
天眼AI——
通用全模态AIGC检测系统

摘要

随着生成式人工智能（AIGC）技术的快速发展，AI 生成内容在诈骗、学术、政务司法及社交媒体等领域的滥用引发了严峻的安全与诚信问题。现有检测系统面临模态覆盖不全、跨模型泛化能力弱、短文本检测效果差等技术瓶颈。本研究提出 “天眼 AI—— 通用全模态 AIGC 检测系统”，旨在构建覆盖文本、图像、视频、音频的全模态检测体系，实现对 AI 生成内容的精准鉴别。

检测系统通过 “预训练特征提取器 + 定制化分类器” 统一框架，冻结 CLIP:ConvNext、RoBERTa-MPU、CLIP:VIT、WavLM 等预训练模型提取通用特征，避免依赖特定生成模型伪影，提升跨域泛化能力。针对不同模态特性设计专属检测链路：文本检测采用 RoBERTa-MPU 融合多尺度语义特征，解决短文本语义稀疏问题；图像检测通过 CLIP:ConvNext 结合 Conformer-MLP 捕捉视觉伪影；音视频检测利用 CrossModal-STF-Fusion 模块，实现“时空-频域——单帧-时序——音频-画面”特征的高效融合与联合建模，实现动态媒体的时序异常检测。

实验基于 DICIX、HC3、Celeb-DF-v2、ASVspoof 等数据集，验证了检测模型在域内与跨域场景的有效性。结果显示，图像、文本、视频检测准确率分别达 98.51%、99.03%、97.11%，较主流方案平均提升 2.98%、3.02%、1.60%；音频等错误率（EER）降至 3.21%，跨域检测性能平均提升 1.13%-1.46%。系统突破传统检测模型依赖的局限，实现 “生成模型无关” 的通用检测，突破了短文本的技术瓶颈。

“天眼 AI” 通过技术创新与场景适配，为政务司法证据核验、学术诚信审查、社交媒体内容审核、AI 换脸诈骗防范等提供了高效解决方案，实现 AIGC 检测从单一模态适配向全模态检测的跨越，具有显著的技术前瞻性与社会应用价值。

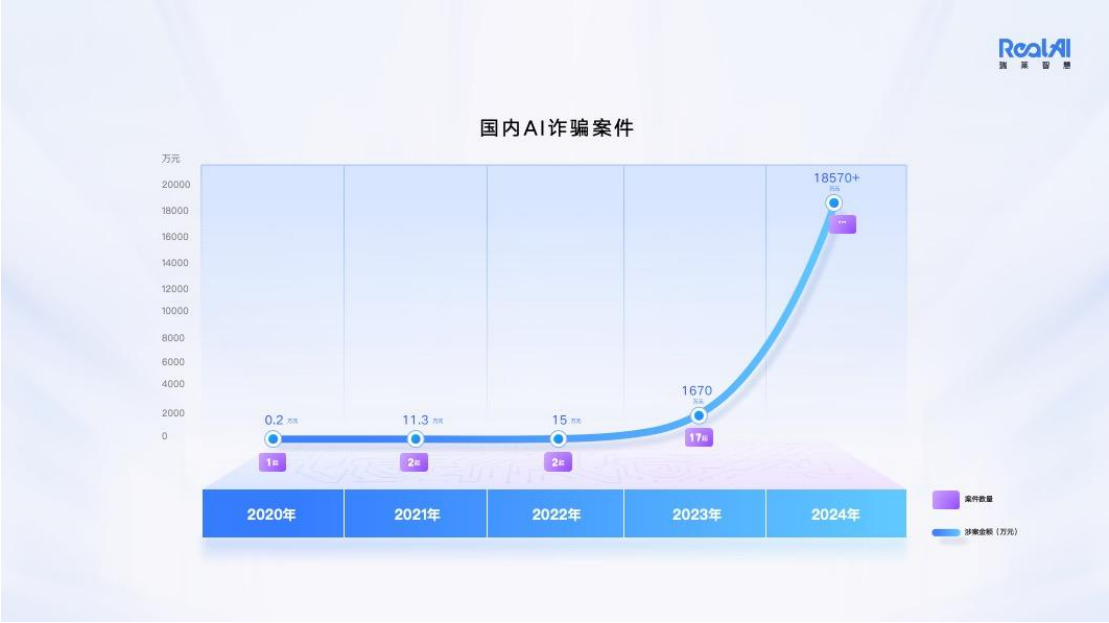
目　录

第 1 章 项目背景和国内外研究现状

1.1 项目背景

1.1.1 AI 技术诈骗挑战加剧

近年来，人工智能（AI）技术的迅猛发展催生了多种新型诈骗手段，给社会带来严重安全隐患。根据瑞莱智慧公司发布的数据，国内 AI 诈骗案件的涉案金额已从2020年的0.2万元激增至2023年的1670万元，年复合增长率高达1928.8%。2024 年上半年，AI 诈骗案件的涉案金额已超过 1.85 亿元，呈现持续上升趋势。



图表 1 国内AI诈骗案件涉案金额

犯罪分子利用深度伪造（Deepfake）技术，合成受害者亲友或熟人的声音和图像，实施精准诈骗。例如，2024 年 3 月，中国香港一名职员在一场“多人视频会议”中被骗走 2 亿港元，事后发现会议中除其本人外，其他参与者均为 AI 换脸冒充。

1.1.2 AI 生成论文造假

随着 AIGC 技术的兴起，学术领域受到了显著影响。在追求学术成果快速产出的压力下，部分学生和研究人员选择利用 AIGC 工具生成论文内容。据调查，在部分学术数据库中，约有 5% 的论文存在 AI 生成的嫌疑。

这种行为严重破坏了学术的公正性和严谨性。学术研究建立在真实、创新和诚信的基础之上，AI 生成论文造假行为不仅违背了学术道德，还干扰了正常的学术评价体系，使得真正有价值的学术成果难以得到公正的认可和推广，阻碍了学术的健康发展。

当前，学术界对于 AI 生成内容的检测需求日益迫切。然而，现有的检测技术在面对多样化的 AIGC 工具时，还存在诸多不足，无法有效满足学术领域对论文真实性检测的需求，这促使研发更高效、精准的 AIGC 检测系统成为当务之急。

1.1.3 AI 生成虚假信息传播

在数字化时代，社交媒体已成为信息传播的核心枢纽，深刻改变了信息的传播模式和公众的信息获取方式。每天，海量的信息在各类社交媒体平台上流通，这些平台不仅连接了全球范围内的用户，还成为了舆论形成和传播的关键场所。随着 AIGC 技术的迅速发展，其在社交媒体领域的应用日益广泛，使得信息传播的生态变得更加复杂和难以把控。

