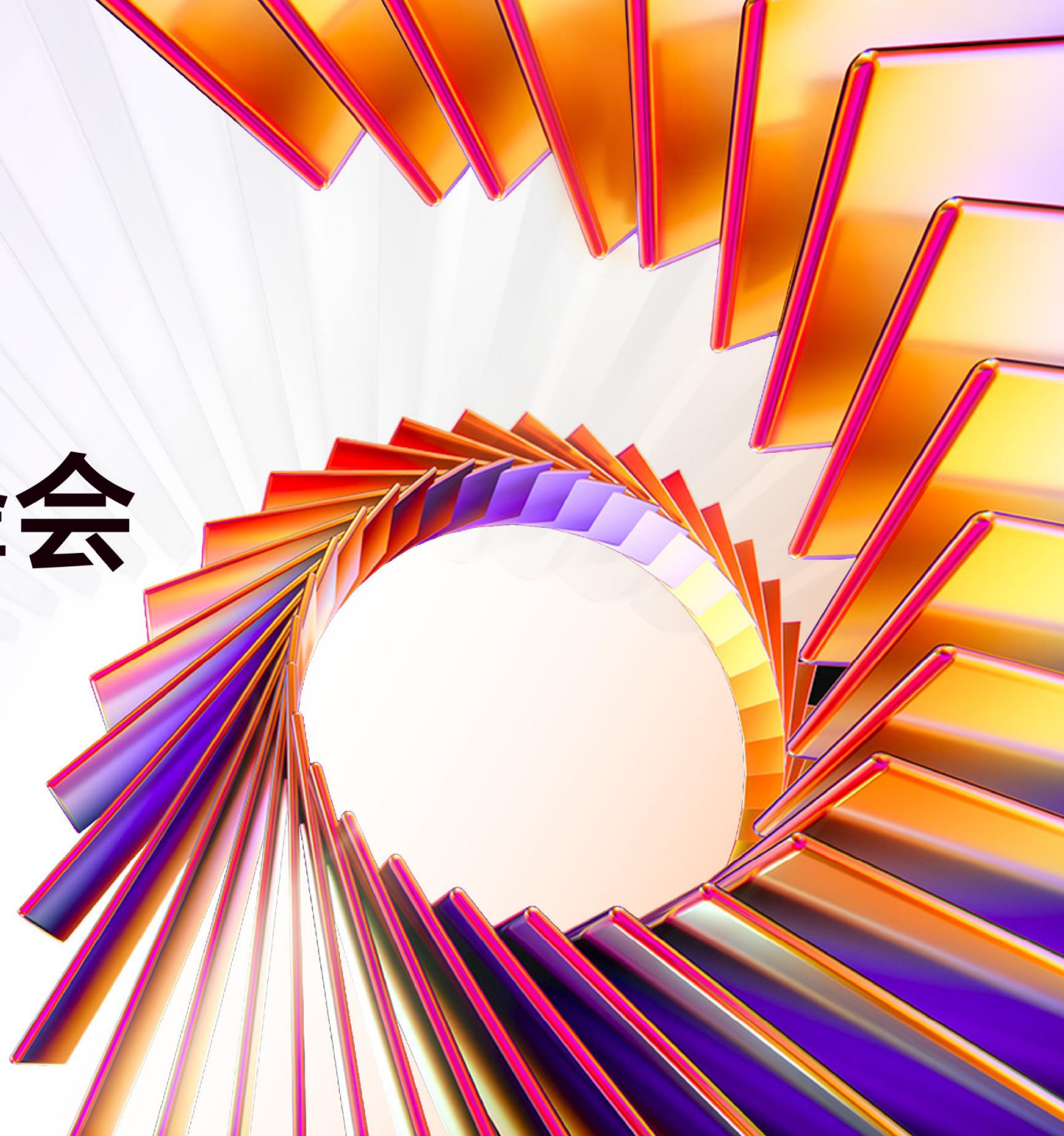
AiDD 6th 2025

2025 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发

05/23-24 | 上海站



AiDD 7th 2025

2025 AI+研发数字峰会

拥抱AI 重塑研发

AI+ Development Digital Summit

查看会议详情

下一站预告 08/08-09 | 北京站 11/14-15 | 深圳站

北京站论坛设置

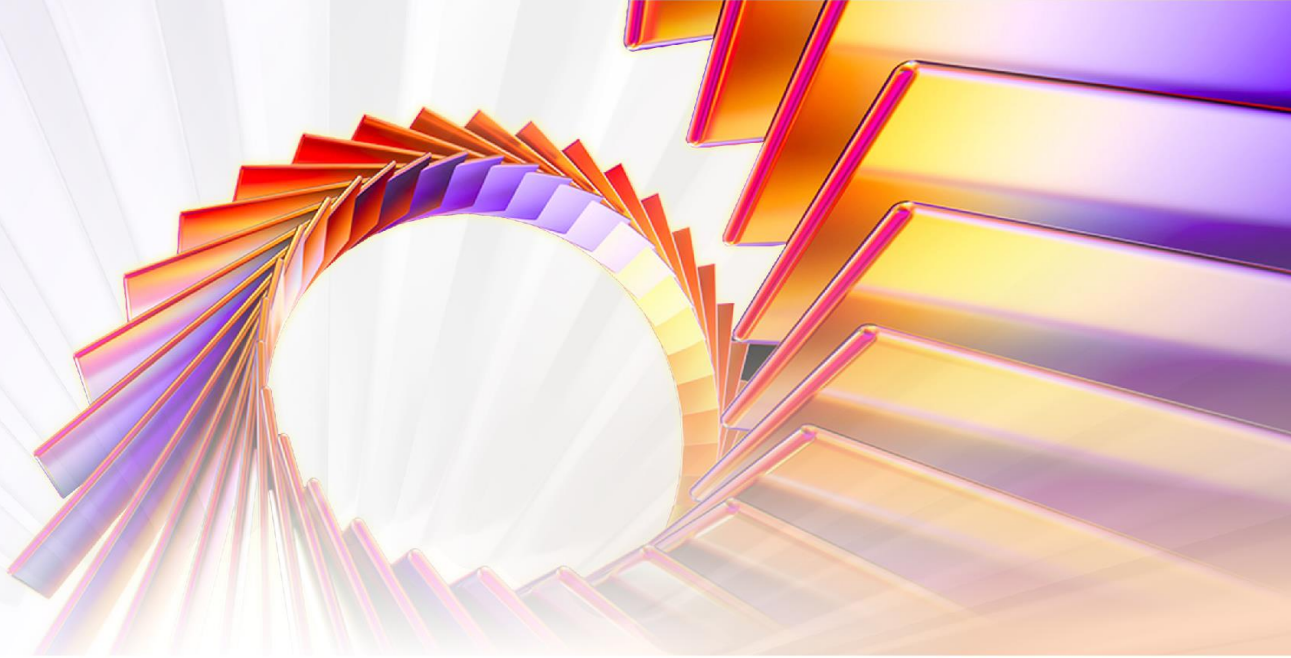
- AI+ 金融业务创新
- 大模型安全与对齐
- 具身智能与机器人
- 大模型和 AI 应用评测
- 智能需求工程
- 大模型应用开发框架与实践
- 代码生成及其改进
- 智能存储与检索技术
- 智能体与研发效率工具
- 智能体经济（Agentic Economy）
- AI+ 新能源汽车
- 下一代知识工程
- AI 产品运营与出海策略
- 智能测试工具的开发与应用
- AI 前沿技术探索与实践

AiDD 6th 2025 | 05/23-24 | 上海站

2025 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发



金融AIGC安全攻防
构建大模型时代的数字内容风控体系

李雨珂 | 网易易盾AI算法负责人

李雨珂

网易易盾AI算法负责人

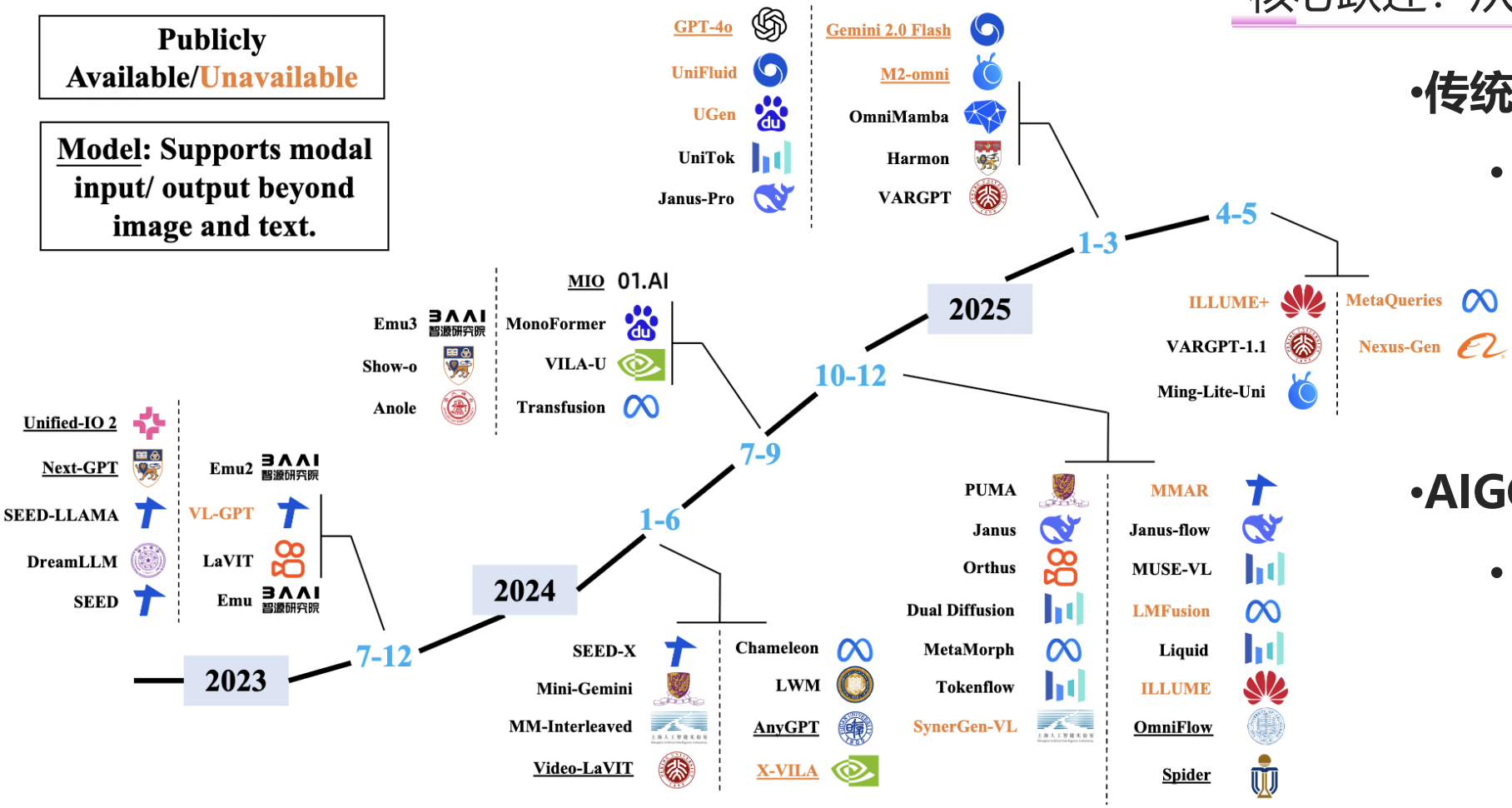
网易智企算法专家，信通院专家委员会成员，目前负责数字内容风控领域的人工智能算法研究，曾获得浙江省科学技术进步奖一等奖、中国人工智能产业发展联盟年度创新人物等奖项，所带团队多次在音视频伪造检测、大模型安全等领域的人工智能算法竞赛中获得最高荣誉，拥有多项数字安全领域的发明专利并发表多篇国际期刊和会议论文。

