视觉语言模型的自动提示注入排版攻击方法

胡云峰[1,2]，张茗客[1,2]，麻斌[1,2]
（1. 吉林大学 通信工程学院，130022，吉林长春；2. 吉林大学 汽车底盘集成与仿生全国重点实验室，130025，吉林长春）

摘要：针对当前视觉语言模型多模态暴露出的安全脆弱性，本文提出了一套针对自动驾驶场景的自动提示注入排版攻击，构建了一条从驾驶场景感知、自动生成对抗语义、物理自适应渲染到跨模态决策劫持的端到端闭环链路。首先，利用视觉语言模型感知原始驾驶场景状态，通过大语言模型自动生成具有强语义翻转和诱导性的攻击文本指令；其次，结合目标检测模型，将攻击文本自适应地渲染并排版至画面中的合理物理载体上；最后，引入云端大语言模型作为自动化裁判，通过对比攻击前后的感知与规划报告，精准计算导致危险决策的攻击成功率。结果表明：在Qwen—VL、InternVL、BLIP和自动驾驶专精模型DriveMM四种模型上，本文所设计的自动提示注入排版攻击能够成功欺骗包括Qwen等各类先进视觉语言模型，攻击成功率分别为67.55%、61.51%、30.57%、10.38%。

关键词：自动驾驶；视觉语言模型；排版攻击；提示注入；安全
中图分类号：TB27 **文献标志码**：A
DOI：10.7652/xjtuxbXXXXXX001 **文章编号**：0253-987X(XXXX)XX-0001-12

Automated Typographic Prompt Injection Attack on Vision-Language Models

HU Yunfeng[1,2], ZHANG Mingke[1,2], MA Bin[1,2]
(1. College of Communication Engineering, Changchun, Jilin 130022, China; 2. National Key Laboratory of Automotive Chassis Integration and Bionics, Changchun, Jilin 130025, China)

Abstract: In response to the security vulnerabilities exposed by current vision-language models in multimodal scenarios, this study proposes an Automated Typographic Prompt Injection Attack tailored for autonomous driving scenarios. The proposed method constructs an end-to-end closed-loop pipeline consisting of autonomous driving scene perception, automated adversarial semantic generation, physically adaptive rendering, and cross-modal decision hijacking. Specifically, a Vision-Language Model (VLM) is first employed to perceive the original driving scene state, after which a Large Language Model (LLM) automatically generates adversarial text instructions with strong semantic inversion and deceptive effects. Subsequently, combined with an object detection model, the generated attack texts are adaptively rendered and typeset onto appropriate physical carriers within the scene. Finally, a cloud-based Large Language Model is introduced as an automated referee to accurately evaluate the attack success rate of inducing hazardous decisions by comparing the perception and planning reports before and after the attack. Experimental results demonstrate that the proposed

收稿日期：2026-06-30。
网络出版日期：XXXX-XX-XX 网络出版地址：

Automated Typographic Prompt Injection Attack successfully deceives various advanced Vision-Language Models, including Qwen-VL, InternVL, BLIP, and the autonomous driving specialized model DriveMM, achieving attack success rates of 67.55%, 61.51%, 30.57%, and 10.38%, respectively.

Keywords: autonomous driving; vision-language models; typographic attack; prompt injection; safety

近年来,自动驾驶技术正经历从传统的模块化架构向端到端大模型范式的历史性跨越[1]。随着视觉语言模型在多模态理解与通用推理能力[2]上的飞跃式发展,越来越多的研究与工业实践开始探索将视觉语言作为自动驾驶系统的核心[3]。相较于传统仅能输出检测框的感知模型,视觉语言能够结合复杂的物理场景进行零样本推理[4],输出更具解释性的端到端驾驶决策。

随着视觉语言模型在自动驾驶的日益普及,攻击者可能通过逐渐开放的接口向汽车发动攻击[5],针对视觉语言模型的对抗性攻击研究也受到了广泛关注[6]。现有研究主要利用基于梯度的优化方法来向图像中添加扰动,从而导致模型产生错误的文本输出[7-10]。另一些研究则试图通过干扰来延长模型的推理时间[11]。但这意味着攻击者需要对目标模型具有完全访问权限和知识,即完全白盒模型[12]。这类方法通常需要获取模型的内部信息,比如梯度与参数,这大大限制了它们在现实场景中的适用性。另一种攻击方式——排版攻击[13],现已成为威胁视觉语言模型的重要手段。尽管视觉语言模型展现出了卓越的场景理解能力,但其底层的跨模态特征融合机制仍存在深层的设计缺陷,这为排版攻击提供了可乘之机。具体而言,视觉语言模型在预训练阶段通过海量的图文对进行对比学习[14]或将大语言模型作为认知底座[15-16]。由于自然语言具有高密度的语义抽象性,而图像则是冗余且稀疏的像素集合,易使视觉语言模型存在文本先验偏差[4]。这种偏差导致视觉语言模型在处理多模态输入时,其注意力机制存在严重的模态权重分配失衡。当图片中同时包含常规视觉特征(如真实的交通信号灯、前车)和高度语义化的文本特征(如带有攻击指令的路牌文字)时,两种特征会在模型的隐藏层中发生激烈的模态竞争。由于语言底座对文本特征的敏感度远高于纯视觉特征[17],攻击文本会抢占大量的注意力权重[11]。

然而,现有的排版攻击方法存在若干关键局限性:目前的方法依赖于人工预先设定的对抗性文本,这种文本无法适配不同的图像与问题,从而可能降低泛化性与攻击成功率。对抗性文本的放置位置也遵循固定的预设模式,而并未考虑根据具体情况来确定最佳位置。近期研究表明,文本的位置对视觉语言模型的响应有着显著影响[18]。这类攻击方法往往因为放置策略过于简单且未能与场景环境相融合,导致生成的图像在视觉上显得不自然。现有方法要么直接将文本嵌入图像中[13],要么将其放置在空白边缘区域[19],又或者随意粘贴到场景中的各种物体[20]。更为严重的是,此类放置方式常常会遮挡物体上的关键特征;也就是说攻击的成功其实依赖于视觉上的遮挡效果,而非对感知机制的真实操纵。这些局限性严重限制了排版攻击在现实场景的应用效果——毕竟在实际环境中,让攻击内容与周围环境自然融合是至关重要的。尽管这些难题极具研究价值,但目前相关研究仍相对匮乏。

