AI在短视频创作与理解上的应用

王仲远 快手

快手技术副总裁，MMU&Y-tech负责人

2021.11.25

QCon+ 案例研习社

扫码学习大厂案例

学习前沿案例，向行业领先迈进

40 个
热门专题
—
行业专家把关内容筹备，
助你快速掌握最新技术发展趋势

200 个
实战案例
—
了解大厂前沿实战案例，
为 200 个真问题找到最优解

40 场
直播答疑
—
40 位技术大咖，每周分享最新
技术认知，互动答疑

365 天
持续学习
—
视频结合配套 PPT
畅学 365 天

快手-国民短视频及直播社区

流量

3.2亿
国内日活用户

5.7亿
国内月活用户

1.8亿+
海外月活用户

内容

数百亿量级
短视频库存

粘性

100min+
日均使用时长

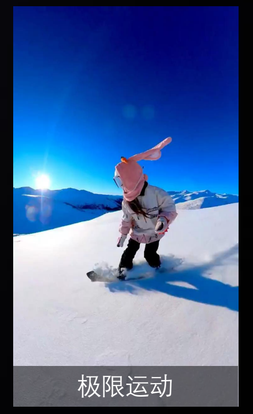
140亿+
对人互相关注

10次
日均访问次数

70%
私域渗透率

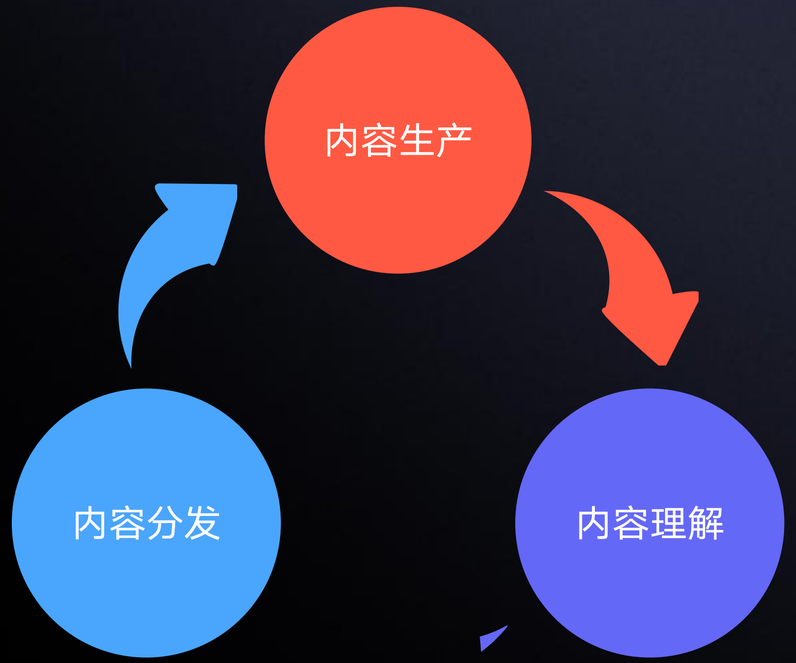
快手

快手-拥抱每一种生活



快手

AI技术在快手的应用



- **内容生产：**在APP中提供炫酷的视觉特效、魔法表情、一键出片、自动字幕等AI工具和玩法，依赖AR引擎、人脸&手势识别、语音转译、智能创作等自助研发技术。

- **内容理解：**基于对社区中视频、图像、音乐、语言语义、主播和创作者的理解，充分结构化解释快手的内容生态，实现了社区海量内容的分类管理、原创保护、安全审核、助力分发等诸多应用。

