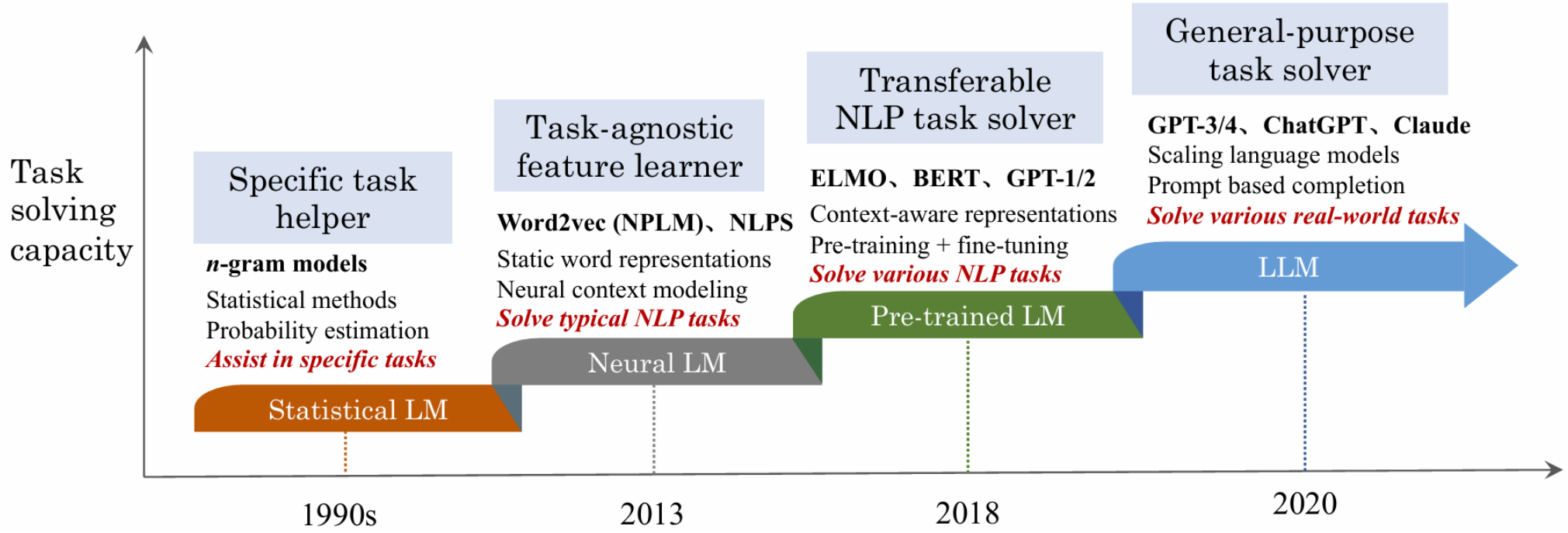
智能驾驶中的大语言模型及算法集成

田炜，同济大学

目录

大语言模型（LLM）

- **语言模型：建模词语之间的概率关系，让机器理解和生成符合人类语言习惯的文本**
- 统计语言模型（1990s）：基于马尔可夫假设的词预测模型；
- 神经语言模型（2013）：通过神经网络预测单词序列的概率；
- 预训练语言模型（2018）：在大规模未标记语料库上进行预训练的Transformer架构模型；
- 大语言模型（2020）：具有大规模参数量和海量训练语料的预训练语言模型，具有**尺度规律**和**涌现能力**。



从最初辅助特定任务，到解决通用NLP问题，再到如今的跨领域通用任务，语言模型的能力不断扩展。

[arXiv:2303.18223]

大语言模型（LLM）

大语言模型发展里程碑

2017年

- 《Attention is All You Need》一文提出**Transformer架构**

  注意力机制高效捕捉长距离依赖关系，并行化计算极大提高训练效率。**为大语言模型奠定基础。**

2018年

- GPT1发布

  采用**Decoder-only架构**，通过**自回归**语言建模进行训练。标志着**预训练语言模型兴起。**

2020年

- GPT3发布

  具有前所未有的**巨大参数和训练数据规模**，展现出强大的少样本学习能力以及**涌现能力。**将“规模”的重要性推向了前所未有的高度。

2022年

- ChatGPT发布

  引入**基于人类反馈的强化学习**，能够进行符合人类交流习惯的**多轮对话。**推进对话式人工智能发展，引发全球AI热潮。

2025年

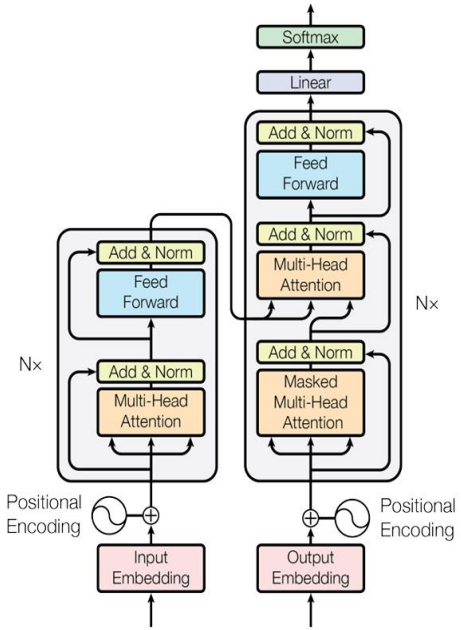
- DeepSeek-R1发布

  提出结合**混合专家**和强化学习等技术，在**成本效益**方面取得了巨大飞跃。为大语言模型的更广泛实际应用提供了重要路径。

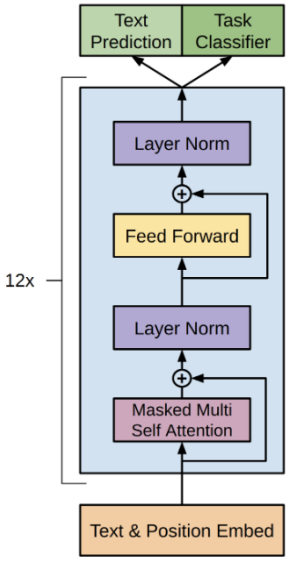
大语言模型（LLM）

大语言模型的网络结构

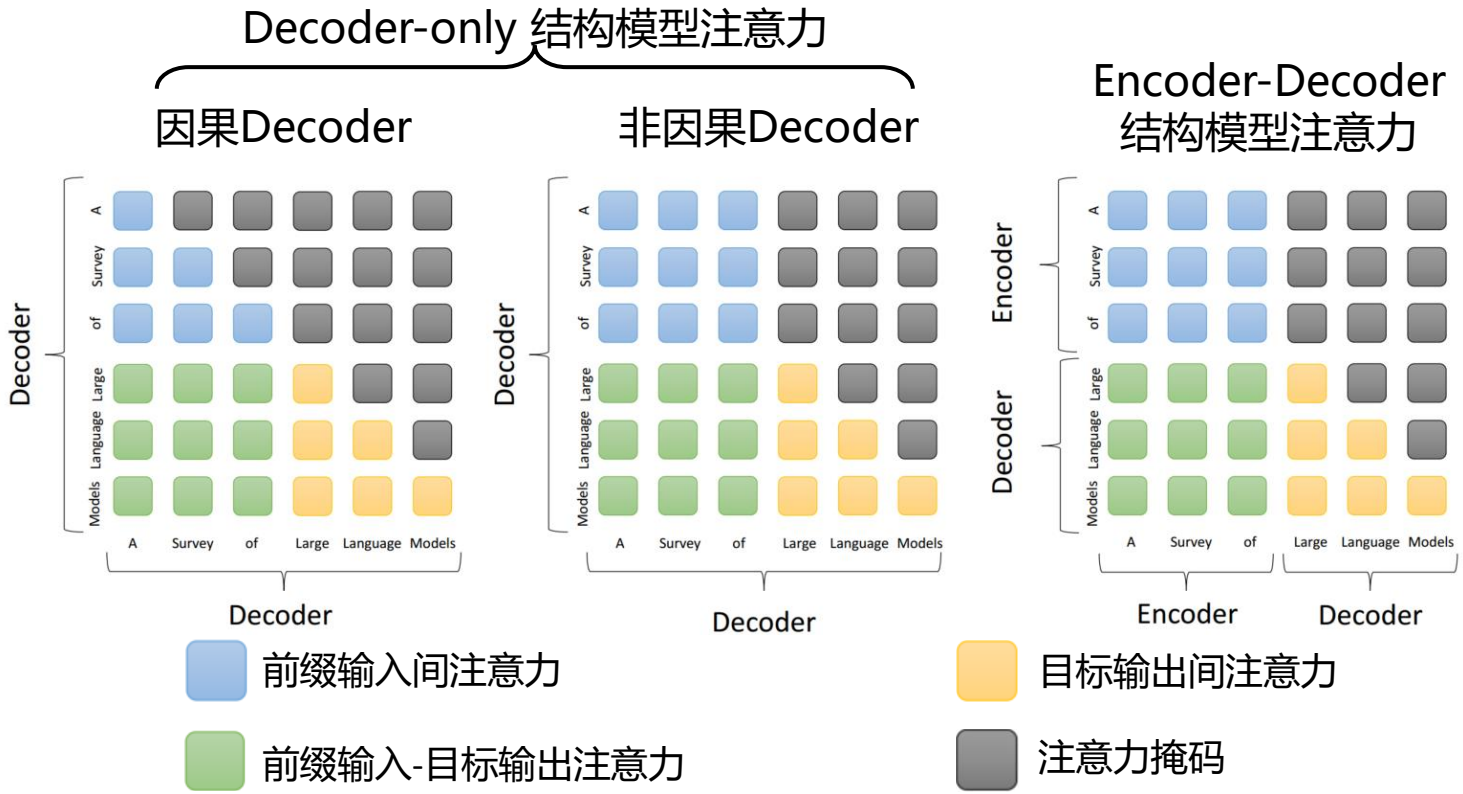
- 由于出色的并行性和容量，Transformer已经成为大模型事实上的标准结构；
- 根据注意力计算类型，可细分为**编码器-解码器**结构、因果**解码器**结构和非因果解码器结构。