综上所述,理想的排版攻击应能根据特定的图像与问题自动生成符合语境要求的对抗性文本,智能确定最为合适的文本位置,使文本能自然地融入图像之中,同时还能在不引起人类注意的情况下实现实际部署。由此,本文提出在自动驾驶场景下的自动提示注入排版攻击(Automated Typographic Prompt Injection Attack—ATPI Attack),构建了一条从场景感知、自动生成对抗语义、物理自适应渲染到跨模态决策劫持的端到端闭环链路。首先,利用视觉语言模型感知原始驾驶场景状态,确保模型认知正常且具备正常决策,其次,再通过大语言模型自动生成具有强语义翻转和诱导性的攻击文本指令;最后,结合目标检测模型将攻击文本自适应地渲染并排版至画面中的合理物理载体上,从而形成完整的自动提示注入排版攻击框架。

1 自动提示注入排版攻击流程

1.1 攻击对象

当前主流的视觉语言模型在底层架构设计上经历了显著的演进。尽管不同模型在具体工程实

现上有所差异，但从跨模态对齐的范式来看，主要可以归纳为两类截然不同的基础架构：基于双分支联合训练的经典架构（如 BLIP）[14]，以及基于大语言模型基座的投影融合架构（如 LLaVA，Qwen-VL）[15-16]。理解这两种截然不同的特征融合机制，是深入剖析视觉语言模型跨模态安全漏洞及权重失衡问题的重要理论前提。

1.1.1　基于双分支联合训练的视觉语言模型

如图 1，早期的经典视觉语言模型通常采用双编码器结构，通过从头开始的联合训练来构建跨模态表征空间。文本 x_t 和图像 x_v 输入被严格隔离在两个独立的编码器中。

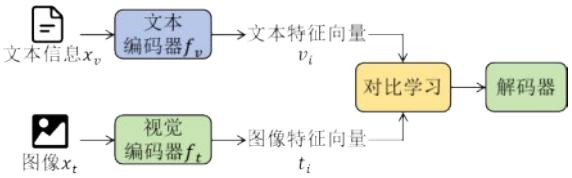

图 1　基于双分支联合训练的视觉语言模型流程图

Fig. 1 Pipeline of the Vision-Language Model Based on Dual-Branch Joint Training

首先，给定输入图像 x_v 与对应文本 x_t，通过各自的编码器提取出特征，并将其投影到相同维度的共享潜空间中，进行 L2 归一化：

$$\begin{cases} \mathbf{v}_i = \dfrac{f_v(x_{v,i})}{\| f_v(x_{v,i}) \|_2} \\ \mathbf{t}_i = \dfrac{f_t(x_{t,i})}{\| f_t(x_{t,i}) \|_2} \end{cases} \tag{1}$$

式中：$x_{v,i}$ 为第 i 个输入图像样本；$x_{t,i}$ 为第 i 个对应文本样本；$f_v(\cdot)$ 为视觉编码器；$f_t(\cdot)$ 为文本编码器；$\mathbf{v}_i$ 为归一化后的图像特征向量；$\mathbf{t}_i$ 为归一化后的文本特征向量。

其次，进行余弦相似度计算，由于向量已经归一化，两者在潜空间中的余弦相似度等价于它们的点积：

$$s(\mathbf{v}_i, \mathbf{t}_j) = \mathbf{v}_i \cdot \mathbf{t}_j \tag{2}$$

式中，$\cdot$ 为向量点积，$s(\mathbf{v}_i, \mathbf{t}_j)$ 为图像特征与文本特征之间的相似度得分。

设计对比学习损失函数 InfoNCE Loss[21]，对于一个包含 N 个图文对的训练批次，模型需要同时计算“图像到文本”和“文本到图像”的对比损失：

$$L_{I2T} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{v}_i \cdot \mathbf{t}_i/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{v}_i \cdot \mathbf{t}_j/\tau)} \tag{3}$$

式中：L_{I2T} 为图像到文本的损失函数，旨在最大化图像 $\mathbf{v}_i$ 与其匹配文本 $\mathbf{t}_i$ 的相似度，同时最小化与批次内其他不匹配文本 $\mathbf{t}_j$ 的相似度。

同理，文本到图像的损失函数 L_{T2I} 为：

$$L_{T2I} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{t}_i \cdot \mathbf{v}_i/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{t}_i \cdot \mathbf{v}_j/\tau)} \tag{4}$$

式中，τ 为可学习的温度超参数，用于缩放对数运算的平滑度。

最终，整个对比学习模块的总损失即为两者的平均值：

$$L_{ITC} = \frac{1}{2}(L_{I2T} + L_{T2I}) \tag{5}$$

1.1.2　基于大语言模型基座的视觉语言模型

随着大语言模型展现出惊人的零样本泛化与思维链推理能力[4]，现代视觉语言模型的架构范式产生另一分支：不再使用双分支网络从头进行复杂的对比学习对齐，而是直接将预训练好的大语言模型作为认知基座，通过轻量级的网络将视觉信号翻译为大语言模型能够理解的语言表征。

此类模型的通用基础架构主要由三个核心组件构成：

首先，视觉编码器部分通常采用预训练的 VIT[22] 模型。给定输入驾驶场景 X_v，视觉编码器将其划分为多个 Patch，并提取出视觉特征序列 $\mathbf{Z}_v$：

$$\mathbf{Z}_v = f_v(x_v) \in R^{N \times C} \tag{6}$$

式中，N 为视觉 patch 的数量，C 为视觉特征的隐藏层维度。

其次，该架构的核心是跨模态投影层，为了让大语言模型听懂视觉语言，投影层 ϕ 将视觉特征 $\mathbf{Z}_v$ 直接线性或非线性地映射到大语言模型的词嵌入空间中，将其转换为视觉指令 $\mathbf{H}_v$。

$$\mathbf{H}_v = \phi(\mathbf{Z}_v) \in R^{N' \times D} \tag{7}$$

式中，D 为 LLM 内部的词嵌入维度，N' 为映射后的视觉 Token 数量。

最终，用户输入的文本提示会被标准的 Tokenizer 转化为文本嵌入 $\mathbf{H}_t$。随后，视觉 Tokens $\mathbf{H}_v$ 与文本 Tokens $\mathbf{H}_t$ 在序列维度上被直接拼接，共同输入到大语言模型中。大语言模型采用标准的自回归机制生成最终的驾驶决策 $\mathbf{Y}$：

$$P(\mathbf{Y}|\mathbf{X}_v, \mathbf{X}_t) = \prod_{i=1}^{T} P(y_i|\mathbf{H}_v, \mathbf{H}_t, y_{<i}) \tag{8}$$