- **内容分发：**推荐是用户与视频的双向匹配，将百亿视频特征和亿万用户特征输入推荐系统，实现精准、个性化的推荐。

AI在内容生产中的应用

快手

一个离不开美颜的时代

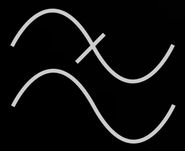
不管是拍照、拍视频还是直播，美颜如今已是大家依赖的基础能力。

人像美化技术现状

现有美化流程和技术

如何达到用户想要美化效果

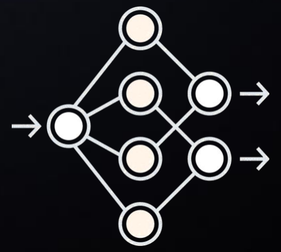
优化现有技术



- 优势：迭代快，性能可控
- 劣势：效果上限不高，自适应能力差

VS

优化引入新变量 - AI能力



- 优势：复杂高级效果，自适应
- 劣势：研发周期长，端上运行慢

人像美化： AI美颜技术

AI一键美颜

一键式磨皮，提升面部立体感肤色自然过渡

AI头发生长

任意短发变成长发模样

AI人像画质增强

修复对焦不准、低端机成像差等等导致的人像画质问题

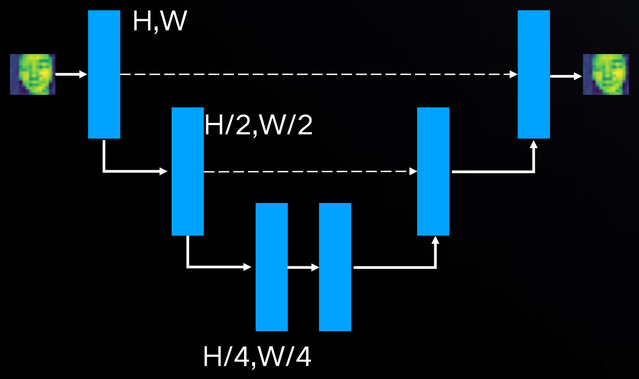
双眼皮生成

自然将单眼皮变为双眼皮

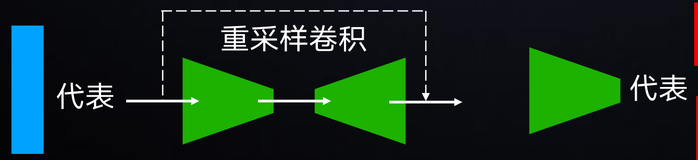
快手

人像美化：一键AI美颜完整技术方案

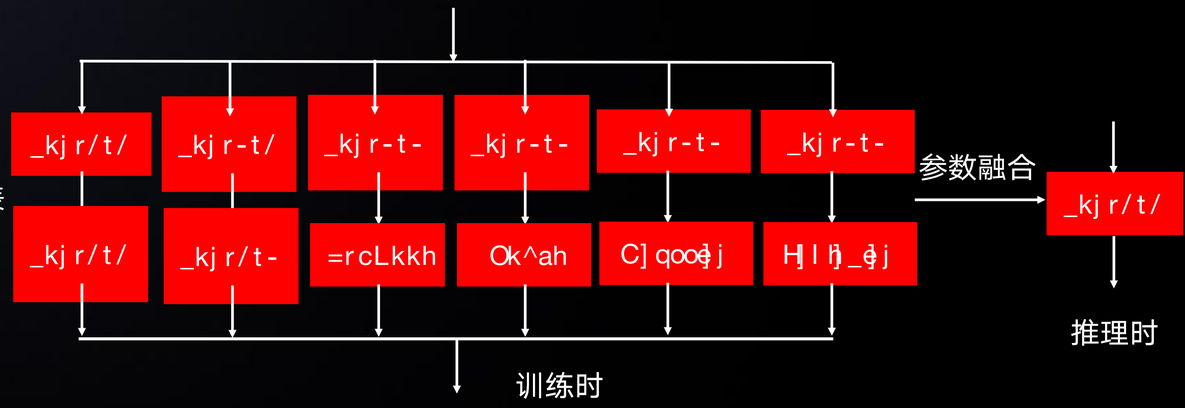
我们在网络设计上使用了encoder-decoder的结构，网络中每个block都使用了先降维度再升维的重采样卷积结构来降低计算量，每个重采样卷积我们创新性的加入了sobel，gaussian等传统图像处理算子在更进一步提升训练时网络容量的情况下不增加推理时的性能开销。