AI 写稿、AI 绘画等 AIGC 技术凭借其高效性和强大的内容生成能力，能够在短时间内创作出大量的文本、图像等内容。从表面上看，这似乎丰富了社交媒体的信息来源和表现形式，但实际上却带来了诸多隐患。其中最突出的问题就是虚假新闻、谣言等不良信息的大量滋生。这些虚假信息往往具有很强的迷惑性，它们可能会巧妙地利用公众关心的热点话题，精心编造看似合理的内容，以吸引用户的注意力。

以 “杭州撤限行令” 的假新闻为例，这一消息经 ChatGPT 编写后，在社交媒体上迅速传播。由于其内容涉及公众出行政策这一敏感话题，很快引发了民众的广泛关注和讨论。许多用户在未核实信息真实性的情况下，便进行了转发和评论，导致该假新闻在短时间内呈指数级扩散。这一事件不仅造成了社会的混乱，让公众对交通政策产生误解，还对政府的公信力造成了负面影响。许多人因为这条假新闻而调整了出行计划，给日常生活带来了不便。而且，这种虚假信息的传播还可能引发社会恐慌，导致公众对政府决策的信任度下降，进而影响社会的稳定发展。


新闻频道 > 法治新闻

杭州3月1日取消限行是假的，警方已介入

来源：浙江之声微博 | 2023年02月17日 10:15:23

原标题：网传杭州3月1日取消限行通稿是假的 假新闻是chatgpt写的

昨天，网络疯传一条关于杭州市政府3月1号取消限行的“新闻稿”。浙江之声记者调查发现，这是一条不实新闻。据了解，昨天下午杭州某小区业主群里讨论chatgpt，一位业主就开玩笑说尝试用它写篇杭州不限行的新闻稿看看。随后该业主在群里直播了chatgpt写作过程，还把文章发群里。其他业主不明所以截图转发了，最后导致错误信息被传播。据了解，警方已介入调查。涉事业主也在群里公开道歉。截至目前，杭州相关政府部门均没有发布此类政策。（浙江之声记者方梦玲）

图表 2 杭州撤限行令假新闻

为了应对社交媒体虚假信息传播的问题，相关部门一直在努力加强对社交媒体内容的管理。他们制定了一系列的规章制度，加大了对虚假信息发布者的处罚力度，同时也加强了对平台内容的审核。然而，由于 AIGC 技术生成的虚假信息具有高度的迷惑性，现有的检测和监管手段面临着巨大的挑战。

此外，社交媒体平台的信息传播速度极快、范围极广，信息一旦发布，往往在短时间内就能扩散到大量用户手中，这使得监管部门难以及时发现和处理所有的虚假信息。即使发现了虚假信息，在采取删除等措施时，也可能因为信息已经广泛传播而难以完全消除其负面影响。因此，迫切需要一种先进的检测技术，能够快速、准确地识别社交媒体上 AIGC 技术生成的虚假信息，从源头上遏制虚假信息的传播，净化社交媒体环境，维护社会公共秩序和公众的知情权。

1.2 国内外研究现状

随着 AIGC 检测技术的发展，众多学者及公司开发出诸多先进的 AIGC 检测模型和系统。这些模型在一定程度上推动了 AIGC 检测技术的进步，为识别和防范 AIGC 相关风险提供了有力支持。然而，不可忽视的是，传统的 AIGC 检测系统依然存在着显著的局限性。

1.2.1 检测模态单一

在当今数字化信息爆炸的时代，数据的形式愈发丰富多样，文本、图像、视频、音频等多种模态的数据广泛存在于各个领域。从社交媒体平台上用户发布的图文动态、视频分享，到学术研究中的文献资料、实验视频，再到司法领域的监控视频、语音证词等，不同模态的数据承载着大量关键信息。然而，现有多数检测系统存在明显的局限性，仅能针对少数模态进行检测。例如，腾讯的朱雀大模型检测只能对图像和文本进行检测，无法对 AI 生成的视频和音频等进行检测。

图表 3 朱雀大模型检测

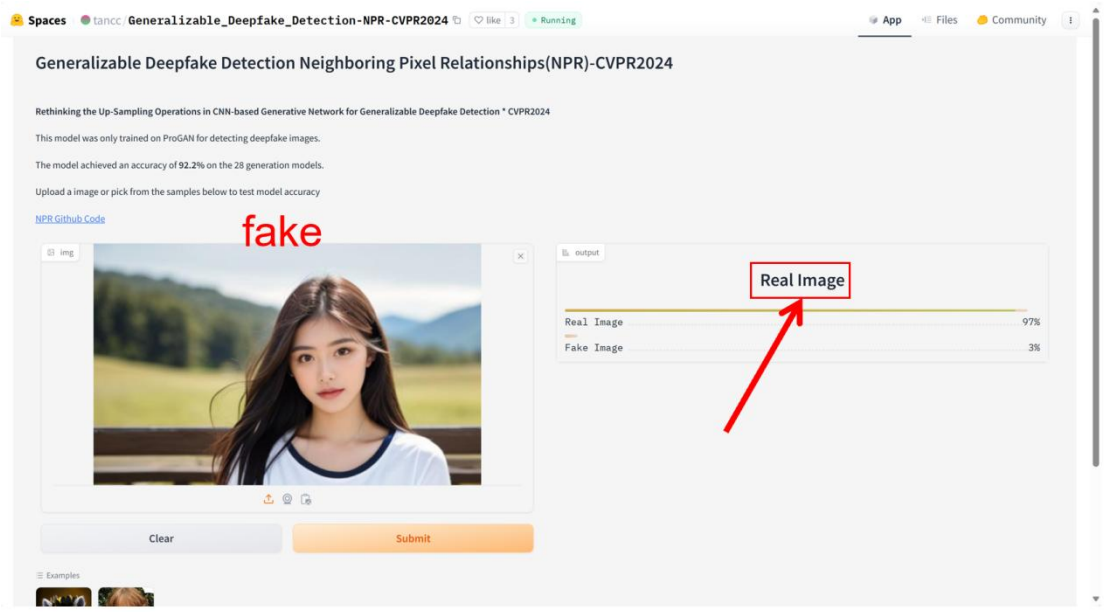
这种单一模态检测的现状，远远无法满足实际应用场景的需求。在防范 AI 诈骗场景中，诈骗分子可能同时利用 AI 换脸技术伪造视频、AI 合成声音进行语音通话以及 AI 写稿编造虚假信息，仅依靠单一模态的检测系统，难以全面识别诈骗行为，导致用户财产安全面临极大威胁。在学术领域，研究成果可能以论文文本、实验图像、讲解视频等多种形式呈现，若不能对这些不同模态的数据进行检测，就无法有效遏制 AI 生成内容在学术领域的造假行为，严重影响学术的公正性和严谨性。