Encoder-Decoder结构



Decoder-only结构



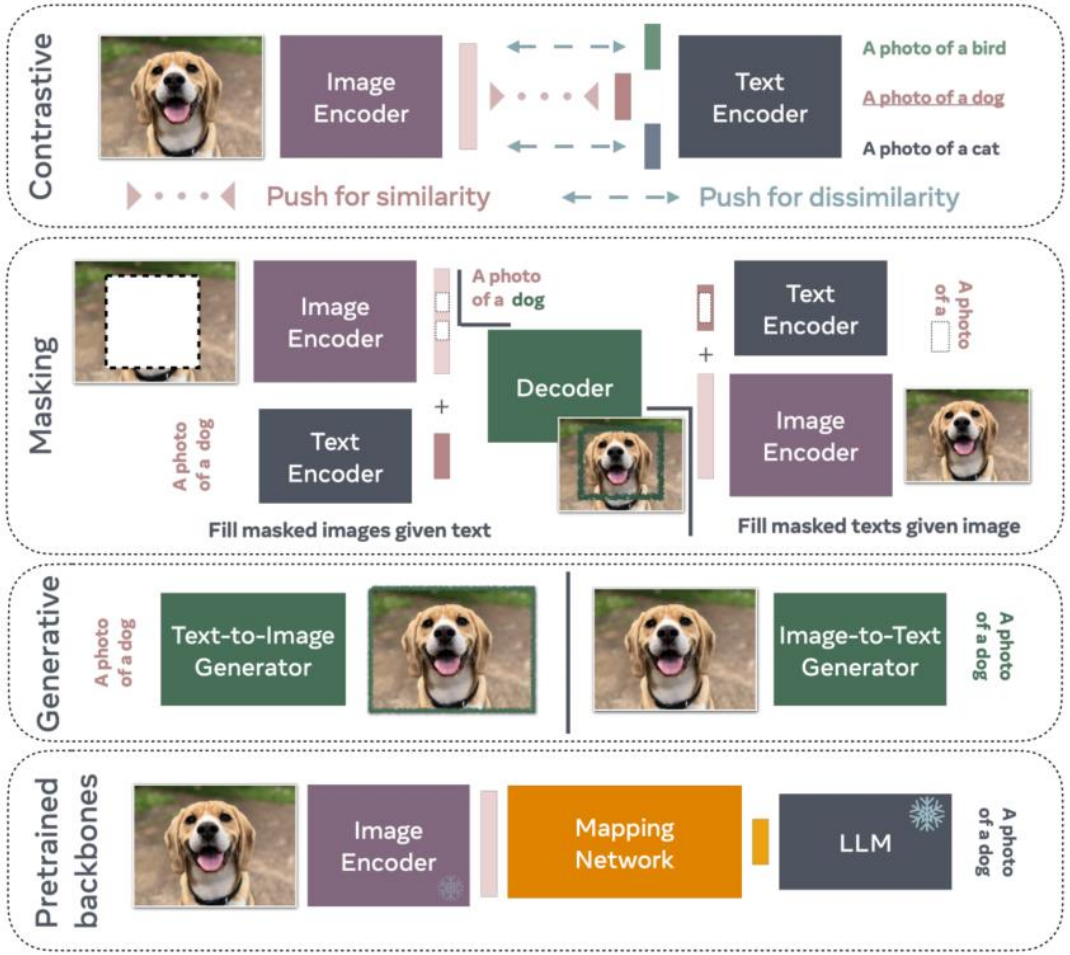
由于在训练效率、泛化能力、工程实现等方面的综合优势，因果Decoder结构被普遍采用

[arXiv:1706.03762, arXiv:2303.18223
https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035]

视觉语言大模型（VLM）

将视觉模态注入大语言模型

桥接视觉-语言域的四类训练方式



对比学习

最大化成对图像文本对的相似度，同时最小化与所有非成对图像文本对的相似度。

掩码建模

随机遮盖图像或文本，要求模型根据上下文信息来预测被遮盖内容。

生成式训练

同时训练一个用于图像理解的编码器和一个用于文本生成的解码器。

预训练骨干

将图像编码器的视觉特征转换成LLM能够理解的token，与文本token一起送入LLM进行推理。

通过预训练骨干和投影网络将视觉模态注入大语言模型成为近期最主要的方式

[arXiv: 2405.17247]

视觉语言大模型（VLM）

投影网络的实现方式

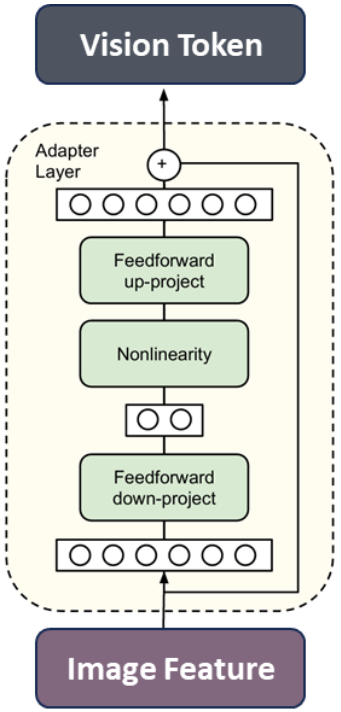
基于对齐的Adapter

假设：预训练的视觉编码器和语言模型功能强大，只需要将视觉特征空间和语言特征空间对齐即可。

结构：多层感知机

原理：直接映射学习

特点：无指令感知，
结构简单、参数少，
训练简单高效

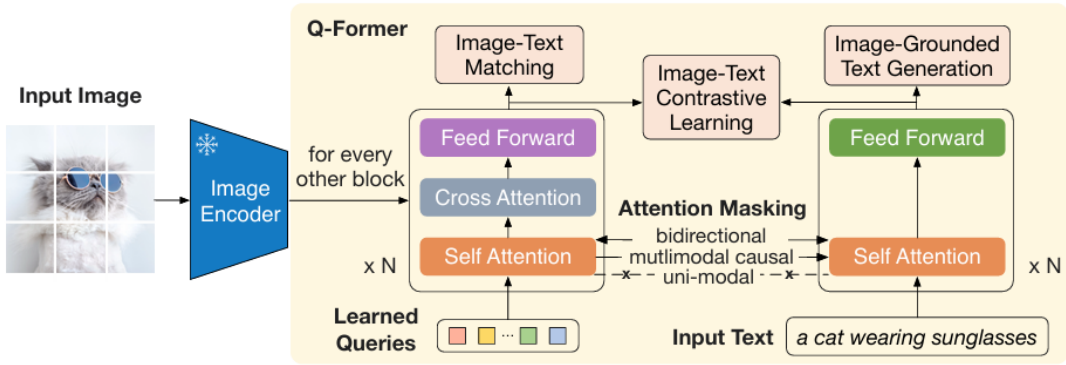


基于提取的Q-Former

结构：小型Transformer

原理：查询交互的图像文本匹配

特点：指令感知，
结构复杂、参数多，
需要多步、多任务训练



[arXiv:1705.08045, arXiv: 2301.12597]

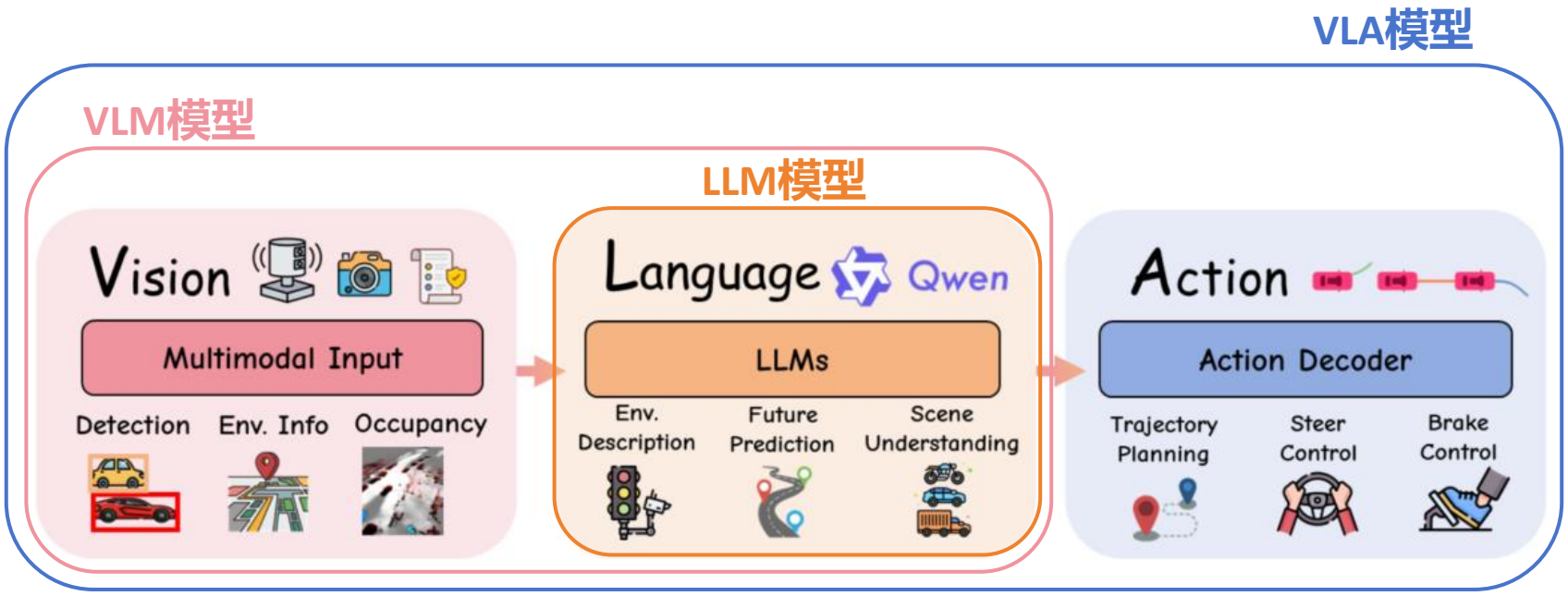
自动驾驶中的大模型

以LLM为中心的“解释者”与“规划者”

“解释者”：由VLM实现，根据视觉和文本输入完成场景描述、解释、用户问答

“规划者”：由VLA实现，根据视觉和文本输入生成可执行的动作或轨迹

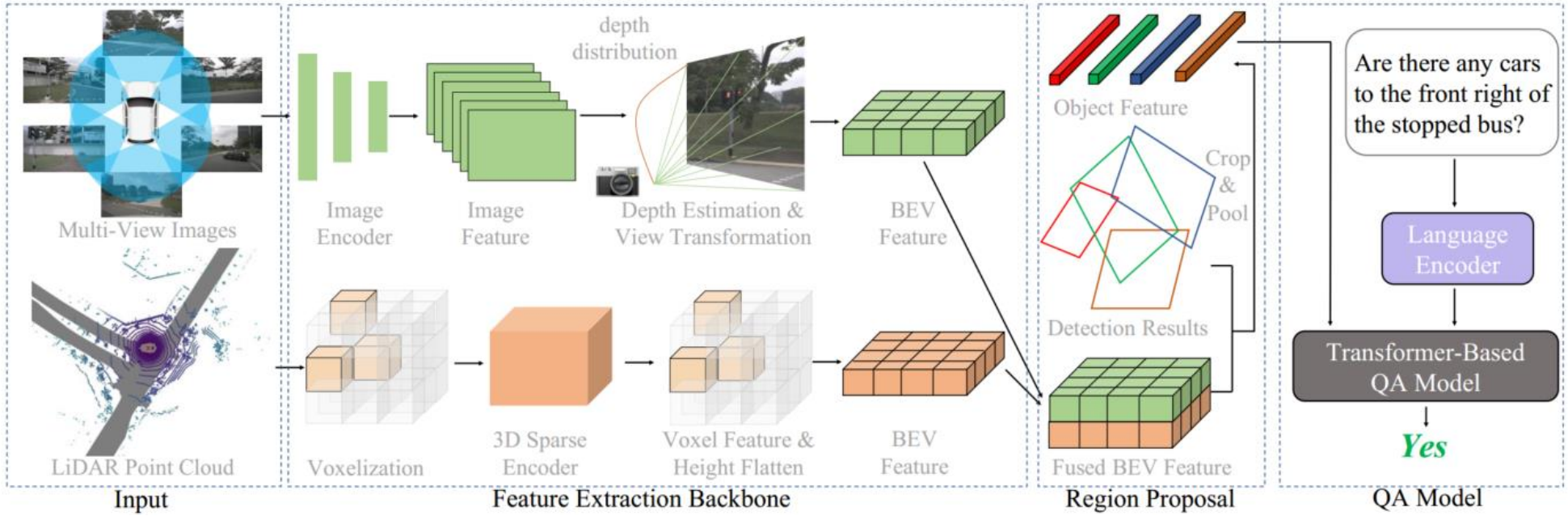
- 在LLM的基础上增加视觉编码器构成VLM
- 在LLM的基础上增加视觉编码器和动作解码器构成VLA



