

mmDefender: 基于多模态多语言对抗训练的大语言模型增强方法

张道娟, 赵宇飞, 张翎, 吴天琦, 陈凯, 戴聿雯

(中国电力科学研究院有限公司电力网络安全防护与监测技术实验室, 北京 102209)

摘要: 多模态多语言文本生成作为自然语言处理的前沿领域, 通过融合视觉与文本输入增强跨语言内容生成能力。这种方法突破了传统纯文本模型的限制, 将视觉线索纳入文本生成过程, 从而实现更丰富且可上下文感知的语言细微差别理解。首先探讨了多模态元素融入多语言文本生成系统时引入的挑战与脆弱性, 重点研究对抗攻击与数据投毒如何通过操纵视觉上下文来影响生成翻译的准确性与可靠性。为应对这些威胁, 提出一种基于噪声对比学习的新型防御方法。该方法通过训练系统区分真实视觉输入与篡改内容, 显著增强了系统的鲁棒性, 有效缓解恶意攻击对翻译过程的干扰。该研究不仅揭示了多模态数据在多语言翻译架构中的潜在风险, 还为构建安全、可靠的跨模态生成系统提供了切实可行的解决方案, 对推动人工智能模型在实际场景中的安全落地具有重要意义。

关键词: 多模态多语言攻击; 多语言对抗训练; 大语言模型

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025049

mmDefender: enhancing large language model with multimodal multilingual adversarial training

ZHANG Daojuan, ZHAO Yufei, ZHANG Peng, WU Tianqi, CHEN Kai, DAI Yuwen

State Grid Laboratory of Power Cyber-Security Protection and Monitoring Technology,
China Electric Power Research Institute, Beijing 102209, China

Abstract: Multilingual and multimodal text generation was recognized as a cutting-edge field in natural language processing, in which cross-lingual content generation was enhanced through the integration of visual and textual inputs. By incorporating visual cues into the text generation process, the limitations of traditional text-only models were overcome, enabling a richer and more context-aware understanding of linguistic nuances. The challenges and vulnerabilities introduced by multimodal elements in multilingual text generation systems were systematically investigated, with a focus on how adversarial attacks and data poisoning could manipulate visual context to compro-

收稿日期: 2025-04-29; 修回日期: 2025-06-13

通信作者: 赵宇飞, zhaoyufei@epri.sgcc.com.cn

基金项目: 国家电网有限公司总部科技项目 (5700-202358708A-3-3-JC)

Foundation Item: Science and Technology Project of State Grid Corporation of China (5700-202358708A-3-3-JC).

引用格式: 张道娟, 赵宇飞, 张翎, 等. mmDefender: 基于多模态多语言对抗训练的大语言模型增强方法[J]. 网络与信息安全学报, 2025, 11(4): 160-172.

Citation Format: ZHANG D J, ZHAO Y F, ZHANG P, et al. mmDefender: enhancing large language model with multimodal multilingual adversarial training[J]. Chinese Journal of Network and Information Security, 2025, 11(4): 160-172.

mise translation accuracy and reliability. To address these threats, a novel defense method based on noise contrastive learning was proposed. The system was trained to distinguish genuine visual inputs from tampered content, which significantly improved its robustness and effectively mitigated the interference of malicious attacks on the translation process. This study not only revealed the potential risks posed by multimodal data in multilingual translation architectures, but also provided practical solutions for building secure and reliable cross-modal generation systems. These contributions are crucial for ensuring the safe deployment of AI models in real-world applications.

Keywords: multimodal multilingual attack, multilingual adversarial training, large language model

0 引言

多模态多语言攻击是一种融合语言操纵与感知欺骗的高级网络安全威胁,其通过跨语言维度和跨通信渠道的协同攻击,突破传统安全协议的防护边界^[1-4]。这类攻击往往利用人类认知偏差,借助欺骗性的文本、图像、声音甚至触觉反馈等多元输入形式,诱导用户泄露敏感信息或执行未经授权操作。从攻击特性来看,其多语言特性体现在威胁行为者可使用多种语言发起攻击,这不仅能扩大潜在影响范围,还能利用语言相关漏洞(例如,针对安全性较低的区域化软件版本,或利用非母语使用者处理信息时警惕性下降的心理弱点)。而多模态特性则通过整合文本、视觉、音频等多维度信息,构建更具迷惑性的攻击场景,使传统单一维度的安全防护体系难以有效识别。

由于多模态多语言攻击要求安全系统和协议不仅需要抵御多种威胁向量,还需具备识别和过滤不同语言、不同格式的恶意内容的能力,因此其检测与防御面临极大挑战。传统网络安全防护手段在应对此类攻击时存在显著局限性,具体表现在以下几个方面。

1) 现有方案主要聚焦单模态、单语言环境下的威胁识别,对多模态多语言攻击的跨维度协同特性认知不足,致使防御机制存在结构性缺陷,难以应对多源异构信息融合的新型复杂攻击。

2) 多数安全工具缺乏先进的机器学习算法与自然语言处理技术支撑,无法有效捕捉跨语言、跨模态的细微特征等隐蔽攻击模式,导致异常检测精度不足。

3) 依赖过时或静态的安全防护策略,难以适配多模态多语言威胁的动态演化特性,在攻击者持续迭代的跨模态对抗策略面前,显得力不从心。

多模态融合机制在引入强大跨模态理解能力的同时,也暴露了独特的对抗攻击风险。攻击者可利用模型在跨模态数据解析与关联机制中的缺陷,诱导多模态模型产生错误分类和翻译失真,甚至生成有害内容等安全问题。例如,在图像或音频等单模态数据中构造对抗样本(如经过细微篡改的视觉特征或人为调制的声学信号),可能导致模型在处理文本翻译等其他模态任务上的性能显著下降。值得警惕的是,针对模态间交互关系设计的跨模态对抗扰动,正对模型的鲁棒性与可靠性构成严重威胁。如图1所示,由文本和图像模态组成的多模态攻击可能导致神经网络系统崩溃。因此,深入剖析并构建有效的防御体系,是打造安全可信的多模态人工智能系统的核心前提。

