAPT，包括 BLEU4 [76], METEOR[77],ROUGE-L [78]和 CIDEr [79]（在后续表格中分别简称为 B4、M、R 和 C）。由于描述生成的定量评估仍是一个开放问题，我们还提供了详细的人工评估结果，以评估生成文本在主观层面上的正确性。消融研究进一步证明了所提出的联合训练框架的有效性。

A. 数据集

BDD-X [68]是一个驾驶领域描述数据集，由近7000个与控制信号配对的视频组成。这些视频和控制信号采集自 BDD100K 数据集 [80]。每个视频的平均时长为40秒，分辨率为 1280×720 ，帧率为30 FPS。每个视频包含1至5种车辆行为，例如加速、右转或并入车道。所有这些行为都配有文本标注，包括动作叙述（例如“汽车停下来了”）和推理（例如“因为交通信号灯是红灯”）。总计约有29000对行为—标注。据我们所知，BDD-X是唯一一个同时配有汽车视频和控制信号的驾驶领域描述数据集。

表 I 视频描述指标与 SOTA 的比较

方法	叙述			推理		
	B4	C	M	B4	C	M
S2VT [42]	30.2	179.8	27.5	6.3	53.4	11.2
S2VT++ [42]	27.1	157.0	26.4	5.8	52.7	10.9
SAA [68]	31.8	214.8	29.1	7.1	66.1	12.2
WAA [68]	32.3	215.8	29.2	7.3	69.5	12.2
Ours	**34.6**	**247.5**	**30.6**	**11.4**	**102.6**	**15.2**

表 II 与 SOTA 在人工评估上的比较

方法	叙述	推理	完整句子
SAA [68]	90.8%	62.4%	-
WAA [68]	**93.5**%	66.0%	-
Ours	90.0%	**90.3**%	82.7%

B. 实现细节

Video Swin Transformer 在 Kinetics-600 [81]，上进行了预训练，而视觉-语言 Transformer 和运动 Transformer 则采用随机初始化。请注意，在我们的实现中，我们没有冻结 Video Swin 的参数，因此 ADAPT 以完全端到端的方式进行训练。输入视频帧会调整大小并裁剪至 224 的空间尺寸。对于叙述和推理，我们使用 WordPiece 嵌入 [75] ，而不是完整单词（例如，“stops”会被拆分为“stop”和“#s”），每个句子的最大长度为 15。训练期间，我们会随机掩码 50% 的标记，以进行掩码语言建模。被掩码的标记有 80% 的概率变为“[MASK]”标记，10% 的概率变为随机单词，另有 10% 的概率保持不变。我们采用 AdamW 优化器，并在训练初期的 10% 步数中进行学习率预热，随后线性衰减。整个训练过程共 40 个周期，在 4 块 NVIDIA V100 GPU 上进行，每块 GPU 的批量大小为 4，耗时约 13 小时。

C. 主要结果

我们在 BDD-X 数据集上将 ADAPT 与最先进方法进行比较。表 I 展示了标准描述指标上的比较结果。我们观察到，ADAPT 的性能显著优于现有方法。具体而言，在 CIDEr 指标上，ADAPT 在动作叙述和推理方面分别比此前最先进的工作[68] 高出 31.7 和 33.1。

除了自动评估指标外，我们还进行人工评估，以衡量输出叙述和推理的主观正确性

表 III 单一字幕生成与动作感知字幕生成

方法	叙述				推理			
	B4	C	M	R	B4	C	M	R
单一	33.2	238.9	29.7	62.0	8.6	89.7	14.1	31.4
单一+	33.9	**248.3**	30.5	**63.1**	9.3	97.2	14.6	31.5
Ours	**34.6**	247.5	**30.6**	62.8	**11.4**	**102.6**	**15.2**	**32.0**

表 IV 控制信号类型分析

信号		叙述			推理		
速度	航向	C	M	R	C	M	R
✓		232.0	29.9	61.5	88.0	15.1	31.0
	✓	218.2	29.3	61.2	88.6	14.1	30.6
✓	✓	**247.5**	**30.6**	**62.8**	**102.6**	**15.2**	**32.0**