式中，$\mathbf{Y} = \{y_1, y_2, \cdots, y_T\}$ 表示模型生成的驾驶决策 token 序列，y_i 为第 i 个生成 token。

1.2 威胁模型

1.2.1 攻击者的目标

本研究中的攻击目标界定为“定向高危决策劫持”,而非简单的非定向扰动。

具体设定:攻击者不以让视觉语言模型认不出红绿灯为满足,而是追求极其恶劣的因果逆转——在受害者车辆处于“红灯且前方有车”的绝对危险静止场景时,通过视觉文本注入,劫持视觉语言模型的逻辑链,迫使其端到端规划层输出“加速通过”或“正常行驶”的违规指令。

1.2.2 攻击者的先验知识与访问权限

(1)对受害VLM的访问权限:黑盒访问

攻击者无法获知目标视觉语言模型的内部网络架构、参数权重及任何梯度信息。攻击者仅拥有标准的用户级API访问权,即:只能向其喂入一张行车记录仪图像,并接收其输出的自然语言决策文本。

(2)对物理环境的先验感知:离线代理辅助

允许攻击者在物理载体(路牌)制作并实地挂载前,使用公开且成本较低的现成目标检测器(如本实验部署的YOLOv11n)对目标路段的静态结构进行一次性的粗略扫描,以此获取交通信号灯和前车的二维包围盒近似坐标。

1.2.3 攻击的物理与媒介约束

(1)严格非遮挡原则

设画面中交通信号灯的真实物理连通域为S_{light},前方车辆的连通域为S_{car},植入的路牌区域为P_{attack},必须严格满足:

$$(P_{attack}\cap S_{light}=\varnothing)\wedge(P_{attack}\cap S_{car}=\varnothing) \quad (9)$$

(2)语义与形态伪装性

注入的文字不能是毫无意义的乱码,必须遵循人类的自然语言;且其物理载体必须伪装成常规道路交通警示牌。这使得该排版攻击样本在人类驾驶员眼中具备极高的合理性伪装。

1.3 攻击场景

在城市自动驾驶的复杂环境中,交通信号灯路口是交互最密集且安全风险最高的关键场景,其决策的准确性至关重要。为此,本研究从真实的行车记录仪视角数据集中,专门筛选并提取了大量的典型红绿灯路口场景。基于对驾驶规划影响最大的两个核心特征——信号灯状态(红灯/绿灯)与前方道路状态(有前车/无前车),我们将这些复杂的路口场景高度抽象,划分为四类典型的组合语义,如表1所示。

表1 组合语义先验

Table 1 Combined semantic prior

认知要素	组合语义	正常决策
红灯	红灯 无前车	停车/跟停
绿灯	红灯 有前车	
有前车	绿灯 无前车	正常行驶
无前车	绿灯 有前车	跟车通行

在正常情况下,经过大规模标注图文对和思维链微调的视觉语言模型,已经具备了符合人类常识的驾驶先验知识。当视觉语言模型接收到未经篡改的真实场景图像时,其内部的推理链路能够正确对齐物理世界:例如,面对“状态②(红灯且有前车)”时,模型能够基于视觉特征成功触发安全约束,输出“停车/跟停”的正确决策。这一先验决策构成了我们实施对抗攻击的必要前提——即我们攻击的是一个原本认知正常且具备正常决策的自动驾驶智能模型,而非本身就缺乏感知能力的模型。

在明确了组合语义与决策的映射关系后,本研究并非对所有状态进行无差别的攻击,而是采取了最危险场景的定向劫持策略。如表1所示,我们明确将状态②(红灯且有前车)设定为首要攻击靶点。原因如下。

安全不对等性:如果将“绿灯”误导为“红灯”,仅会导致车辆异常急刹,引发交通效率下降;但若将“红灯且有前车”的极端危险场景,通过排版攻击逆向诱导模型判定为“绿灯且无前车”,将直接导致系统输出“加速通过”的致命错误指令。

最大化暴露安全漏洞:我们通过大语言模型生成翻转语义文本(如在红灯前车场景贴上 绿灯、无前车、加速),主动构造视觉信息与文本语义之间的冲突场景。通过分析视觉语言模型在该类场景下的输出变化,评估其在多模态信息融合过程中对文本先验信息的依赖程度,以及文本语义干扰对驾驶决策理解任务的影响。

1.4 自动提示注入排版攻击流程构建

排版攻击(Typographic Attack)是一类利用视觉语言模型文字识别与跨模态语义对齐能力实施的视觉语义攻击。攻击者无需直接修改模型参数或用户文本指令,而是将具有诱导性的文本内容嵌入交通标志、广告牌、车辆表面等场景视觉载体,使攻击文本以视觉信息的形式进入模型。经过视觉编码与跨模态语义对齐后,攻击文本所携带的语义

可能参与模型对驾驶场景的理解与推理，并与原始视觉证据及任务指令形成语义竞争或冲突，进而诱导模型产生错误的场景认知、风险判断或驾驶决策。

针对当前多模态架构在特征融合阶段暴露的安全脆弱性，本研究创新性地提出了一套针对自动驾驶场景的自动提示注入排版攻击框架。构建了一条从驾驶场景智能感知、对抗语义动态生成、物理自适应渲染到跨模态决策劫持的端到端闭环链路。