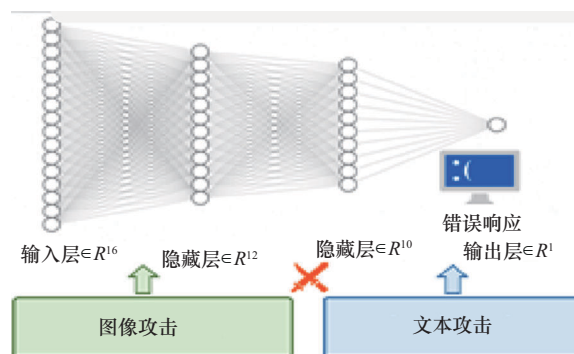


图1 针对神经网络的多模态攻击
Figure 1 Multimodal attacks on neural networks

为此,本文提出一种多模态多语言对抗训练的大语言模型增强方法 mmDefender。作为人工智能领域的前沿技术,该方法特别适用于自然语言处理与计算机视觉交叉研究场景。其创新核心在于通过学习文本、图像、音频等多模态多语言混合数据,强化模型对各类噪声干扰的鲁棒性。具体而言, mmDefender 采用对比学习框架训练模型精准区分多模态输入中的相似样本(正样

本)与非相似样本(负样本)。这一机制不仅能提升模型对单一语言语义的理解能力,还能实现跨语言的语义对齐,从而在跨语言图文检索、多语言文档分类及机器翻译等任务中显著增强模型的泛化能力,使其在复杂真实场景中保持稳定表现。通过聚焦噪声环境下的对比学习范式,该方法致力于构建更具环境适应性的人工智能系统,以应对全球数字化时代人类语言与跨模态交互的复杂性挑战。

本文所提方法在多语言翻译任务中展现出卓越性能,即使在 102 种语言的大规模翻译场景中也能保持高效的处理能力。有监督翻译实验数据表明,相较传统纯文本模型及现有多模态多语言方法,该方法实现了 BLEU (bilingual evaluation understudy) 分数 1 至 4 分的提升。通过 InstrMulti102 数据集的进一步验证,多模态多语言对比学习的核心价值得到充分证实:分析显示,多语言文本与视觉模态对齐机制是驱动模型性能提升的关键因素。

1 相关工作

(1) 视觉语言生成模型

最新的视觉语言(vision-language, VL)生成模型通过深度整合大语言模型与计算机视觉的前沿技术,有效弥合了文本与视觉模态之间的鸿沟。这种多模态整合主要得益于多模态预训练框架(如 CLIP (contrastive language-image-pretraining) [5])的创新,该框架通过在超大规模图文配对数据集上进行对比学习,成功地将视觉与语言表征结合起来。实验结果表明,此类模型在图像字幕生成^[6]、视觉问答^[7]和文本到图像生成^[8]多项任务上实现了性能的显著提升。

以 CLIP 和 DALLE 为代表的跨模态模型充分展现了自然语言处理与计算机视觉技术融合的巨大潜力,它们不仅能生成语义连贯的多模态内容,还能保持严格的上下文一致性。这些模型普遍采用 Transformer 架构。该架构不但在处理大规模数据集、捕获跨视觉语言输入的复杂依赖关系方面展现出显著优势,而且在零样本学习场景下展现出卓越的泛化能力。例如, DALLE 可直接根据文本提示生成图像,无须依赖特定任务的训练数据。尽管大语言模型在文本生成与理解领

域已取得显著成功,但在多模态多语言的实际应用场景中,如何实现其与视觉数据的有效集成,仍是当前研究面临的关键挑战。

(2) 多模态多语言安全

多模态与多语言安全领域致力于应对日益复杂的网络攻击,这些攻击利用多语言文本、听觉和视觉数据进行渗透。随着全球数字平台的广泛采用,开发能够有效检测和抵御此类攻击的先进防御机制变得愈发迫切。

该领域的研究致力于开发融合自然语言处理、计算机视觉和语音识别的多模态威胁检测系统。当前主要研究方向包括:①基于文本语义增强的视觉防御框架,通过文本描述信息强化图像特征的鲁棒性^[9];②跨模态对抗样本生成技术利用音频特征辅助构建更复杂的攻击样本^[10];③多语言威胁检测模型聚焦于解决不同语言的恶意内容识别难题,以提升系统的泛化能力^[11-12]。值得注意的是,最新研究特别关注文化语境和上下文细微差别,这对于提升不同语言环境下威胁检测的准确性至关重要^[13]。这一高度跨学科的研究领域综合了网络安全、计算语言学 and 人工智能等技术,其终极目标是构建具有自适应能力的智能防御系统,能够有效应对日益复杂的多模态网络威胁^[14]。

(3) 大语言模型的对抗训练

近年来,大语言模型的对抗训练研究聚焦于推理增强、连续空间优化、多模态鲁棒性提升和架构优化等方向。2025 年, OpenAI 提出的 o1 模型^[15]通过动态调整推理时间并优化算力资源分配,在不依赖预训练或先验攻击知识的情况下,显著提升了模型的对抗鲁棒性。在自动驾驶场景的测试中,该模型针对 Many-shot Attack 和 Soft Token Attack 等复杂攻击的防御表现尤为突出,使系统决策准确率提升了 23.5%。文献[16]提出的 C-AdvGLUE 算法突破了传统离散符号对抗攻击的计算瓶颈,通过直接在神经网络连续嵌入空间进行对抗训练,将计算资源消耗降至传统方法的 1/300。实验结果表明,该算法在 Gemma 和 Phi-3 等模型上能显著降低对抗攻击成功率(达 41.2%),同时有效规避防御过度问题。文献[17]提出的文本引导注意力机制解决了多模态场景下的对抗防

御问题。该方法通过对比分析对抗样本与干净样本的注意力分布差异,在CLIP模型微调过程中实现了显著的性能提升(在Tiny-ImageNet数据集上的零样本对抗鲁棒性较基线方法提高了37.19%)。

2 本文方法

2.1 架构概述

本文提出的模型mmDefender由跨语言编码器、视觉Transformer编码器和多语言解码器构成。具体而言,给定第 k 个图文对 (x^k, y^k) 及其对应图像 z^k ,首先使用跨语言预训练语言模型对源语句进行编码(在此过程中,将目标语言符号作为前缀添加至源语句前端,以标识翻译方向)。同时,将输入图像 $z^k \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times \mathcal{C}}$ ($\mathcal{H}$ 为高度, $\mathcal{W}$ 为宽度, $\mathcal{C}$ 为通道数)重塑为扁平化的图像块序列,并利用视觉Transformer编码器提取视觉上下文 s^k 。为缩小语言间的语义差异,模型将图像视为“中心语言”,通过多模态多语言对比学习机制,显式地将不同语言的表示映射至一个共享语义空间。随后,模型将包含 U 个标记的语言编码器状态 $s^k = \{s_u^k\}_{u=1}^U$ 与包含 V 个标记的辅助视觉编码器状态 $h^k = \{h_v^k\}_{v=1}^V$ 进行融合,生成条件化视觉语言记忆作为最终的多模态编码状态 $\{e_u^k\}_{u=1}^U$,将其输入多语言解码器 $\mathcal{D}$,以实现目标语言的翻译预测 y^k 。