。整个评估过程分为三个部分：（1）叙述；（2）推理；（3）完整句子。在第一部分中，人工评估者判断预测叙述是否符合车辆的行为。在第二部分中，我们同时展示真实叙述和预测推理，并要求人工评估者判断推理是否正确。最后一部分则展示预测叙述和预测推理。表 II 显示，ADAPT 的推理准确率优于以往工作，同时在叙述评估中保持较高准确率，证明了 ADAPT 的有效性。

D. 消融研究

我们进行了全面的消融研究，以分析 ADAPT 设计的各个方面。

动作感知联合训练的效果 为研究联合训练中的动作感知对 ADAPT 的影响，我们移除 ADAPT 的控制信号预测（CSP）头，训练一个单一字幕生成模型，并将其称为“单一”。如表 III 所示，ADAPT 的表现优于单一训练，在 CIDEr 指标上，叙述和推理分别提升了 15.9 和 7.2。这表明，来自另一项任务的线索有助于正则化共享视频表征，并提升文本生成任务的性能。

此外，从图 2(a) 中可以看出，在 ADAPT 中，字幕和控制信号数据分别用于两条流。一个有趣的问题是：我们能否直接将控制信号传递给多模态变换器，以获得最终的

表 V 叙述与推理交互的影响

方法	叙述			推理		
	B4	C	M	B4	C	M
无交叉注意力	32.7	234.8	30.0	10.7	96.6	15.1
使用交换注意力	28.3	180.4	28.7	9.3	97.7	14.3
Ours	**34.6**	**247.5**	**30.6**	**11.4**	**102.6**	**15.2**
仅叙述	32.9	240.4	29.6	-	-	-
仅推理	-	-	-	8.1	94.4	13.3

表 VI 视频帧的影响(T)

方法	叙述				推理				成本（分钟）
	B4	C	M	R	B4	C	M	R	
2	33.4	227.7	28.7	61.0	8.7	62.9	15.1	29.8	294
4	32.9	225.7	29.0	60.9	9.9	81.3	14.9	31.1	382
8	32.6	236.1	29.3	61.8	8.4	83.7	13.4	30.6	447
16	32.5	231.0	29.5	61.9	8.7	91.5	13.8	32.0	528
32	**34.6**	**247.5**	**30.6**	**62.8**	**11.4**	**102.6**	**15.2**	**32.0**	797

字幕预测？因此，我们设计了这样一种架构：以视频令牌、控制信号令牌（由可学习嵌入层生成）和掩码文本令牌作为输入，并生成对掩码令牌的预测结果，该架构被称为“Single+”。结果如表 III 第二行所示。可以看出，所提出的 ADAPT 仍然取得了最佳结果，尤其是在推理片段方面，这证明了多任务学习优于同时将视频和控制信号作为输入，尽管后者是一种直观的设置。

不同控制信号类型的影响在我们的实现中，我们利用控制信号（例如航向）作为 CSP 任务的监督信号。在这项分析中，我们研究了不同监督信号类型对 ADAPT 的影响。实验中的基础信号是速度和航向。我们首先通过移除其中一个信号进行实验，结果如表 IV 的前两行所示。然后在第三行中同时使用速度和航向，这与之前的实验相同。我们观察到，移除任一信号都会导致性能下降。例如，在没有速度输入的情况下，叙述的 CIDEr 指标下降了 29.3，推理的 CIDEr 指标下降了 14.0。这是可以理解的，因为了解速度和航向有助于网络学习对叙述和推理有用的表征，而缺少任一信号都可能导致视频表征产生偏差。

叙述与推理的交互与通用字幕任务相比，驾驶描述任务会生成两个句子：动作叙述和推理。本节中，我们通过控制注意力掩码或两个句子的顺序，探讨这两个部分如何相互作用。

具体而言，如图 2(c) 右侧所示，我们为每个句子使用因果自注意力掩码，使词令牌只能关注已有的输出标记，并对视频令牌采用稀疏注意力 [63] 。推理片段对叙述片段具有完全注意力，这称为交叉注意力，定义了推理对叙述的依赖关系。本节中，我们首先进行了未使用交叉注意力或交换交叉注意力（通过交换叙述与推理的顺序）的实验。结果如表 V 所示。与默认设置（记为“我们的方法”）相比，未使用交叉注意力时两个句子的表现都较差，这表明以叙述片段为条件来训练推理片段是有益的。交换交叉注意力后，表现也有所下降，尤其是叙述部分，这进一步证明了这种依赖关系