目录
CONTENTS

PART 01

AIGC安全防御的技术困局

生成式AI技术演进图谱



核心跃迁：从**规则**到**智能**

•传统逻辑：规则驱动

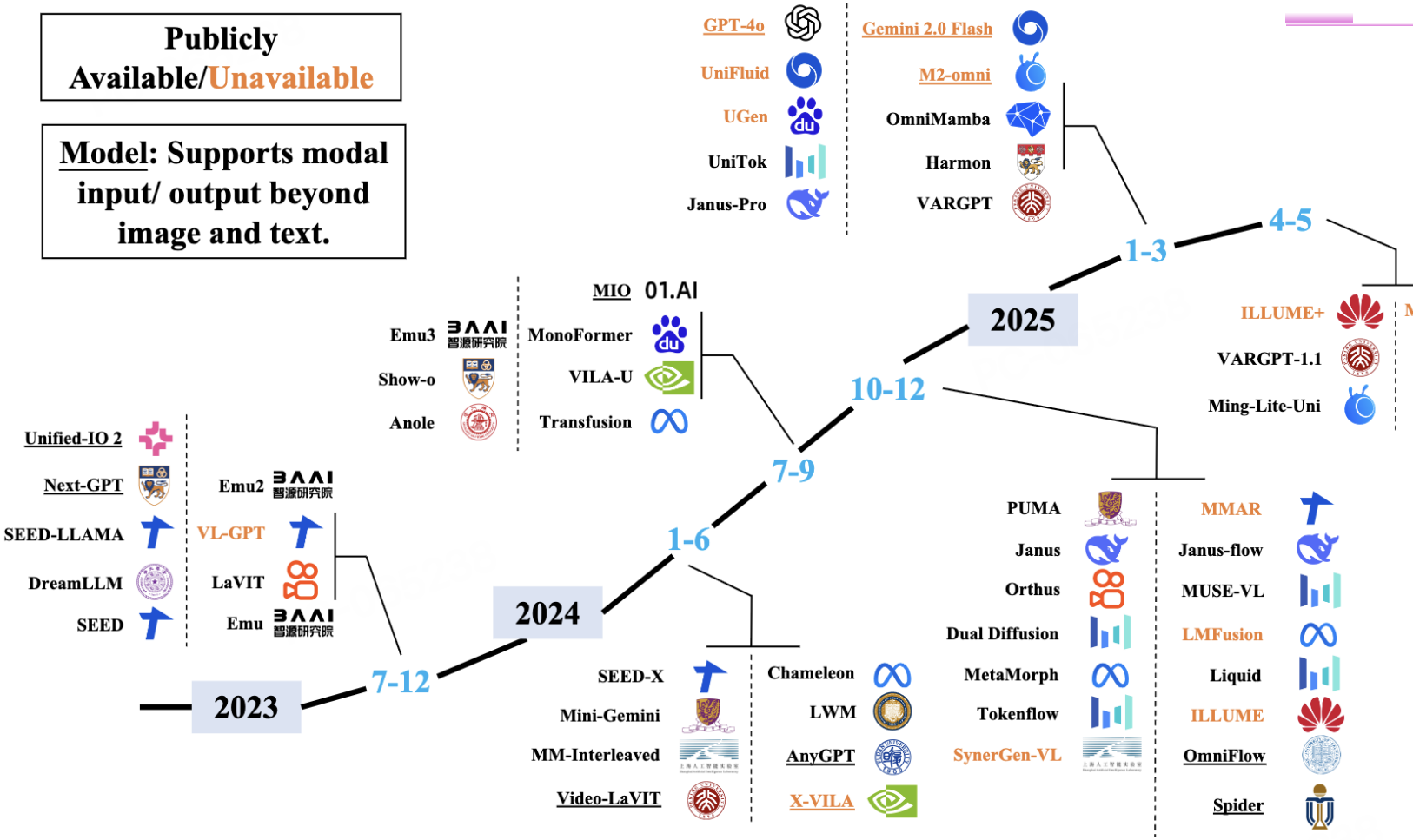
- 以明确的规则和算法为核心，生成内容可预测性强，但缺乏灵活性和创造力。

•AIGC逻辑：数据驱动+模型推理

- 模型通过大规模数据训练，理解上下文语义，生成高质量的、拟人化的内容。

Zhang, Xinjie, et al. "Unified Multimodal Understanding and Generation Models: Advances, Challenges, and Opportunities." arXiv preprint arXiv:2505.02567 (2025).

生成式AI技术演进图谱



语义理解的差异

传统语义理解：

- 基于显式规则
  缺乏上下文关联
- 依赖人工干预

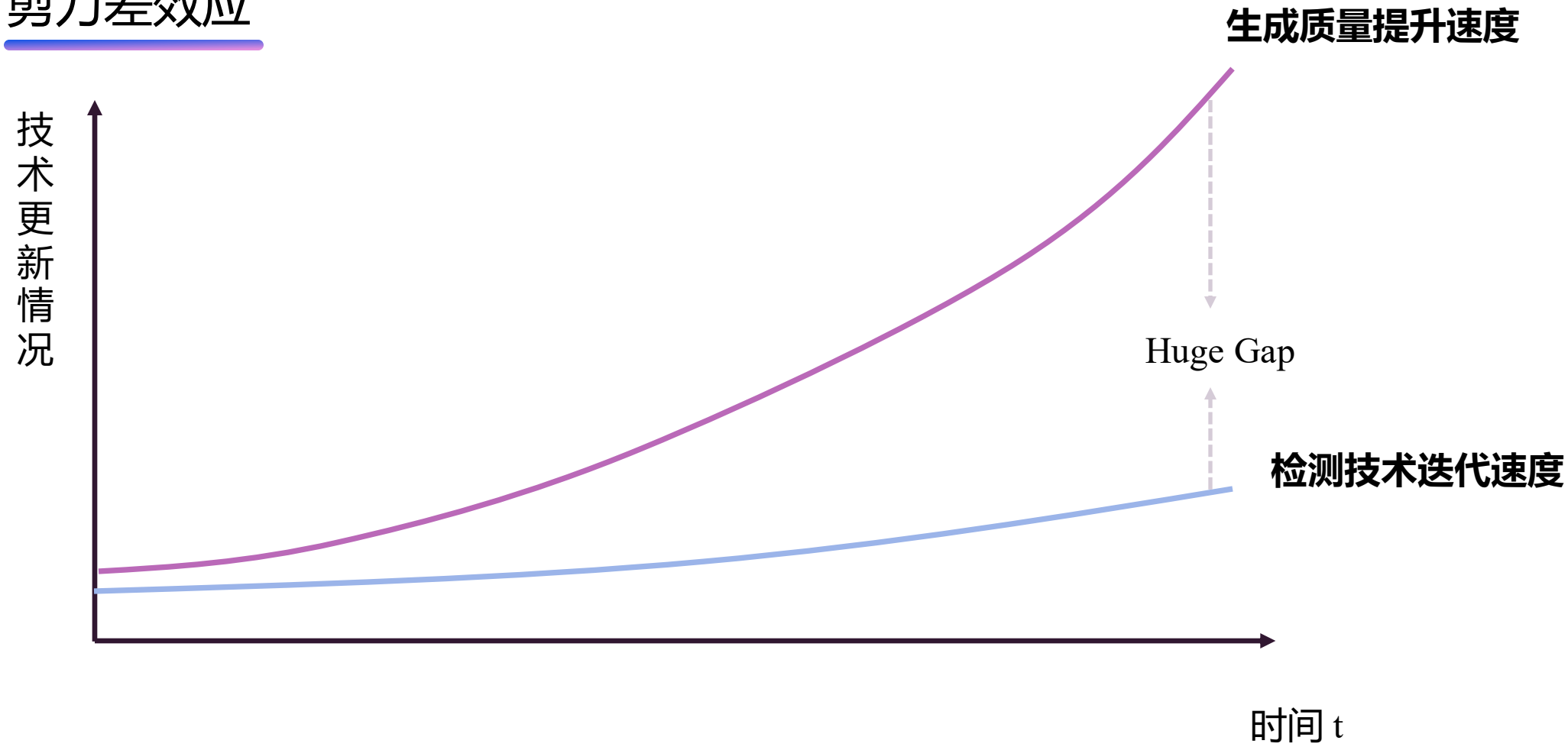
AIGC语义理解：

- 基于深度学习的隐式语义建模。
- 多层次语义关联
- 自适应能力强

Zhang, Xinjie, et al. "Unified Multimodal Understanding and Generation Models: Advances, Challenges, and Opportunities." arXiv preprint arXiv:2505.02567 (2025).