网络结构



每个block结构



每个重采样卷积结构

AI在人脸属性变换中的应用

人脸属性编辑：对含有人脸的图像，进行人脸属性变化，可返回各种处理效果，效果真实自然。

Demo体验

原图

变少年

变老

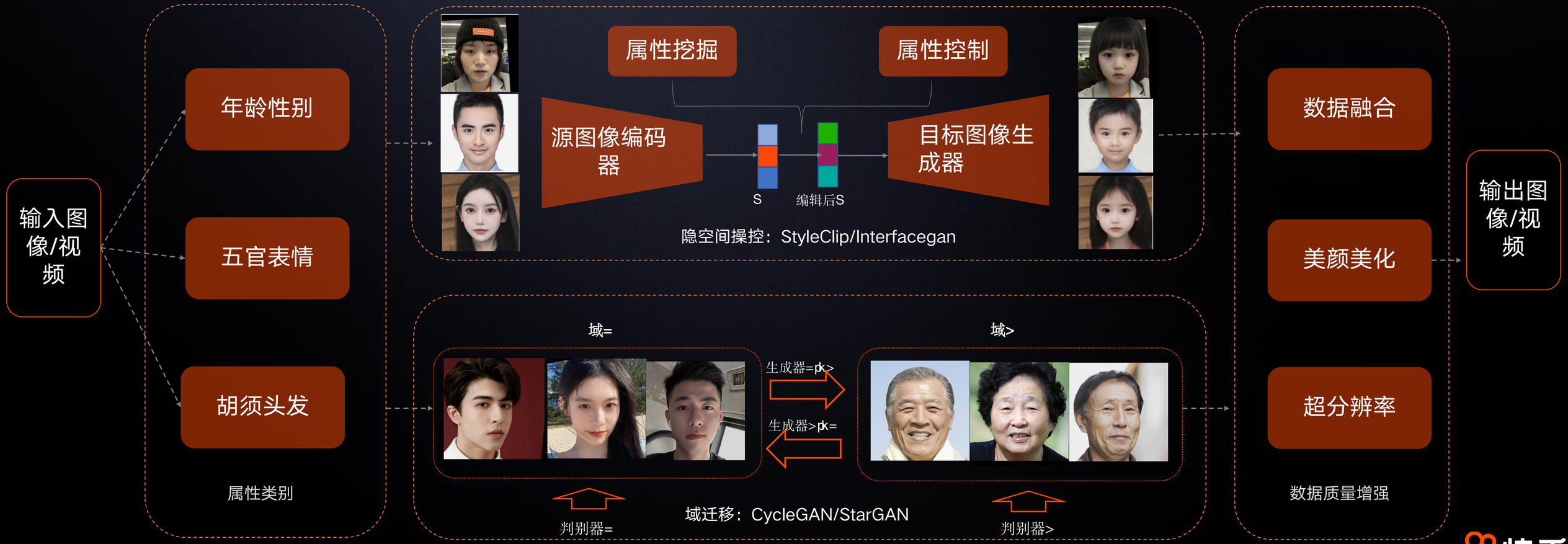
变性别

变胖

快手

AI在人脸属性变换中的应用

人脸属性编辑：人脸属性编辑项目主要包含属性的类别判断，属性的编辑操控及一些数据质量增强方法，通过串联这些技术点，最终实现从输入图像到输出图像的属性变换的效果。其中最核心的为属性操控，主要采用两种思路，分别为基于domain transfer的学习，和基于隐空间的操控。



魔法表情：人像风格化

生成式人像风格化：人像风格化是指通过AI算法，将输入的人像图片转换成具有目标风格的人像图片。转换后的图片既保留输入人像图片的五官和外形等特征，又具备目标风格的美学和艺术效果。

