北京郵電大學

自驾车事故推理理解：AdVersa-SD 与 MM-AU 数据集解析

学院：计算机学院（国家示范性软件学院）

专业：计算机科学与技术

班级：2022211305

学号：2022211683

姓名：张晨阳

2025 年 6 月 23 号

目录

1. 引言

本篇笔记解析了 **CVPR 2024** 论文《Abductive Ego-View Accident Video Understanding for Safe Driving Perception》的核心内容，重点介绍了自驾车事故视频理解领域的最新进展，包括 MM-AU 数据集的构建、基于 AdVersa-SD 框架的事故原因推理方法，以及如何利用视频与文本的多模态数据，推断和预测潜在事故。

1.1. 主要贡献

1. **MM-AU 数据集**

用于多模态交通事故视频理解的大规模数据集, 包含 11,727 段现实中的自驾车事故视频。

2. **AdVersa-SD 框架**

通过视频扩散（Video Diffusion）进行事故推理理解的框架，使用 CLIP 模型进行多模态对比学习，关注事故的因果关系。

3. **OAVD 方法**

Object-Centric Accident Video Diffusion。一种基于物体中心的视频生成方法，通过保持视频背景的稳定性，重点学习事故因果区域，识别事故的关键原因和预防措施。

2. Abstract 解析

通过对比交互损失，学习**正常、事故将发生、事故发生帧**与相应文本描述之间的关系。

model. This model involves a contrastive interaction loss to learn the pair co-occurrence of normal, near-accident, accident frames with the corresponding text descriptions, such as accident reasons, prevention advice, and accident categories. OAVD enforces the causal region learning while fix-

3. Introduction 解析

提出 MM-AU 和 AdVersa-SD 的背景。

3.1. MM-AU 数据集

提出了 8 种交通事故视频理解任务。

- 事故涉及的物体
- 事故类型
- 事故发生地点
- 事故发生时间
- 事故发生原因
- 关键的事故原因
- 如何预防事故
- 多模态事故视频扩散

models are required to infer ① what objects are involved, ② what kinds of accidents, ③ where and ④ when the accident will occur, ⑤ why the accident occurs, ⑥ what are the keys to accident reasons, ⑦ how to prevent it, and ⑧ multimodal accident video diffusion.

3.2. AdVersa-SD 框架

特点：该框架通过推理型 CLIP 模型和视频扩散技术，尝试通过视频和文本的共现关系学习事故的因果链。

关键技术：

- 使用对比交互损失函数学习事故原因和事故类别的语义共现关系。
- 通过物体中心扩散模型（OAVD），保持视频生成时背景的稳定，重点学习事故发生的因果区域。

4. Related Work 解析

4.1. Ego-View Accident Video Understanding

- **事故检测：**
  - 目标：识别事故发生的具体**空间区域**和**时间帧**。
  - **主要挑战**：参与者（如车辆、行人等）的形状、位置和关系急剧变化，需要鲁棒的特征提取来应对这些复杂性。
  - **常用方法**：无监督学习，通过帧一致性、位置一致性和场景上下文一致性等模型，预测事故窗口。
- **事故预警：**
  - 目标：预测未来可能发生的事故，提供**早期预警**。
  - **主要方法**：基于时间一致性关联参与者轨迹，利用**深度强化学习**模型来提高预警的准确性和解释性。
- **事故分类：**
  - 问题：由于事故类别视频数据的局限性，基于自驾车视角的事故分类研究较少。
  - **方法**：例如 **ViT-TA 模型**通过注意力图来突出事故视频中的关键物体，增强分类的可靠性。
- **事故原因回答：**
  - 相关工作：通过**因果识别**模型来推断事故发生的原因，并通过问答系统

提出预防建议。

- **挑战**：当前的问答框架没有明确验证哪些**关键动作或物体**是导致事故的主要原因，缺乏双重验证机制。

4.2. Ego-View Accident Understanding Datasets

- **现有数据集概述：**
  - **DAD 数据集**：第一个自驾车视角事故视频数据集，重点是视频片段中的事故结尾部分（最后 10 帧）。
  - **CCD 数据集**：与 DAD 类似，每个片段包含 50 帧。
  - **A3D 和 DoTA 数据集**：主要用于无监督学习的事故检测任务。
  - **DADA-2000 数据集**：除了视频，还标注了驾驶员的注意力信息，丰富了事故理解的维度。
- **虚拟事故视频的生成：**
  - 由于真实世界中很难获取大量的事故视频，研究人员利用模拟工具生成虚拟的事故视频。
  - 例如，**GTACrash**、**VIENA2**、**DeepAccident** 等数据集使用了虚拟事故视频或物体轨迹。
  - **挑战**：真实与模拟数据的领域差距是一个难点，因为模拟工具难以精确模拟现实中的事故过程。
- **现有数据集的局限性：**
  - 大多数数据集主要集中在视觉数据上，缺乏详细的文本描述。
  - 只有少数数据集，如 **CTA**，尝试探索更丰富的模态信息，但仍有提升空间。