表 VII 控制信号预测准确率比较

方法	航向						速度					
	均方根误差（度）↓	$A_{0.1}$ ↑	$A_{0.5}$ ↑	$A_{1.0}$ ↑	$A_{5.0}$ ↑	$A_{10.0}$ ↑	均方根误差(m/s)↓	$A_{0.1}$ ↑	$A_{0.5}$ ↑	$A_{1.0}$ ↑	$A_{5.0}$ ↑	$A_{10.0}$ ↑
单一	**6.3**	8.3	84.7	**90.5**	97.2	98.7	3.4	5.0	25.5	37.8	86.8	98.7
Ours	6.4	**62.2**	**85.5**	89.9	**97.2**	**98.8**	**2.5**	**11.1**	**28.1**	**45.3**	**94.3**	**99.5**

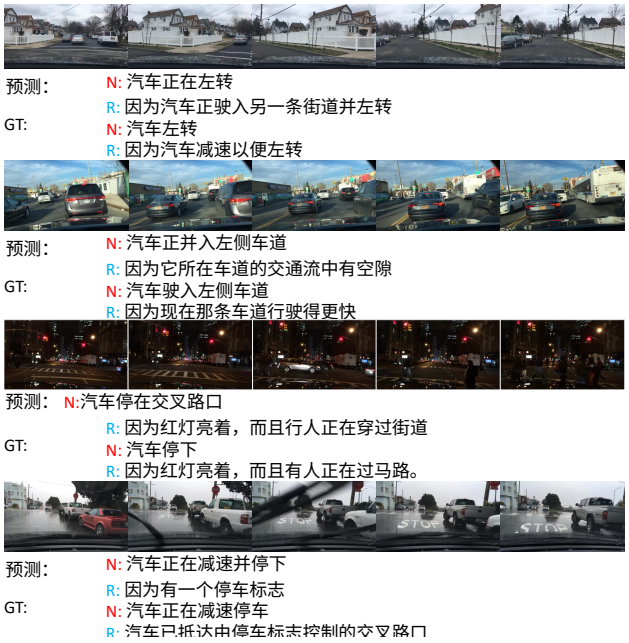

图3. 定性分析：生成的叙述能够正确描述当前动作，并进行合理推理。

即推理依赖于叙述，而非反过来。

此外，我们还进行了仅使用一个句子的实验，分别称为“仅叙述”和“仅推理”。表 V 显示，同时使用两个句子进行训练能够提升性能，尤其是推理片段的性能，这表明叙述与推理的交互有助于提升完整描述任务中各个组成部分的表现。

不同采样率的影响 在之前的实验中，我们从给定视频中均匀采样 $T = 32$ 帧，并获取相同时间戳对应的控制信号数据。在本研究中，我们通过改变采样帧数来探究采样率的影响。具体而言，如表 VI 所示，我们从可变长度的视频中均匀采样 $T = 2$、4、8、16、32 帧。随着采样帧数增加，ADAPT 的性能稳步提升，因为更多帧意味着视觉内容缺失更少。这表明，密集采样帧和控制信号能够提升描述结果。表 VI 还列出了训练时间成本。我们希望这项消融实验能为机器人领域从业者提供有关驾驶描述准确率与效率权衡的参考。

E. 控制信号预测分析

虽然驾驶描述任务的主要目标是生成句子，但我们也研究了控制信号预测任务的性能。我们采用均方根误差（RMSE）和容差准确率（A_σ）来衡量最终性能。容差准确率是指先使用两个阈值确定控制信号偏差的范围，并对其进行截断。例如，我们将预测航向$\hat{c}$ 的截断值定义为：

$$c_\sigma = \begin{cases} 1, & -\sigma < \hat{c} - c < \sigma \\ 0, & otherwise \end{cases} \tag{3}$$

其中，c 为真实航向，σ 为容差阈值。然后，航向的 A_σ 表示 c_σ 的准确率，以百分比记录；速度的 A_σ 定义类似。结果见表 VII。我们观察到，我们的联合训练框架能够进一步提升控制信号预测的性能，表明联合训练具有优势。