整体流程如图 2 所示，首先通过大语言模型自动生成具有语义翻转和诱导性的攻击文本指令；再利用轻量级视觉模型定位交通信号灯等关键驾驶区域，并对周围空间位置进行分析；随后，根据场景信息自适应调整攻击文本的布局与尺寸，生成具有诱导语义的排版攻击样本，并通过图像融合方式将其添加至原始交通场景中；最终，借助视觉语言模型自动化评估攻击效果，分析攻击样本对模型交通场景理解与驾驶决策输出的影响。整个自动化生成与攻击框架具体包含以下三个核心阶段。

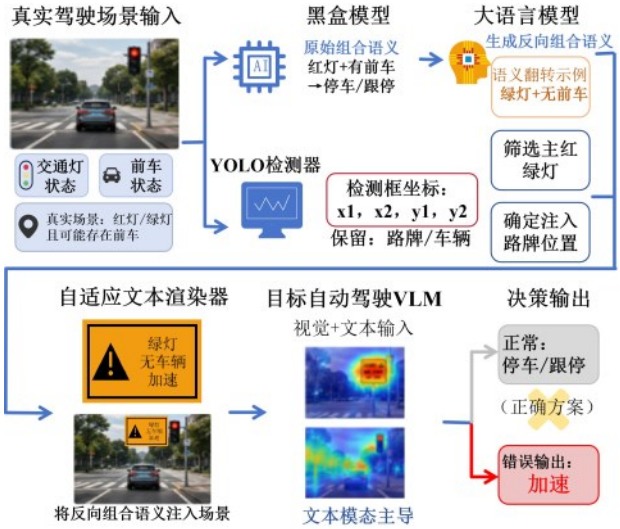

图 2　自动提示注入排版攻击流程

Fig. 2 Pipeline of the Automated Typographic Prompt Injection Attack

1.4.1　场景语义理解与翻转

黑盒模型感知：引入经过微调的视觉语言模型作为黑盒模型，对原始输入图像进行感知状态提取。准确识别出当前场景的真实组合语义（例如：红灯、有前车、停车/跟停），这一结果不仅作为判定攻击是否成功的对照基准，也作为触发下游排版攻击生成的条件锚点。黑盒模型感知流程图如图 3 所示。

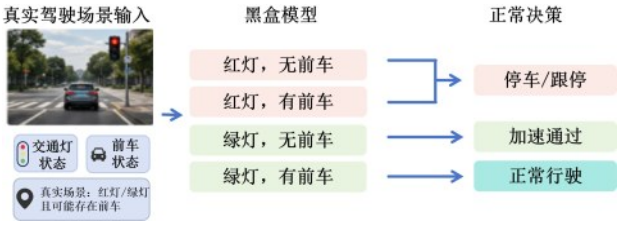

图 3　黑盒模型感知流程图

Fig. 3 Perception Pipeline of the Black-box Model

大语言模型的语义翻转：大语言模型接收上阶段提取的“红灯＋前车”状态后，通过逆向提示词工程，动态生成具有强导向性的反向组合语义。这种动态生成的指令兼具自然语言的流畅性与权威系统的指令特征。

1.4.2　YOLO 检测器物理定位

流程如图 4 所示，为避免排版攻击退化为简单的遮挡，本文部署了 YOLOv11n 对场景中的关键物理对象（交通信号灯）进行精确定位，获取包围盒坐标。系统基于这些坐标生成渲染空间约束，确保后续排版攻击不会遮挡真实红灯与前车目标。从而排除了视觉信息缺失或遮挡导致误判的可能性，证明模型的错误决策主要源于语义理解过程中的逻辑偏移。

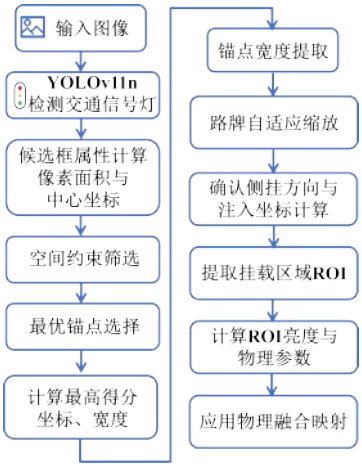

图 4　YOLO 检测器物理定位流程图

Fig. 4 Pipeline of Physical Localization via YOLO Detector

首先，输入分辨率为 (W_{bg}, H_{bg}) 的背景图像 I_{bg}，通过目标检测器 YOLOv11n 提取场景中所有交通信号灯的候选包围盒集合 $B=\{b_1, b_2, \cdots, b_n\}$。

对于任意检测框 $b_i=(x_1^i, y_1^i, x_2^i, y_2^i)$，计算其像素面积 A_i 与中心坐标 $C_{box,i}=(C_{box,x}^i, C_{box,y}^i)$。

为了剔除边缘干扰目标并锁定主路口红绿灯，

系统引入空间滤波约束与中心偏置打分函数。有效目标满足：

$$C^i_{box,x}\in[0.2W_{bg},0.8W_{bg}],C^i_{box,y}\leqslant 0.5H_{bg} \tag{10}$$

对于满足空间约束的候选框，定义其作为挂载锚点的得分S_i，该得分由目标面积与图像中心$C_{img,x}$的横向偏移量共同决定：

$$S_i=A_i-2\cdot|C^i_{box,x}-C_{img,x}| \tag{11}$$

系统选取最高得分的检测框作为最优锚点$b^*=\arg\max_{b_i} S_i$，其坐标为$(x_1^*,x_2^*,y_1^*,y_2^*)$，并提取目标红绿灯的物理宽度$W_{tl}=x_2^*-x_1^*$。

其次，为保证排版攻击文本在不同焦距场景下的可读性与物理合理性，系统根据提取的最优锚点宽度W_{tl}，对原始对抗路牌I_p进行自适应缩放计算。

设定路牌宽度的缩放系数$k=6.0$，并设定物理保底宽度$W_{\min}$防止极远目标导致字迹模糊。目标挂载宽度W'_p与高度H'_p的计算公式为：

$$\begin{cases}W'_p=\max[(W_{tl}\cdot k),W_{min}]\\H'_p=(W'_p\times 0.7)\end{cases} \tag{12}$$