[arXiv:2506.24044]

自动驾驶中的解释者

- **NuScenes-QA：构建了首个面向自动驾驶场景的多模态视觉问答基准**
- 全局输入：多模态视觉输入、问题文本
- 中间表示：多模态融合BEV特征、目标检测结果
- VLM输入：目标特征
- VLM /全局输出：对驾驶场景理解问题的回答



[arXiv:2305.14836]

自动驾驶中的解释者

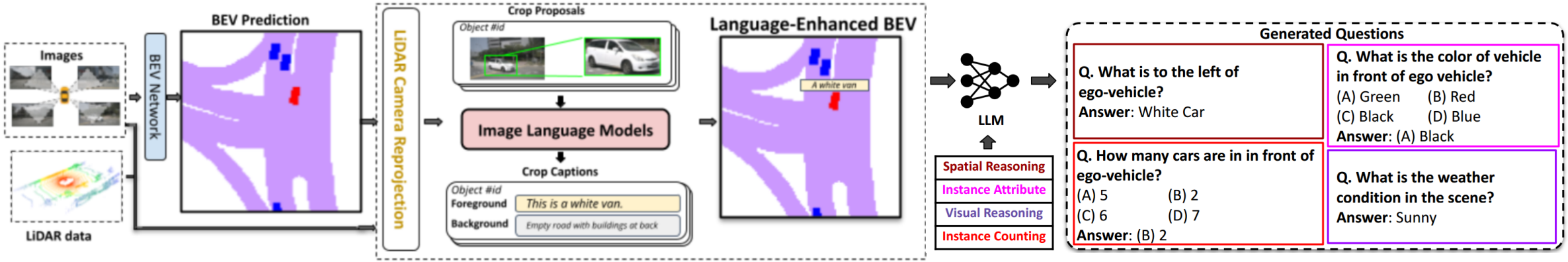
Talk2BEV：基于VLM的语义增强型BEV地图及问答系统，增强了场景交互理解能力

地图构建阶段

- 全局输入：多模态视觉输入
- 中间表示：BEV地图的图像化
- VLM输入：目标的局部BEV图像
- VLM输出：目标的文本描述
- 全局输出：语义增强型BEV地图

地图问答阶段

- VLM/全局输入：语义增强型BEV地图、问题文本
- VLM/全局输出：视觉问答



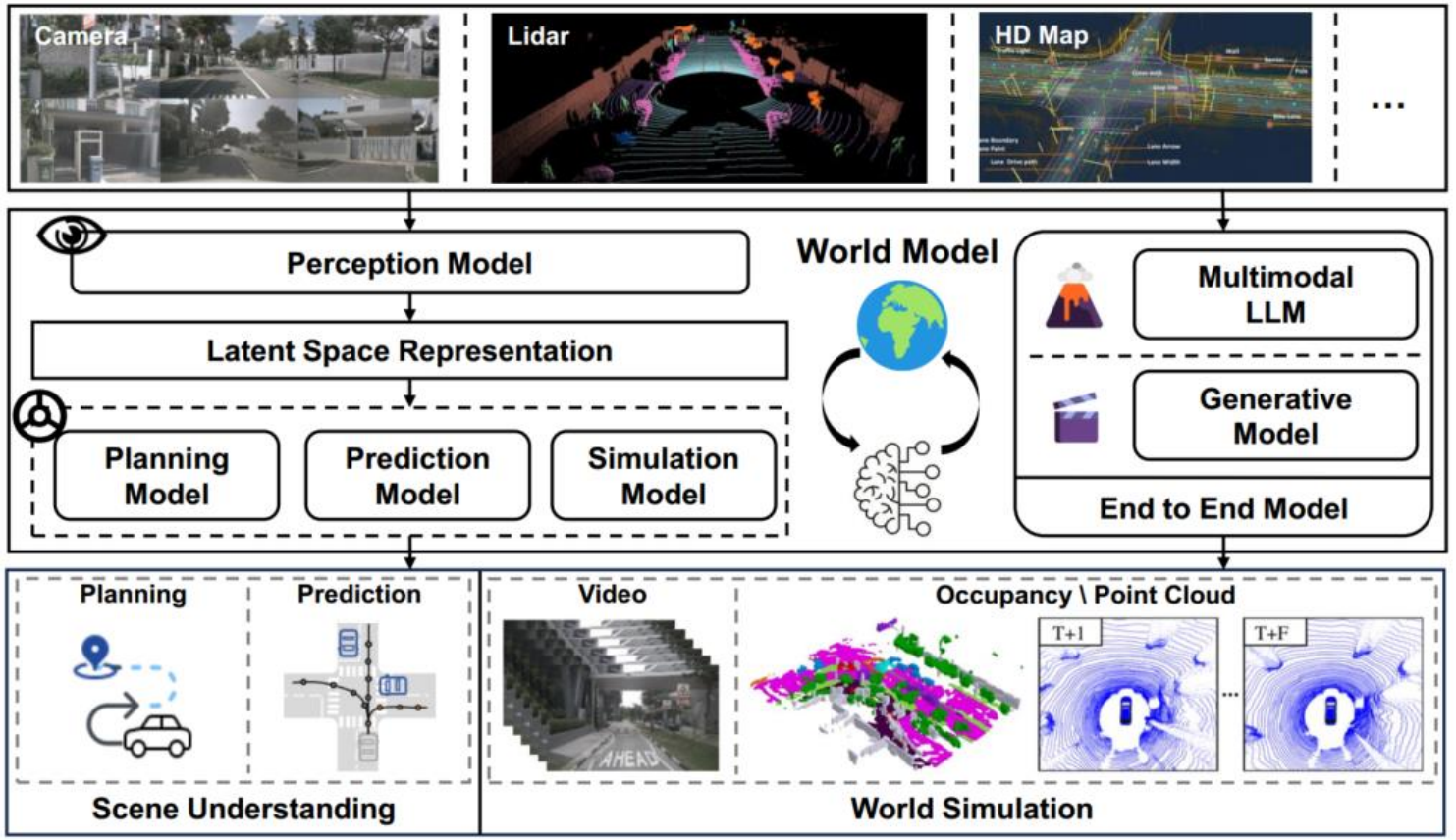
早期工作中，大模型和其他网络结构的结合较为松散，作为解释者对驾驶场景问题做出推理，回答通常仅限于固定格式，比如选择、是非判断、简单的词组回答。

[arXiv:2310.02251]

结合世界模型的自动驾驶大模型

自动驾驶任务：学习车辆外部环境的表示，并进行预测和规划

- 大模型：为自动驾驶系统注入常识和认知智能，提升系统的场景理解和推理能力
- 世界模型：为自动驾驶系统构建一个能够预测环境动态变化的模型



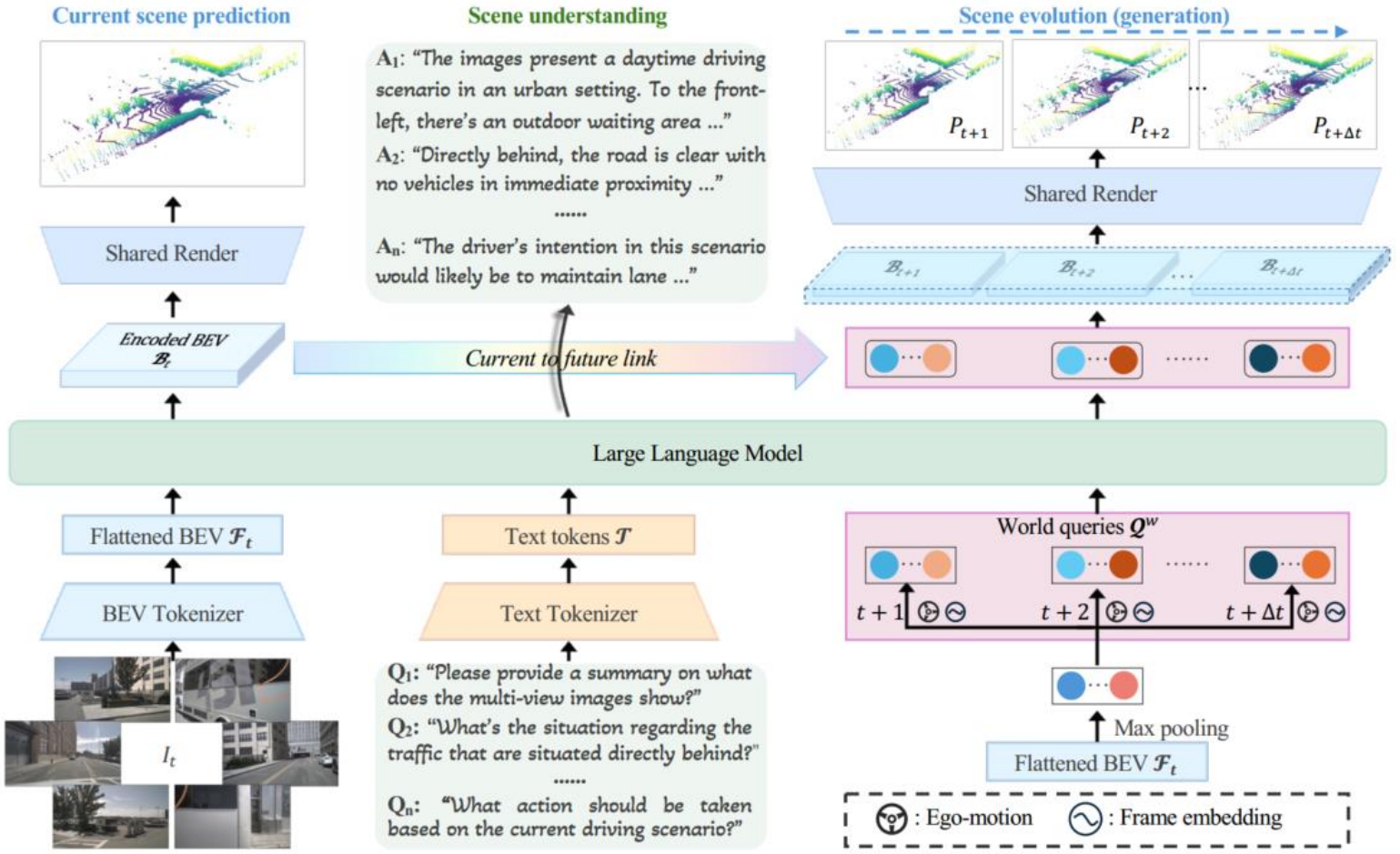
	大模型	世界模型
核心机制	概率统计 模式匹配	动态模拟 因果推理
表示方式	离散token	低维压缩向量
推理方式	基于上下文的概率预测	基于物理规则的模拟推演
物理理解	隐含于数据中，不可解释	显式建模物理规律