F. 在自主系统中的部署

我们进一步开发了一套流程，用于将 ADAPT 部署到模拟器环境（例如 CARLA[82]）和现实世界中。系统以原始车辆视频为输入，并实时生成动作叙述和推理解释。具体而言，我们首先记录由前视摄像机拍摄的帧。然后，将最近几秒的帧作为输入传入 ADAPT，以生成当前步骤的动作叙述和推理。此外，我们还利用文本转语音技术将生成的句子转换为语音叙述，使普通乘客使用起来更加方便、交互性更强（尤其有助于视障乘客）。

V. 结论

基于语言的可解释性对于自动驾驶车辆获得社会认可至关重要。我们提出 Adapt（动作感知驾驶字幕变换器），这是一种全新的端到端、基于变换器的框架，可为自动驾驶车辆生成动作叙述和推理。ADAPT 利用多任务联合训练，缩小驾驶动作描述任务与控制信号预测任务之间的差异。在 BDD-X 数据集上进行的实验，结合标准描述指标和人工评估，证明了 ADAPT 相较于最先进方法的有效性。我们进一步开发了可部署流程，使 ADAPT 能够应用于模拟器环境和现实世界。

参考文献

[1] S. Behere 和 M. T¨orngren, 《自主驾驶的功能架构》, 载于《第一届汽车软件架构国际研讨会论文集》, 2015 年, 第 3–10 页。[2] E. Yurtsever、J. Lambert、A. Carballo 和 K. Takeda, 《自动驾驶综述：常见实践与新兴技术》, IEEE Access, 第 8 卷, 第 58 443–58 469 页, 2020 年。[3] A. Sauer、N. Savinov 和 A. Geiger, 《城市环境驾驶的条件可供性学习》, 载于《机器人学习会议》。PMLR, 2018 年, 第 237–252 页。[4] Y. Xiao、F. Codevilla、C. Pal 和 A. M. L´opez, 《面向自动驾驶的基于动作的表征学习》, arXiv 预印本 arXiv:2008.09417, 2020 年。[5] D. Chen、V. Koltun 和 P. Kr¨ahenb¨uhl, 《在轨道世界中学习驾驶》, 载于《IEEE/CVF国际计算机视觉会议论文集》, 2021 年, 第 15 590–15 599 页。[6] M. Toromanoff、E. Wirbel 和 F. Moutarde, 《利用隐式可供性进行城市驾驶的端到端无模型强化学习》, 载于《IEEE/CVF计算机视觉与模式识别会议论文集》, 2020 年, 第 7153–7162 页。[7] J. Wang、Y. Wang、D. Zhang、Y. Yang 和 R. Xiong, 《面向自动驾驶的分层行为学习与运动规划》, 载于《2020 IEEE/RSJ智能机器人与系统国际会议（IROS）》。IEEE, 2020 年, 第 2235–2242 页。[8] M. Bojarski、D. Del Testa、D. Dworakowski、B. Firner、B. Flepp、P. Goyal、L. D. Jackel、M. Monfort、U. Muller、J. Zhang 等, 《自动驾驶汽车的端到端学习》, arXiv 预印本 arXiv:1604 07316, 2016 年。.[9] F. Codevilla、M. M¨uller、A. L´opez、V. Koltun 和 A. Dosovitskiy, 《通过条件模仿学习实现端到端驾驶》, 载于《2018 IEEE国际机器人与自动化会议（ICRA）》。IEEE, 2018 年, 第 4693–4700 页。[10] M. M¨uller、A. Dosovitskiy、B. Ghanem 和 V. Koltun, 《通过模块化和抽象实现驾驶策略迁移》, arXiv 预印本 arXiv:1804.09364, 2018 年。[11] D. A. Pomerleau, 《ALVINN：神经网络中的自主陆地车辆》, 《神经信息处理系统进展》, 第 1 卷, 1988 年。[12] B. Wei、M. Ren、W. Zeng、M. Liang、B. Yang 和 R. Urtasun, 《感知、关注与驾驶：学习空间注意力以实现安全自动驾驶》, 载于《2021 IEEE国际机器人与自动化会议（ICRA）》。