言情手绘

主站-我的手绘脸

东方国漫

主站-神仙拜年

国风风格

主站-新春画中人

国风风格

主站-国风美人

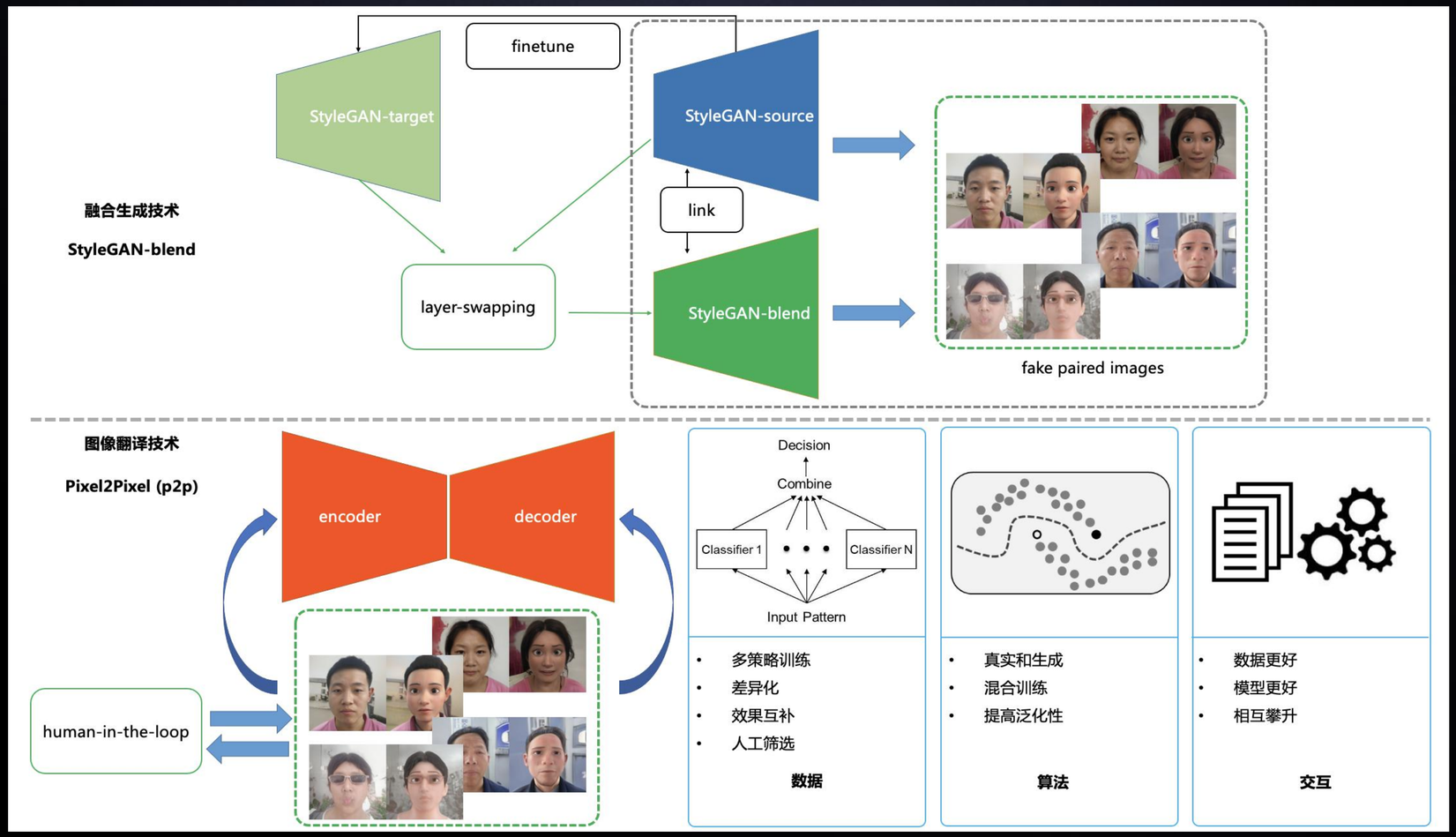
韩漫风格

主站-综艺大咖

快手

魔法表情：人像风格化

生成式人像风格化：我们提出了一个高效的人像风格化落地方案，包括基于SpuraC=J)^haj`ėc等技术的小成本的高质量风格数据生成，以及采用了半监督混合训练框架和dqi]j)ė)pda)hkkl 策略，以适配于快手场景的风格模型训练与迭代。