1.2.2 跨生成模型泛化弱

随着人工智能技术的快速发展，生成模型的种类日益繁多，如 GAN（生成对抗网络）、扩散模型、Deepfake 等。不同的生成模型在生成内容的方式、特点和质量上存在差异。传统的检测模型往往依赖特定生成模型的指纹特征来进行检

测，例如，针对 GAN 生成图像的检测模型，主要通过识别 GAN 生成图像中特有的周期性模式、频率模式异常等低级伪影来判断图像是否为 AI 生成。

然而，当面对其他模型生成的内容时，这种依赖特定指纹特征的检测模型就显得力不从心。如图 4 所示，CVPR2024 中一篇论文提出的检测方法，在面对未知生成模型生成的图像时，无法有效识别这些图像是否为 AI 生成。



图表 4 跨域检测错误

在实际应用中，这一问题尤为突出。在社交媒体内容审核场景中，若检测系统只能检测特定模型生成的虚假信息，而对其他模型生成的内容毫无察觉，那么大量的 AI 生成虚假信息就会在平台上肆意传播，误导公众，引发社会舆论混乱。在司法领域，如果无法对各类生成模型生成的篡改证据进行准确检测，将严重影响司法公正和案件的正确裁决。因此，提升检测模型跨生成模型的泛化能力迫在眉睫，这样才能确保检测系统在面对不断涌现的新生成模型时，依然能够保持高效、准确的检测性能。

1.2.3 短文本检测缺陷

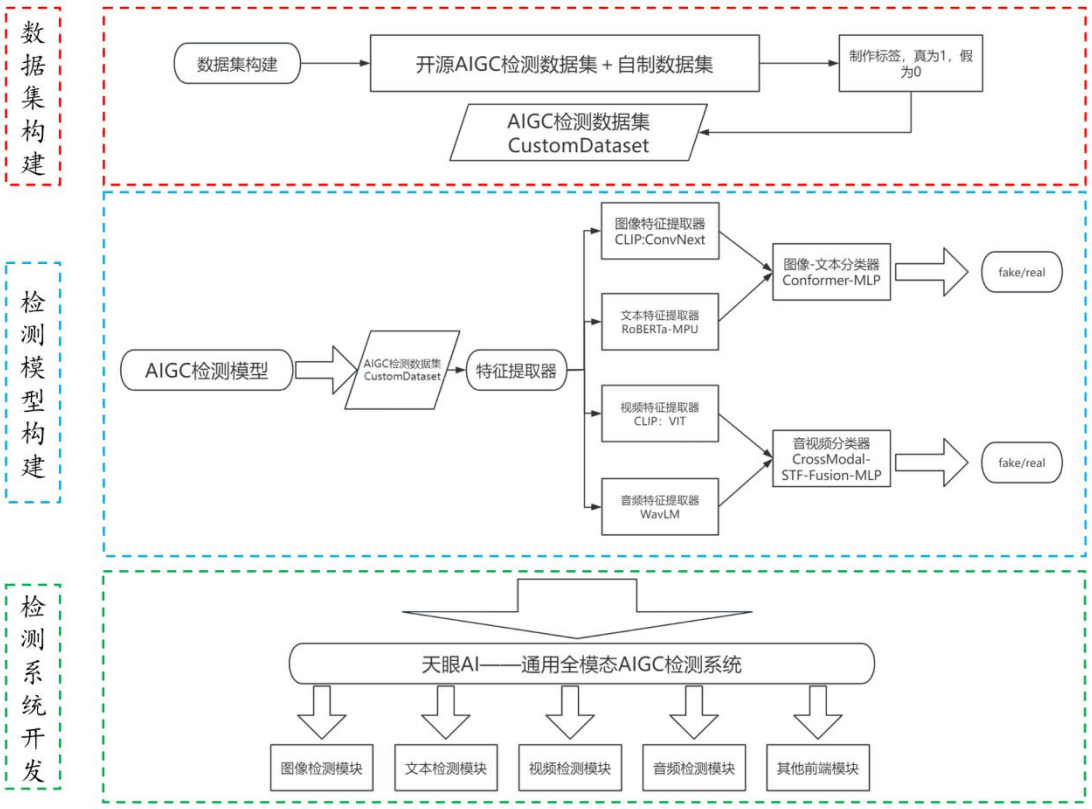
社交媒体、在线评论、即时通讯等众多场景中，短文本广泛存在且发挥着重要作用。例如，微博上的推文、电商平台的商品评论、论坛中的简短留言等，这些短文本往往能够快速反映用户的观点、情感和需求。然而，现有文本检测模型在处理短文本时，面临着诸多挑战，表现出性能显著下降的问题，甚至有些模型根本不支持短文本检测。例如腾讯的朱雀大模型检测，无法检测小于 350 字的文

本内容。

图表 5 短文本检测局限

在实际应用中，这一问题给内容管理和风险防范带来了极大困扰。在社交媒体平台上，大量短文本形式的虚假信息、广告营销、恶意评论等难以被有效识别和过滤，影响了平台的内容质量和用户体验。在舆情监测领域，无法准确检测短文本中的 AI 生成内容，可能导致对舆情的误判，错过最佳的应对时机。因此，解决短文本检测性能差的问题，是提升检测系统全面性和准确性的关键环节。

第 2 章 技术路线



图表 10 技术路线

如技术路线图所示，本项目主要从数据集的构建、检测模型的构建以及检测系统的开发三个方面进行了相关研究。

2.1 数据集的构建

第一阶段，在 GitHub、kaggle、谷歌等网站收集 AIGC 检测数据集，得到不同生成模型的各类 AIGC 数据集；最后整合搜集的各类 AIGC 数据，分为图像、文本、视频、音频四部分。设置真实数据标签为 0，AI 生成数据标签为 1，构建 AIGC 检测数据集。