IEEE, 2021 年, 第 4875–4881 页。[13] W. Zeng、W. Luo、S. Suo、A. Sadat、B. Yang、S. Casas 和 R. Urtasun, 《端到端可解释神经运动规划器》, 载于《IEEE/CVF计算机视觉与模式识别会议论文集》, 2019 年, 第 8660–8669 页。[14] S. Casas、A. Sadat 和 R. Urtasun, 《MP3：用于建图、感知、预测与规划的统一模型》, 载于《IEEE/CVF计算机视觉与模式识别会议论文集》, 2021 年, 第 14 403–14 412 页。[15] D. Chen、B. Zhou、V. Koltun 和 P. Kr¨ahenb¨uhl, 《通过作弊学习》, 载于《机器人学习会议》。PMLR, 2020 年, 第 66–75 页。[16] A. Filos、P. Tigkas、R. McAllister、N. Rhinehart、S. Levine 和 Y. Gal, 《自动驾驶车辆能否识别分布偏移、从中恢复并适应它? 》, 载于《国际机器学习会议》。PMLR, 2020 年, 第 3145–3153 页。[17] A. Behl、K. Chitta、A. Prakash、E. Ohn-Bar 和 A. Geiger, 《面向自动驾驶的标签高效视觉抽象》, 载于《2020 IEEE/RSJ智能机器人与系统国际会议（IROS）》。IEEE, 2020 年, 第 2338–2345 页。[18] A. B¨uhler、A. Gaidon、A. Cramariuc、R. Ambrus、G. Rosman 和 W. Burgard, 《穿越幽灵：带假阳性的行为克隆》, 载于《 2020 IEEE/RSJ智能机器人与系统国际会议（IROS）》。IEEE, 2020 年, 第 5431–5437 页。[19] F. Codevilla 、 E. Santana 、 A. M. L´opez 和 A. Gaidon, 《探索行为克隆在自动驾驶中的局限性》, 载于《IEEE/CVF国际计算机视觉会议论文集》, 2019 年, 第 9329–9338 页。[20] Z. Huang、C. Lv、Y. Xing 和 J. Wu, 《基于多模态传感器融合的深度神经网络：用于场景理解的端到端自动驾驶》, 《IEEE传感器杂志》, 第 21 卷, 第 10 期, 第 11 781–11 790 页, 2020 年。[21] E. Ohn-Bar、A. Prakash、A. Behl、K. Chitta 和 A. Geiger, 《情境驾驶学习》, 载于《IEEE/CVF计算机视觉与模式识别会议论文集》, 2020年, 第11 296–11 305页。[22] A. Prakash、A. Behl、E. Ohn-Bar、K. Chitta 和 A. Geiger, 《探索基于视觉的城市自动驾驶策略学习中的数据聚合》, 载于《IEEE/CVF计算机视觉与模式识别会议论文集》, 2020年, 第11 763–11 773页。[23] Y. Xiao、F. Codevilla、A. Gurram、O. Urfalioglu 和 A. M. L´opez, 《多模态端到端自动驾驶》, 《IEEE智能交通系统汇刊》, 2020年。[24] J. Kim 和 J. Canny, 《通过可视化因果注意力实现自动驾驶汽车的可解释学习》, 载于《IEEE国际计算机视觉会议论文集》, 2017年, 第2942–2950页。[25] A. Hu、Z. Murez、N. Mohan、S. Dudas、J. Hawke、V. Badrinarayanan、R. Cipolla 和 A. Kendall, 《Fiery：基于环视单目摄像机的鸟瞰视角未来实例预测》, 载于《IEEE/CVF国际计算机视觉会议论文集》, 2021年, 第15 273–15 282页。[26] J. Philion 和 S. Fidler, 《Lift, Splat, Shoot：通过将图像隐式反投影到3D，对任意相机装置采集的图像进行编码》, 载于《欧洲计算机视觉会议》。