小成本数据生成

- 科学的数据分布与场景覆盖
- 海量的虚拟/真实数据收集
- 基于少量样本的高质量数据生成

模型训练与分级部署

- 高效的模型训练框架
- SOTA的风格迁移模型效果
- CPU/GPU/NPU/DSP支持与性能优化
- 不同算力下的最优效果展示

多种玩法灵活支持

- 服务端全图风格化
- 客户端实时多人风格化
- 客户端实时单人多风格化

智能语音

智能语音在快手普遍被应用，自动字幕、语音合成、智能RAP为用户提供更方便、快捷、有趣的工具和玩法

自动字幕（ASR）

智能语音合成（TTS）

智能RAP

在生产环节创造的价值：

- 降低创作视频的门槛，配合推荐在配乐、视频制作上提供更多工具，让每个人都可以止制作更精致的视频
- 提供更多玩法，让视频更有趣

智能说唱配音

智能说唱配音可以识别视频主题，自动编写与主题匹配的说唱歌词并演唱，可丰富视频配乐的玩法手段。

图像

自动视频理解

可爱的小猫

NLP

自动歌词生成

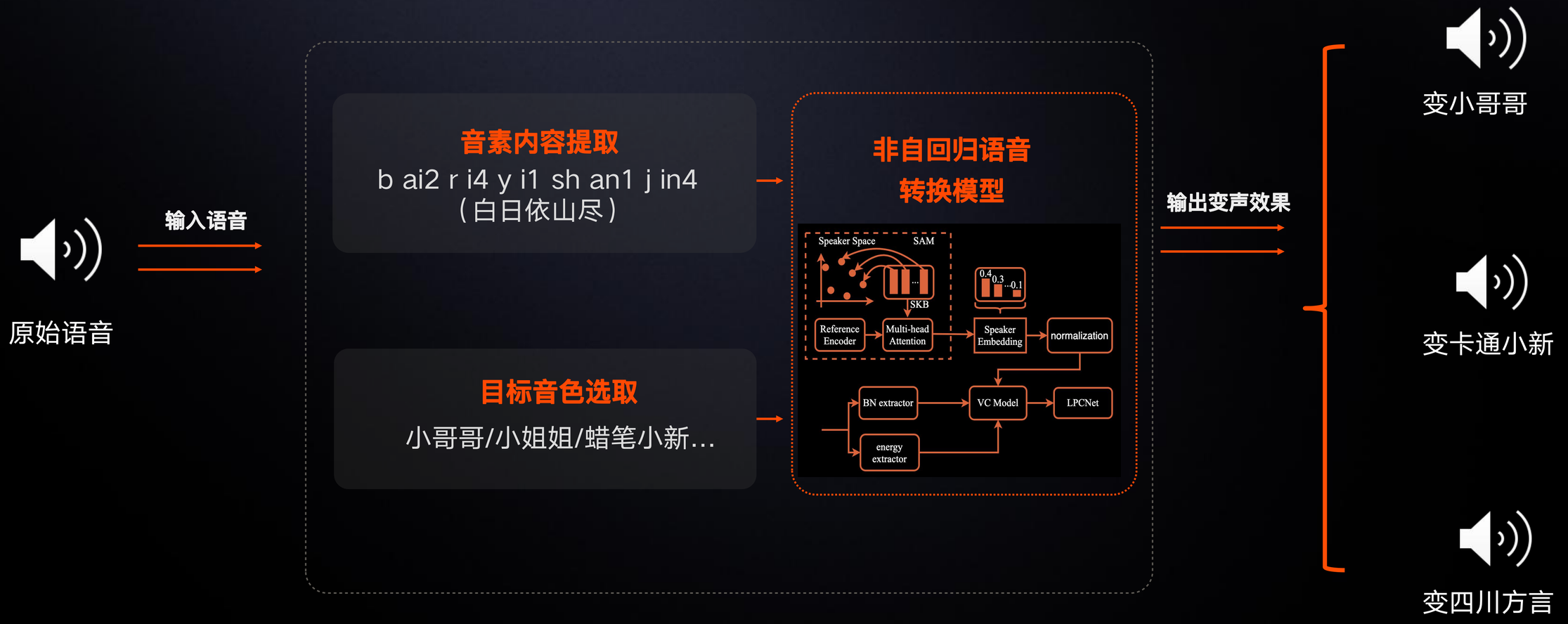
| 可爱猫咪 好听话
打扮起来 本领大 |
| 小小爪子 是活动家
吐司头套 给你抓 |