然而,这种复杂架构也存在安全隐患:攻击者可通过在文本输入、视觉输入或二者协同输入中精心构造对抗扰动,诱导模型产生错误翻译。针对这一问题,本文提出一种对抗训练方法,通过在语言编码器的隐藏状态、辅助视觉编码器的输出表示及文本解码器的输入中分别注入少量可控扰动,开展对抗训练,有效提升了模型的鲁棒性。

2.2 多模态多语言生成

给定一个包含图像的 M 个双语平行语料库 $D_{\text{all}} = \{D_m\}_{m=1}^M$,其中 M 表示涵盖 N 种语言 $L_{\text{all}} = \{L_n\}_{n=1}^N$ 的训练语料数量(L_n 为第 n 种语言),每个图文双语语料 $D_m = \{x^k, y^k, z^k\}_{k=1}^K$ 由源语句、目标语句及对应图像构成。多模态多语言翻译的训练目标可形式化表述为:

$$\mathcal{L}_m = - \sum_{m=1}^M \mathbb{E}_{x^k, y^k, z^k \in D_m} [\log P(y^k | x^k, z^k; \Theta)] \quad (1)$$

其中,多模态多语言模型在所有翻译方向上采用完全共享的参数 Θ 。模型采用Transformer作为语言和视觉编码的骨干架构。具体而言,使用预训练的多语言模型XLM-R^[18]初始化语言编码器,并使用预训练的CLIP模型初始化视觉编码器。目标语言符号(如[En]或[De])作为前缀被添加在源语句前,以明确翻译方向。

对于给定的源语句 x^k 、图像 z^k 及其目标语言翻译 y^k ,需要构建适用于多模态提示的纯解码器模型(如Llama2)的完整输入,其中 L_i 和 L_j 分别表示源语言与目标语言。使用视觉模型将原始图像 z^k 编码为 V 个视觉标记 $h^k = \{h_v^k\}_{v=1}^V$ 。

在编码器-解码器架构中,模型将源语句的标记 x^k 输入文本编码器,同时将图像标记 z^k 输入视觉编码器,两者分别进行独立编码。

2.3 多模态编码

在编码器-解码器架构中,给定文本提示,模型通过语言Transformer编码器对由 U 个标记组成的串联序列进行编码,获得语言表征 s^k 。

$$s^k = \mathcal{S}(x^k) \quad (2)$$

其中, $\mathcal{S}$ 表示语言编码器, x^k 为第 k 个源语句。

对于输入图像 $z^k \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times \mathcal{C}}$,将其重塑为一个扁平化的图像块序列 $h \in \mathcal{R}^{\mathcal{V} \times (\mathcal{P}^2 \times \mathcal{C})}$ 。其中, $\mathcal{P}$ 表示每个图像块的分辨率; $\mathcal{V}$ 表示图像块的总数, $\mathcal{V} = \frac{\mathcal{H} \times \mathcal{W}}{\mathcal{P}^2}$ 。基于Transformer的视觉编码器 $\mathcal{H}$ 提取原始图像 z^k 的视觉表征过程可表示为

$$h^k = \mathcal{H}(z^k) \quad (3)$$

其中, $\mathcal{H}$ 表示Transformer视觉编码器, h^k 为生成的视觉表征向量。

2.4 多模态多语言融合

为有效融合多语言文本与视觉特征,可将图像视作通用语言,以弥合不同语言间的表征差异。基于此,本文提出多模态多语言对比学习机制,通过强化图文对齐及多语言文本对齐能力,实现多模态语义空间的统一建模。该机制采用InfoNCE目标函数^[19]学习图文对应关系,具体通过最小化两项多模态对比损失之和实现。

$$\mathcal{L}_c = \sum_{(x^k, y^k, z^k) \in D_{\text{all}}} (f(x^k, y^k) + f(x^k, z^k)) \quad (4)$$

其中, D_{all} 为包含多语言图文对的训练数据集, $f(x^k, y^k)$ 与 $f(x^k, z^k)$ 分别表示文本-文本相似度与文本-图像相似度的对比损失。具体计算公式如下:

$$f(x^k, z^k) = -\log \left(\frac{e^{\frac{x^k \cdot z^k}{\tau}}}{\sum_{z^j \in \{z^-\}} e^{\frac{x^k \cdot z^j}{\tau}}} \right) \quad (5)$$

式中, τ 为温度超参数, x^k 是与图像片段嵌入 z^k 配对的正例文本片段嵌入。计算文本-文本对比损失时, z^- 表示训练批次 B 中其他文本片段隐式构成的负例嵌入; 计算文本-图像对比损失时, z^- 表示图像片段嵌入。

为构建训练批次 B 中的多语言文本并平衡多个双语语料库, 本文采用基于温度的采样方法, 通过采样概率 $\{q_1, \dots, q_M\}$ 在单批次中汇集不同语言的句子。

2.5 注入式多模态多语言

本文聚焦于利用多模态多语言对比目标中的困难样本, 对多语言图文对齐关系展开建模。具体而言, 在训练阶段, 采用掩码语言建模与掩码图像建模技术, 构建语义近似但实际为负例的困难样本。

不过, 困难样本的引入和多模态数据固有的复杂性, 使模型面临潜在的对抗攻击风险。攻击者能够通过微调文本或图像输入生成对抗样本, 误导模型产生错误的对齐或翻译结果。这要求开发鲁棒的防御机制, 以有效检测并消除此类对抗扰动。

在图像增强方面, 本文运用函数对原始图像进行裁剪、尺寸调整、旋转、遮挡、色彩变换、高斯模糊和 Sobel 滤波等增强处理。随后, 将处理后的图像划分为规则的非重叠区块, 并基于均匀分布采样进行区块掩码, 以此作为掩码图像建模的组成部分。值得警惕的是, 这些增强技术可能衍生新的安全漏洞, 攻击者可借由增强过程植入难以察觉的扰动, 这些扰动会对模型的性能造成显著影响。