Springer, 2020年, 第194–210页。[27] H. Wang、P. Cai、Y. Sun、L. Wang 和 M. Liu, 《通过光流蒸馏，为自动驾驶学习可解释的端到端基于视觉的运动规划》, 载于《2021年IEEE国际机器人与自动化会议（ICRA）》。IEEE, 2021年, 第13 731–13 737页。[28] A. Cui、S. Casas、A. Sadat、R. Liao 和 R. Urtasun, 《Lookout：面向自动驾驶的多样化多未来预测与规划》, 载于《IEEE/CVF国际计算机视觉会议论文集》, 2021年, 第16 107–16 116页。[29] W. Zeng、S. Wang、R. Liao、Y. Chen、B. Yang 和 R. Urtasun, 《Ds-dnet：深度结构化自动驾驶网络》, 载于《欧洲计算机视觉会议》。Springer, 2020年, 第156–172页。[30] A. Sadat、S. Casas、M. Ren、X. Wu、P. Dhawan 和 R. Urtasun, 《感知、预测与规划：通过可解释的语义表示实现安全运动规划》, 载于《欧洲计算机视觉会议》。Springer, 2020年, 第414–430页。[31] A. Saha、O. Mendez、C. Russell 和 R. Bowden, 《将图像转换为地图》, 载于《2022年国际机器人与自动化会议（ICRA）》。IEEE, 2022年, 第9200–9206页。[32] S. Guadarrama、N. Krishnamoorthy、G. Malkarnenkar、S. Venugopalan、R. Mooney、T. Darrell 和 K. Saenko, 《Youtube2text：利用语义层次结构和零样本识别来识别并描述任意活动》, 载于《IEEE国际计算机视觉会议论文集》, 2013年, 第2712–2719页。[33] P. Hanckmann、K. Schutte 和 G. J. Burghouts, 《针对范围广泛的视频事件及48种人类动作的自动文本描述》, 载于《欧洲计算机视觉会议》。Springer, 2012年, 第372–380页。[34] A. Kojima、T. Tamura 和 K. Fukunaga, 《基于动作概念层次结构的从视频图像中生成人类活动的自然语言描述》, 《国际计算机视觉期刊》, 第50卷, 第2期, 第171–184页, 2002年。[35] M. Rohrbach、W. Qiu、I. Titov、S. Thater、M. Pinkal 和 B. Schiele, 《将视频内容转换为自然语言描述》, 载于《IEEE国际计算机视觉会议论文集》, 2013年, 第433–440页。[36] S. Chen、J. Chen、Q. Jin 和 A. Hauptmann, 《借助多模态潜在主题的视频描述生成》, 载于《第25届ACM国际多媒体会议论文集》, 2017年, 第1838–1846页。[37] L. Gao、Y. Lei、P. Zeng、J. Song、M. Wang 和 H. T. Shen, 《用于视频描述生成和视频问答的辅助任务层次表示网络》, 《IEEE图像处理汇刊》, 2022年。[38] W. Ji 和 R. Wang, 《一种用于视频描述生成的多实例多标签双重学习方法》, 《ACM多媒体计算、通信与应用汇刊》, 第17卷, 第2s期, 第1–18页, 2021年。[39] Y. Pan、T. Mei、T. Yao、H. Li 和 Y. Rui, 《联合建模嵌入与翻译，以连接视频与语言》, 载于《IEEE计算机视觉与模式识别会议论文集》, 2016年, 第4594–4602页。[40] Y. Pan、T. Yao、H. Li 和 T. Mei, 《利用迁移语义属性的视频描述生成》, 载于《IEEE计算机视觉与模式识别会议论文集》, 2017年, 第6504–6512页。[41] R. Pasunuru 和 M. Bansal, 《结合视频与蕴含生成的多任务视频描述生成》, arXiv预印本 arXiv:1704.07489, 2017年。