[arXiv:2411.14499]

结合世界模型的自动驾驶解释者

HERMES：3D场景理解与未来预测的统一模型

- 全局输入：环视图像、问题文本
- 中间表示：BEV Token、世界查询
- LLM输入：BEV Token、文本Token、世界查询
- LLM输出：BEV Token、世界查询、对驾驶场景理解问题的回答
- 全局输出：未来帧预测



世界模型实现

- 基于BEV特征和世界查询的编码
- 共享渲染器生成未来预测

[arXiv:2501.14729]

目录

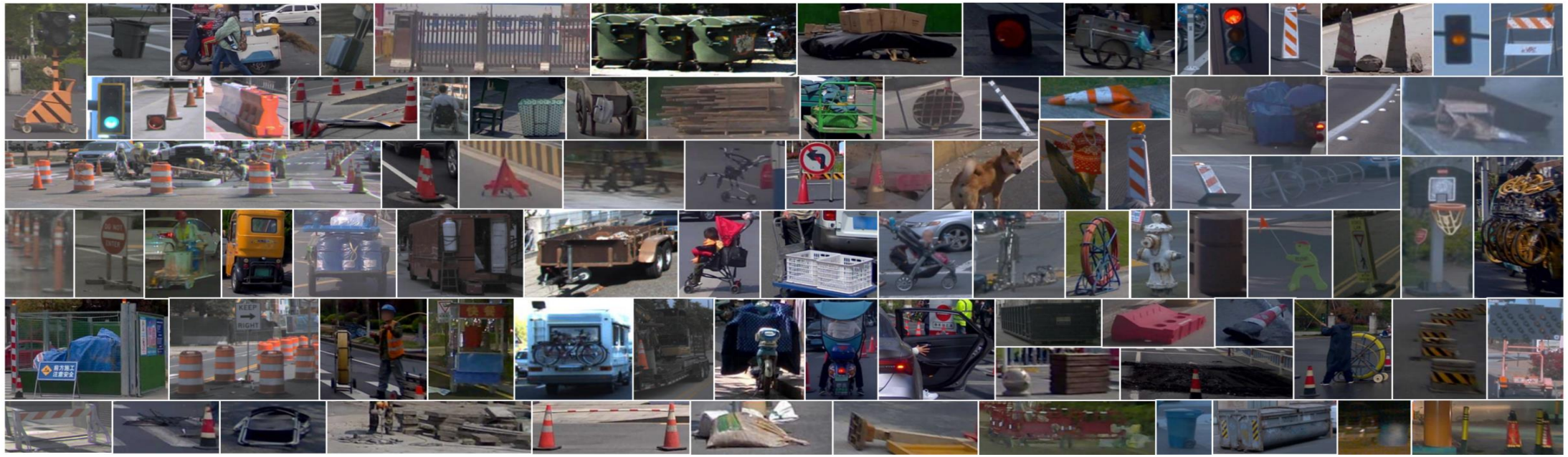
现实场景需求驱动

➢对新型和罕见物体识别的依赖

现实道路场景复杂多变，会出现训练集中未涵盖的物体，如**新型交通工具、特殊路障**，传统目标检测方法难以识别这些物体。

➢减少标注依赖，快速适应新场景

收集和标注大量自动驾驶数据成本**高昂且耗时**，迫切需要一种在遇到新的道路标志或物体时，无需大量新标注数据，依靠已有的**预训练知识**和少量样本即可进行检测的方法。



开放词汇概念引入

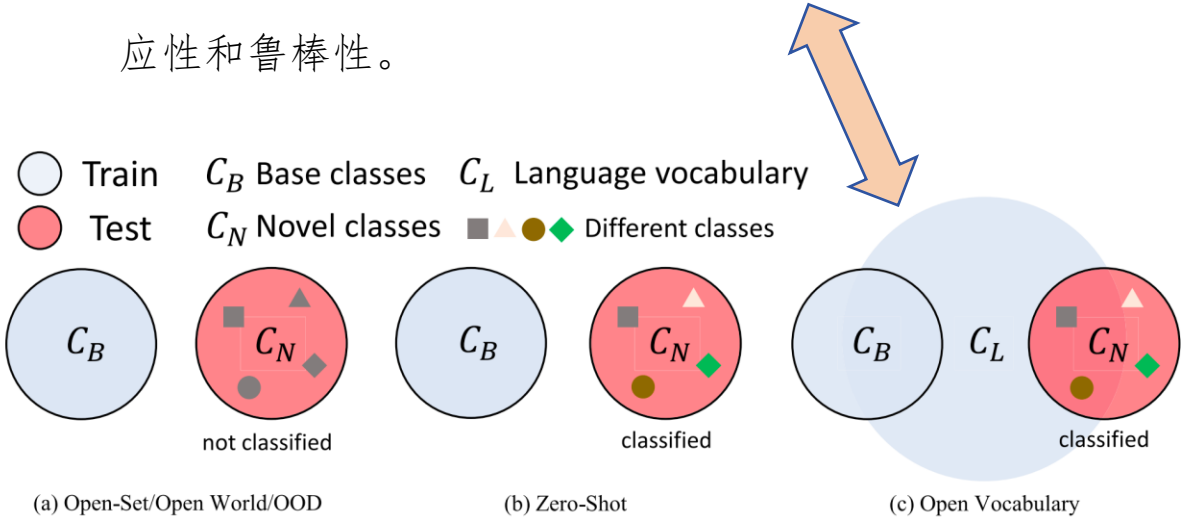
为解决新类别拓展问题，研究者提出了**开集学习**（Open-Set Learning）、**零样本学习**（Zero-Shot Learning）等新概念，最终推动了开放词汇学习的诞生。

任务类型	核心目标	是否依赖LLVK
开集 / 开放世界 / OOD	仅需识别出新类，无需对新类进行进一步分类	否
零样本学习	需将新类分类到具体类别，但无相关语言知识辅助	否
开放词汇学习	可借助大规模语言词汇知识对新类进行分类	是

LLVK表示Large language vocabulary knowledge

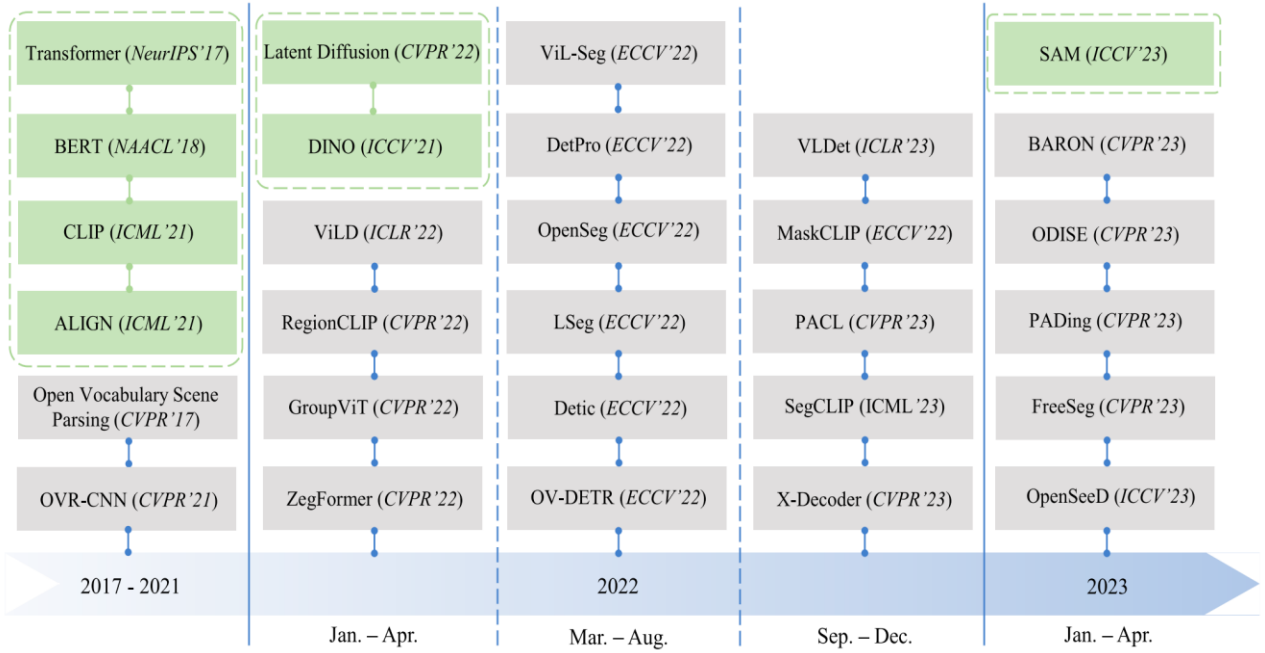
➢开放词汇目标检测任务定义：

开放词汇目标检测（Open-Vocabulary Object Detection）是一种目标检测技术，它能够在训练阶段未见过的类别上进行准确的**检测和识别**。这种技术的核心在于利用有限的标注数据和先验知识，使模型能够泛化到新的、未见过的类别，从而在开放世界场景中具备更强的适应性和鲁棒性。



2D开放词汇目标检测任务

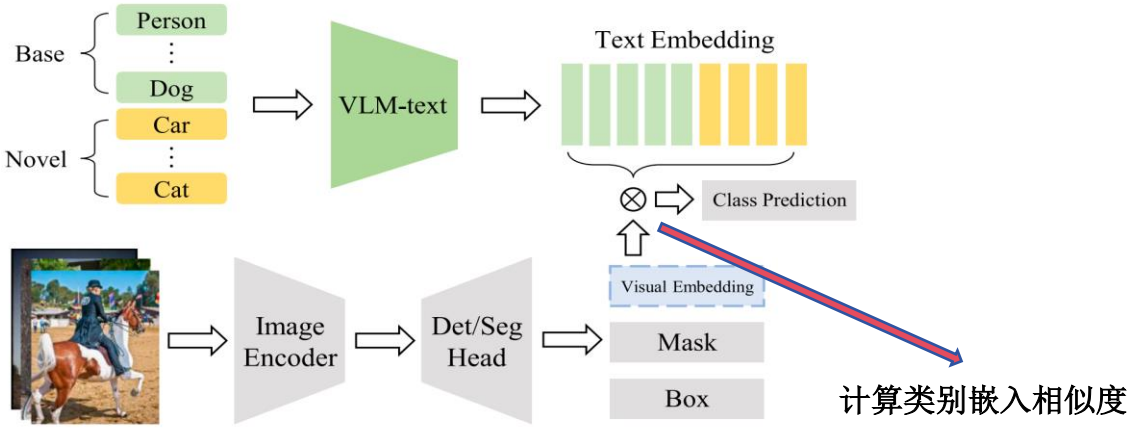
从视觉基础模型与视觉语言模型出发



在开放词汇学习领域，众多研究利用SAM等预训练视觉基础模型和CLIP等视觉语言模型习得的知识。

➢核心思想

视觉模型为每个边界框预测**类别嵌入向量**，这些嵌入向量通过**点积运算**与CLIP等视觉语言模型**文本模块**生成的**类别嵌入向量**集进行比对。得分最高的类别将被选作该物体的预测标签。