针对多语言文本, 本文随机屏蔽连续的标记

跨度。对每个句子, 本文采用多语言数据增强 $\mathcal{T}(\cdot)$ 对不同语言的原始句子进行增强。增强后的源句子与经过多模态多语言增强的图像 $\{\mathcal{T}(x^k), \mathcal{T}(z^k)\}$, 共同用于强化对比学习, 进而学习特定的表征不变性。但这一过程也为攻击创造了新途径, 在训练过程中, 对手可能通过在文本或图像增强环节制造扰动来欺骗模型, 从而削弱模型的鲁棒性和可靠性。因此, 开发能在增强和学习阶段有效检测并缓解对抗操作的强大防御机制至关重要。

为优化文本处理, 本文引入了“反翻译”方法。针对目标句子, 先将其翻译为源语言, 再翻译回目标语言。在此过程中, 翻译机制能够消除原始数据中的部分扰动, 保障数据和模型的鲁棒性。

此外, 本文设计了文本-图像联合增强方法。在多模态多语言任务中, 若仅对单一模态(如图像或文本)进行独立增强, 可能会破坏原始数据中固有的跨模态关联。因此, 在增强过程中, 本文尝试同步修改两种模式(图像与文本)的数据, 并对其联合增强操作。例如, 当对图像实施剪切处理时, 同步删除文本描述中与剪切区域对应的内容。这种同步操作确保了增强后的文本-图像对在语义上保持一致, 有效避免了单模态增强导致的语义偏差。同时还引入了解抗样本生成技术, 通过在图像特征空间与文本嵌入空间中同步注入微小扰动, 实现跨模态的协同数据增强, 进一步强化多模态表征的鲁棒性。

2.6 多任务学习

本文方法的任务可分解为文本翻译与图像描述两大任务。

(1) 文本翻译任务

为有效训练文本编码器与解码器, 本文设计了基于源语言输入的单模态目标词预测机制。

$$y_t^k = \mathcal{D}(y_{1:t-1}^k, s^k; \theta) \quad (6)$$

式中, $\mathcal{D}$ 表示标准的 Transformer 解码器, y_t^k 表示基于历史 $t-1$ 个目标词 $y_{1:t-1}^k$ 生成的第 t 个目标词。

(2) 图像描述任务

为实现图像与目标语言的对齐, 本文采用基于纯视觉上下文的图像描述训练目标。

$$y_t^k = \mathcal{D}(y_{1:t-1}^k, h^k; \theta) \quad (7)$$

给定融合了语言和视觉信息的条件化视觉语言记忆 $e^k = \{e_u^k\}_{u=1}^U$, 使用标准 Transformer 解码器进行序列化目标词预测。

$$y_t^k = \mathcal{D}(y_{1:t-1}^k, e^k; \theta) \quad (8)$$

本文采用多模态丢弃网络 (MDropNet) 训练策略, 具体如下。

- 1) 25%时间单独训练文本翻译目标。
- 2) 25%时间单独训练图像描述目标。
- 3) 50%时间联合训练多模态目标。

2.7 训练目标

在训练过程中, mmDefender 通过联合最小化公式的多模态多语言对比学习目标与翻译目标进行优化。

$$\mathcal{L}_{\text{all}} = \mathcal{L}_m + \lambda \mathcal{L}_c \quad (9)$$

其中, 超参数 λ 用于平衡翻译损失项 $\mathcal{L}_m$ 与多模态对比损失项 $\mathcal{L}_c$ 的权重。

同时需警惕训练过程对抗性攻击的脆弱性。攻击者可能通过注入细微扰动误导模型学习, 导致性能下降。针对以上风险, 本文通过对抗性训练, 以提高 mmDefender 在实际应用中的鲁棒性与可靠性。

3 实验

3.1 实验设置

本文在 Multi30K^[20] 数据集上开展实验。训练集和验证集分别包含 29 000 个和 1 000 个句子。该数据集涵盖英语 (En)、德语 (De)、法语 (Fr) 和捷克语 (Cs) 4 种语言, 每个句子对都有相应的图像。模型性能测试在 Flickr2016 (即 Multi30K 2016 版) 数据集^[21-22] 上开展。

3.2 对比基线

在纯文本方法中, XLM-R、CLIP、M3P^[23] 和 DAS-CL^[24] 均采用基于 Transformer 的架构, 但各自具有独特的训练策略。XLM-R 专注于单一翻译方向的训练; CLIP 采用联合训练方式处理多语言数据, 并通过添加目标语言标识符来区分不同翻译任务; M3P 通过多语言预训练和跨语言对比学习来增强语义对齐能力; DAS-CL 则创新性地采用动态自适应采样策略, 在训练过程中重

点优化难样本以提高模型鲁棒性。在多模态处理方面, 这些模型通过门控融合或特征连接等技术有效地整合视觉上下文信息。值得注意的是, 为确保模型在多语言和多模态任务上的一致性, 所有模型的文本编码器都统一使用 XLM-R 和 CLIP 的视觉编码器参数进行初始化, 这种策略显著提升了模型的跨模态理解能力和泛化性能。

3.3 实现细节

在编码器-解码器设置中, 模型包含一个由跨语言预训练编码器 XLM-R 初始化的语言编码器, 以及一个由 CLIP 初始化的视觉编码器, 并采用 Adam^[25] ($\beta_1 = 0.8$, $\beta_2 = 0.9$) 优化器训练多语言模型。在仅解码器设置中, 本文采用 LLAMA2 执行文本生成任务, 同时通过 CLIP 提取视觉特征。

学习率设定为 5×10^{-4} , 热身步长为 4 000 步。模型训练采用平滑系数为 0.1 的标签平滑交叉熵损失函数。本文所提模型包含视觉编码器、语言编码器和语言解码器, 三者均由 12 层构成, 隐藏层维度为 768, 且共享同一嵌入矩阵。