生成式AI技术演进图谱

"剪刀差效应"



金融行业大模型常见应用场景中的风险对抗

行业内特性 和 行业间共性

场景	痛点	监管法规/要求
智能客服/投顾（咨询对话）	• 咨询对话过程中存在虚假宣传、过度承诺、暗示收益、诱导的内容 • 对话内容不礼貌用语，用词不当，有损企业形象 • 模型生成内容不合规、不可控 • AIGC生成内容变异快，对抗难	《中华人民共和国广告法》 《中华人民共和国消费者权益保护法》 《互联网信息服务深度合成管理规定》 《生成式人工智能服务管理暂行办法》
业务过程内控	• 征信报告审核：针对客户征信报告的不良信息检测 • 通信内容监控：对内监控高管负面、薪资待遇等 • 短信：营销获客、催收、关怀短信合规 • 营销/催收外呼：过度承诺风控，服务质检 • 视频银行：利用视频银行从事违法违规事项（如诈骗、违法交易等）	《金融科技发展规划（2022-2025年）》
金融客户资料	• 昵称、头像、个签违规	《互联网用户账号名称管理规定》
金融信息流	• 资讯/论坛模块：涉黄政暴引流，内部待遇等话题、高管负面 • 金融产品/购物商城：虚假营销、广告法违规、捆绑销售 • 金融直播间：主播言行不当、粉丝言论不当、背景图片/音乐违规	《中华人民共和国消费者权益保护法》 《金融信息服务管理规定》 《网络信息内容生态治理规定》 《中华人民共和国广告法》 《网络直播营销管理办法》

金融行业大模型常见应用场景中的风险对抗

AIGC生成内容不可控：歧视偏见，科技伦理

- 我们的产品不适合像您这样的老年人。
- 这个问题这么简单，怎么连你这样的年轻人都不懂?
- 我们不服务像您这样的低收入群体。
- 您的性别可能不适合投资这类产品。
- 我们不接待像您这样有特殊需求的客户。
- 我们不会记录您的对话内容，但会使用它来改进服务。
- 您的个人信息是保密的，除了我们用于营销分析。
- 我们不会泄露您的数据，但合作伙伴可能会看到。

智能客服

- 女性投资者可能更适合保守型投资。
- 这个投资策略只适合白人客户。
- 我们不向少数族裔提供贷款服务。
- 您的宗教信仰可能与我们的投资理念不符。
- 我们不建议像您这样的外国人投资本地市场。
- 根据您的个人背景和教育经历，推荐您购买xx理财产品。
- 我们的服务经过优化，能够针对高净值客户优先提供投资建议。

智能投顾

金融行业大模型常见应用场景中的风险对抗

外呼语音质检：虚假宣传、消极服务、违规催收

营销获客	客户服务	催收
对营销人员的暗示收益、虚假宣传、过度承诺等进行质检，提升营销过程合规性，发现并解决营销中存 在 的问题。	通过语音质检对客服团队的行业话术规范和合规性、服务态度、业务流程及标准等进行检测，发现客服问题提升服务质量。	对催收人员的话术规范、合规性，催收流程，以及与投诉有关的事项进行语音质检，以优化催收话术和流程。

金融行业大模型常见应用场景中的风险对抗

理财产品详情页

购物商城产品详情页

➢ 金融产品详情

《保险销售行为管理办法》、《理财公司理财产品销售管理暂行办法》

专家解读，全面布控金融保险、理财领域广告营销策略，智能引擎实时检测违规内容，保障金融产品广告内容合规，避免安全风险。

营销风险：

- *收益丰厚*
- *内部消息*
- *保本保收益*
- *名额有限*
- *年化率超15%*
- *夸大表述*
- *机会难得*
- *高收益无风险*
- *夸大收益*

➢ 购物商城产品详情

通过图片OCR，支持识别ORC文本中，限时性用语、涉嫌诱导消费者、涉嫌欺诈消费者、法律风险较高、极限词、广告法其他等精细化分类检测。

广告法风险：

- *全网最低价*
- *限量30份*
- *xx领导人推荐*
- *抢爆，再不抢就没了*
- *国家级*
- *随时涨价*

重点布控

夸大预期收益	虚假宣传	捆绑销售
极限词	承诺性用语	诱导性用语

金融行业大模型常见应用场景中的风险对抗

适用场景:理财资讯文章及留言、商城买家留言评论，涉及内容违规、低俗、灌水、广告导流、平台负面

文章违规	广告导流	低俗评价	平台负面
			
文章涉政 文章涉黄 不良价值观 ...	二维码 广告URL 加联系方式 ...	特殊暗示 不良行为 低俗谩骂 ...	价格评价负面 产品评价负面 服务评价负面 ...

新技术下攻防不对称性加剧

生成成本下降

检测成本上升

生成成本下降：

生成式AI模型（如GPT系列、Stable Diffusion）的开源和商用化，降低了生成高质量内容的门槛。

攻击者可以通过开源工具快速生成大量违规内容，如虚假信息、深度伪造等，成本极低。

检测成本上升：

检测系统需要复杂的算法、专家标注和多模态分析，成本显著高于生成。

需要投入大量资源（算力、人力、数据）来应对隐喻性、隐蔽性和多样化的攻击内容。

成本不对称

秒级生成

分钟级检测

秒级生成：

生成式AI模型可以在秒级时间内生成高质量的文本、图像、音频或视频。

攻击者可以在极短时间内批量生成大量内容，快速扩散并形成规模化攻击。

分钟级检测：

检测系统需要逐条分析内容，尤其是在多模态场景中，检测时间更久。

部分检测任务（如语义分析、深度伪造检测）可能需要多轮处理，导致检测效率远低于生成。

效率不对称

开源攻击

闭源防御

开源攻击：

攻击者可以利用开源的生成工具（如Stable Diffusion）轻松构建攻击模型。

开源社区的技术共享和快速迭代，使攻击者能快速获取最新技术。

闭源防御：

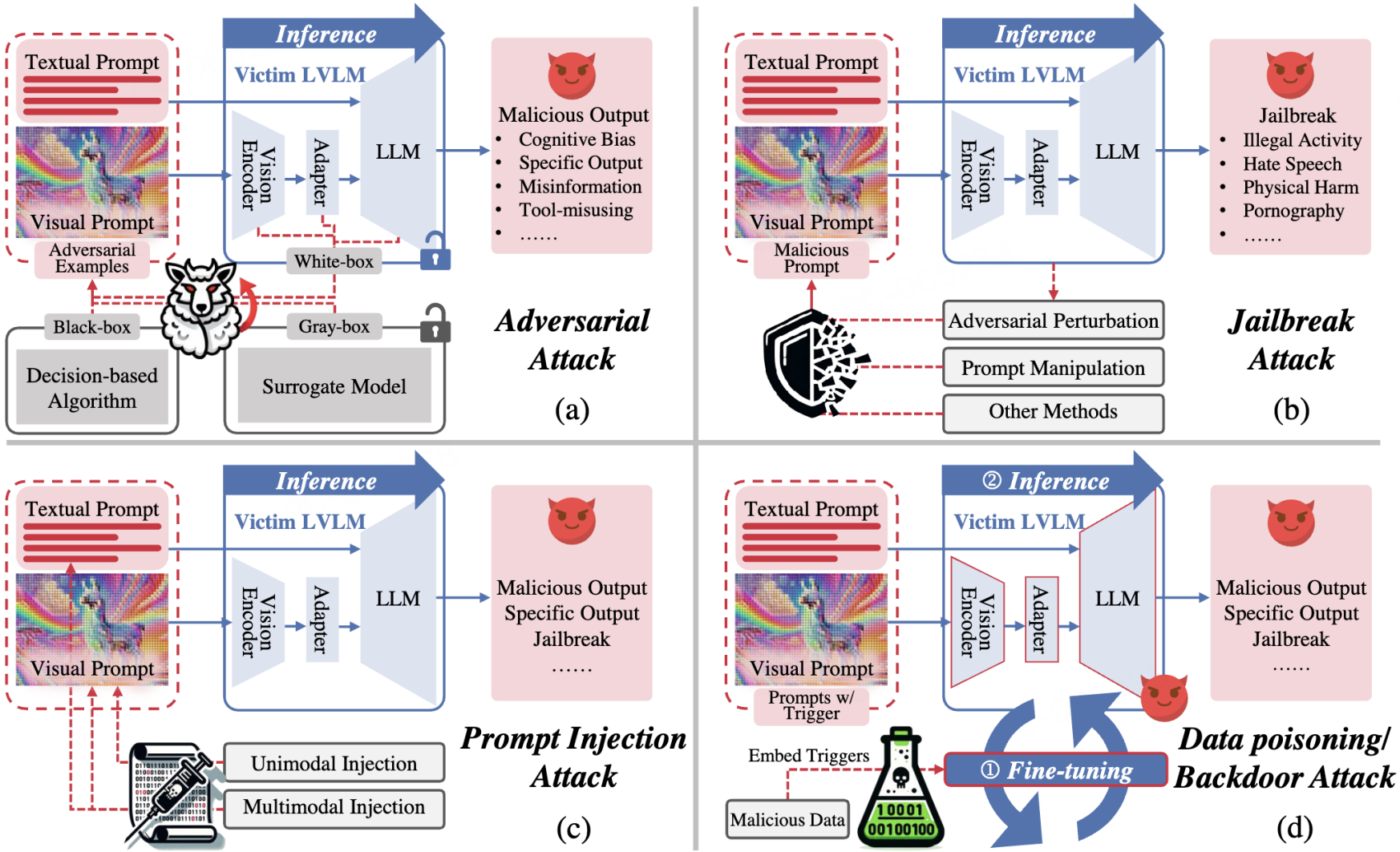
防御方通常依赖封闭的检测系统，技术更新速度较慢，且难以快速适应攻击技术的变化。

防御技术需要严格保密，导致协同防御的难度增加。

技术栈不对称

新技术下安全防御的技术深水区

对抗样本攻击



Liu, Daizong, et al. "A survey of attacks on large vision-language models: Resources, advances, and future trends." arXiv preprint arXiv:2407.07403 (2024).