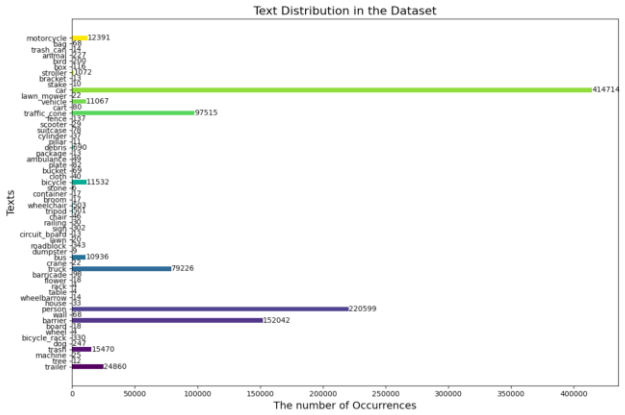


2D开放词汇目标检测一般范式架构示意图

3D开放词汇目标检测工作的不足

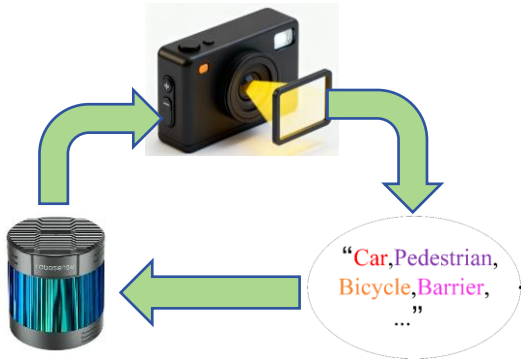
现有的面向开放词汇的3D目标检测算法（Open3DWorld，OV-Uni3DETR等）在识别未知类别物体方面取得了一定进展，然而仍然存在以下问题：

高成本3D标注数据获取



点云-文本数据对的采集与标注难度颇高，标注过程繁杂，成本非常高，难以进行大面积的标注。

多模态特征融合困难



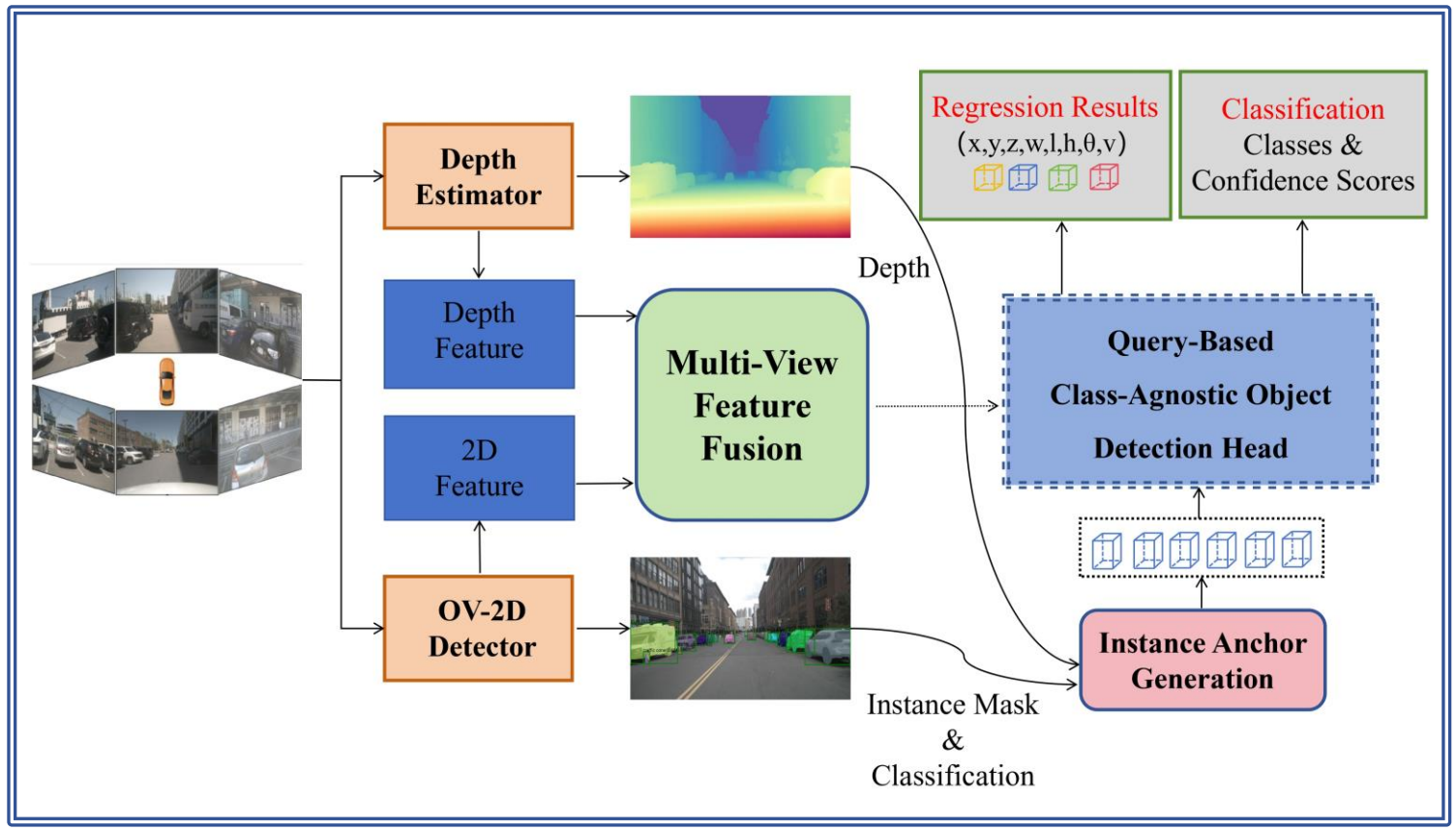
现有方法难以有效将3D几何信息与2D语义信息融合，致使在处理开放词汇时难以维持鲁棒性。

小目标检测精度差

小目标特征稀疏，易受噪声干扰与背景掩盖，现有方法难以有效捕捉其细节信息。

OVSparse

OVSparse：基于强先验的多视图3D开放词汇目标检测框架



OVSparse架构图

三个方面的贡献：

- 提出一种面向自动驾驶场景的开放词汇多视图3D目标检测方法；
- 设计基于强先验的实例锚点初始化方法，提升开放词汇类别的检测性能；
- 构建类别无关的检测头，通过解耦分类与回归任务，为开放词汇场景下的类别扩展提供自然支持。

基于视觉基础大模型的先验助力

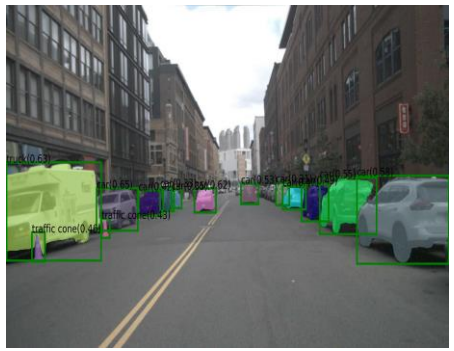
三大视觉基础模型（Grouding DINO、SAM、UniDepth V2）

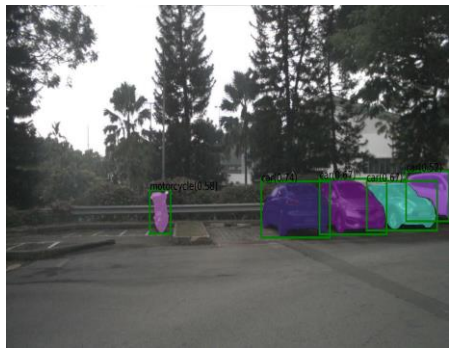
由于缺乏3D的类别标注信息，使用2D开放词汇检测器作为2D先验，如Grouding DINO。