在多语言训练中, 硬件配置为 8×Tesla V100 GPU, 批处理大小为每批 2 048 个令牌。评估指标采用区分大小写的 SacreBLEU 分数。

3.4 结果分析

(1) 双语训练

在双语训练中, 模型针对单一语言对 (如 En→Fr 或 Fr→En) 开展专项训练, 学习在两种语言之间进行互译。这一训练机制通过聚焦特定翻译方向, 确保模型能够深度适配语言对的特性。数据集的划分遵循一致性原则: 训练集包含两种语言的平行句对, 验证集与测试集同样基于该语言对构建, 从而保证评估场景与训练目标的匹配性。

如表 1 所示, 在双语训练环境下, mmDefender 在不同语言对的翻译性能上呈现出显著优势。具体而言, 该模型在几乎所有语言对中均取得最优结果, 性能超越 XLM-R、CLIP、Gated Transformer 及 Unified Transformer 等基准方法。例如, 在 En→Fr 翻译任务中, mmDefender 的得分为 68.0 分, 显著优于次优方法 DAS-CL 的

63.7 分；类似的优势在其他语言对中同样明显，如在 En→Cs 和 En→De 翻译任务中的得分分别为 35.0 分和 41.0 分。mmDefender 在 6 种翻译任务中的平均得分为 48.3 分。这一结果表明，mmDefender 通过其创新方法（得益于先进的多模态数据集成策略），有效提升了特定语言对的翻译性能。

（2）多语言训练

多语言训练进一步扩展了双语训练的理念，通过在训练过程中同步融入多种语言数据实现能

力升级。模型在 En、De、Fr、Cs 共 4 种语言上开展训练，目标是学习不同语言对之间的跨语言迁移模式，从而具备处理多向翻译任务的能力。

从表 1 的数据可以看出，mmDefender 在多语言训练任务中的表现依然出色。尽管相较于双语训练，其单项任务的性能提升幅度有所收窄，但 mmDefender 在整体上仍保持了显著的竞争优势。例如，在 En→Fr 翻译任务中，mmDefender 的得分为 65.2 分，高于 XLM-R（64.2 分）和 CLIP（63.8 分）等方法；在 En→De 和 De→En 翻

表 1 在 Flickr2016 数据集上的实验结果
Table 1 Experimental results on the Flickr2016 dataset

训练策略	方法	En → Fr	En → Cs	En → De	Fr → En	Cs → En	De → En	平均
双语训练	XLM-R	62.1	32.2	37.2	52.1	42.2	43.3	44.9
	CLIP	60.2	31.5	35.3	51.5	42.3	41.3	43.7
	Gated Transformer	62.5	33.4	38.0	53.1	43.5	44.2	45.8
	Unified Transformer	62.7	33.2	38.1	53.0	43.5	44.0	45.8
	M3P	62.9	33.4	39.2	53.1	43.7	44.1	46.1
	DAS-CL	63.7	34.7	39.6	54.2	44.1	45.2	46.9
	mmDefender（本文方法）	68.0	35.0	41.0	55.0	45.0	46.0	48.3
多语言训练	XLM-R	64.2	35.3	38.1	53.4	43.0	45.1	46.5
	CLIP	63.8	34.8	37.8	52.9	42.9	44.8	46.2
	Gated Transformer	64.5	35.7	38.2	53.5	43.8	46.1	47.0
	Unified Transformer	64.2	34.8	38.0	52.8	42.7	45.4	46.3
	M3P	64.2	35.4	38.8	54.1	44.0	45.5	47.0
	DAS-CL	64.1	35.7	39.9	55.2	44.8	46.7	47.8
	mmDefender（本文方法）	65.2	35.1	40.9	54.9	45.1	46.0	47.9
自我训练	XLM-R	65.3	35.8	39.5	54.8	45.4	46.3	47.8
	CLIP	64.8	34.8	37.5	53.9	44.8	45.8	46.9
	Gated Transformer	63.0	34.9	39.2	54.5	45.4	46.2	47.2
	Unified Transformer	63.5	35.1	38.9	54.8	45.8	45.5	47.3
	M3P	65.2	36.1	40.1	55.2	45.0	45.2	47.8
	DAS-CL	65.7	36.0	41.2	55.8	45.8	45.7	48.4
	mmDefender（本文方法）	66.2	35.8	41.5	56.0	46.0	46.2	48.6
反向翻译训练	XLM-R	66.3	35.8	39.3	56.2	43.4	44.5	47.6
	CLIP	66.1	35.2	39.1	56.1	43.8	45.6	47.7
	Gated Transformer	66.0	36.1	39.1	57.0	44.2	46.4	48.1
	Unified Transformer	65.5	37.5	40.5	55.8	44.8	45.8	46.2
	M3P	66.2	35.9	40.1	56.5	44.2	45.3	46.7
	DAS-CL	67.8	36.5	41.1	55.8	45.3	46.5	47.2
	mmDefender（本文方法）	66.8	36.2	42.1	57.3	46.3	46.3	47.7

译任务中, mmDefender 的得分分别为 40.9 分和 46.0 分。这进一步印证了其在复杂多语言场景中处理双向翻译的鲁棒性。从综合表现来看, mmDefender 的平均分为 47.9 分, 这一结果凸显了其在多语言环境中的一致性与适应性。各项指标的增量提升表明, mmDefender 能够有效利用多语言训练数据中的跨语言关联信息, 确保模型在不同语言对之间实现良好泛化。

(3) 自我训练

自我训练是一种利用模型自身的预测结果扩充训练数据的技术。在完成初始训练后, 模型会对未标注数据进行翻译预测, 并将这些预测结果整合至训练过程中, 通过迭代方式持续优化自身性能。

如表 1 所示, 整合自我训练数据后, mmDefender 的性能得到进一步提升, 在多个语言对翻译任务中均取得更高分数, 突显了自我训练在提升翻译质量上的显著优势。具体来看, 在 En→Fr 翻译任务中, mmDefender 的得分提升至 66.2 分, 在所有对比方法中位居榜首; 在 Fr→En 翻译任务中, 其得分为 56.0 分, 同样领先于其他方法。从综合表现来看, mmDefender 的平均分为 48.6 分, 这一结果证明了整合自我训练数据在增强模型准确性与鲁棒性方面的有效性, 表明 mmDefender 能够充分利用自我训练数据提供的额外信息优化翻译结果。