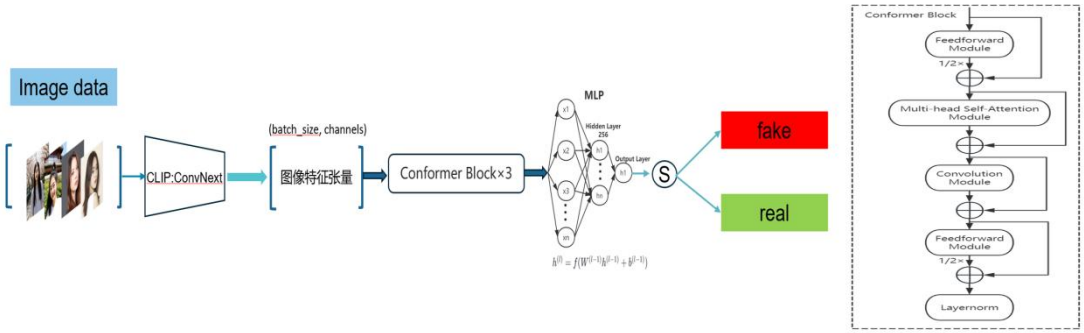
2.2 设计检测模型

第二阶段，构建通用 AIGC 检测模型。为了解决传统模型学习虚假特征的不良行为，避免模型过拟合特定生成模型的伪影，我们采用冻结的预训练模型作为特征提取器，并在特征空间中进行训练分类，实现高精度、高泛化性的检测。并

基于第一阶段构建的 AIGC 检测数据集，从图像、文本、视频、音频四大模态切入，构建全模态 AIGC 检测模型体系。

（一）图像检测模型：

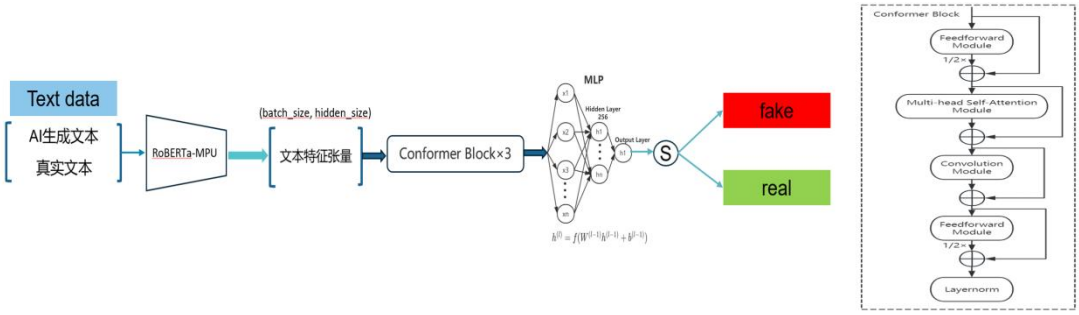
采用 “CLIP:ConvNext 特征提取器 + Conformer-MLP 分类器” 架构。首先，利用 CLIP:ConvNext 在 4 亿图文对预训练中形成的全局语义理解与局部细节捕捉能力，对输入图像进行特征抽取。然后，将提取的特征输入 Conformer-MLP 分类器：通过深度可分离卷积层进一步细化局部特征，利用 Transformer 自注意力机制建模图像长距离依赖关系，最后经多层感知器（MLP）融合特征，输出图像为 “fake”（AI 生成）或 “real”（真实）的概率值，实现对 GAN、扩散模型等生成图像的精准判别。



图表 11 图像检测模型网络图

（二）文本检测模型：

构建 “RoBERTa-MPU 特征提取器 + Conformer-MLP 分类器” 框架。针对短文本检测难题，RoBERTa-MPU 融合 RoBERTa 预训练模型的语义表征能力与多尺度正负无标签（MPU）框架，通过多窗口滑动机制提取字级、词级、句级多尺度语义特征，解决传统模型在短文本场景下存在的问题。搭配 Conformer-MLP 分类器，基于自注意力机制动态聚焦文本关键语义片段，从语法连贯性、语义逻辑性、语言特征规律性等维度综合分析，强化对短文本的检测能力，输出文本生成属性判别结果。



图表 12 文本检测模型网络图

（三）音视频跨模态检测模型：

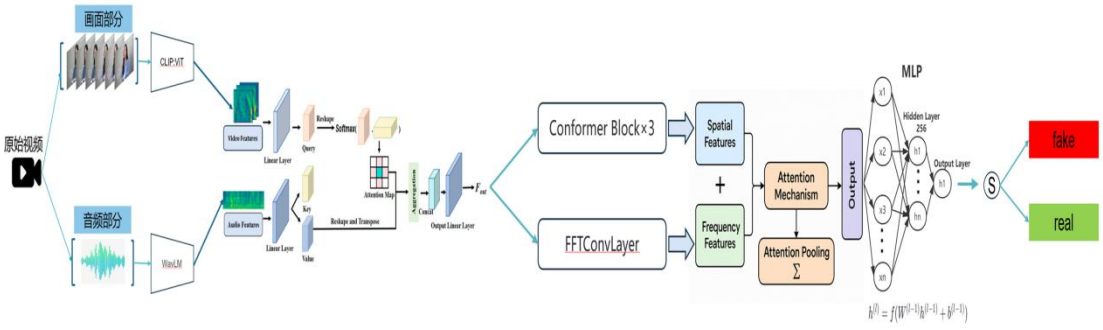
设计 “（CLIP:ViT + WavLM）特征提取器 + 跨模态时空频域融合模块（CrossModal-STF-Fusion）” 体系。CLIP:ViT 基于大规模互联网图文数据预训练，提取视频帧的视觉语义特征；WavLM 通过改进的 Transformer 架构提取音频信号的长时上下文依赖特征。后续接入的 CrossModal-STF-Fusion 模块包含四重处理机制：

跨模态注意力交互层：以视频特征为查询（Query），音频特征为键值对（Key/Value），通过多头注意力机制计算模态间关联权重，实现视觉与听觉特征的细粒度对齐与互补融合，生成跨模态特征表征。