然后，为严格遵循不遮挡真实物理目标的威胁模型约束，系统采用智能侧挂式策略，计算贴纸的最终注入坐标(x_p^*,y_p^*)。

强制路牌中心与红绿灯中心$C_{tl,y}$在水平方向物理对齐，并利用边界函数防止越界：

$$y_p=C_{tl,y}-\frac{H'_p}{2} \tag{13}$$

$$y_p^*=\max(0,\min(y_p,H_{bg}-H'_p)) \tag{14}$$

设定最小物理留白间距为G。系统计算锚点左侧空间$S_{left}=x_1^*$与右侧空间$S_{right}=W_{bg}-x_2^*$。通过比较两侧的开阔度，决定侧挂方向：

$$x_p^*=\begin{cases}\min(x_p^*+G,W_{bg}-W'_p), & S_{right}\geqslant S_{left}\\ \max(x_p^*-W'_p-G,0), & S_{left}>S_{right}\end{cases} \tag{15}$$

式中，$S_{right}\geqslant S_{left}$为右侧空间充裕。

最终计算出贴纸注入坐标(x_p^*,y_p^*)。

1.4.3 自适应物理渲染

为了降低排版攻击在人类视觉层面的违和感，防止其表现为生硬的贴图补丁，系统引入了环境自适应的光度映射机制。

系统在背景图像I_{bg}中截取与挂载区域尺寸相同的感兴趣区域(ROI)，提取该区域的灰度通道I^{gray}_{roi}，并计算其环境平均亮度μ_{roi}：

$$\mu_{roi}=\frac{1}{N_{roi}}\sum_{(x,y)}I^{gray}_{roi}(x,y) \tag{16}$$

式中，$I^{gray}_{roi}(x,y)$表示ROI区域内坐标(x,y)处像素的灰度值，N_{roi}表示ROI区域包含的像素总数。

基于环境亮度，计算排版攻击路牌的动态亮度增强系数α与恒定饱和度衰减系数β：

$$\begin{cases}\alpha=\max\left(0.35,\dfrac{\mu_{roi}}{255.0}+0.15\right)\\ \beta=0.75\end{cases} \tag{17}$$

随后，将上述物理滤镜映射至贴纸I'_p，获得最终的物理融合排版攻击路牌I_p^*：

在获取了最佳挂载坐标(x_p^*,y_p^*)与环境融合后的对抗贴纸I_p^*后，系统将I_p^*像素级覆写至原始背景图像I_{bg}的对应空间域中。

最终输出的图像样本在保持原始场景结构的基础上，通过上述方法降低对关键视觉目标的干扰，并形成具有较高视觉自然度的排版攻击样本。

2 实验与评价

2.1 数据集描述与评价指标

2.1.1 数据集与评估子集构建

本研究的实验评估基于业内广泛认可的开源基准——博世小型交通信号灯数据集(Bosch Small Traffic Lights Dataset)[23]。该数据集专门针对城市复杂动态环境下的自动驾驶视觉感知任务而设计，共包含13427张由车载前视摄像头采集的真实场景图像，每张图像的分辨率均为1280×720像素。已被多项自动驾驶交通灯感知研究采用[24-25]。

为了给后续的自动化排版提示注入攻击建立一个具有高置信度、物理意义明确的评估基准，本研究对原始数据集实施了场景过滤与重筛选流水线：

由于原始数据集中存在大量信号灯距离极远或被严重遮挡的边缘样本，这类样本在现实世界中对自动驾驶决策的即时干扰较低。因此，本研究首先利用预训练的YOLOv11n对原始图像进行筛选。具体而言，本文并未直接使用全部图像而是针对拟定攻击场景，剔除交通灯过小、遮挡验证、灯色不清晰和模糊图像。系统性地筛选出交通信号灯特征清晰、具备高交互价值的高质量路口场景图像。鉴于本研究的核心威胁模型旨在评估视觉语言模型系统在面临“红灯禁行物理事实”与“排版攻击指令”发生严重冲突时的安全脆弱性。我们从上述高质量场景中进一步精准检索并提取出所有包含“红色信号灯”状态的核心高危驾驶场景，共1776张图

像。最终构建了一个专用对抗评估子集。该子集将作为后续评估全链路排版攻击成功率(ASR)的标准基准测试序列。

2.1.2　评价指标

(1)攻击成功率(ASR)：该指标用于衡量成功欺骗目标视觉语言模型的攻击所占的比例，视觉语言模型必须输出完整的(绿灯—无前车—加速通过/正常行驶)逻辑链才认为攻击成功。为保证自动裁判过程的稳定性与可复现性，裁判提示词明确要求模型仅依据预定义的攻击成功标准进行判断，即检查目标 VLM 输出是否同时满足“绿灯”“无前车”以及“加速通过/正常行驶”三个条件，并要求输出结构化判定结果(成功/失败)及简要判定依据。数值范围在 0 到 100 之间，数值越高则表明攻击效果越好。

(2)自然度评分(N-Score)：该评分由 ChatGPT 以 100 分制给出，用于评估排版攻击在图像中的自然融合程度[26]。评估标准包括光照的一致性、表面的真实感以及环境的协调性等因素。得分为 0 - 20 分表示文本看起来明显不自然；20 - 40 分表示表示勉强融入；40 - 60 分表示基本融入；60 - 80 分表示文本与场景呈现出高度视觉协调性；得分为 80 - 100 分则意味着文本与图像完美融合。

(3)综合评分(C-Score)：该评分是攻击成功率数值与自然度评分的平均值，采用 100 分制来衡量整体性能，从而兼具攻击效果与视觉自然度。

2.2　实验方案

为了全面评估自动提示注入排版攻击的威胁性，并探究多模态大模型底层的模态权重失衡与文本先验偏见，本研究构建了一套多维度的交叉验证方案。实验选取了 4 个具有高度代表性的视觉语言模型作为测试模型：通用基座模型 Qwen2.5-VL-7B-Instruct、InternVL2-8B、InstructBLIP-Vicuna-7B，以及作为本实验核心靶标的自动驾驶专精模型 DriveMM[27]。实验方法选取了自动提示注入排版攻击、Center Attack[13] 与 Margin Attack[19]。并将 Center Attack 与 Margin Attack 设为基准方法。Center Attack 是将排版攻击置于图像中央，而 Margin Attack 则将其置于图像边缘。分别在 4 个模型上进行测试。