(4) 反向翻译训练

反向翻译是指借助辅助翻译模型将目标语言句子回译为源语言, 生成合成平行句对, 并将其

用于主模型进一步训练的技术。通过提供额外的训练样本, 该技术已被证实能有效提升机器翻译性能。

如表 1 所示, 整合反向翻译数据后, mmDefender 的翻译性能实现新突破。在 En→Fr 和 Fr→En 翻译任务中, mmDefender 的得分分别为 66.8 分和 57.3 分, 超越 DAS-CL 和 M3P 等其他方法; 在 Cs→En 翻译任务中, mmDefender 的得分为 46.3 分, 展现出优异的跨语言转换能力。从综合表现来看, mmDefender 的平均分为 47.7 分, 在同类方法中保持竞争力。这些结果表明, 反向翻译数据通过构建丰富的语言语境、减少语义歧义, 显著提升了翻译准确性。实验结果证实, mmDefender 能够高效利用反向翻译数据, 持续优化其在多语言对间的翻译性能。

(5) 对抗性训练

对抗性训练是一种旨在提升模型鲁棒性的技术。其核心机制是将模型暴露于对抗性样本(这类样本通过对原始输入施加扰动而生成, 专门用于诱导模型产生错误预测)。通过主动引入此类样本进行训练, 模型能够学习抵抗扰动, 从而增强对异常或恶意输入的适应能力。

如表 2 所示, mmDefender 通过逐步引入文本编码器噪声、图像编码器噪声和文本解码器噪声, 系统性地验证了对抗性训练的有效性。最终模型在 6 组翻译任务上的平均性能达到 47.6 BLEU, 较基线 XLM-R 提升 1.5 BLEU, 其中对低资源语言(如 En→Cs) 的改善尤为显著(+1.2 BLEU)。这表明对抗性扰动能有效增强模型对输入噪声的鲁

表 2 不同模块在 Flickr2016 上的消融实验(mmDefender 为最终模型)

Table 2 Ablation experiments of different modules on Flickr2016 (mmDefender: final model)

方法	En → Fr	En → Cs	En → De	Fr → En	Cs → En	De → En	平均
XLM-R	63.0	33.8	38.8	53.3	43.6	44.0	46.1
CLIP	63.1	33.9	39.0	53.5	43.8	44.3	46.3
Gated Transformer	63.3	34.0	39.5	53.7	44.0	44.6	46.6
文本编码器噪声	63.4	34.1	39.8	53.8	44.2	44.8	46.8
图像编码器噪声	63.5	34.2	40.0	54.0	44.4	45.0	47.0
文本解码器噪声	63.6	34.3	40.3	54.1	44.6	45.2	47.2
对比学习	63.7	34.4	40.5	54.2	44.7	45.3	47.4
多任务学习	63.8	34.5	40.7	54.3	44.8	45.5	47.5
mmDefender (本文方法)	64.0	35.0	41.0	54.5	45.0	45.8	47.6

棒性，尤其在多模态场景下，结合对比学习 (+0.2 BLEU) 和多任务学习 (+0.1 BLEU) 后，模型对跨语言和跨模态扰动的抵抗能力得到协同提升。

4 安全性分析

4.1 攻击方法

(1) 攻击文本编码器

在视觉语言模型中，文本编码器负责将自然语言输入映射为密集向量表示，其可靠性直接影响多模态对齐效果。针对文本编码器的攻击往往通过干扰输入文本来实施，具体手段包括单词替换、字符级操作或句法重排序等，这些干扰可能导致模型生成错误的语言表示。

为文本编码器构造的对抗性示例，能够在表示层面误导模型，进而使视觉内容与文本查询之间形成不相关或不准确的关联。目前，可用于破坏文本编码器所学表示的技术多种多样，如词嵌入扰动、基于梯度的对抗性攻击（如 FGSM 或 PGD），以及基于同义词的攻击等。这类攻击会直接导致下游任务（如图像-文本检索、图像字幕生成等）失效。

(2) 攻击图像编码器

图像编码器的核心功能是将原始像素数据转化为高级视觉特征。针对图像编码器的攻击，通常沿用计算机视觉领域中成熟的图像扰动方法，具体技术路径包括：通过对抗攻击算法（如快速梯度符号法、投影梯度下降法）生成难以察觉的像素级扰动，或采用更加复杂的攻击手段（如 Carlini-Wagner 攻击、DeepFool 攻击）。这类扰动会欺骗图像编码器生成误导性的特征表示，直接导致视觉语言模型对视觉场景产生误判，进而输出错误的文本描述或响应。此外，部分对抗技术侧重于在关键视觉区域通过添加噪声或引入局部变化，刻意扭曲视觉语言系统的注意力机制，最终破坏其将视觉内容与文本信息正确关联的能力。

(3) 攻击文本解码器

文本解码器的核心功能是基于文本与视觉输入的编码表示生成或合成文本输出。针对文本解码器的攻击旨在干扰其生成逻辑，导致输出文本出现不连贯、不相关或错误等问题。具体攻击路

径可分为以下几类。

1) 注意力机制干扰：通过对抗性攻击操纵模型的注意力机制，对视觉区域与文本生成之间的对齐关系产生偏差或误导。

2) 潜在嵌入空间篡改：采用基于梯度的方法修改潜在嵌入空间，使看似正常的输入触发幻觉或无意义的输出。

3) 解码器稳定性破坏：利用基于梯度的对抗性样本、隐藏状态扰动或输入依赖的文本提示等技术，直接破坏解码器的稳定运行，导致其生成有缺陷的描述、标题或响应。

4) 关联偏见利用：通过强化视觉与文本标记间的意外关联，借助训练数据中的固有偏见，在受控对抗性输入下诱导模型生成有毒或攻击性文本，尤其在脆弱环境中更易生效。

(4) 跨模态协同攻击

跨模态协同攻击是指攻击者通过同时对模型的文本输入和图像输入施加扰动，诱导模型在视觉特征和文本特征的识别上产生误判。相较单一模态攻击，防御这种攻击更具挑战性，因为它要求模型能够同时抵御来自两种不同模态的干扰。在跨模态协同攻击中，攻击者常采用以下两种典型策略实现多模态扰动的协同作用。

1) 文本扰动触发视觉特征误判：攻击者在文本输入中植入细微扰动（如同义词替换、拼写错误等），通过干扰模型对文本语义的理解，间接误导视觉编码器生成错误的视觉特征表示。