频域特征挖掘分支：Conformer 模块通过深度可分离卷积捕捉单帧局部特征，结合 Transformer 自注意力机制建模视频帧序列的时序依赖。

频域特征挖掘分支：FFTConvLayer 模块对跨模态特征进行快速傅里叶变换，将特征映射至频域，挖掘音视频信号的周期性频谱异常。

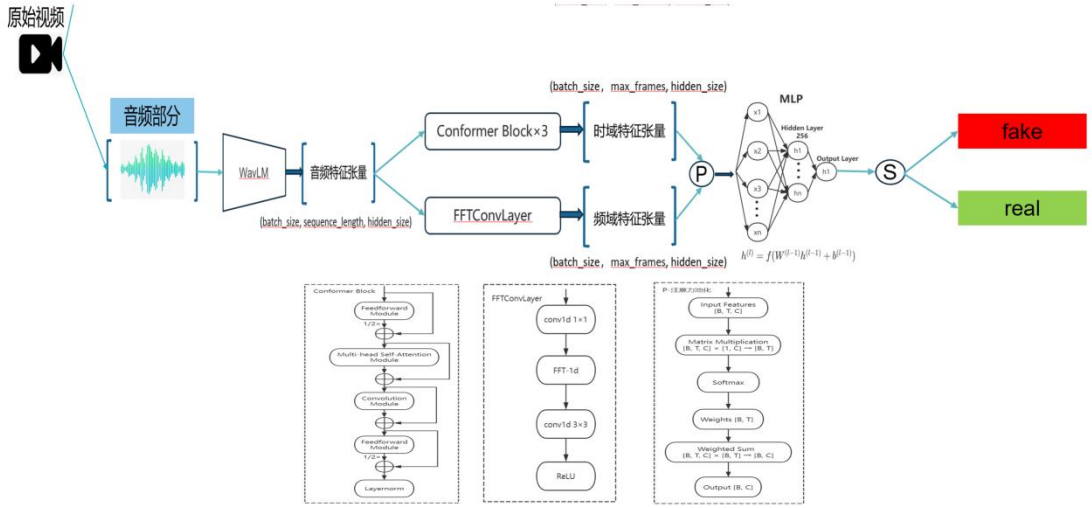
特征聚合与判别层：通过注意力池化动态聚合时空域与频域特征，抑制冗余信息，突出关键特征，使模型能自适应选择对当前视频最具判别性的特征维度，避免对特定生成模型的过拟合。最后通过 MLP 输出鉴别结果。



图表 13 音视频跨模态检测模型网络图

（四）音频检测模型：

搭建 “WavLM 特征提取器 + Conformer-FFTConv-P 分类器” 架构。WavLM 基于改进的 Transformer 架构，在大规模音频数据预训练中高效捕捉语音信号长时上下文依赖，提取高精度语音特征。结合的 Conformer-FFTConv-P 分类器，先通过 Conformer 模块分析音频时域波形特征，再利用 FFTConv 层对频域能量分布进行解析，捕捉生成音频特有的频率响应模式，最终通过特征融合与分类决策，实现对 AI 合成语音、声音伪造内容的精准检测。



图表 14 音频检测模型网络图

2.3 检测系统开发

第三阶段，基于构建的模型进行系统开发，通过 React 开发多模态检测界面，实现包括图像检测模块、文本检测模块模块、视频检测模块、音频检测模块、其他前端模块等内容。并采用 Flask 搭建微服务，实现数据预处理、模型调用、结果缓存及 API 接口管理。

第 3 章 项目研究内容与系统实现

3.1 数据集构建

数据收集：通过从 GitHub、Kaggle、谷歌等渠道获取开源 AIGC 数据集，覆盖不同生成模型生成的图像、文本、视频、音频数据，确保数据多样性。

数据处理：按模态分类整理数据，图像统一分辨率、文本清洗特殊字符、视频抽帧标准化、音频归一化采样率；采用人工标注方式，标记数据真实 / 生成属性，最终形成结构化 AIGC 检测数据集 CustomDataset。

3.2 模型开发与训练

设计统一的 AIGC 检测框架，采用预训练特征提取器+分类器结构，实现高精度、高泛化性的检测。

（一）图像检测模型实现：

采用 “CLIP:ConvNext 特征提取器 + Conformer-MLP 分类器” 架构。首先，利用 CLIP:ConvNext 在 4 亿图文对预训练中形成的全局语义理解与局部细节捕捉能力，对输入图像进行特征抽取。然后，将提取的特征输入 Conformer-MLP 分类器训练。

训练超参数：

```
params = {
    'hidden_dims': [2048, 1024, 512, 256],
    'dropout_rate': 0.3,
    'learning_rate': 0.0001,
    'l2_lambda': 0.0001,
    'batch_size': 128,
    'num_epochs': 50,
    'num_heads': 8,
    'num_conformer_blocks': 2
}
```

图表 15 图像检测模型训练超参数

（二）文本检测模型实现：

构建 “RoBERTa-MPU 特征提取器 + Conformer-MLP 分类器” 框架。使

用 RoBERTa-MPU 提取文本特征，输入到 Conformer-MLP 中进行训练。

训练超参数：

```
batch_size = 16
epochs = 5
learning_rate = 2e-4
```

图表 16 文本检测模型训练超参数

（三）音视频跨模态检测模型实现：

设计 “（CLIP:ViT + WavLM）特征提取器 + 跨模态时空频域融合模块（CrossModal-STF-Fusion）” 体系。CLIP:ViT 基于大规模互联网图文数据预训练，提取视频帧的视觉语义特征；WavLM 通过改进的 Transformer 架构提取音频信号的长时上下文依赖特征。输入到 CrossModal-STF-Fusion 模块中进行训练。

训练超参数：

```
"max_frames": 120,
"num_workers": 4,
"prefetch_factor": 2,
"batch_size": 4,
"num_epochs": 50,
"learning_rate": 1e-4,
"dropout_rate": 0.1,
```

图表 17 音视频跨模态检测模型训练超参数

（四）音频检测模型实现：

搭建 “WavLM 特征提取器 + Conformer-FFTConv-P 分类器” 架构。WavLM 基于改进的 Transformer 架构，在大规模音频数据预训练中高效捕捉语音信号长时上下文依赖，提取高精度语音特征。结合的 Conformer-FFTConv-P 分类器进行训练。

训练超参数：

```
'dropout_rate': 0.3,
'learning_rate': 0.0001,
'l2_lambda': 0.0001,
'batch_size': 64,
'num_epochs': 50
```

图表 18 音频检测模型训练超参数

3.3 系统开发与部署

（一）后端系统实现：

该系统使用 Flask 框架构建，负责处理前端请求，并与多个预训练模型进行交互来检测不同类型的 AI 生成内容。后端提供了四个主要接口：文本、图像、视频和音频检测。

（二）前端系统实现：

前端使用 React 框架构建，主要功能是为用户提供上传文件的界面，并将文件发送给后端进行处理。前端支持图像、文本、视频和音频文件的上传，用户可以通过拖拽或点击上传文件。

1.文件上传。前端提供了一个文件上传组件，用户可以选择图像、文本、视频或音频文件进行上传。上传后，文件会被读取并显示在界面上（图像和视频会显示预览，文本会显示文件名和类型）。对于文本文件，支持 .docx 和 .txt 格式；图像文件支持常见的图片格式（jpg、png 等）；视频文件支持 .mp4 和 .mov 格式；音频文件支持 .mp3、.wav、.flac、.m4a 和 .aac 格式。