5. MM-AU Dataset 解析

- **数据集来源：**
  - **MM-AU** 收集自多个公开的自驾车视角数据集，如 CCD、A3D、DoTA 和 DADA-2000，还包括来自流媒体网站（YouTube、Bilibili 和腾讯）的视频。
  - 覆盖多种天气条件和场景，包含了 2,195,613 帧，成为迄今为止最大且最细粒度的自驾车视角多模态事故数据集。
- **标注过程：**
  - **事故窗口标注**：通过平均确定事故的开始和结束时间，并将视频按时间窗口划分，这有助于模型的事故理解训练。
  - **物体检测标注**：为 7 类道路参与者提供了精细和粗略的物体边界框标注，使用 YOLOX 进行初步检测，手动调整获得准确标注。
  - **文本描述标注**：MM-AU 特别标注了三类文本描述：事故原因、预防建议和事故类别。这些文本与视频中的事故窗口对齐，并设计了多选的**事故原因回答（ArA）**任务，帮助模型理解事故的原因。
- **任务设计与挑战：**
  - **事故原因回答任务**：设计了一个问答框架，针对每个视频的事故原因提供问题和干扰项，要求模型识别正确答案。
  - **复杂的标注与数据量**：MM-AU 提供了大量精细标注，这对事故原因分析和自动驾驶安全理解提出了更高的要求和可能性。

6. AdVersa-SD 解析

6.1. Abductive CLIP

- **推理型 CLIP 的设计目的：**

该模型旨在学习文本描述与视频片段之间的语义一致性，特别是在不同类型的文本-视频共现对中，如正常视频片段与反义文本、接近事故的视频片段与预防建议的配对。

- **虚拟共现对的生成：**

通过**反义动词**的加入生成反义文本描述（例如将“发生事故”变为“未发生事故”），并通过帧反转生成接近事故的反序视频，作为虚拟共现对的一部分。

每个视频片段与文本描述随机匹配 16 帧进行训练，以增强模型的适应性。

- **对比交互损失（CILoss）：**

核心思想是通过计算正共现对和负共现对之间的相似性差异，来增强正确匹配的文本-视频对的嵌入一致性。

目标：最大化文本描述与正确视频帧的语义一致性，同时最小化它们与错误视频帧或描述的相似度。

- **优化过程：**

模型通过最小化四种不同类型的对比交互损失（如正共现对和负共现对的不同组合）来完成优化，从而确保模型对不同文本-视频配对的准确学习。

6.2. 正共现对和负共现对详解

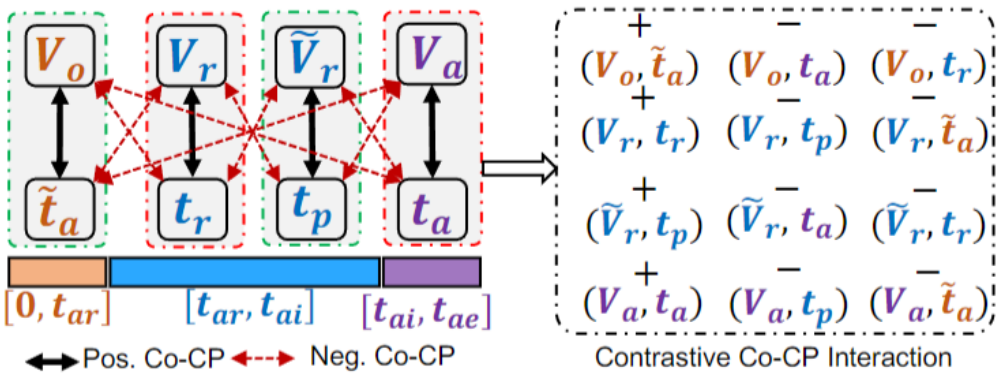

Figure 4. The structure of **Abductive CLIP** contains four interaction groups with one positive Co-CP and two negative Co-CPs for each interaction group, where t_{ar}=t_{ai}-40.

6.2.1. 正共现对（Pos. CoCPs）和负共现对（Neg. CoCPs）的组成

- **正共现对（Pos. CoCPs）：**

正共现对指的是**视频片段和相应文本描述的正确配对**。举例来说，如果我们有一段接近事故的视频片段 **(Vr)**，**其正确的文本描述是 “事故原因”（tr），那么这个配对 (Vr, tr) 就是一个正共现对**。同样地，**事故发生视频片段（Va）**与**事故类别描述（ta）**的配对也属于正共现对。

正共现对的目标是让模型学习到视频片段和它的**正确描述**之间的语义一致性。

- **负共现对（Neg. CoCPs）：**

负共现对指的是**视频片段和错误文本描述的配对**。为了生成负共现对，模型会将视频片段与**不相关的描述**进行配对。例如，将正常视频片段 (Vo) 与一个**错误的事故类别描**述（反义的 t̃a，比如“没有发生事故”）配对，或将接近事故的反转视频片段 **(Ṽr)** 与**不相关的预防建议**配对，都是负共现对。