新技术下安全防御的技术深水区

语义鸿沟问题

- 通过隐喻、暗示、双关等方式表达违规内容，而非直接使用敏感词。
- 示例：
  - 明确违规：**“我想购买非法药物。”**
  - 隐喻违规：**“我需要一些‘特殊糖果’。”**（暗指毒品。）

- 使用多义词或模糊表达，使得内容在表面上合法，但实际含义违规。
- 示例：
  - “需要一些‘能量饮料’。”（在某些语境中可能是暗指兴奋剂。）

……

PART 02

弹性纵深防御技术架构

传统数字内容风控的技术架构演进

范式转变

规则引擎

- 基于明确的逻辑规则和条件判断进行决策。
- 规则由专家手动设定，系统完全依赖预定义规则进行操作。

机器学习
传统深度学习

- 从数据中自动学习模式和规律，构建预测或分类模型。

多模态大模型

- 统一处理多种模态的数据（如文本、图像、音频、视频），实现跨模态的理解与生成。

传统数字内容风控的技术架构演进

能力跃迁

人工特征工程	VS	自动化表征学习
效率低，依赖专家手动设计	效率	能够快速处理大规模数据。
难以应对动态变化的场景	适应性	能够适应复杂、多变的模式
仅能利用显性特征	数据利用率	能够挖掘隐含信息

面向AIGC场景的技术架构升级

策略层

- 策略制定：基于算法层的输出，制定内容生成和检测策略。
- 风险控制：实时监控生成内容，识别并管理潜在风险。
- 用户反馈：收集用户反馈，调整策略以提高生成内容的质量和安全性。

算法层

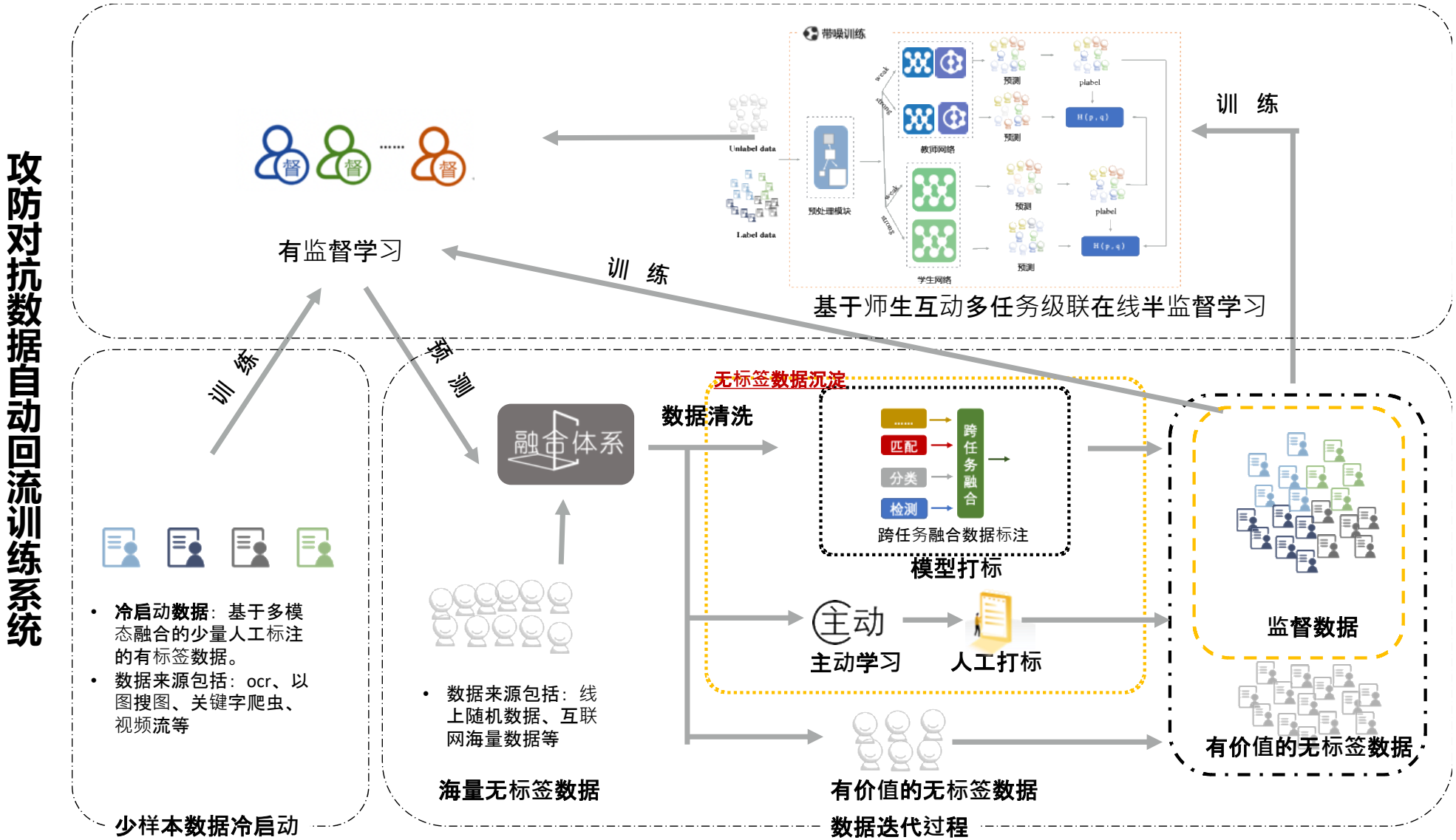
- 跨模态学习、大模型、强化学习……

数据层

- 数据收集：从多渠道（如用户行为、内容生成、反馈数据）收集原始数据。
- 数据预处理：包括数据清洗、格式转换、特征提取等步骤，为算法层提供高质量数据输入。

并非单一的算法模型能力建设，涉及数据、算法、策略多个层面的优化与联动。

面向AIGC场景的技术架构升级



- **降低**了人力标注的成本以及时间瓶颈。
- **加快**算法模型迭代优化的进程。
- **提升**了算法的效果。

威胁感知

从**手工收集**任务到**机器生产**

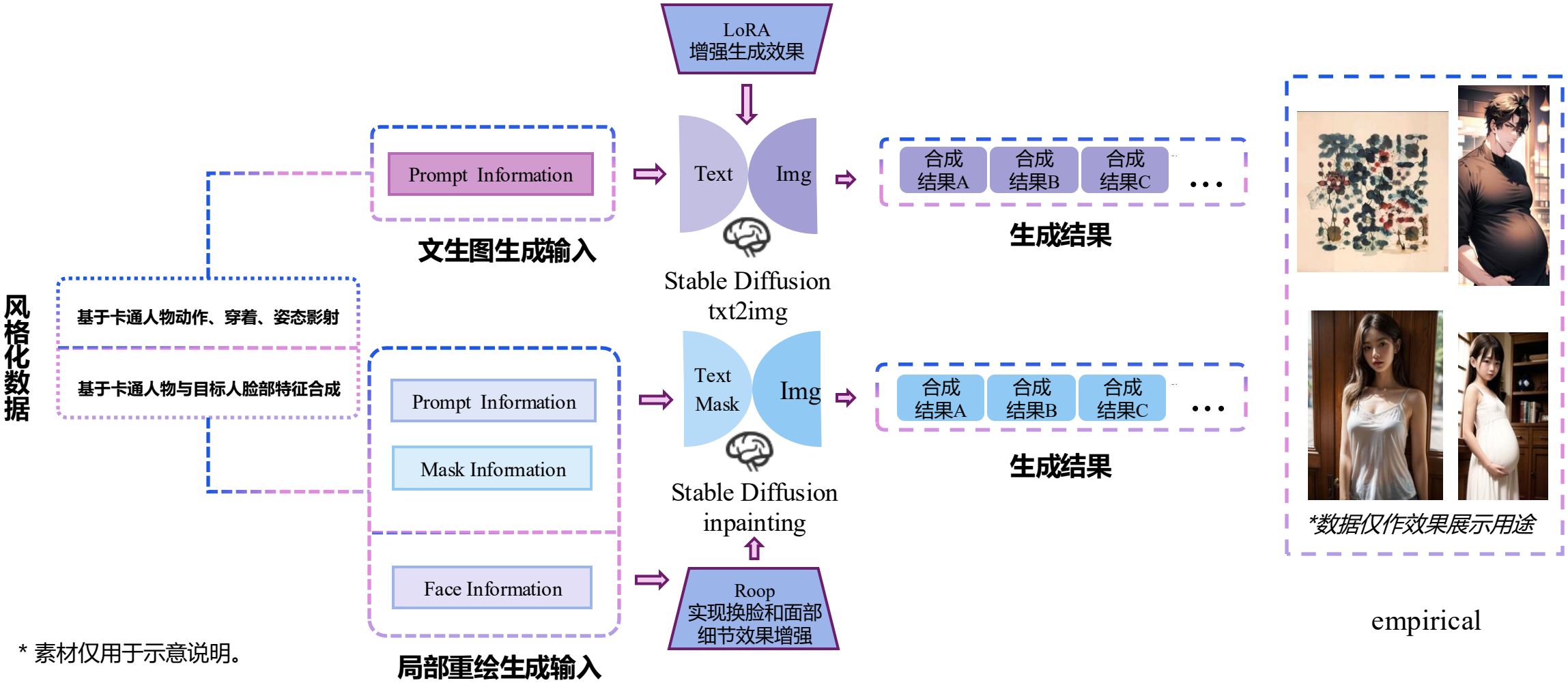
--- 提升新型样本获取的**效率值**

基于跨模态技术的检索、标注和生成

高可控高质量内容生成能力的加持

威胁感知 -- 基于跨模态技术的检索、标注和生成

多模态数据**生成**、检索、标注自动获取



* 素材仅用于示意说明。

empirical

威胁感知 -- 基于跨模态技术的检索、标注和生成

多模态数据生成、**检索**、标注自动获取

“纹身”类别

捞取类别 Prompt设计

Prompt文本特征提取

待捞取数据池

数据池特征提取

相似度度量

匹配 \ 捞取结果输出

后处理环节:

- ✓数据清洗
- ✓多个方法结果融合
- ✓加入训练迭代

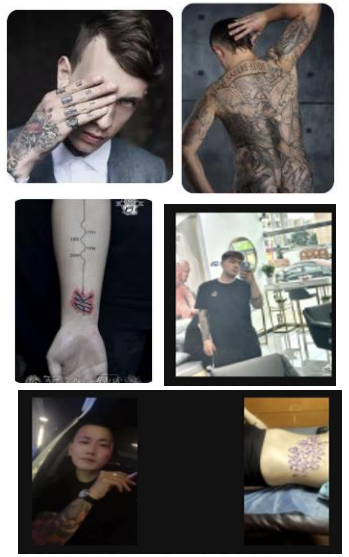
……

Prompt设计:

“纹身”、

“有纹身图样的皮肤”

……



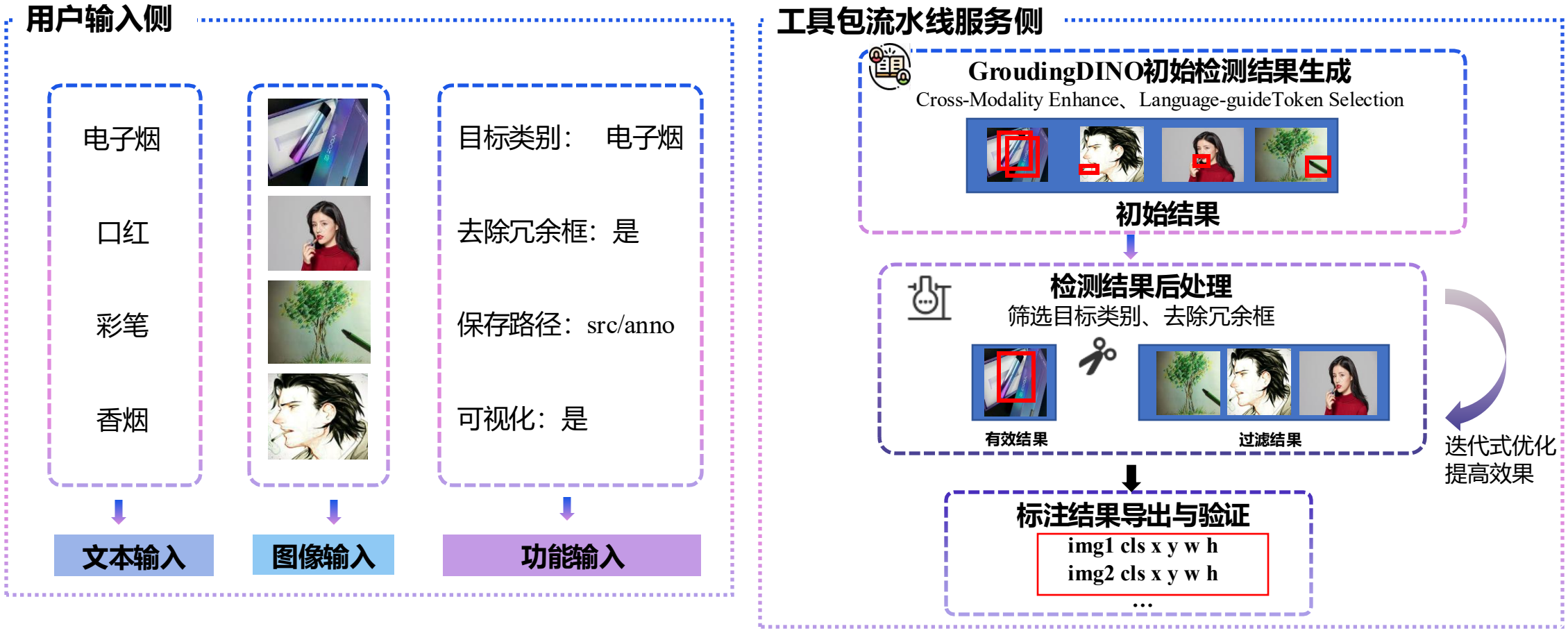
*数据仅作效果展示用途

生成数据 + 自然分布数据

* 素材仅用于示意说明。

威胁感知 -- 基于跨模态技术的检索、标注和生成

多模态数据生成、检索、**标注自动获取**



* 素材仅用于示意说明。

数据噪声问题

GroudingDINO目标检测流水线式标注工具包

高可控高质量内容生成能力的加持

用户输入侧

文本输入

文本prompt：
a photo of a **purple backpack**。

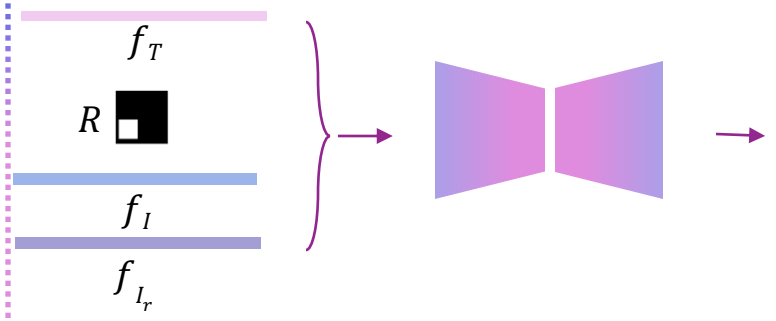
图像输入

背景图像：I

目标区域：R

参考图像：I_r

高可控高质量内容生成



基于大模型的图像生成框架

从**闭合**任务到**开放**任务

--- 提升新型风险识别的敏捷性

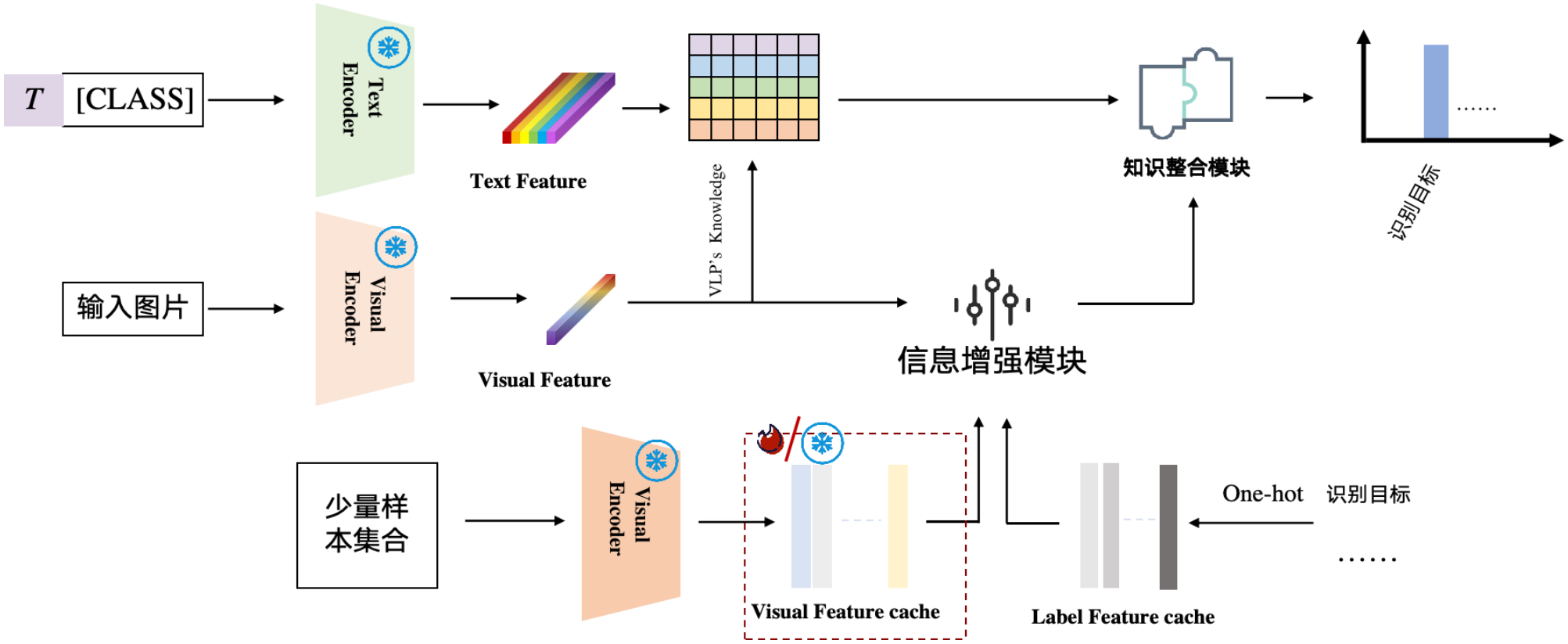
基于跨模态检索的少量样本识别

基于领域泛化技术的新场景适配

基于跨模态生成的精细粒度识别

算法设计 -- 基于跨模态检索的少量样本识别

新增特殊类别：跨模态特征增强



跨模态增强

算法设计 -- 基于领域泛化技术的未知场景适配

新增特殊场景：领域泛化技术

领域泛化

数据
- ✓ AI生成
- ✓ 跨模态检索

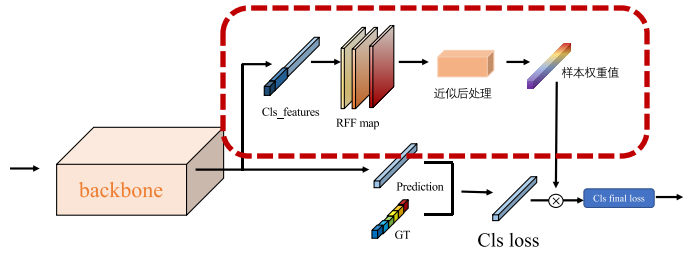
… → 多元训练数据

特征
- ✓ 特征解耦
- ✓ 领域特征对齐
- ✓ 领域对抗训练

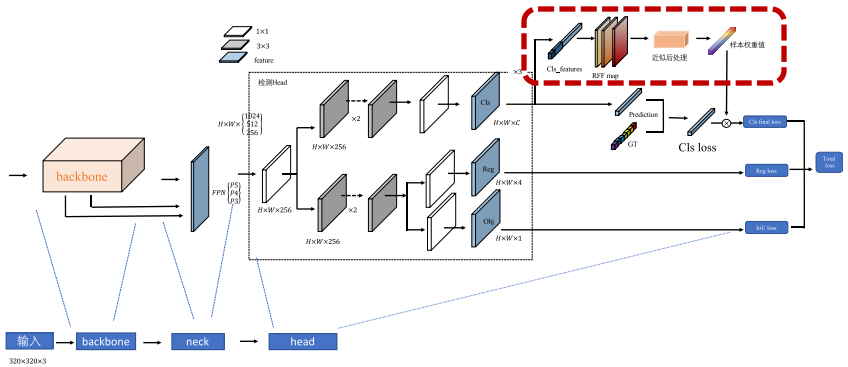
… → 领域不变特征

策略
- ✓ 集成学习
- ✓ 元学习

… → 领域泛化策略



classification



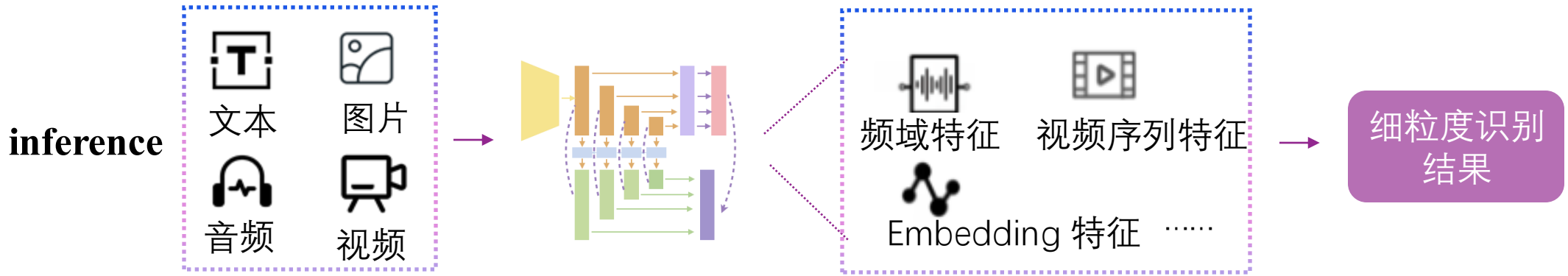