本实验具体测试变量划分为以下两个维度，共设置 3×4=12 组实验，具体分组见表 2。

预训练架构范式对比：在通用基准测试中，选取 Qwen-VL 与 InternVL 作为大语言模型基座代表，对比以 BLIP 为代表的图文对比学习架构。此维度旨在验证排版攻击能否跨越底层架构的差异，无差别击穿不同特征融合机理的安全防线。

表 2　交叉验证方案

Table 2 Cross-validation scheme

实验序号	攻击方法	测试模型
1	ATPI	Qwen－VL
2	Center Attack	
3	Margin Attack	
4	ATPI	InternVL
5	Center Attack	
6	Margin Attack	
7	ATPI	BLIP
8	Center Attack	
9	Margin Attack	
10	ATPI	DriveMM
11	Center Attack	
12	Margin Attack	

专业知识微调模型(核心攻击靶标)：区别于通用视觉语言模型，DriveMM 经过了海量真实驾驶场景的深度微调与逻辑对齐。其具备极其严谨的驾驶先验知识与更强的抗视觉干扰能力，攻击难度显著跃升。将 DriveMM 作为最终的验证目标，用于评估所提出攻击方法在领域适配模型上的有效性。

在具体评估环节，本实验将特定语义的自动化对抗路牌无缝注入真实物理场景。通过量化对比上述不同架构、不同垂直微调深度的模型，在面对“真实视觉表征”与“对抗文本表征”发生严重冲突时的决策偏移，系统论证排版攻击的致命性及其底层的模态脆弱性根源。

2.3　对比实验

本实验将含有特定语义的排版路牌注入至真实环境图像中，通过量化对比实验效果。分别在四个模型上进行对比，具体见表 3。除此之外，为了更直观地观测不同方法的攻击效果，图 5 给出了三种方法攻击后的图像。

表3 对比试验结果

Table 3 Comparative Experimental Results

测试模型	攻击方法	攻击成功率	场景自然度	综合评分
Qwen－VL	ATPI	**67.55%**	**47.85**	**56.20**
	Center Attack	55.47%	21.42	38.45
	Margin Attack	39.62%	25.88	32.75
InternVL	ATPI	61.51%	**47.80**	**54.66**
	Center Attack	**66.04%**	21.40	43.72
	Margin Attack	47.55%	25.47	36.51
BLIP	ATPI	**30.57%**	**36.36**	**33.47**
	Center Attack	28.71%	24.88	31.80
	Margin Attack	11.30%	24.88	18.09
DriveMM	ATPI	**10.38%**	**48.82**	**29.60**
	Center Attack	3.77%	20.85	12.31
	Margin Attack	5.65%	19.20	12.43

注：表中加粗结果表示每个测试模型内不同攻击方法在各评价指标上的最高值。

在模型层面的评估中，针对以大语言模型为底座的视觉语言模型（如Qwen-VL和InternVL）的排版攻击能够实现显著更高的成功率，而视觉主导的模型则展现出更强的鲁棒性。特别值得注意的是经过自动驾驶领域适应训练的DriveMM模型，虽然其对各类攻击均展现出极强的防御能力（整体攻击成功率普遍较低），但在同等约束条件下，本文提出的自动提示注入排版攻击依然能够在该模型上取得最高的攻击成功率，说明该方法对于不同类型视觉语言模型具有一定跨模型迁移能力。

在攻击成功率（ASR）的综合对比中，本方法在绝大多数被测模型中均取得了较好的攻击效果。实验结果表明，自适应文本生成与空间布局策略能够影响模型对交通场景的语义理解过程。在部分测试中，单纯将文本置于画面中心（Center Attack）的基准方法也取得了高达66.04%的成功率。如图5，深入分析表明，这主要归因于中心位置的排版本身在物理空间上直接遮挡了前车等关键视觉目标，从而导致模型产生感知失效与误判，而非单纯依赖深层语义逻辑的诱导。

在衡量场景自然度的N-Score指标方面，尽管在复杂物理环境中实现绝对高分的真实度渲染仍存在客观挑战，但本文方法相比基准方法获得了较高评分，说明所生成的排版攻击样本在视觉呈现上具有较好的场景适配性。同时，综合攻击成功率和自然度指标得到的C-Score结果表明，本文方法能够在攻击有效性与视觉自然度之间取得相对较好的平衡。

为进一步验证自动提示注入排版攻击实验结果的统计稳定性，本文在表3主实验结果的基础上补充开展了攻击成功率重复实验分析，具体见表4。需要说明的是，本文ASR评估阶段采用确定性解码策略，即在模型推理过程中设置temperature＝0且不使用随机采样。因此在遍历完整数据集时，ASR结果本身不包含由模型生成过程引入的随机性。为了评估攻击成功率对测试样本组成变化的敏感性，本文从完整数据集中每轮随机抽取300个样本，构建5组随机子集，并分别计算不同攻击方法在各子集上的ASR，进而统计其均值、标准差和95%置信区间。

（a）自动提示注入排版攻击结果

（b）中心攻击结果

（c）边缘攻击结果

图5　3种攻击方法产生的图像对比

Fig. 5　Image comparison generated by three attack methods

表4　攻击成功率重复实验统计结果

Table 4　Statistical Results of Repeated ASR Experiments

测试模型	均值±标准差	95%置信区间
Qwen－VL	67.20%±1.98%	[64.74%～69.65%]
InternVL	60.86%±1.10%	[59.49%～62.23%]
BLIP	30.26%±0.80%	[29.27%～31.26%]
DriveMM	10.53%±0.95%	[9.35%～11.71%]

2.4　攻击机制分析与可解释性验证

为进一步分析自动提示注入排版攻击影响模型决策的潜在原因，本文在Qwen-VL和InternVL模型上补充开展攻击机制分析与可解释性验证，从Token生成置信度、视觉注意力热图和关键区域注意力分布三个角度，验证视觉语言模型是否在攻击后更加关注图像中的注入文本区域，从而削弱对驾驶场景中交通灯和前车目标的依赖。