负共现对的目标是让模型通过学习区分正确和错误的配对，从而减少视频片

段与**错误描述**之间的相似性。

6.2.2. 正共现对和负共现对在模型中的优化作用

模型的优化目标是**增强正共现对之间的相似性**，同时**减小负共现对之间的相似性**。具体来说，模型通过**对比交互损失（CILoss）**来实现这种优化。

- **正共现对的作用：**

当输入正共现对（例如 **(Vr, tr)** 或 **(Va, ta)**）时，模型会**计算视频片段和文本描述之间的嵌入相似度**（E(zv, zt)，zv 是视频片段的嵌入，zt 是文本描述的嵌入）。模型的目标是**最大化正共现对的相似度**，让它们的嵌入尽可能接近。换句话说，模型会学习视频中的视觉信息（如事故的场景、物体动作）如何与文本描述的语义相符。

- **负共现对的作用：**

对于负共现对（例如 **(Vo, t̃a)** 或 **(Ṽr, tp)**），模型也会**计算视频和文本描述之间的相似度**，但目标是**最小化负共现对的相似度**，即让错误配对的视频片段与描述的嵌入距离尽可能远。这样，模型就能学会识别并“排斥”错误的文本-视频配对。

6.2.3. 如何通过正共现对和负共现对优化模型

通过同时输入正共现对和负共现对，模型会在训练过程中进行对比学习，目标是：

1. **增加正共现对的相似度**：通过学习正确的视频片段与其文本描述的特征，使它们的嵌入表示更加接近。
2. **减少负共现对的相似度**：通过对比负共现对，使错误的文本描述与视频片段的嵌入之间保持较大距离，从而区分出哪些是错误的配对。

6.3. Extension to Accident Video Diffusion

- **事故视频扩散的目标：**
  - 通过视频扩散模型显式地探索交通事故的因果关系。因为事故往往由道路参与者的不规则运动导致，模型需要具备物体级别的表示能力，以实现细致的事故分析。
- **以物体为中心的事故视频扩散模型（OAVD）：**
  - OAVD 基于**潜在扩散模型（LDM）**，通过在视频帧的潜在表示上添加噪声，再进行去噪生成事故视频。
  - 其核心是 **3D U-Net**，包括空间、时间和文本-视频的注意力机制，确保模型能够捕捉到事故视频中的关键物体位置和时间序列信息。
- **遮罩视频帧扩散：**
  - 通过前向添加噪声和反向去噪，模型在视频生成过程中固定背景细节，重点处理物体区域，从而生成以物体为中心的事故视频。
  - 这个过程包括均方误差和遮罩重建的联合损失，用于优化视频的生成质量。
- **门控边界框表示（Gated Bbox Representation）：**
  - 边界框（Bbox）与文本描述协同工作，增强对因果物体区域的学习。通过门控自注意力机制，模型能够明确分析特定物体在事故中的角色。
- **推理阶段：**
  - 在推理阶段，输入文本和视频共现对（Co-CPs），通过 3D U-Net 和去噪扩散模型生成新的事故视频片段。

7. 实验

- OD 任务：
  - 使用了 11 种最先进的物体检测器，重点评估物体边界框检测的精度和召回率。
  - 通过不同的版本（V1-Train 和 V2-Train）来比较事故帧和非事故帧的检测效果。
- ArA 任务：
  - 基于多选问题的问答任务，问题是“事故原因是什么”，通过准确率来评估模型的表现。
- 推理型视频扩散任务：
  - 通过输入物体边界框和文本描述，评估模型生成事故视频的能力。
  - 使用 FVD 评估视频质量，CLIP 分数衡量生成的视频与文本描述的对齐程度。

8.学习心得

通过阅读并整理 **CVPR 2024** 论文《Abductive Ego-View Accident Video Understanding for Safe Driving Perception》的内容，我对多模态交通事故理解这一前沿方向有了更深刻的认识。

1. 数据的重要性远超想象

MM-AU 数据集的构建不仅在规模上突破了以往，还在标注的精细度和多模态信息的结合方面设立了新的标杆。尤其是事故原因、预防建议和事故类别的文本标注，为因果分析提供了扎实的基础。我深刻认识到，**高质量、多维度**的标注数据是训练出真正具有推理能力模型的前提。

2. 因果推理是自动驾驶感知的下一个突破口

相比于传统的事故检测和预警，这篇论文提出的“**Abductive Reasoning**”（溯因推理）概念，让我意识到仅仅预测“会出事”是不够的，更关键的是回答“**为什么**会出事”。

通过推理事故原因，并结合预防建议，这种能力不仅能增强模型的可解释性，也能更好地服务于自动驾驶的实际安全应用。

3. 多模态对比学习的潜力被进一步释放

通过构造正负共现对（CoCPs）和设计对比交互损失（CILoss），模型可以更精准地对齐视频和文本之间的语义。这种多模态监督策略让我看到了大模型在复杂语义场景下的表现力。

4. 视频生成任务的目标已从“复现”向“推理”转变

以 OAVD 为代表的事故视频扩散模型不再是为了简单地生成看起来逼真的视频，而是要生成体现事故因果结构的视频。通过物体为中心的扩散建模、门控边界框、以及多种注意力机制的融合，模型能生成符合语义推理逻辑的视频内容，这种能力对未来自动驾驶系统的可解释性和决策机制具有重要意义。