2) 视觉扰动触发文本生成错误：攻击者借助 FGSM、PGD 等算法在图像输入中添加像素级对抗性扰动，试图误导文本解码器生成错误的文本输出。

4.2 攻击效果和防御性能分析

在实验中，针对文本编码器选用 TextFooler 攻击，针对图像编码器选用快速梯度符号法攻击，针对文本解码器则采用自定义的“生成欺骗”攻击。本文评估了在不同条件下这些攻击对模型性能的影响，并通过 mmDefender 进行防御，相关结果如表 3 所示。

4.2.1 准确率评估

(1) 文本编码器攻击

通过替换同义词和引入拼写错误干扰输入文

表3 攻击方法与防御性能分析

Table 3 Analysis of attack methods and defense performance

攻击方法	准确率				
	基线	攻击后	防御后	攻击影响	防御效果
文本编码器攻击	92.5%	77.0%	89.0%	-15.5%	+12.0%
图像编码器攻击	93.8%	68.2%	91.5%	-25.6%	+23.3%
文本解码器攻击	94.2%	70.5%	91.8%	-23.7%	+21.3%
文本扰动触发视觉特征误判攻击	92.5%	77.0%	88.5%	-15.5%	+15.5%
视觉扰动触发文本生成错误的攻击	93.8%	68.2%	90.5%	-25.6%	+22.3%

本。在未应用 mmDefender 时,模型的准确率下降约 15.5%;应用 mmDefender 后,准确率提升至 89.0%,提升了 12.0%。

(2) 图像编码器攻击

采用快速梯度符号法对图像施加对抗性扰动,导致模型性能显著下降,准确率降低 25.6%。不过,mmDefender 显著增强了模型的鲁棒性,使准确率恢复至 91.5%,较无防御状态提升 23.3%。

(3) 文本解码器攻击

通过模拟“生成欺骗”攻击迫使解码器生成错误内容,导致生成文本准确率下降 23.7%。应用 mmDefender 后,准确率恢复至 91.8%,提升 21.3%。

(4) 跨模态协同攻击

文本扰动触发视觉特征误判攻击的攻击:采用基于梯度的字符级 HotFlip 攻击(通过最小化模型对分类标签的置信度,选择字符替换、插入或删除操作),旨在误导视觉编码器。未应用 mmDefender 时,模型的准确率下降约 15.5%;应

用 mmDefender 后,准确率提升至 88.5%。

视觉扰动触发文本生成错误的攻击:通过 AutoAttack (APGD) 对图像施加对抗性扰动,导致模型性能显著下降,准确率降低 25.6%。应用 mmDefender 后,准确率提升至 90.5%,较无防御状态提升 22.3%。

4.2.2 生成质量评估

除准确性外,本文还评估了生成内容的质量(涵盖流畅性、一致性和逻辑连贯性)。表 4 呈现了攻击前、攻击后及应用 mmDefender 后的生成质量评分(1~5 分,5 分为最佳)。

在文本编码器攻击场景中,生成质量从 4.7 分降至 3.2 分,降幅为 31.9%;应用 mmDefender 后,质量回升至 4.5 分,回升率达 40.6%。

在图像编码器攻击场景中,生成质量从 4.8 分大幅降至 2.9 分,降幅为 39.6%;经 mmDefender 防御后,质量恢复至 4.6 分,相当于原始质量的 58.6%。

在文本解码器攻击场景中,生成质量同样受到显著冲击,从 4.6 分降至 3.1 分,降幅为 32.6%;

表4 生成质量对比

Table 4 Comparison of generation quality

攻击方法	生成质量评分				
	基线	攻击后	防御后	攻击影响	防御效果
文本编码器攻击	4.7	3.2	4.5	-31.9%	+40.6%
图像编码器攻击	4.8	2.9	4.6	-39.6%	+58.6%
文本解码器攻击	4.6	3.1	4.4	-32.6%	+41.9%
文本扰动触发视觉特征误判攻击	4.7	3.2	4.4	-31.9%	+37.5%
视觉扰动触发文本生成错误的攻击	4.8	2.9	4.5	-43.8%	+55.2%

应用 mmDefender 后, 质量提升至 4.4 分, 升幅为 41.9%。

在跨模态协同攻击实验中, 文本扰动触发视觉特征误判攻击导致生成文本质量从 4.7 分降至 3.2 分, 降幅为 31.9%, 应用 mmDefender 后回升至 4.4 分, 恢复率达 37.5%; 视觉扰动触发文本生成错误的攻击 (即针对图像编码器的攻击) 使生成质量从 4.8 分急剧降至 2.9 分, 降幅高达 43.8%, 经 mmDefender 防御后提升至 4.5 分, 恢复率达 55.2%。

上述结果表明, mmDefender 在防御跨模态协同攻击方面具有显著优势, 能有效提升模型的鲁棒性、保障生成质量, 并显著降低攻击对模型性能的影响。

4.2.3 训练时间和推理效率

模型在不同攻击条件下的训练时间与推理效率数据如表 5 所示。

在训练时间方面, 引入 mmDefender 后, 训练时间增加了 1~1.5 h, 这一额外开销主要源于训练过程中噪声对比学习带来的计算负载。

在推理效率方面, mmDefender 使单句推理时间增加了 0.03~0.05 s。考虑到其实现的实质性防御效果, 这种时间开销处于可接受范围, 表明该方法具有较高的实用价值。

4.3 防御策略

多模态元素的集成在提升翻译质量的同时, 也为模型引入了新的攻击漏洞。攻击者可通过操纵文本与视觉输入, 诱导模型生成错误的翻译。例如, 图像 z^k 或源语句 x^k 中难以察觉的扰动便可能引发显著的翻译偏差。此外, 数据中毒攻击会污染多模态训练数据, 严重损害模型的完整性与可靠性。

为应对上述安全挑战, 构建全面的防御机制至关重要, 具体包括以下 3 个核心策略。

1) 对抗性训练: 在模型训练过程中主动纳入对抗性样本, 通过模拟攻击场景迫使模型学习扰动特征, 从而增强模拟抵御各类对抗性攻击的鲁棒性。

2) 异常检测: 部署针对性的异常检测算法, 精准识别输入数据中可能预示攻击的异常模式。

3) 持续监控: 建立实时性能监控系统, 及时检测并预警由攻击引发的性能偏移, 为快速响应提供依据。

综合应用上述防御策略, 有效保障多模态多语言翻译系统的安全性与可靠性, 显著提升其抵御潜在攻击的综合能力。

5 结束语

本文提出的 mmDefender 方法通过多模态多语言对抗训练, 成功增强了大型语言模型在跨语言场景下的鲁棒性和翻译性能。实验结果表明, 该方法不仅能有效融合视觉与文本信息以提升翻译质量, 还能通过噪声对比学习显著降低对抗攻击的影响。mmDefender 的创新之处在于将多模态对齐与对抗训练相结合, 为构建安全可靠的多语言生成模型提供了新思路。