2.文件检测。当用户上传文件后，前端会向后端发送请求，要求进行内容检测。前端根据用户选择的文件类型调用不同的后端接口（如 /predict_text、/predict_image、/predict_video 或 /predict_audio）。检测结果会显示在界面上，用户可以看到相应内容的 AI 生成概率。

3.用户交互。前端界面提供了动态的反馈功能，包括：

①上传进度和错误信息提示。

②检测结果显示，告诉用户内容是否可能由 AI 生成。

③用户可以根据需要清空已上传的内容或选择新的文件进行检测。

（三）天眼 AI 系统：

系统通过集成多种深度学习模型，能够高效地检测文本、图像、视频和音频内容是否由 AI 生成。后端使用了强大的模型来处理不同类型的数据。前端提供了直观的用户界面，支持多种文件格式的上传和实时检测。系统界面如图所示。

图表 19 天眼 AI 系统界面

第 4 章 实验测试分析

4.1 数据介绍

图像数据：采用第六届 DICIX（全球校园人工智能算法精英挑战赛）赛题和 Towards Universal Fake Image Detectors that Generalize Across Generative Models 论文中的数据集进行训练并测试。

文本数据：采用 HC3-EN 和 HC3-ZH 数据集进行训练并测试。HC3（Human ChatGPT Comparison Corpus）是第一个人类与 ChatGPT 对比的语料库。该数据集由 Hello-SimpleAI 团队提出，我们对数据集进行改进，采用短文本内容进行训练并测试。同时我们使用 deepseek 进行拓展，生成了跨域测试使用的拓展数据集。

视频数据：采用 Celeb-DF-v2、DH、DFDC 数据集进行测试并训练。Celeb-DF-v2 数据集是一个用于深度伪造取证的大规模挑战性数据集，该数据集由吴恩达教授领导的斯坦福大学团队于 2020 年发布，其核心目标是解决现有数据集在真实性和多样性方面的不足；DH（Homologous_deepfake_dataset）数据集是西安电子科技大学杜建超教授发布的首个全为中国人的 deepfake 数据集；DFDC（Deepfake Detection Challenge）数据集是一个用于深度伪造检测的大规模视频数据集，DFDC 数据集是通过多种深度伪造、基于 GAN 的面部交换方法以及非学习方法生成的。

音频数据：采用 ASV-2019 和 ASV-2021 数据集进行训练并测试。ASVspoof 2019 是一个大规模的公开数据库，包含合成、转换和重放的语音数据。ASVspoof 2021 是一项旨在促进研究欺骗和设计对策以保护自动说话人验证系统免受操纵的挑战。ASVspoof 2021 引入深度伪造语音数据集。本实验主要采用其中的语音合成和深度伪造语音数据集。

4.2 实验测试

实验分为域内测试与跨域测试。域内测试采用 1：5 的比例划分训练集与测试集；跨域测试采用一种生成模型生成的数据进行训练，采用另一种生成模型的数据进行测试。

图像测试：从实验数据中可以看出，在域内测试中，本项目模型较开源先进模型平均 ACC 提升 2.98%；在跨域测试中，相比开源先进模型平均 ACC 提升 1.32%。

Detection method	Variant	第六届DIGIX赛题			Total
		A	B	C	Avg-acc
CLIP:VIT+	NN, k = 1	96.2	89.98	79.63	88.60
	NN, k = 3	96.24	88.74	76.67	87.22
	NN, k = 5	96.06	87.11	72.96	85.38
	NN, k = 9	95.78	85.33	65.93	82.35
	LC	97.72	96.46	92.41	95.53
Ours	VIT+MLP	98.13	96.13	94.44	96.23
	ConvNeXt+CMLP	**99.87**	**98.26**	**97.41**	**98.51**

图表 20 图像检测域内测试数据

Detection method	Variant	Generative Adversarial Networks						Deepfakes	Guided	LDM			Glide			DALL-E	Total
		Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN			200steps	200w/CFG	100steps	100-27	50-27	100-10		mAP
Trained deep network	Blur+JPEG (0.1)	**100**	93.47	84.5	**99.54**	89.49	98.15	89.02	73.72	70.62	71	70.54	80.65	84.91	82.07	70.59	83.88
	Blur+JPEG (0.5)	**100**	96.83	88.24	98.29	98.09	95.44	66.27	68.57	66	66.68	65.39	73.29	78.02	76.23	65.93	80.22
	ViT:CLIP(B+J0.5)	99.98	93.32	83.63	88.14	92.81	84.62	67.23	55.74	52.52	54.51	52.2	56.64	61.13	56.64	62.74	70.79
Patch classifer	ResNet50-Layer1	98.86	72.04	68.79	92.96	55.9	92.06	60.18	70.05	87.84	84.94	88.1	74.54	76.28	75.84	77.07	78.36
	Xception-Block2	80.88	72.84	71.66	85.75	65.99	69.25	76.55	75.03	87.1	86.72	86.4	85.37	83.73	78.38	75.67	78.75
Co-occurence		99.74	80.95	50.61	98.63	53.11	67.99	59.14	70.2	91.21	89.02	92.39	89.32	88.35	82.79	80.96	79.63
Freq-spec	CyeleGAN	55.39	**100**	75.08	55.11	66.08	**100**	45.18	57.72	77.72	77.25	76.47	68.58	64.58	61.92	67.77	69.92
CLIP:VIT+	NN, k = 1	**100**	98.14	94.49	86.68	99.26	99.53	93.09	79.31	95.84	79.84	95.97	93.98	95.17	96.05	88.51	93.06
	NN, k = 3	**100**	98.13	94.46	86.67	99.25	99.53	93.03	79.26	95.81	79.78	95.94	93.94	95.13	94.6	88.47	92.93
	NN, k = 5	**100**	98.13	94.46	86.66	99.25	99.53	93.02	79.25	95.81	79.78	95.94	93.94	95.13	94.6	88.46	92.93
	NN, k = 9	**100**	98.13	94.46	86.66	99.25	99.53	91.67	79.24	95.81	79.77	95.93	93.93	95.12	94.59	88.45	92.84
	LC	**100**	99.46	**99.59**	97.24	99.98	99.6	82.45	87.77	99.14	92.15	99.17	94.74	95.34	94.57	**97.15**	95.89
Ours	VIT+MLP	**100**	99.89	99.47	99.21	99.97	99.54	**93.26**	91.69	98.84	92.76	98.98	94.43	95.98	94.69	97.1	97.05
	ConvNeXt+CMLP	**100**	99.53	99.01	99.03	**99.99**	99.31	79.86	**95.77**	**99.74**	**96.57**	**99.61**	**97.61**	**98.31**	**97.64**	96.1	**97.21**