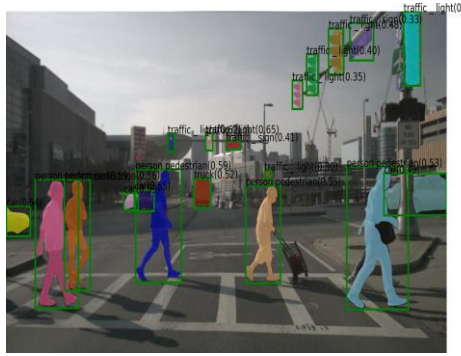
获取深度信息

分割实例掩码

结合SAM，UniDepth V2

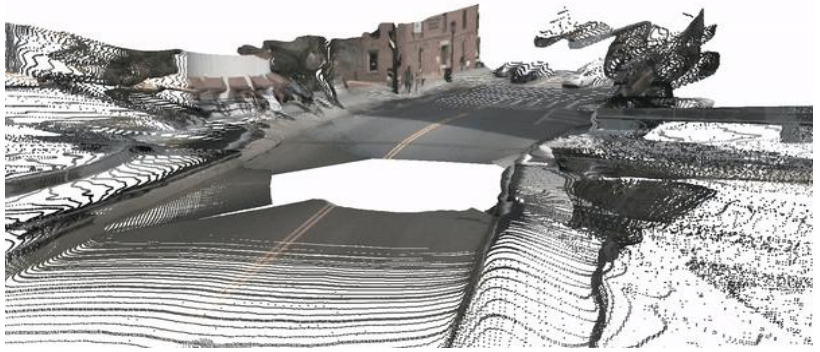






该步骤可精准获取所有实例的2D位置信息（像素级掩码）与类别标签，为后续3D目标检测提供关键数据支撑。

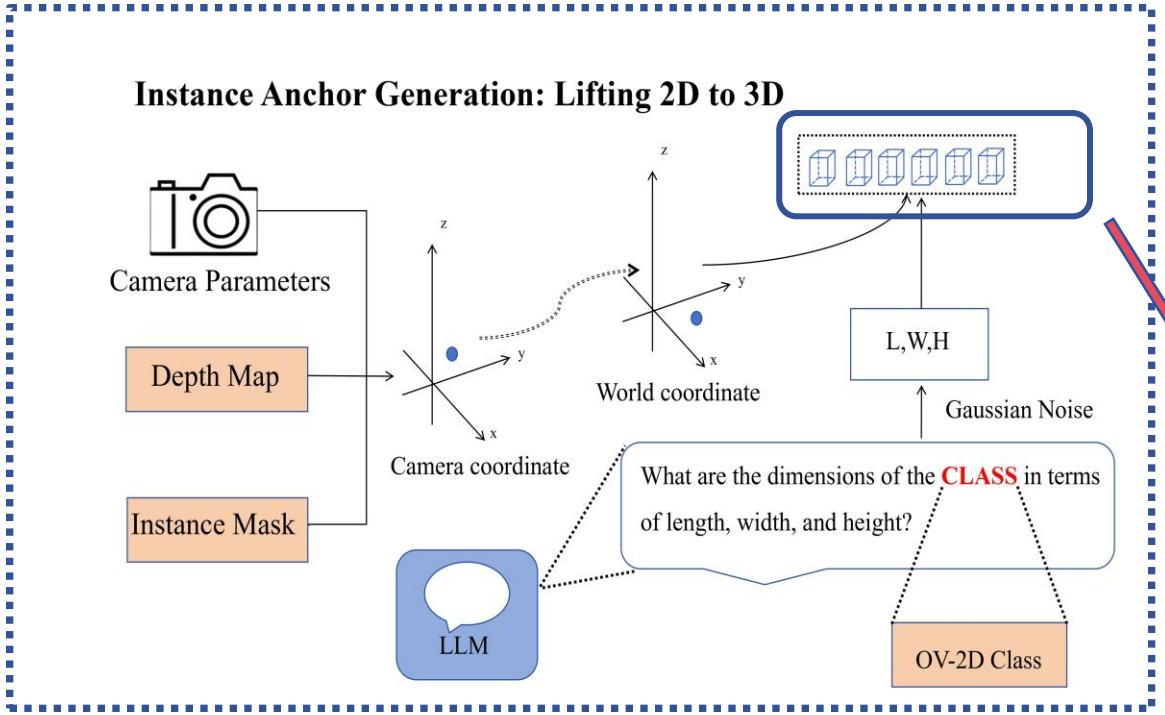
使用UniDepth V2在nuScenes数据集上的深度估计结果



使用UniDepth V2在nuScenes数据集上的场景重建结果

基于大语言模型的知识先验

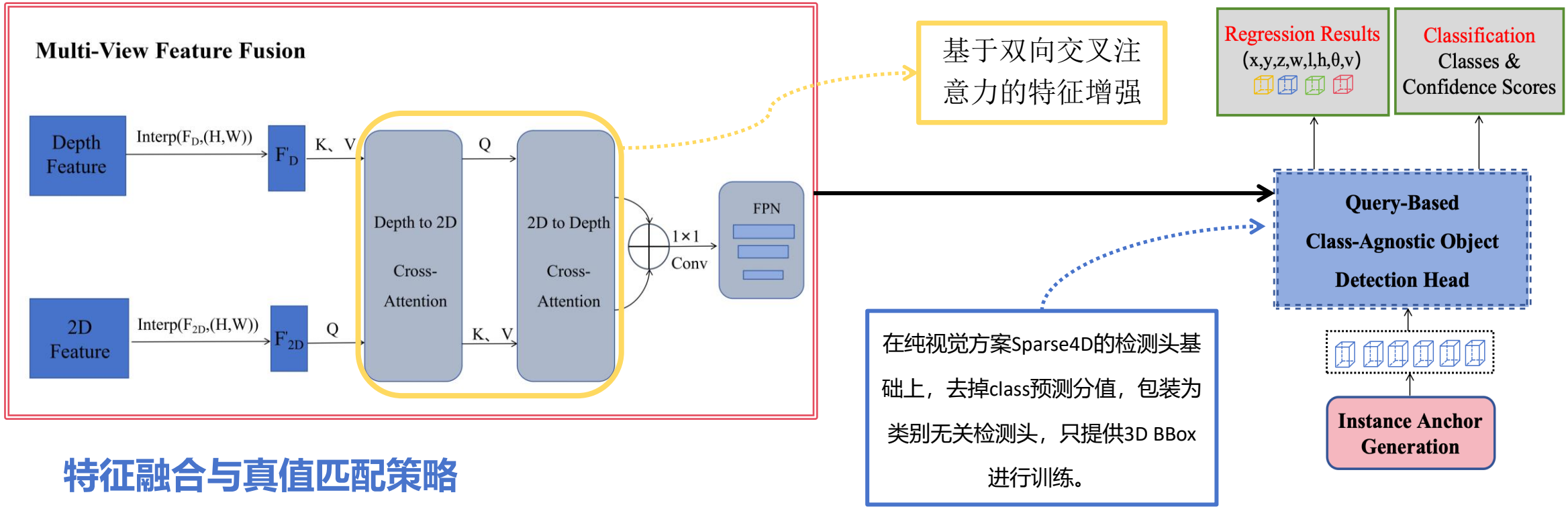
从大语言模型的尺寸先验到实例Anchor的生成



基于强先验的Anchor初始化生成结果

- 通过查询**大语言模型**，获取每个类别的尺寸，并据此生成**尺寸模板**；
- 结合**相机内外参矩阵**，完成Anchor的几何属性初始化与坐标系转换工作。

特征融合与类别无关检测头



特征融合与真值匹配策略

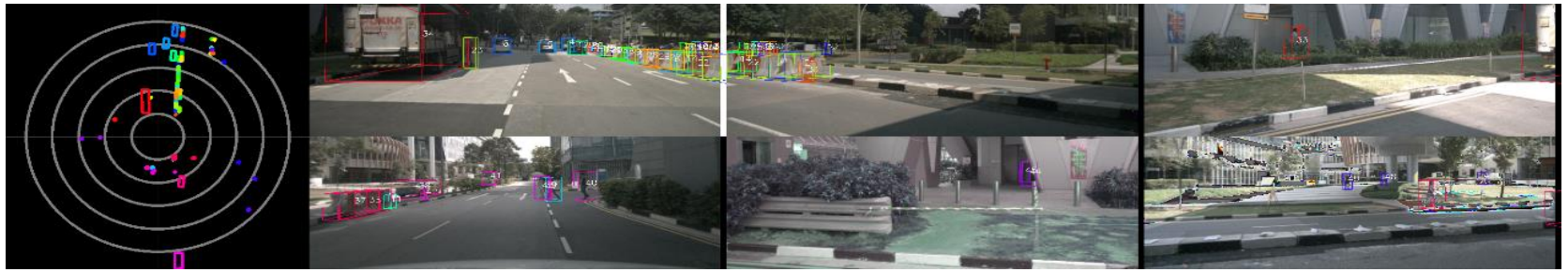
- 深度特征-2D图像特征的**双向交叉注意力**， 以互为查询的形式增强显著的2D图像特征以及深度特征，最后通过综合特征的形式输入后续网络；
- 在稀疏特征**初始化阶段**直接建立特征与真实目标之间的映射关系，让模型从训练的**第一次迭代**即可获得稳定的监督信号。

特征融合与类别无关检测头

匹配策略调整前的匹配框

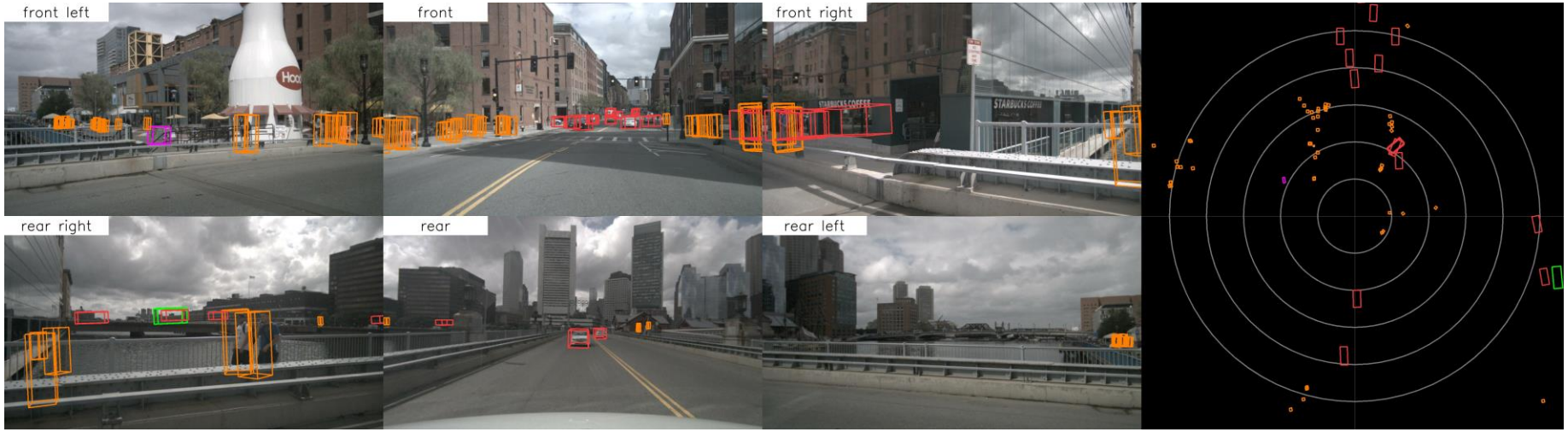
匹配策略调整后的匹配框

真值框



调整匈牙利匹配策略可有效对应预测框与真值框的位置。

对比实验



OVSparse检测结果可视化

TABLE III
COMPARISON OF OVSPARSE WITH OTHER METHODS.

Method	Modality	mAP (%)	NDS (%)
OV-PointCLIP [21]	L	0.1	5.0
OV-PointCLIP V2 [22]	L	2.8	10.9
CLIP-3D [12]	L	7.2	14.2
OV-3DET [13]	L	5.7	12.0
CoDA [14]	L	10.3	16.1
OV-Uni3DETR [23]	L	15.48	15.61
OpenSight [24]	L	23.5	24.0
OVMONO3D-GEO [25]	C	8.56	10.1
OVMONO3D-LIFT [25]	C	14.39	12.0
OV-Uni3DETR [23]	C	12.54	14.67
OpenAD-Ens [26]	C	12.36	14.02
OVSparse (Ours)	C	21.74	22.56

- **与激光雷达（Lidar）的方法相比：**

OVSparse的mAP（21.74）和NDS（22.56）仅次于当前最优的Lidar方法OpenSight（23.5 mAP/24.0 NDS）。

- **与相机图像（Camera）的方法相比：**

OVSparse 展现出当前最先进的性能，其 mAP 比排名第二的 OVMONO3D-LIFT 高出 7.35%，NDS 领先 10.56%。

目录

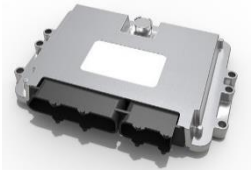
研究背景

人体行为识别(Human Behavior Recognition, HAR)主要通过计算机视觉技术，自动识别人体动作与行为模式，在智能安防、人机交互、医疗康复、智能交通等领域具有广泛应用前景。