未来的研究工作将聚焦于进一步优化跨模态对齐机制, 探索更高效的对抗训练策略, 并致力于提升模型在低资源语言和复杂攻击场景下的表现。同时, 如何平衡模型性能与计算效率, 以及如何更好地保护数据隐私, 也将成为后续研究的重点方向。这些工作将推动多模态大语言模型在实际应用中的安全落地, 为人工智能技术的发展注入新的活力。

表 5 模型计算性能对比

Table 5 Comparison of computational performance among different models

攻击方法	攻击前训练时间/h	攻击后训练时间/h	防御后训练时间/h	推理时间/s	防御后推理时间/s
文本编码器攻击	10.5	10.8	12.0	0.35	0.38
图像编码器攻击	11.2	11.5	12.3	0.42	0.45
文本解码器攻击	10.8	11.0	12.1	0.38	0.41
文本扰动触发特征视觉误判	11.0	11.3	12.5	0.40	0.43
视觉扰动触发文本生成错误	10.9	11.4	12.6	0.41	0.44

参考文献:

- [1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. 2017: 5998-6008.
- [2] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models[EB]. 2023.
- [3] LIU Z L, LI Y W, SHU P, et al. Radiology-LLaMA2: best-in-class large language model for radiology[EB]. 2023.
- [4] XHONNEUX S, SORDONI A, GÜNNEMANN S, et al. Efficient adversarial training in LLMs with continuous attacks[EB]. 2024.
- [5] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 8738-8753.
- [6] XU K, BA J, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention[EB]. 2015.
- [7] ANTOL S, AGRAWAL A, LU J S, et al. VQA: visual question answering[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 2425-2433.
- [8] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[C]//Proceedings of International Conference on Machine Learning. New York: PMLR, 2021: 8821-8831.
- [9] 魏宣宣, 刘万平, 卢玲. 基于多模态特征融合的对抗样本防御方法研究[J]. 网络与信息安全学报, 2025, 11(2): 175-188.
WEI X X, LIU W P, LU L. Research on adversarial examples defense method based on multi-modal feature fusion[J]. Chinese Journal of Network and Information Security, 2025, 11(2): 175-188.
- [10] 田梓琳, 龙云飞, 田野, 等. 针对音频分类的跨表征可迁移对抗样本生成方法[J]. 网络与信息安全学报, 2025, 11(2): 125-135.
TIAN Z L, LONG Y F, TIAN Y, et al. Cross-representation transferable adversarial examples generation for audio classification[J]. Chinese Journal of Network and Information Security, 2025, 11(2): 125-135.
- [11] LIU Y H, YE H T, WEISSWEILER L, et al. A crosslingual investigation of conceptualization in 1335 languages[EB]. 2023.
- [12] ZHANG B, BAPNA A, JOHNSON M, et al. Multilingual document-level translation enables zero-shot transfer from sentences to documents[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 4176-4192.
- [13] KSHETRI N, ALCANTARA L L. Cyber-threats and cybersecurity challenges[M]. London: Routledge, 2015: 285-293.
- [14] JI H Y, YANG J, CHAI L Z, et al. SEvenLLM: benchmarking, eliciting, and enhancing abilities of large language models in cyber threat intelligence[EB]. 2024.
- [15] OpenAI. Improving robustness of large language models via inference - time enhancement[EB]. 2025.
- [16] FAN X J, TAO C L. Towards resilient and efficient LLMs: a comparative study of efficiency, performance, and adversarial robustness[EB]. 2024.
- [17] YU L, ZHANG H Y, XU C S. Text-guided attention is all you need for zero-shot robustness in vision-language models[EB]. 2024.
- [18] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 8440-8451.
- [19] VAN-DEN-OORD A, LI Y Z, VINYALS O, et al. Representation learning with contrastive predictive coding[EB]. 2018.
- [20] ELLIOTT D, FRANK S, SIMA'AN K, et al. Multi30K: multilingual english-german image descriptions[C]//Proceedings of the 5th Workshop on Vision and Language. Stroudsburg: ACL, 2016: 70-74.
- [21] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[M]//Computer Vision - ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.
- [22] BARRAULT L, BOUGARES F, SPECIA L, et al. Findings of the third shared task on multimodal machine translation[C]//Proceedings of the Third Conference on Machine Translation: Shared Task Papers. Stroudsburg: ACL, 2018: 304-323.
- [23] YANG J, GUO H, YIN Y, et al. M3P: towards multimodal multilingual translation with multimodal prompt[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. Torino: ELRA and ICCL, 2024: 10858-10871.
- [24] CHENG X X, ZHU Z H, LI Y W, et al. DAS-CL: towards multimodal machine translation via dual-level asymmetric contrastive learning[C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. New York: ACM, 2023: 337-347.
- [25] KINGMA D P, BA J. Adam: a method for stochastic optimization[C]//Proceedings of the 3rd International Conference on Learning Representations. 2015.

[作者简介]



张道娟 (1989-), 女, 山东临沂人, 中国电力科学研究院有限公司高级工程师, 主要研究方向为人工智能安全、智能攻防和威胁检测。



赵宇飞 (1990-), 男, 山西太原人, 中国电力科学研究院有限公司中级工程师, 主要研究方向为网络流量分析、软件逆向工程和人工智能安全。



张翱 (1981-), 男, 北京人, 中国电力科学研究院有限公司高级工程师, 主要研究方向为人工智能安全、智能攻防和威胁检测。



陈凯 (1985-), 男, 天津人, 中国电力科学研究院有限公司高级工程师, 主要研究方向为工控系统安全、人工智能安全。



吴天琦 (1996-), 女, 河南三门峡人, 中国电力科学研究院有限公司中级工程师, 主要研究方向为工控系统安全、人工智能安全。



戴聿雯 (1982-), 女, 安徽合肥人, 中国电力科学研究院有限公司高级工程师, 主要研究方向为智能攻防、人工智能安全。