[42] S. Venugopalan、M. Rohrbach、J. Donahue、R. Mooney、T. Darrell 和 K. Saenko，《序列到序列视频转文本》，载于《IEEE国际计算机视觉会议论文集》，2015年，第4534–4542页。[43] S. Venugopalan、H. Xu、J. Donahue、M. Rohrbach、R. Mooney 和 K. Saenko，《利用深度循环神经网络将视频转换为自然语言》，arXiv预印本 arXiv:1412.4729，2014年。[44] L. Yao、A. Torabi、K. Cho、N. Ballas、C. Pal、H. Larochelle 和 A. Courville，《利用时间结构描述视频》，载于《IEEE国际计算机视觉会议论文集》，2015年，第4507–4515页。[45] H. Yu、J. Wang、Z. Huang、Y. Yang 和 W. Xu，《使用分层循环神经网络生成视频段落描述》，载于《IEEE计算机视觉与模式识别会议论文集》，2016年，第4584–4593页。[46] B. Zhao、X. Li、X. Lu 等，《利用管状特征的视频描述生成》，载于 IJCAI，2018年，第1177–1183页。[47] B. Pan、H. Cai、D.-A. Huang、K.-H. Lee、A. Gaidon、E. Adeli 和 J. C. Niebles，《利用知识蒸馏进行视频描述生成的时空图》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2020年，第10 870–10 879页。[48] Z. Zhang、Y. Shi、C. Yuan、B. Li、P. Wang、W. Hu 和 Z.-J. Zha，《结合教师推荐学习的视频描述生成对象关系图》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2020年，第13 278–13 288页。[49] S. Liu、Z. Ren 和 J. Yuan，《Sibnet：用于视频描述生成的孪生卷积编码器》，载于《第26届ACM国际多媒体会议论文集》，2018年，第1425–1434页。[50] T. Jin、S. Huang、M. Chen、Y. Li 和 Z. Zhang，《Sbat：采用稀疏边界感知Transformer的视频描述生成》，arXiv预印本 arXiv:2007.11888，2020年。[51] W. Wang、H. Bao、L. Dong、J. Bjorck、Z. Peng、Q. Liu、K. Aggarwal、O. K. Mohammed、S. Singhal、S. Som 等，《图像作为外语：面向所有视觉与视觉语言任务的Beit预训练》，arXiv预印本 arXiv:2208.10442，2022年。[52] K. He、X. Chen、S. Xie、Y. Li、P. Doll´ar 和 R. Girshick，《掩码自编码器是可扩展的视觉学习器》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第16 000–16 009页。[53] Z. Liu、Y. Lin、Y. Cao、H. Hu、Y. Wei、Z. Zhang、S. Lin 和 B. Guo，《Swin Transformer：使用移位窗口的分层视觉Transformer》，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第10 012–10 022页。[54] T. Zhang、X. Chen、Y. Wang、Y. Wang 和 H. Zhao，《Mutr3d：通过3D到2D查询实现的多摄像头跟踪框架》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第4537–4546页。[55] M. Han、Z. Zhang、Z. Jiao、X. Xie、Y. Zhu、S.-C. Zhu 和 H. Liu，《利用功能性物体进行场景重建，实现机器人自主性》，《国际计算机视觉期刊》，第130卷，第12期，第2940–2961页，2022年。[56] S. Huang、Z. Wang、P. Li、B. Jia、T. Liu、Y. Zhu、W. Liang 和 S.-C. Zhu，《3D场景中的基于扩散的生成、优化与规划》，arXiv预印本 arXiv:2301.06015，2023年。[57] Y. Li、Y. Tu、X. Chen、H. Zhao 和 G. Zhou，《利用聚焦注意力进行距离感知遮挡检测》，《IEEE图像处理汇刊》，第31卷，第5661–5676页，2022年。[58] B. Jin、B. Tian、H. Zhao 和 G. Zhou，《语言引导的3D室内场景语义风格迁移》，载于《第1届多媒体实验用照片级图像与环境合成研讨会论文集》，2022年，第11–17页。[59] P. Li、B. Tian、Y. Shi、X. Chen、H. Zhao、G. Zhou 和 Y.-Q. Zhang，《Toist：采用名词-代词蒸馏的任务导向实例分割Transformer》，arXiv预印本 arXiv:2210.10775，2022年。[60] X. Chen、H. Zhao、G. Zhou 和 Y.-Q. Zhang，《PQ-transformer：联合解析点云中的3D物体与布局》，《IEEE机器人与自动化快报》，第7卷，第2期，第2519–2526页，2022年。[61] H. Zhao、M. Lu、A. Yao、Y. Chen 和 L. Zhang，《学习绘制视线》，《国际计算机视觉期刊》，第128卷，第1076–1100页，2020年。[62] H. Zhao、M. Lu、A. Yao、Y. Guo、Y. Chen 和 L. Zhang，《语义迁移特征上的物理启发式优化：一种房间布局估计的替代方法》，载于《IEEE