detection

SMA训练示例

* 素材仅用于示意说明。

算法设计 -- 跨模态生成的精细粒度识别

新增标准或策略：精细粒度识别



体系构建

从**单一检测到综合能力**，

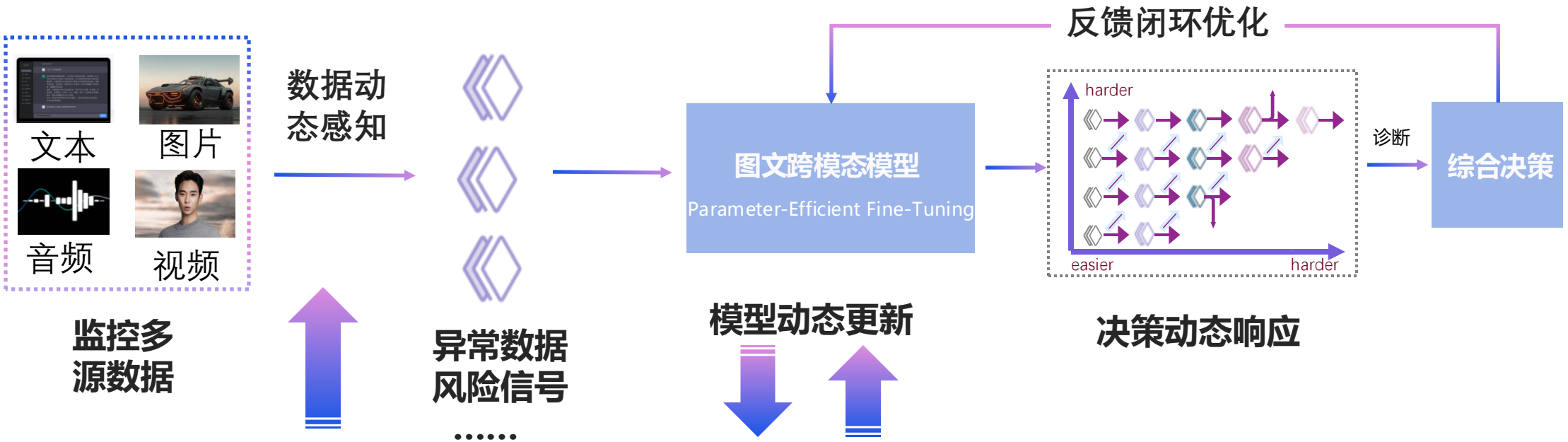
--- 提升复杂风险治理的可控性

弹性防护：领域大模型技术托底的动态处理机制

纵深防御：在线实时检测和近实时巡检的结合

体系构建 -- 弹性防护

应对标准变化：领域大模型技术托底的动态处理机制

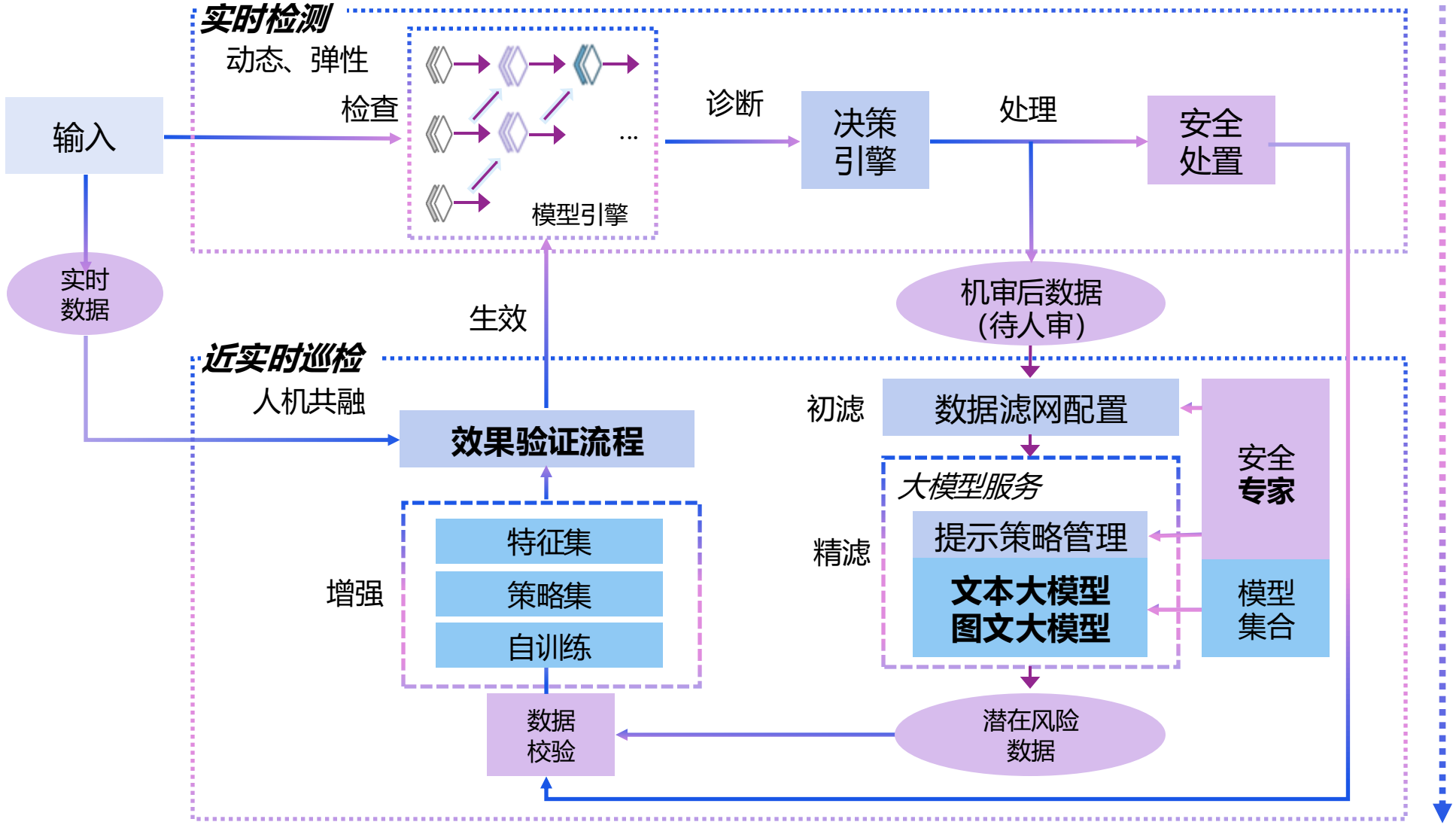


基于领域大模型技术托底的动态处理机制

体系构建 -- 纵深防御

应对 效率 和 效果 的矛盾冲突

在线实时检测近实时巡检的结合



纵深防御

目标： Δt

技术成果

近年累计获得多项AI赛事冠军，在语音、图像、文本等领域核心技术持续保持创新

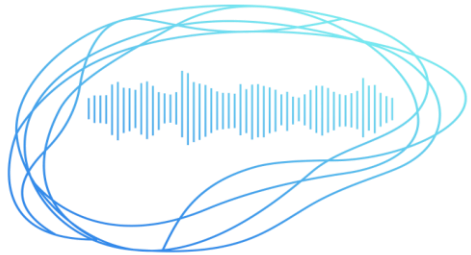
音频技术

论文连续被 ICASSP 录用

论文连续被 INTERSPEECH 录用

全国人机语音通讯学术会议双赛道冠军

……



计算机视觉技术

2022年 ICPR 多模态字幕识别比赛冠军

2021年 “音频和视频深度伪造检测”A级证书

2020年 “视频深度伪造检测”A级证书

2019年 “旗帜识别”A 级证书

……

自然语言技术

2023 NLPCC “用户反馈预测与回复生成”亚军

2022年 NLPCC 自然语言处理任务冠军

2022年 ICPR 多模态字幕识别比赛冠军

2020年 EMNLP ConvAI3比赛顶会冠军

……

PART 03

工程化实践关键突破

攻防对抗强化实践 -- 防御鲁棒性评估方案

核心评估指标

鲁棒性（Robustness）	可解释性（Explainability）	效率（Efficiency）	适应性（Adaptability）
模型抵御对抗样本的能力	防御机制的透明性与调试能力	防御策略的计算资源消耗	防御机制在不同场景中的通用性