语音

自动说唱生成

智能变声

智能变声能够将一个人的声音变成另外一个人的音色，同时保持说话内容不变，可用于视频创作、直播等场景。首先通过一个音素内容提取模块从原始语音中提取内容信息，然后根据选择的目标音色id生成目标说话人表征向量，将这两者信息通过语音转换模型进行耦合，生成含有目标人音色，原始语音内容信息的变声语音。

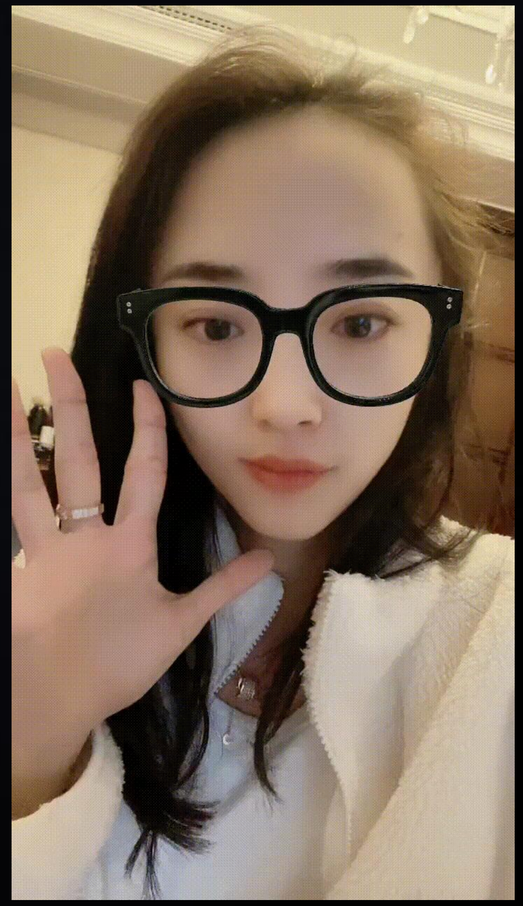


AI歌手


2019年12月
搭建基础BaseLine，采用传统BLSTM语音合成框架
2020年5月
采用改进后的自回归网络+LPCNet声码器
2020年8月
优化声码器高音预测音质差的问题
2020年11月
改用PW-GAN声码器方案，音色更加自然，且提升了音质及采样率
2020年12月
优化声学模型结构，改善发音及跑调，提升合成表现力
NeuralNet声码器版本迭代
传统Source-Filter声码器版本迭代
采用WORLD版本，保证音准
优化发音及自然度
增加谱后处理优化