计算机视觉与模式识别会议，2017，第10–18页。[63] K. Lin、L. Li、C.-C. Lin、F. Ahmed、Z. Gan、Z. Liu、Y. Lu和L. Wang，《Swinbert：采用稀疏注意力进行视频描述的端到端Transformer》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022，第17 949–17 958页。[64] J. Janai、F. G¨uney、A. Behl、A. Geiger等，《自动驾驶车辆的计算机视觉：问题、数据集与最新进展》，《计算机图形学与视觉基础与趋势》，第12卷，第1–3期，第1–308页，2020。[65] A. Tampuu、T. Matiisen、M. Semikin、D. Fishman和N. Muhammad，《端到端驾驶综述：架构与训练方法》，《IEEE神经网络与学习系统汇刊》，2020。[66] L. Yuan、X. Gao、Z. Zheng、M. Edmonds、Y. N. Wu、F. Rossano、H. Lu、Y. Zhu和S.-C. Zhu，《情境中的双向人机价值对齐》，《科学·机器人》，第7卷，第68期，文章编号eabm4183，2022。[67] M. Edmonds、F. Gao、H. Liu、X. Xie、S. Qi、B. Rothrock、Y. Zhu、Y. N. Wu、H. Lu和S.-C. Zhu，《两个解释的故事：通过解释机器人行为增强人类信任》，《科学·机器人》，第4卷，第37期，文章编号eaay4663，2019。[68] J. Kim、A. Rohrbach、T. Darrell、J. Canny和Z. Akata，《自动驾驶车辆的文本解释》，载于《欧洲计算机视觉会议（ECCV）论文集》，2018，第563–578页。[69] X. Chen、T. Liu、H. Zhao、G. Zhou和Y.-Q. Zhang，《Cerberus Transformer：联合语义、可供性与属性解析》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022，第19 649–19 658页。[70] C. Feichtenhofer、A. Pinz和A. Zisserman，《先检测后跟踪，先跟踪后检测》，载于《IEEE国际计算机视觉会议论文集》，2017，第3038–3046页。[71] D. Frossard、E. Kee和R. Urtasun，《DeepSignals：通过视觉信号预测驾驶员意图》，载于2019年国际机器人与自动化会议（ICRA）。IEEE，2019，第9697–9703页。[72] W. Luo、B. Yang和R. Urtasun，《极速响应：使用单一卷积网络实时进行端到端三维检测、跟踪与运动预测》，载于《IEEE计算机视觉与模式识别会议论文集》，2018，第3569–3577页。[73] S. Casas、W. Luo和R. Urtasun，《IntentNet：学习从原始传感器数据预测意图》，载于机器人学习会议。PMLR，2018，第947–956页。[74] Z. Liu、J. Ning、Y. Cao、Y. Wei、Z. Zhang、S. Lin和H. Hu，《Video Swin Transformer》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022，第3202–3211页。[75] J. Devlin、M.-W. Chang、K. Lee和K. Toutanova，《BERT：用于语言理解的深度双向Transformer预训练》，arXiv预印本，arXiv:1810.04805，2018。[76] K. Papineni、S. Roukos、T. Ward和W.-J. Zhu，《BLEU：一种机器翻译自动评估方法》，载于计算语言学协会第40届年会论文集，2002，第311–318页。[77] S. Banerjee和A. Lavie，《METEOR：一种与人工判断相关性更高的机器翻译评估自动指标》，载于机器翻译和/或摘要的内在与外在评估指标ACL研讨会论文集，2005，第65–72页。[78] C.-Y. Lin和F. J. Och，《使用最长公共子序列和跳跃二元组统计自动评估机器翻译质量》，载于计算语言学协会第42届年会论文集（ACL-04），2004，第605–612页。[79] R. Vedantam、C. Lawrence Zitnick和D. Parikh，《CIDEr：基于共识的图像描述评估》，载于《IEEE计算机视觉与模式识别会议论文集》，2015，第4566–4575页。[80] F. Yu、H. Chen、X. Wang、W. Xian、Y. Chen、F. Liu、V. Madhavan和T. Darrell，《BDD100K：用于异构多任务学习的多样化驾驶数据集》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2020，第2636–2645页。[81] J. Carreira、E. Noland、A. Banki-Horvath、C. Hillier和A. Zisserman，《关于Kinetics-600的简短说明，01340，2018 1808》，arXiv预印本，arXiv: . .[82] A. Dosovitskiy、G. Ros、F. Codevilla、A. Lopez和V. Koltun，《CARLA：开放式城市驾驶模拟器》，载于机器人学习会议。PMLR，2017，第1–16页。