视觉信息


行为识别模型

行为分类


整车控制单元


车内驾驶员行为识别


车外行人穿行行为识别

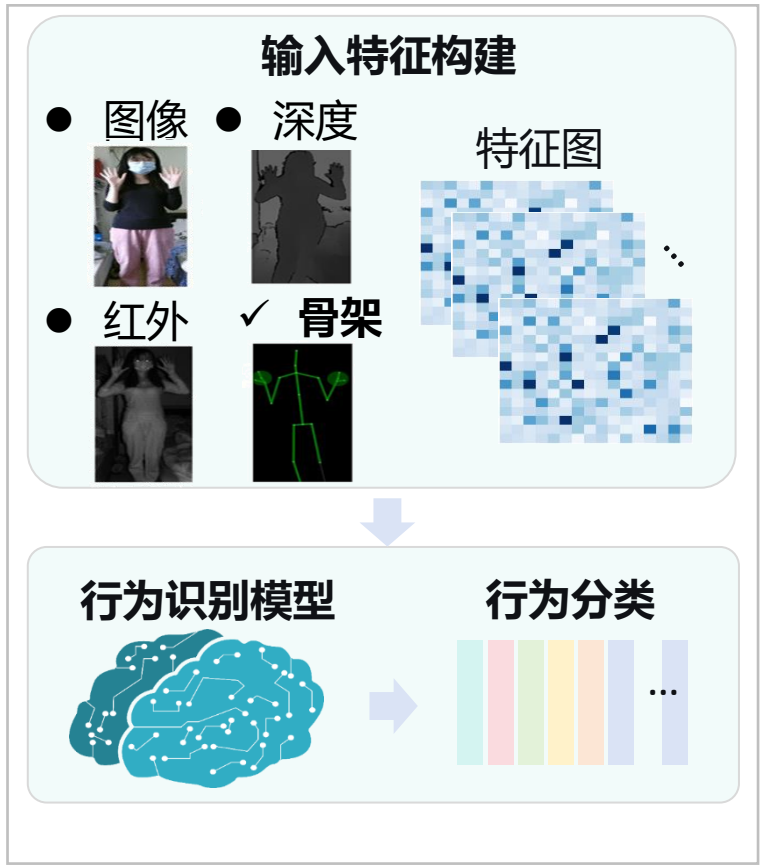
预警提示

整车控制

对驾驶员、车外行人等交通场景下的人体行为进行识别，对提高智能交通安全性具有重要意义

研究背景

基于骨架信息的人体行为识别是主流方法，通过监督学习在各大基准测试中取得了领先性能，并对人体外观与背景环境的变化具备出色的鲁棒性。



挑战1：骨架信息忽略局部细节，部分行为易混淆

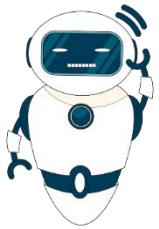
开车喝水

开车吃零食

挑战2：数据集制作难度大(涉及隐私、标注成本高等)，未知行为泛化能力弱

数据集制作涉及隐私

数据标注成本高

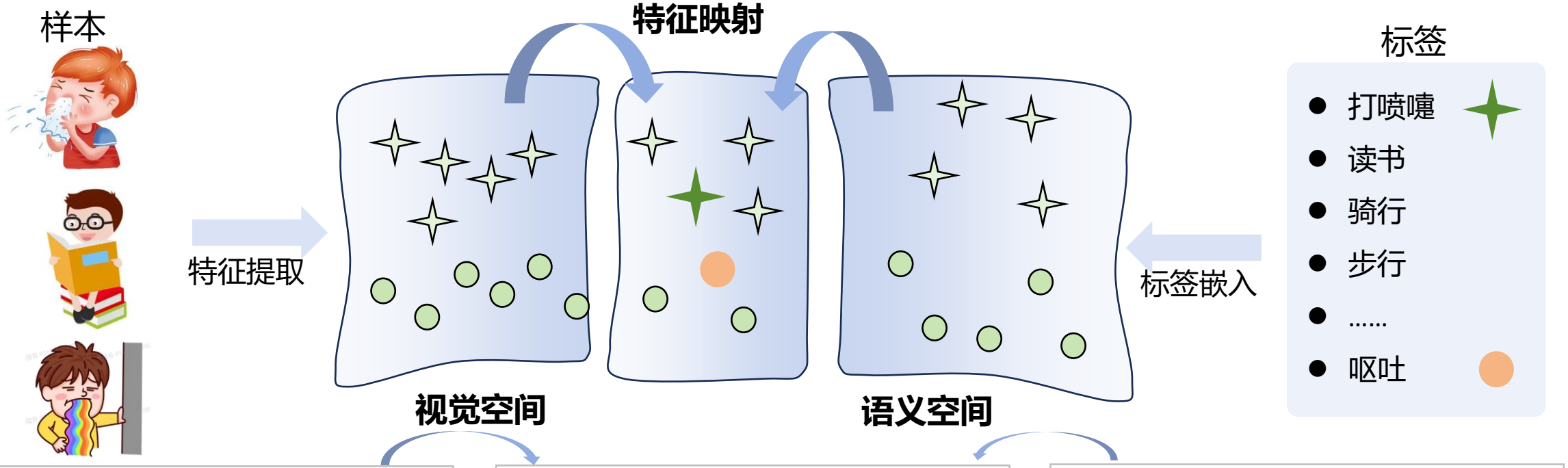


未知行为无法识别

迫切需要提高模型特征提取能力以及未见行为类别泛化能力

研究背景

零样本学习（Zero-Shot Learning, LSL）旨在识别在训练中未出现过的人体行为，其核心是借助**语义属性**，将已学知识从**已知行为迁移至未知行为**，进而实现未知行为识别。



视觉特征

基于骨架数据，利用图卷积网络提取空间姿态特征

特征映射

将视觉特征与语义特征映射至共享潜在空间，以实现表征对齐

文本特征

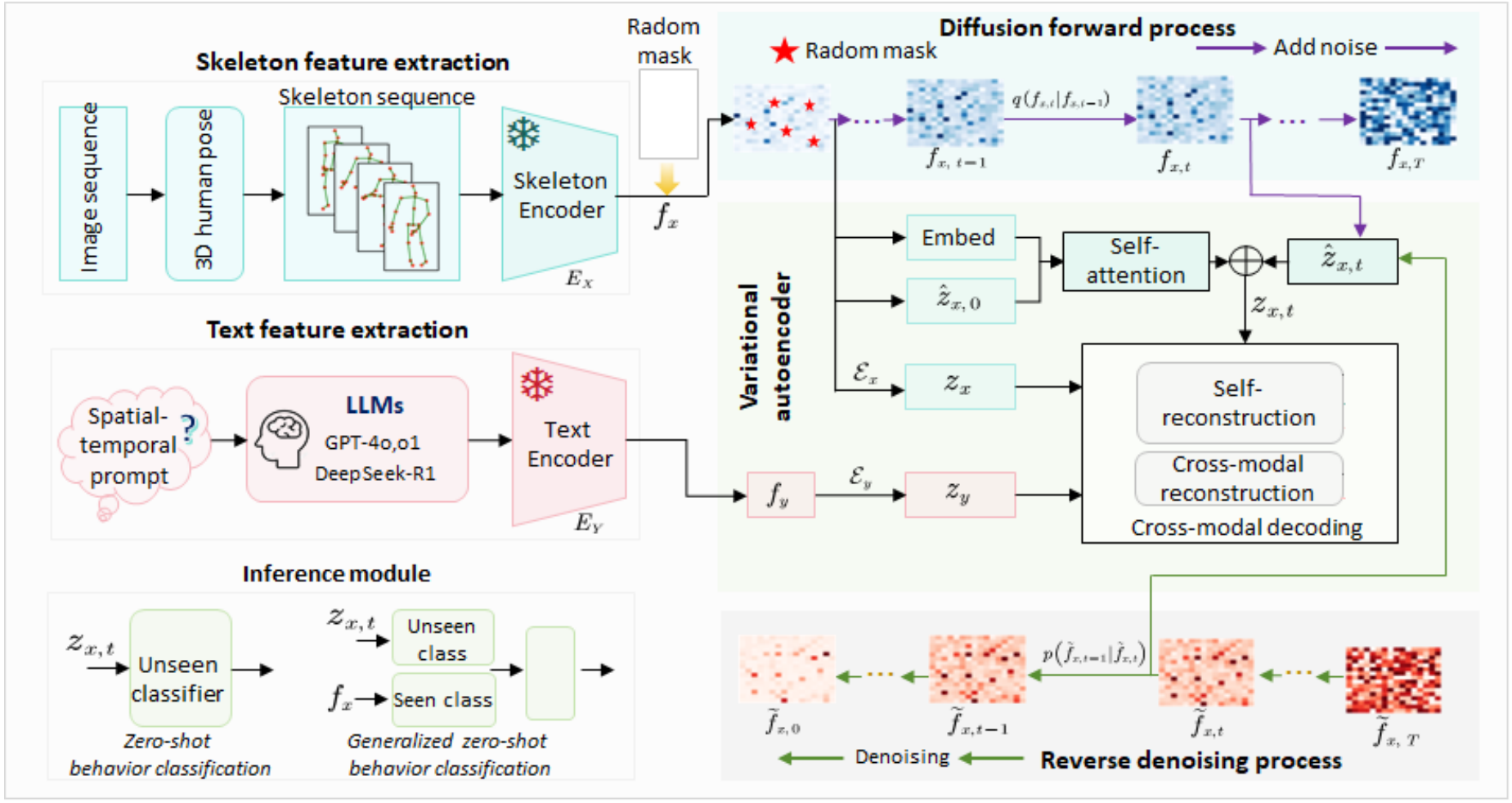
借助行为的语义描述，通过编码转化为特征表示

文本特征的获得——大语言模型通过海量语料训练，不仅具有**生成人体行为描述的“丰富经验”**，还能够将语言描述**映射到代表其语义信息分布的特征空间**

所提方法

方法框架：基于跨模态扩散模型的零样本人体行为识别

- **跨模态特征提取**：骨架特征提取模块、文本特征提取模块、随机掩码
- **跨模态映射对齐**：扩散模型建模、变分自编码器
- **行为分类推理**：零样本行为分类、广义零样本(全样本)行为分类



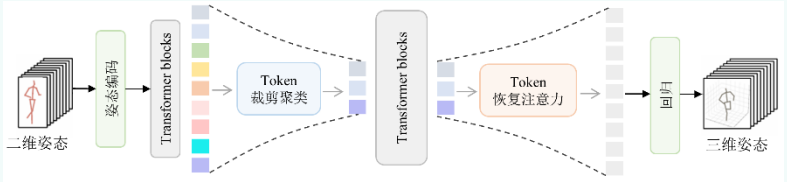
所提方法

跨模态特征提取：骨架特征提取编码（随机掩码）+文本特征提取编码

骨架模态特征提取

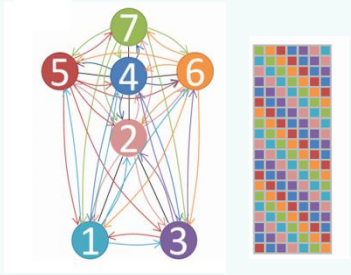
人体3D姿态获取

Kinect V2相机



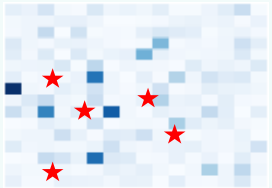
沙漏型标记器三维人体姿态估计

3D姿态编码



Shift-GCN

随机掩码



$$f_{x,0} = (\mathbf{1} - \mathbf{M}_\alpha) \odot f_x + \mathbf{M}_\beta \odot \epsilon$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

缓解信息冗余

文本模态特征提取

行为本文描述获取

- **身体解剖划分**：头部、手部、躯干、腿部
- **时间阶段划分**：开始、中间、结束
- **全局行为描述**
- **环境语境描述**

- **LLM**：GPT-4o, GPT-o1, Deepseek

文本特征编码

All-MiniLM-L6-v2编码器

$$f_y = [\mathbf{t}_t^{(1)}, \ldots, \mathbf{t}_t^{(9)}] \in \mathbb{R}^{9 \times 384},\ \mathbf{t}_t^{(k)} = \mathbf{z}(S_t^{(k)}) \in \mathbb{R}^{384}$$

所提方法

基于GPT-o1对易混淆“打喷嚏”“呕吐”行为生成的文本描述示例

Q1: For < Behavior X>, please provide movement descriptions of four parts: head, hands, trunk and legs.

	C41 sneeze	C48 throw up
Head	Moves forward abruptly, often turning away or downward to avoid spreading droplets.	Bows forward or over a receptacle, face contorted in discomfort.
Hands	Covers mouth and nose with one hand or uses a tissue to shield expelled air.	Covers mouth or braces on a nearby surface to catch expelled contents.
trunk	Contracts slightly with the force of the sneeze or cough.	Contracts in a heaving motion, often doubling over.
legs	May brace slightly to maintain balance during the reflexive movement.	Might tremble or have knees slightly bent to maintain stability.