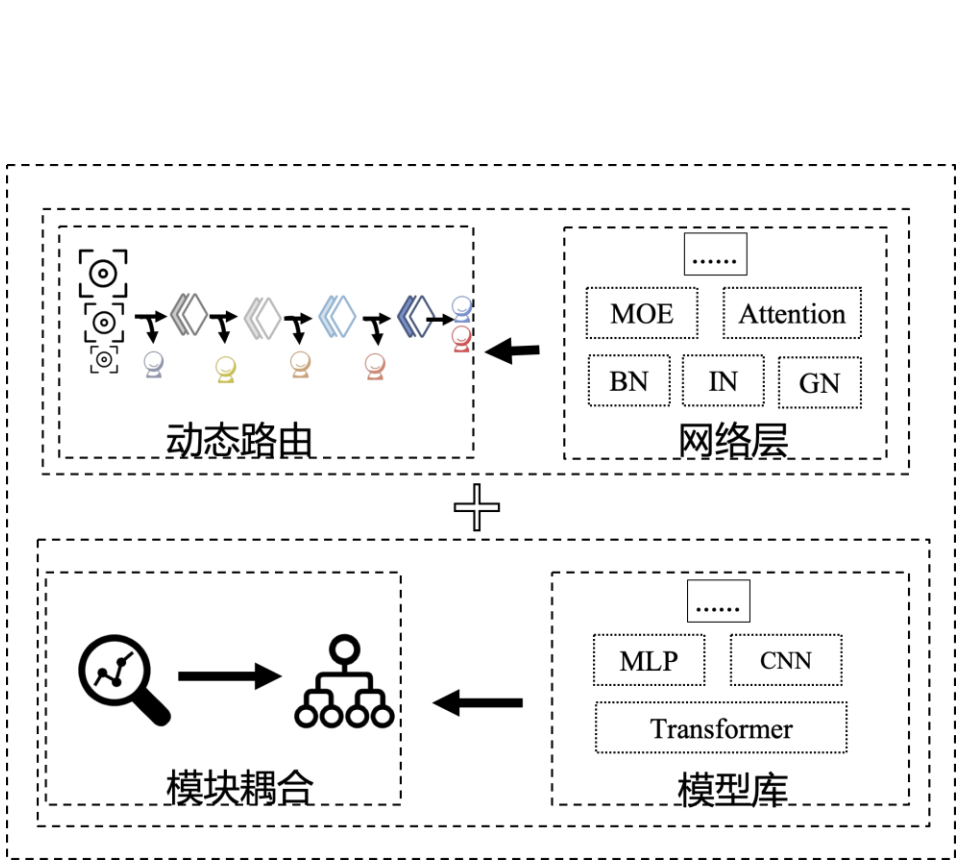
评估流程

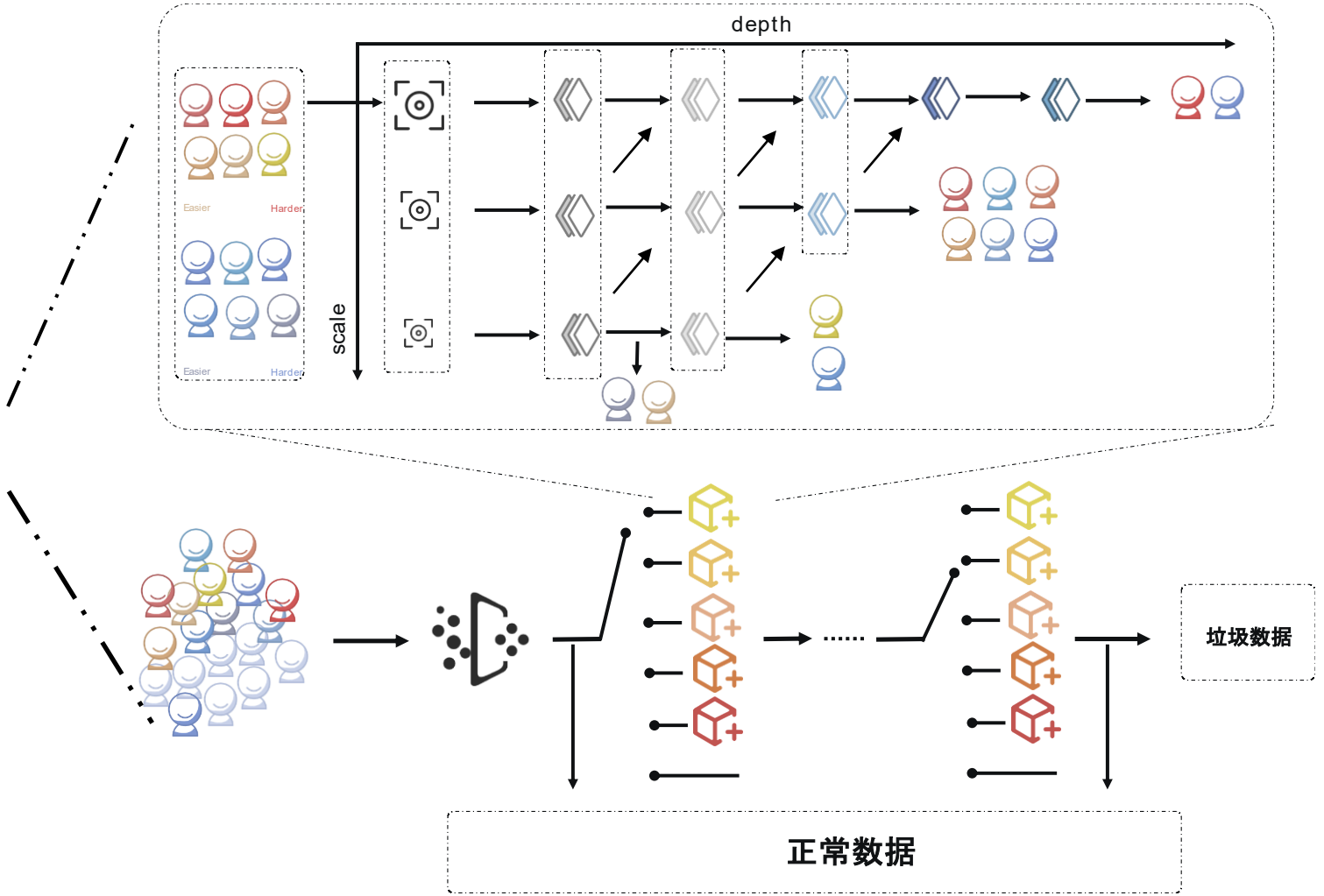
威胁建模	→	防御机制部署	→	鲁棒性测试	→	性能分析
模拟真实攻击场景（如对抗样本生成、伪造内容生成等）		针对不同威胁场景，采用多种防御机制（如对抗训练、检测算法等）。		使用多种攻击方法（如白盒攻击、黑盒攻击）测试防御效果。		根据核心指标（鲁棒性、效率、适应性等）对防御机制进行量化评估。

测试场景

对抗样本攻击　伪造内容检测　滥用防护　.....

攻防对抗强化实践 -- 动态场景下的自适应调节策略





PART 04

金融场景应用与前沿展望

AIGC数字内容风控模型在金融行业中的实践

某证券

内部社区：内容安全机审+人工审核服务+舆情服务+阶段性历史扫描

内容合规痛点

- 平台中荐股刷量广告泛滥，直播黑屏、卡顿，弹幕、评论、资讯文章环节存在涉黄、涉政、低俗的不良内容，面临着垃圾变种信息审核的挑战。
- 纯人工审核成本高，效率低难以应对复杂多变的审核需求。

业务风险场景

文章资讯

直播

昵称

领域	关键审核点
荐股吹票识别	精准识别大肆吹嘘个股价值，直接提示买卖点信号，明显个人情绪导向，缺乏研 究支撑的个股推荐
虚假夸大识别	客观陈述事实，不得博眼球，避免夸大，不出现“震惊世界”“领先全球”等用语
市场负面识别	不得激烈描述上市公司负面信息，或监管单位，金融机构的负面观点
风险提示识别	视频显著标识投资有风险，观点仅供参考，视频内容仅代表作者个人观点等免责 声明不得缺失
投顾身份识别	人脸识别知名财经人物，识别投顾资格证编号等信息，核实身份有效性。
数据来源识别	涉及具体财经数据和图表时需明确数据来源及统计周期，不得采用非平台认可数 据源数据