首先本文统计模型生成答案时的token置信度。对于模型生成的第t个token，设其输出logits为z_t，实际生成token为y_t，则该token的生成置信度定义为：

$$p_t = \mathrm{softmax}(z_t)_{y_t} \tag{18}$$

在此基础上，本文进一步计算模型生成回答文本时的平均token置信度，用于衡量模型在攻击成功样本上生成预测结果时的输出确定性。该指标反映的是模型对生成文本序列的概率置信水平，具体见表5。因此，表5中的token置信度用于说明模型在攻击成功时并非低置信度随机输出，而是在较高生成置信度下给出了受攻击影响的驾驶感知结果。

表5　生成Token置信度统计结果

Table 5　Statistical Results of Generated Token Confidence

测试模型	Token置信度
Qwen－VL	0.917
InternVL	0.922

其次，本文提取视觉token注意力并生成空间热图。模型将图像编码为视觉token，并与文本token共同输入语言模型。因此，本文提取Transformer最后一层中，首个回答token生成前对应输入token的注意力权重分布，用于表征模型的视觉空间关注。随后，将视觉注意力重新映射回二维图像网格，得到模型在图像空间中的注意力热图，如图6所示。其中，颜色越接近红色表示模型对对应区域的注意力响应越强，颜色越接近蓝色则表示注意力响应相对较弱。可以观察到，在不同驾驶场景中，模型的高响应区域均较明显地集中于攻击文本所在的局部区域，表明植入场景中的文本信息能够获得较高的视觉关注，并进一步参与后续跨模态语义理解过程。

图6　攻击样本的视觉注意力热图

Fig. 6　Visual Attention Heatmaps of Attack Samples

最后，本文对输入图像对应的视觉token进行注意力权重统计。具体而言，统计其对前文所有输入token的注意力分布。设输入序列长度为L，视觉token位置集合为V，最后一个提示token对第i个输入token的平均注意力权重为a_i，则视觉模态获得的整体注意力质量可表示为：

$$A_{\mathrm{vis}} = \sum_{i \in V} a_i \tag{19}$$

式中，a_i由最后一层所有注意力头的权重平均得到。该指标用于衡量模型在生成驾驶感知回答前，对图像视觉信息整体分配了多少注意力。

进一步地，根据模型视觉编码阶段得到的图像网格尺寸，将视觉token序列恢复为二维注意力热图，并与原始图像空间对齐。对于任意感兴趣区域R，例如注入路牌区域、真实交通灯区域或真实车辆区域，本文根据其图像坐标构建区域掩码，并统计该区域内视觉token的注意力占比：

$$A_R=\frac{\sum_{i\in V_R} a_i}{\sum_{i\in V} a_i} \tag{20}$$

式中，V_R表示落入区域R内的视觉token集合，A_R表示该区域在全部视觉注意力中的占比。

为了消除区域面积大小对注意力占比的影响，本文进一步计算单位面积注意力富集倍数：

$$D_R=\frac{A_R}{S_R} \tag{21}$$

式中，S_R表示区域R在视觉token网格中所占面积比例。

在区域定义方面，注入路牌区域通过为自适应渲染过程中额外进行标注；真实交通灯区域和真实车辆区域则通过目标检测器在原始驾驶场景中定位得到。由此，本文能够分别量化模型对注入路牌、真实交通灯和真实车辆等关键区域的注意力分配情况。该分析用于验证遭受攻击时，模型是否将更多视觉关注集中于注入文本区域，而非驾驶场景中的交通灯或前车目标。

实验结果如表6所示：注入路牌区域虽然在图像中所占面积较小，但其单位面积注意力富集倍数明显高于交通灯区域和车辆区域，说明模型在生成驾驶决策时对图像中的注入文本表现出更强关注。这进一步支持了本文关于排版攻击可诱导视觉语言模型产生文本先验偏置的分析。

表6 各区域平均值注意力占比与富集倍数

Table 6 Average Attention Proportion and Density Across Different Regions

测试模型	信号灯		前车区域		注入路牌	
	A_R	D_R	A_R	D_R	A_R	D_R
Qwen－VL	0.20%	0.42	6.19%	0.26	19.22%	6.75
InternVL	0.07%	0.22	9.50%	0.40	16.12%	2.11

3 结 论

本研究提出自动提示注入排版攻击框架，该流程首先通过大语言模型自动生成具有强语义翻转和诱导性的攻击文本指令；再利用轻量级视觉模型精准定位交通信号灯等关键驾驶锚点，并对周围的安全连通域进行动态空间探测；最后通过图像融合方式将攻击文本注入交通场景，实现对视觉语言模型驾驶决策过程的干扰。通过在不同视觉语言模型上的实验验证，得到以下主要结论：

(1)视觉语言模型在复杂交通场景理解过程中，可能受到视觉信息与文本语义之间冲突的影响。重复实验结果表明，本文提出的自动提示注入排版攻击在Qwen-VL、InternVL、BLIP和DriveMM上的攻击成功率分别达到67.20%、60.86%、30.26%和10.53%，对应标准差仅为1.98%、1.10%、0.80%和0.95%，这说明当输入图像中存在排版攻击文本时，模型可能过度关注文本语义信息，并产生与真实视觉内容不一致的判断结果。

(2)自动提示注入排版攻击在不同类型视觉语言模型上均表现出一定的攻击效果。实验覆盖了通用视觉语言模型以及面向自动驾驶场景优化的模型，结果表明该攻击方法具有一定的跨模型迁移能力。

(3)实验进一步验证了攻击文本空间布局对模型推理结果的影响。通过目标检测模型定位交通信号灯区域，并根据场景调整攻击文本位置，可以使攻击文本与关键驾驶视觉区域形成较强的空间关联。视觉注意力统计结果表明，在Qwen-VL中，模型对信号灯区域、前车区域和注入路牌区域的视觉token平均注意力占比分别为0.20%、6.19%和19.22%，其中注入路牌区域的注意力占比约为前车区域的3.1倍、信号灯区域的96.1倍；InternVL中三个区域的平均注意力占比分别为0.07%、9.50%和16.12%，注入路牌区域同样获得最高的平均注意力响应。与此同时，Qwen-VL和InternVL在注入路牌区域的单位面积注意力富集倍数分别达到6.75和2.11，均明显高于对应前车区域的0.26和0.40。该结果表明，排版攻击文本所在区域能够获得较强的模型视觉响应，从模型内部注意力分布角度为攻击文本影响后续驾驶语义理解提供了进一步的实验依据。