图表 21 图像检测跨域测试数据

文本测试：从实验数据中可以看出，在域内测试中，本项目模型较开源先进模型平均 ACC 提升 3.02%；在跨域测试中，相比开源先进模型平均 ACC 提升 1.15%。

Detection method	Variant	HC3			Total
		HC3-EN	HC3-ZH	HC3-ALL	Avg-acc
华为的检测模型	RoBERTa-MPU	95.73	96.19	96.12	96.01
ours	**RoBERTa-MPU+CMLP**	**98.97**	**99.08**	**99.05**	**99.03**

图表 22 文本检测域内测试数据

Detection method	Variant	HC3		Total
		HC3-ZH	HC3-拓展	Avg-acc
华为的检测模型	RoBERTa-MPU	训练集	87.32	87.32
ours	**RoBERTa-MPU+CMLP**		**88.47**	**88.47**

图表 23 文本检测跨域测试数据

视频测试：从实验数据中可以看出，在域内测试中，本项目模型较开源先进模型平均 ACC 提升 1.60%；在跨域测试中，相比开源先进模型平均 ACC 提升 1.46%。

Detection method	Variant	CelebDFV2	HD	DFDC	Total
					Avg-acc
F3-Net		95.97	97.32	93.25	95.51
ours	**CLIP：VIT+C-F-MLP**	**97.08**	**98.5**	**95.76**	**97.11**

图表 24 视频检测域内测试数据

Detection method	Variant	HD	CelebDFV2	DFDC	Total
		训练集			Avg-acc
F3-Net			74.32	76.84	75.58
ours	**CLIP：VIT+C-F-MLP**		**76.14**	**77.93**	**77.04**

图表 25 视频检测跨域测试数据

音频测试：从实验数据中可以看出，在域内测试中，本项目模型较开源先进模型平均 EER（等错误率）降低了 1.59%；在跨域测试中，相比开源先进模型平均 EER 降低 1.13%。

Detection method	Variant	ASV-19LA	ASV-21LA	ASV-21DF	Total
					Avg-EER
SafeEar		3.1	7.22	6.43	5.58
ours	**WavLM+C-F-MLP**	**2.82**	**4.94**	**4.22**	**3.99**

图表 26 音频检测域内测试数据

Detection method	Variant	ASV-21DF	ASV-19LA	ASV-21LA	Total
		训练集			Avg-EER
SafeEar			13.42	12.17	12.80
ours	WavLM+C-F-MLP		**11.26**	**12.08**	**11.67**

图表 27 音频检测跨域测试数据

4.3 结果分析

本实验通过多维度测试表明，本项目模型在检测精度、跨域泛化能力、模态覆盖及短文本处理方面均显著优于现有方案。在域内测试中，图像、文本、视频、音频模态检测准确率分别较开源先进模型提升 2.98%、3.02%、1.60%，音频等错误率降低 1.59%；跨域测试中，各模态检测性能均保持领先，图像、文本、视频跨域准确率分别提升 1.32%、1.15%、1.46%，音频跨域等错误率降低 1.13%。

系统通过 “冻结预训练特征提取器 + 定制化分类器” 架构，突破传统检测的 “模态单一” 局限，实现文本、图像、视频、音频四模态的高效检测；借助 Conformer 系列模块的多尺度特征融合，解决了 “模型依赖” 问题；依托 RoBERTa-MPU 与 Conformer-MLP 的协同，攻克 “短文本盲区”。实验结果验证了系统核心技术创新的有效性，为全部模态 AIGC 的通用检测提供了可靠的技术方案。

第 5 章 项目创新点

5.1 全模态检测能力

“天眼 AI” 突破传统检测系统的模态割裂局限，构建了覆盖文本、图像、视频、音频的全维度检测体系，实现全模态数据的 “分模态专属处理 + 统一框架集成”。系统针对不同模态的数据特性与检测难点，设计了专属的技术链路：文本检测依托 RoBERTa-MPU 与 Conformer-MLP，解决短文本语义稀疏问题；图像检测通过 CLIP:ConvNext 与 Conformer-MLP，捕捉生成图像的视觉语义与局部伪影；音视频跨模态检测利用 CLIP:VIT+WavLM 与 CrossModal-STF-Fusion，提取音视频内容的多维特征表征；音频检测借助 WavLM 与 Conformer-FFTConv-P，提取语音信号的长时上下文依赖与频域特征。

通过统一界面支持用户对单模态数据的独立检测需求，从根本上解决了传统系统 “检测模态单一” 的问题，实现四大模态 “全覆盖、高精度、一站式” 检测的通用系统，为政务司法证据核验、内容平台多格式审核、教育学术多形式成果检测等场景提供了 “即插即用、精准适配” 的技术支撑。

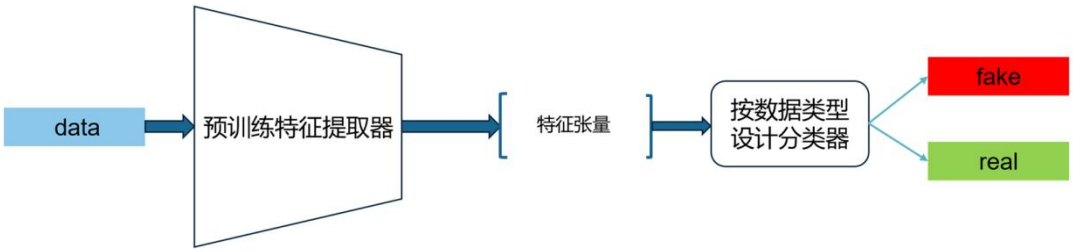
图表 28 支持检测模态数量对比

5.2 AIGC 统一检测框架

“天眼 AI” 构建了 “预训练特征提取器 + 定制化分类器” 的统一架构，突破了传统检测模型依赖特定生成模型伪影的技术瓶颈。框架采用冻结的预训练模型提取数据的通用特征，避免模型在训练过程中过拟合 GAN、扩散模型等特