Q2: For < Behavior X>, please provide movement descriptions at the beginning, middle and end time stages.

	C41 sneeze	C48 throw up
Start	Prepares to sneeze or cough by inhaling deeply or feeling an urge.	Feels queasy or nauseous, preparing to vomit.
Middle	Expels air forcefully from the lungs, releasing droplets from the mouth and nose.	Contracts the abdominal muscles to expel the stomach contents.
End	Recovers posture, possibly wiping the face or using a tissue after the sneeze or cough.	Reaches for a receptacle like a trash can or restroom, then recovers posture after vomiting.

Q3: For < Behavior X>, please provide description of the overall behavior

	C41 sneeze	C48 throw up
Whole 	Initiates a forceful expulsion of air from the lungs through the nose and mouth to clear irritants or respond to an urge to sneeze or cough.	Contracts abdominal muscles to expel stomach contents forcefully through the mouth in response to nausea or gastrointestinal distress

Q4: For < Behavior X>, please provide the environment in which this behavior is executed.

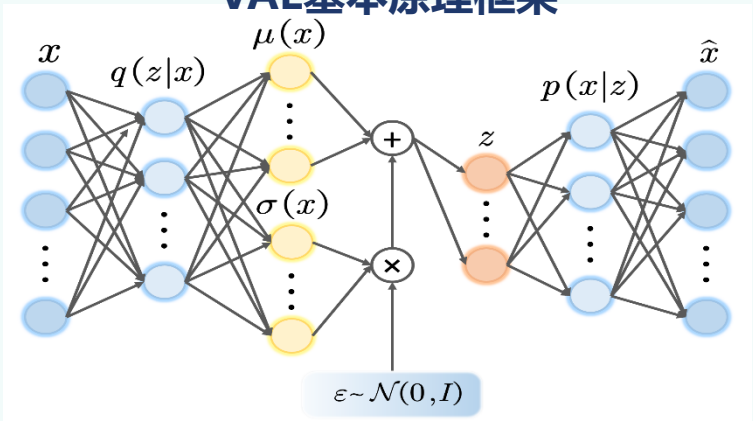
Environment 	Occurs in response to irritants in the respiratory system, allergies, viral infections, or as a reflex to clear the airways from mucus and foreign particles.	Occurs as a bodily response to illness, motion sickness, indigestion, or emotional distress, aimed at expelling harmful substances from the stomach.

所提方法

跨模态映射对齐： 变分自编码器+扩散模型建模

变分自编码器

VAE基本原理框架



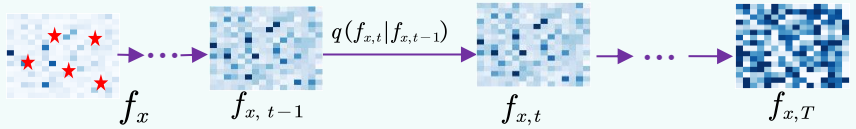
- 双向重建目标来加强跨模态重建一致性
- 隐空间的交互实现跨模态语义对齐

$$f'_x, f'_y = \mathcal{D}_{xs}(z_x), \mathcal{D}_{ys}(z_y),$$
$$f''_x, f''_y = \mathcal{D}_{yc}(z_y), \mathcal{D}_{xc}(z_x),$$
$$\mathcal{L}_{\text{srec}} = \|f_x - f'_x\|_2 + \|f_y - f'_y\|_2,$$
$$\mathcal{L}_{\text{crec}} = \|f_x - f''_x\|_2 + \|f_y - f''_y\|_2.$$

自重建损失
跨模态重建损失

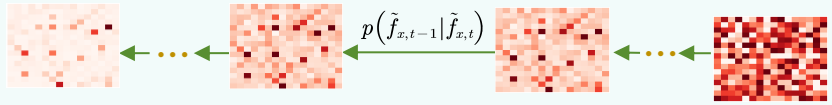
扩散模型建模

正向扩散过程



$$q(f_{x,t} \mid f_{x,0}) = \mathcal{N}\left(f_{x,t}; \sqrt{\bar{\alpha}_t} f_{x,0}, (1-\bar{\alpha}_t)\mathbf{I}\right),$$
$$\alpha_t = (1-\beta_t), \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$$

逆向去噪过程



$$p(f_{x,0}) = N(f_{x,t-1}; \mu_\theta(f_{x,t}), \sigma_t^2 I)$$
$$f_{x_{t-1}} = \sqrt{\bar{\alpha}_t} f_\theta(f_{x_t,t}) + \sqrt{1-\bar{\alpha}_{t-1}} \epsilon_\theta(f_{x,t}, t)$$

扩散重建损失 $\mathcal{L}_{\text{diff}} = \|f_x - \hat{f}'_x\|_2 + \|f_y - \hat{f}''_y\|_2.$

实验结果

■ 与主流方法性能对比

- **领先的SOTA性能**：在NTU RGB+D 60与120数据集的多项标准划分下，**H-score均达到最优**，尤其在更具挑战性的48/12与96/24划分中优势明显。
- **较强的泛化能力**：模型在未见类别数量增加时（如从55/5到48/12），仍具有较高的精度，证明了其优异的语义对齐与跨类别泛化鲁棒性。

NTU RGB+D 60 results

Methods	55/5				48/12			
	ZSL	GZSL			ZSL	GZSL		
	Acc	S	U	H	Acc	S	U	H
ReViSE	53.91	74.22	34.73	47.32	17.49	62.36	20.77	31.16
JPoSE	64.82	64.44	50.29	56.49	28.75	60.49	20.26	30.75
CADA-VAE	76.84	69.38	61.79	65.37	28.96	51.32	27.03	35.41
SynSE	75.81	61.27	56.93	59.02	33.30	52.21	27.85	36.33
SMIE	77.98	-	-	-	40.18	-	-	-
GZSSAR	83.63	71.73	66.15	68.83	49.19	58.80	40.00	47.61
PURLS	79.23	-	-	-	40.99	-	-	-
SA-DVAE	82.37	62.8	70.8	66.3	41.38	50.2	36.9	42.6
STAR	81.40	69.00	69.90	69.40	45.10	62.70	37.00	46.60
Neuron	**86.90**	69.10	73.80	71.40	**62.70**	61.60	**56.80**	59.10
InfoCPL	85.91	-	-	-	53.32	-	-	-
Ours	84.37	**85.78**	**76.03**	**80.61**	62.11	**82.64**	56.36	**67.02**

NTU RGB+D 120 results

Methods	110/10				96/24			
	ZSL	GZSL			ZSL	GZSL		
	Acc	S	U	H	Acc	S	U	H
ReViSE	55.04	48.69	44.84	46.68	32.38	49.66	25.06	33.31
JPoSE	51.93	47.66	46.40	47.05	32.44	38.62	22.79	28.67
CADA-VAE	59.53	47.16	49.78	48.44	35.77	41.11	34.14	37.31
SynSE	62.69	52.51	57.60	54.94	38.70	56.39	32.25	41.04
SMIE	65.74	-	-	-	45.30	-	-	-
GZSSAR	71.20	46.84	68.30	55.57	59.73	56.84	48.61	52.40
PURLS	71.95	-	-	-	52.01	-	-	-
SA-DVAE	68.77	61.10	59.75	60.42	46.12	58.82	35.79	44.50
STAR	63.30	59.90	52.70	56.10	44.30	51.20	36.90	42.90
Neuron	71.50	67.60	59.50	63.30	57.10	67.50	44.40	53.60
InfoCPL	74.81	-	-	-	**60.05**	-	-	-
Ours	**78.47**	**77.02**	**76.02**	**76.52**	57.15	**75.60**	**50.34**	**60.44**

实验结果

■ 消融实验

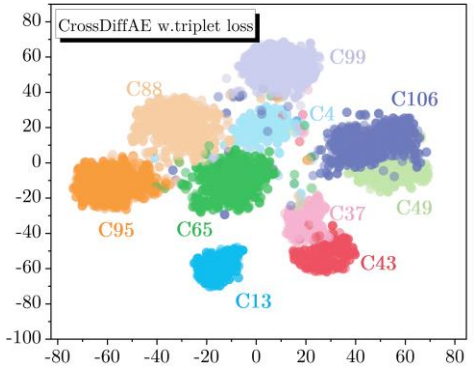
- □ 扩散步数：在所有类别划分下，性能均显著优于SA-DVAE基线模型；即使仅使用**单步采样**，也展现出卓越的有效性。
- □ 三元损失：引入三元组损失后，通过**减小类内方差、增大类间分隔**，有效增强了特征的可判别性，零样本行为识别准确率上升显著，在复杂场景中，它能显著缓解不同类别间特征分布重叠的问题。

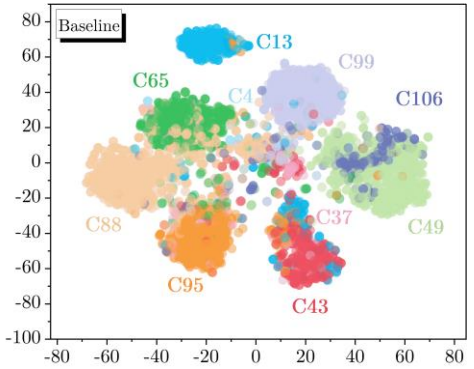
Different diffusion steps

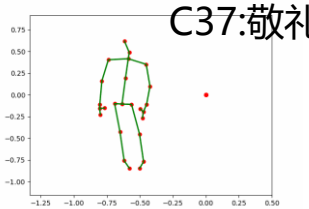
Methods	NTU-60				NTU-120			
	55/5		48/12		110/10		96/24	
	Acc	Time	Acc	Time	Acc	Time	Acc	Time
SA-DVAE	82.82	-	54.86	-	76.61	-	52.25	-
Ours (s1)	**84.37**	8.2e-6	**62.11**	5.7e-6	**78.47**	2.2e-5	57.15	1.0e-5
Ours (s10)	**84.37**	4.9e-5	61.56	4.8e-5	78.43	4.9e-5	**57.51**	4.9e-5
Ours (s20)	84.00	9.3e-5	60.98	9.4e-5	78.31	8.e-5	57.27	8.9e-5

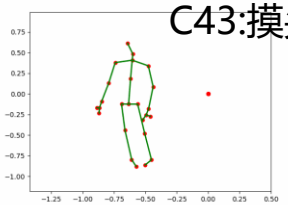
With/without triplet loss

Triplet loss	NTU-60		NTU-120	
	55/5	48/12	110/10	96/24
w.	84.37	62.11	78.47	57.15
w/o	84.15	61.13	73.76	54.04









总结

- 大模型在自动驾驶应用中以LLM为中心，通过添加多模态视觉编码器和动作解码器组成了VLM和VLA，分别承担着“解释者”和“规划者”的角色。
- （视觉）大语言模型传统闭集检测任务扩展到开集词汇检测任务时，可以提供丰富的先验信息，使模型在不依赖标注数据的情况下，显著提升泛化能力。
- 大语言模型具有强大的解释（reasoning）能力，通过有效设计描述语言和数据模态在特征空间中的映射和对齐机制，有助于改善对易混淆行为的识别能力。