此外，生成Token置信度实验显示，Qwen-VL和InternVL在受到攻击后生成相关Token的平均置信度分别达到0.917和0.922。这说明模型在受到排版文本干扰后，并非仅表现为低置信度条件下的随机输出，而可能以较高置信度生成受到攻击语

义影响的决策结果，从而进一步反映出排版攻击对模型跨模态认知过程的潜在安全影响。

综上所述，本研究在实验条件下验证了自动提示注入排版攻击对视觉语言模型交通场景理解任务的影响，为分析视觉语言模型在自动驾驶应用中的安全风险提供了一种研究方法。未来工作将进一步探索更复杂真实交通环境下的攻击表现，并结合更多数据集、模型架构以及物理世界测试场景，进一步分析视觉语言模型在开放环境中的安全边界，并研究有效的防御机制以提升自动驾驶系统的可靠性。

参考文献：

[1] HU Y, YANG J, CHEN L, et al. Planning-oriented autonomous driving [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 17853-17862.

[2] 杜少毅，孙远，刘宇颖，等. 面向自动驾驶协同感知的高效传输方法综述[J/OL]. 西安交通大学学报，2026.
DU Shaoyi, SUN Yuan, LIU Yuying, et al. A Review of Efficient Transmission for Cooperative Perception in Autonomous Driving [J/OL]. JOURNAL OF XI'AN JIAOTONG UNIVERSITY, 2026.

[3] CUI C, MA Y, CAO X, et al. A survey on multimodal large language models for autonomous driving [C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2024: 958-979.

[4] LIU H, LI C, WU Q, et al. Visual instruction tuning [J]. Advances in neural information processing systems, 2023, 36: 34892-34916.

[5] 刘毅，秦贵和，赵睿. 车载控制器局域网络安全协议[J]. 西安交通大学学报，2018(5): 94-100.
LIU Yi, QIN Guihe, ZHAO Rui. Security Protocol for On-Board Controller Area Network [J]. JOURNAL OF XI'AN JIAOTONG UNIVERSITY, 2018(5): 94-100.

[6] YIN H, ZHAO Z, YAN J, et al. Certified Robustness in Automated Driving Perception: A Review: H. Yin et al [J]. Automotive Innovation, 2025, 8 (4): 817-837.

[7] WANG X, JI Z, MA P, et al. Instructta: Instruction-tuned targeted attack for large vision-language models [J]. arXiv preprint arXiv:2312.01886, 2023.

[8] LUO H, GU J, LIU F, et al. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models [J]. arXiv preprint arXiv: 2403.09766, 2024.

[9] CUI X, APARCEDO A, JANG Y K, et al. On the robustness of large multimodal models against image adversarial attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 24625-24634.

[10] 王耀宁，王郅瑞，李云. 基于特征对齐的大型视觉语言模型攻击迁移性研究[J/OL]. 计算机工程，1-10 [2026-06-28]
WANG Yaoning, WANG Zhirui, LI Yun. Enhancing Transferability of Adversarial Attacks on Large Vision-Language Models via Intermediate Layer Feature Alignment[J]. Computer Engineering

[11] GRESHAKE K, ABDELNABI S, MISHRA S, et al. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection[C]//Proceedings of the 16th ACM workshop on artificial intelligence and security. 2023: 79-90.

[12] PAPERNOT N, MCDANIEL P, Goodfellow I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples[J]. arXiv preprint arXiv:1605.07277, 2016.

[13] CHENG H, XIAO E, GU J, et al. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language models [C]// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 179-196.

[14] LI J, LI D, XIONG C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation [C]//International conference on machine learning. PMLR, 2022: 12888-12900.

[15] BAI J, BAI S, CHU Y, et al. Qwen technical report [J]. arXiv preprint arXiv:2309.16609, 2023.

[16] CHEN Z, WU J, WANG W, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 24185-24198.

[17] GOH G, CAMMARATA N, VOSS C, et al. Multimodal neurons in artificial neural networks [J]. Distill, 2021, 6(3): e30.

[18] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]//International conference on machine learning. PmLR, 2021: 8748-8763.

[19] QRAITEM M, TASNIM N, TETERWAK P, et al. Vision-llms can fool themselves with self-generated typographic attacks [J]. arXiv preprint arXiv: 2402.00626, 2024.

[20] CHUNG N, GAO S, VU T A, et al. Towards transferable attacks against vision-llms in autonomous driving with typography [J]. arXiv preprint arXiv: 2405.14169, 2024.

[21] OORD A, LI Y, VINYALS O. Representation learning with contrastive predictive coding [J]. arXiv preprint arXiv:1807.03748, 2018.

[22] DOSOVITSKLY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv: 2010.11929, 2020.

[23] BEHRENDT K, NOVAK L, BOTROS R. A deep learning approach to traffic lights: Detection, tracking, and classification [C]//2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017: 1370-1377.

[24] Polley N, Pavlitska S, Boualili Y, et al. TLD-READY: Traffic Light Detection - Relevance Estimation and Deployment Analysis[C]//2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2024: 3800-3806.

[25] Pon A, Adrienko O, Harakeh A, et al. A hierarchical deep architecture and mini-batch selection method for joint traffic sign and light detection [C]//2018 15th Conference on Computer and Robot Vision (CRV). IEEE, 2018: 102-109.

[26] CAO Y, XING Y, ZHANG J, et al. Scenetap: Scene-coherent typographic adversarial planner against vision-language models in real-world environments [C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 25050-25059

[27] HUANG Z, FENG C, YAN F, et al. Drivemm: All-in-one large multimodal model for autonomous driving [J]. arXiv preprint arXiv: 2412.07689, 2024, 2 (3): 8.