定生成模型的低级伪影。

针对不同模态设计专属分类器：Conformer-MLP 用于图文检测，通过深度可分离卷积与 Transformer 自注意力实现多尺度特征融合；CrossModal-STF-Fusion 用于音视频检测，整合跨模态-时空域-频域特征进行联合建模。这种设计使模型能够学习真实与生成内容的本质差异（如语义一致性、结构合理性），而非记忆单一模型的表面特征。显著增强了对未知生成模型的泛化能力，实现 “生成模型无关” 检测的通用框架。



图表 29 统一框架网络图

5.3 短文本 AI 生成检测能力

针对短文本语义稀疏、特征模糊的检测难题，“天眼 AI” 通过双重技术联合实现了对短文本场景的精准覆盖。在特征提取层，采用 RoBERTa-MPU 融合 RoBERTa 预训练模型与多尺度正负无标签（MPU）框架。

在分类决策层，Conformer-MLP 分类器通过动态注意力机制聚焦短文本中的关键 token，如疑问词、情感词或异常重复片段，通过自注意力权重分配强化对 “逻辑断层”“语义跳跃” 等生成文本典型特征的识别。为社交媒体评论审核、即时通讯安全预警、AI 生成论文检测等场景提供了高效解决方案。

图表 30 天眼 AI 与朱雀大模型短文本检测对比

5.4 跨模态时序融合音视频检测模块

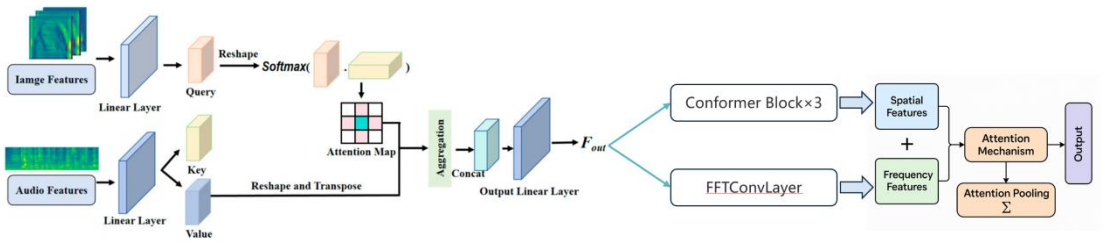
针对视频与音频数据的模态特征与时序动态特性，“天眼 AI” 创新设计了 CrossModal-STF-Fusion 模块，实现“时空-频域——单帧-时序——音频-画面”特征的高效融合与联合建模。

模块核心包含三重处理机制：首先，通过跨模态注意力机制，以视频特征为查询、音频特征为键值对，动态计算模态间关联权重，实现视觉与听觉特征的细粒度对齐与互补融合。显著增强多模态特征的表征能力，为检测提供更全面的信息基础。

其次，设计 Conformer-FFTConv-P 模块，通过 Conformer Block 提取视频帧序列的时序依赖关系，同时利用 FFTConvLayer 将特征转换至频域，挖掘音视频信号的周期性伪影。时空 - 频域特征的双重建模，可有效捕捉 AIGC 生成内容在时序逻辑和频域分布上的漏洞，提升对动态伪造痕迹的敏感度。

最后，融合跨模态交互特征、时序动态特征与频域统计特征，形成多层次判别信号。经注意力池化操作后，模型可自适应聚焦关键特征，抑制冗余信息。使

模型能自适应选择对当前视频最具判别性的特征维度，避免对特定生成模型的过拟合。成为应对动态媒体内容伪造的核心技术组件，有效实现多维度特征融合。



图表 31 CrossModal-STF-Fusion 模块网络图

第 6 章 项目应用前景与社会价值

6.1 政务司法领域深度对接

与政府公共安全部门、司法机构合作，提供 AI 生成内容检测服务：

电子政务核验：为政务平台的文本公告、政策解读视频、政务音频文件提供真实性检测，防范 AI 伪造的虚假政策信息传播；

司法证据鉴定：针对涉 AI 诈骗、深度伪造的案件，提供图像、视频、音频的篡改检测，辅助司法机关快速鉴别证据真伪，降低跨模态证据核验成本。

6.2 企业级内容安全解决方案

面向互联网平台、媒体机构、教育企业推出定制化产品：

社交媒体审核：为微博、抖音、小红书等平台提供 API 接口，实时检测用户生成内容中的 AI 文本、图像、视频及音频，过滤虚假信息、诈骗内容，提升平台内容安全治理效率；

学术诚信审查：对接知网、万方等学术数据库，嵌入论文检测环节，识别 AI 生成的学术文本及伪造实验图像 / 视频，助力高校与科研机构构建全模态学术成果打假体系；

金融风控辅助：为银行、保险等金融机构提供 AI 合成声音检测服务，防范电话诈骗中利用 AI 模仿他人声纹的风险。

第 7 章 项目存在的问题及改进方向

7.1 项目存在的问题

轻量化不足：项目当前模型较大，需要依赖大型预训练模型提取特征，在移动端算力有限的设备上难以实时检测。

生成内容无法溯源：当前项目仅能检测内容是否为 AI 生成，无法准确溯源生成的 AI 模型。

7.2 改进方向

轻量化与移动端适配：针对移动端设备算力限制，通过模型蒸馏、动态量化与结构剪枝技术，实现手机、智能摄像头等边缘设备的低功耗部署。构建 “端云协同” 架构，边缘端通过轻量模型快速过滤常规内容，云端对疑似样本进行全精度检测，平衡效率与准确性。开发跨平台 SDK，支持 iOS/Android 及微信小程序等轻量化场景，提供 “即传即检” 的便捷服务，满足短视频平台实时审核、社交 APP 即时检测的低延迟需求。

生成内容溯源技术：构建多模态生成特征指纹库，提取不同模型在文本句法、图像频域、视频时序等维度的独特伪影，结合区块链技术对检测结果进行存证，形成不可篡改的 “数字指纹” 链。开发可视化溯源模块，向用户展示内容的生成模型类型（如 “Stable Diffusion 生成图像” “GPT-4 撰写文本”）及特征异常点，为政务文件核验、司法证据鉴定提供技术溯源支撑。针对专业领域，开放 API 级溯源接口，实现检测结果与电子政务、司法取证系统的无缝对接，从 “内容检测” 延伸至 “源头追踪”，构建 “检测 - 溯源 - 取证” 闭环。

参考文献

[1] Towards Universal Fake Image Detectors that Generalize Across Generative Models. In CVPR,2023.

[2] MULTISCALE POSITIVE-UNLABELED DETECTION OF AI-GENERATED TEXTS.In ICLR,2024.

[3] Conformer: Local Features Coupling Global Representations for Visual Recognition.In ICCV,2021.