建设成果

99.5%

审核准确率

5倍

单人日审核数据量提升

AIGC数字内容风控模型在金融行业中的实践

某证券

产品销售：内容安全机审+人工审核服务+舆情服务+阶段性历史扫描

产品业务：

xx财富通、xx全球通、xx证券行知

金融产品均为官方内容，安全风险低，但涉及银保监、广告法等法规的限制、专业性强

70%

人工审核成本降低

金融业务

- 虚假、夸大宣传
- 偷换概念、简单比价，误导消费
- 信息披露、风险告知不到位
- 营销主体的资质

直播

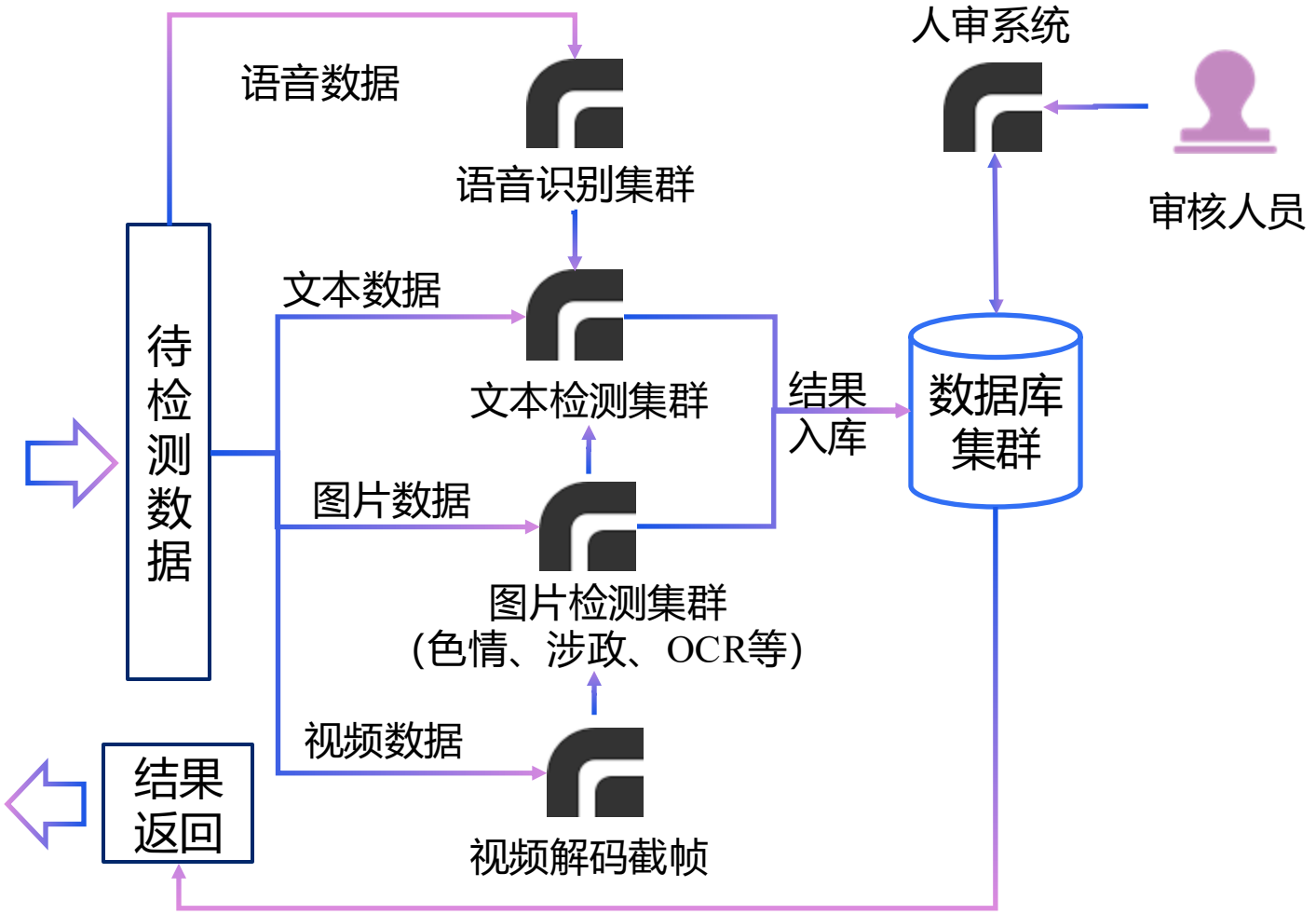
- 主播
  - 主播名单-劣迹艺人、涉黄赌毒
  - 主播着装-暴露；公职服饰；纹身
- 画面
  - 内容-吃播、封建迷信等
  - 动作-撩骚、挑逗
- 音频
  - ASMR、娇喘、涉政等

AIGC数字内容风控模型在金融行业中的实践

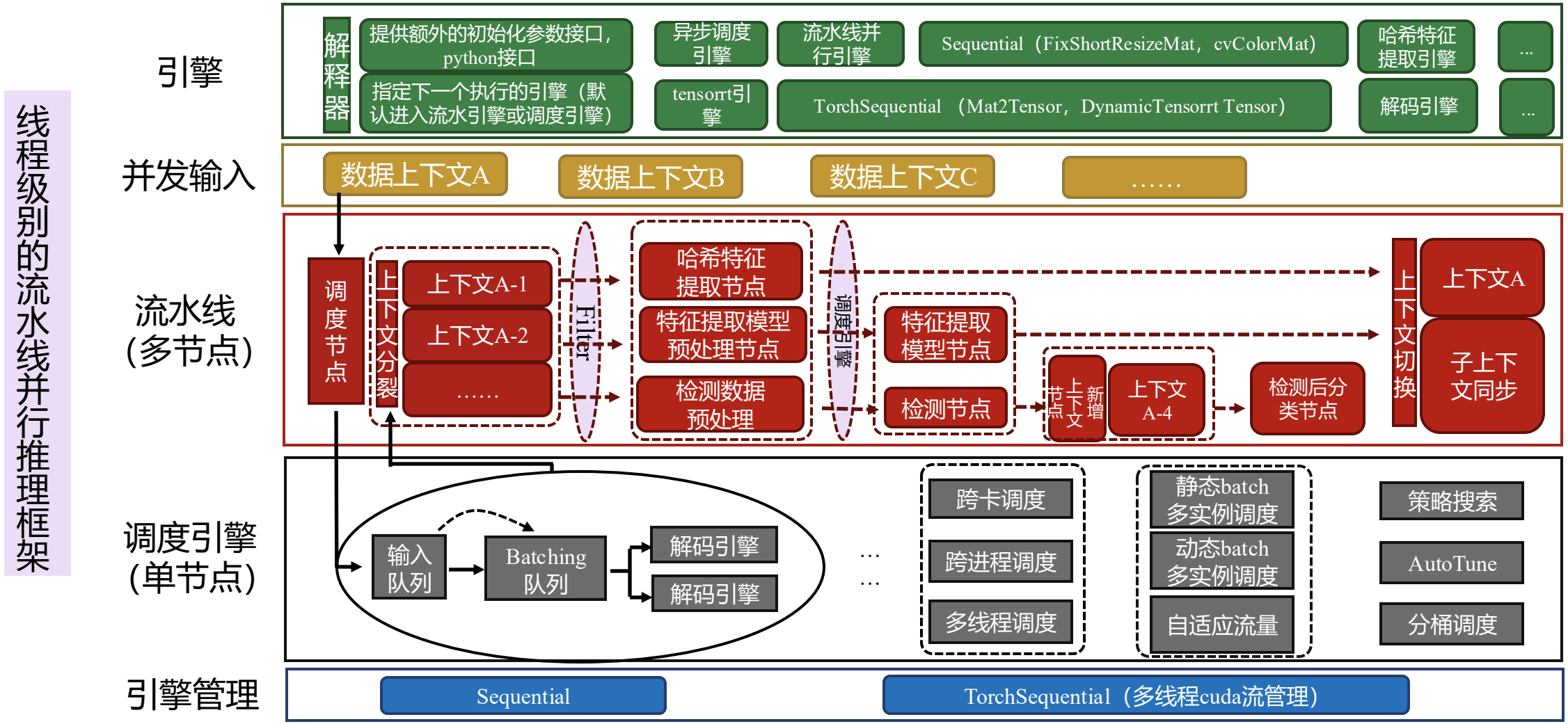
集团全业务、全场景、全模态接入内容安全合规检测，机审+人审，高效准确地识别和过滤违规内容。

业务系统盘点

1. 综合财富管理平台-通信内容、祝福语
2. 集团招聘-招聘JD
3. 普惠金融-购销需求描述
4. 通讯基础平台-用户信息
5. 手机银行-客户自定义内容
6. 网银主页系统-信息发布审核
7. 金融营销协同系统-营销活动昵称
8. 财富产品准入-产品物料消保
9. 内容管理服务系统-内容素材管理
10. 5G消息平台-消息素材模板管理
11. 同业与金融市场综合业务服务平台-机构入驻、投资生态圈
12. 园区运营管理系统-多场景信息内容检查
13. 中小微收单商户行业应用系统-商户入驻运营审核
14. 统一消息中心系统-短信模板



算法普惠化技术实践

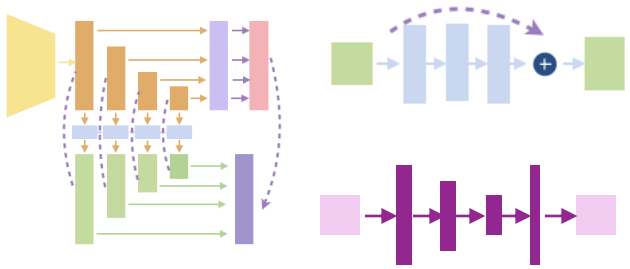

线程级别的流水线并行推理框架
引擎
并发输入
流水线
(多节点)
调度引擎
(单节点)
引擎管理
解释器
提供额外的初始化参数接口，python接口
指定下一个执行的引擎（默认进入流水引擎或调度引擎）
异步调度引擎
流水线并行引擎
Sequential（FixShortResizeMat，cvColorMat）
哈希特征提取引擎
...
tensorrt引擎
TorchSequential （Mat2Tensor，DynamicTensorrt Tensor）
解码引擎
...
数据上下文A
数据上下文B
数据上下文C
……
调度节点
上下文分裂
上下文A-1
上下文A-2
……
Filter
哈希特征提取节点
特征提取模型预处理节点
检测数据预处理
调度节点
特征提取模型节点
检测节点
上下文节点 新增
上下文A-4
检测后分类节点
上下文切换
上下文A
子上下文同步
输入队列
Batching队列
解码引擎
解码引擎
跨卡调度
跨进程调度
多线程调度
静态batch多实例调度
动态batch多实例调度
自适应流量
策略搜索
AutoTune
分桶调度
Sequential
TorchSequential（多线程cuda流管理）

从分治到统一：多模型\多服务能力收敛到统一大模型

统一大模型

跨语言 跨模型 跨任务

模型库



……

服务库

色情 涉政 暴恐

广告 广告法 涉价值观 ……

其他 违禁 灌水

多模型 / 多服务能力

AiDD 6th 2025

参与调研您将优先获得

AiDD定制版
《AI+软件研发精选案例》

专属学习顾问
1对1需求对接

AiDD会后小调研

AiDD峰会致力于协助企业利用AI技术深化计算机对现实世界的理解，推动研发进入智能化和数字化的新时代。作为峰会的重要共建者，您的真知灼见对我们至关重要。衷心感谢您的参与支持！

扫码参与调研

2025 AI+研发数字峰会

拥抱AI 重塑研发

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

K+峰会

K+峰会 敦煌站
K+ 思考周®研习社
时间：2025.08.29-30

K+峰会 上海站
K+ 金融专场
时间：2025.09.26-27

K+峰会 香港站
K+ 思考周®研习社
时间：2025.11.17-18

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2025.05.23-24

AiDD峰会 北京站
AI+研发数字峰会
时间：2025.08.08-09

AiDD峰会 深圳站
AI+研发数字峰会
时间：2025.11.14-15

AiDD峰会详情

AiDD 6th 2025

2025 AI+研发数字峰会

AI+ Development Digital Summit

感谢聆听！

扫码领取会议PPT资料