普通魔法表情

魔法表情使视频制变得更具有趣味性，提升了大家的创作热情。



魔法表情：混合现实

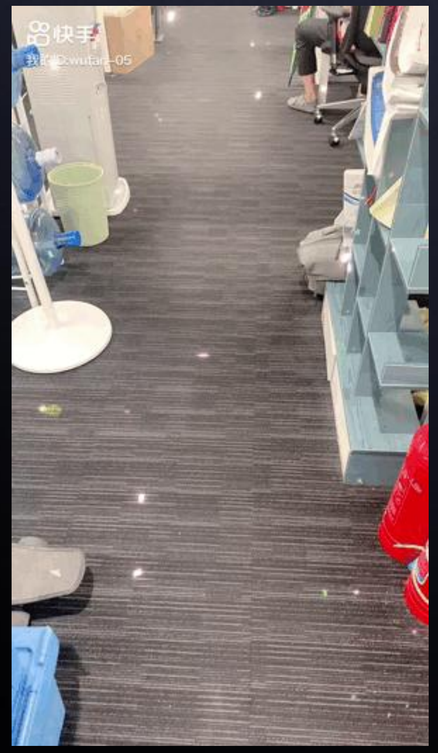
混合现实技术： 通过AR及其拓展技术，实现虚实融合和交互。

地标AR

主站-太古里熊猫

地面AR

主站-地面开花



空间AR

主站-许愿魔表

建筑物平面AR

主站-七夕投影

自研流体特效

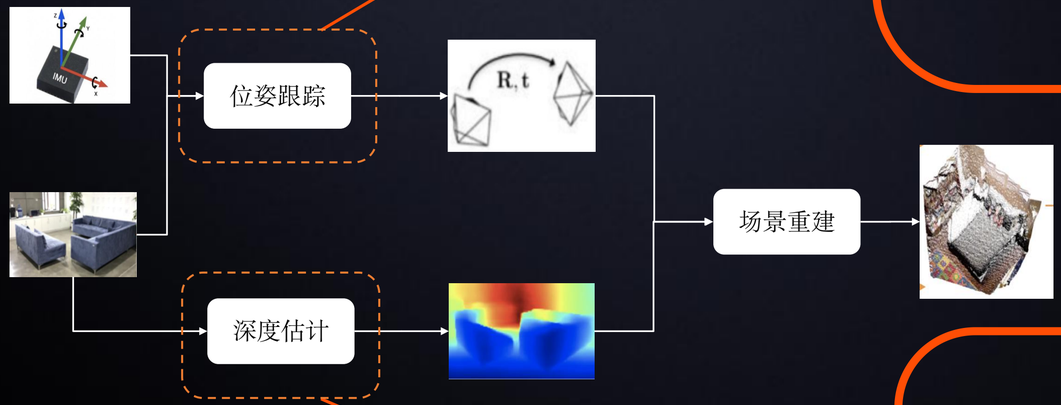
主站-别哭鸭

快手

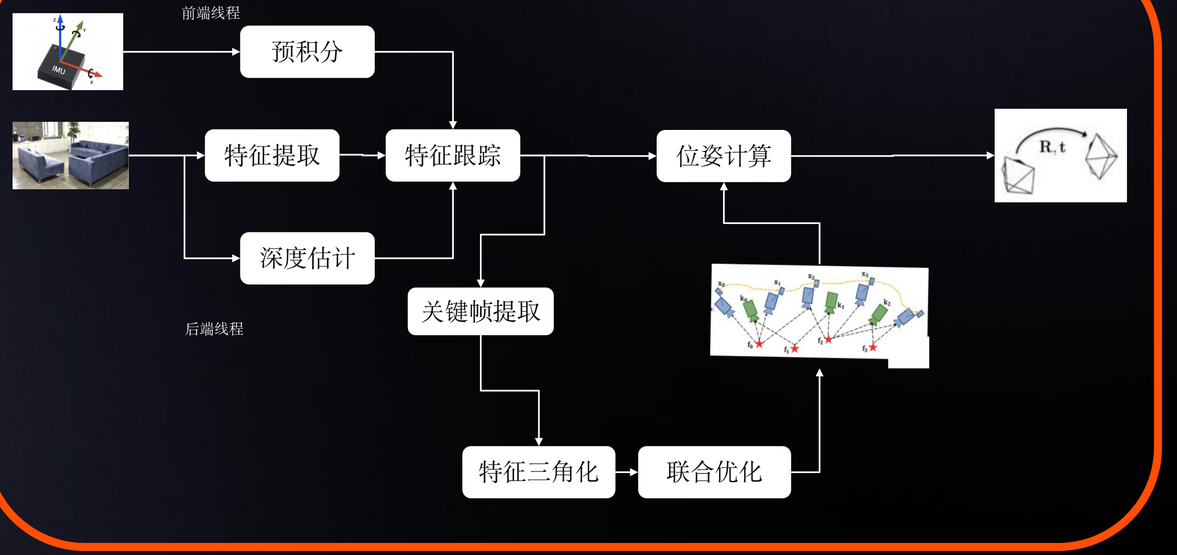
魔法表情：混合现实系统框架

混合现实技术： 混合现实系统主要分为三大模块，包括位姿跟踪、深度估计和场景重建。

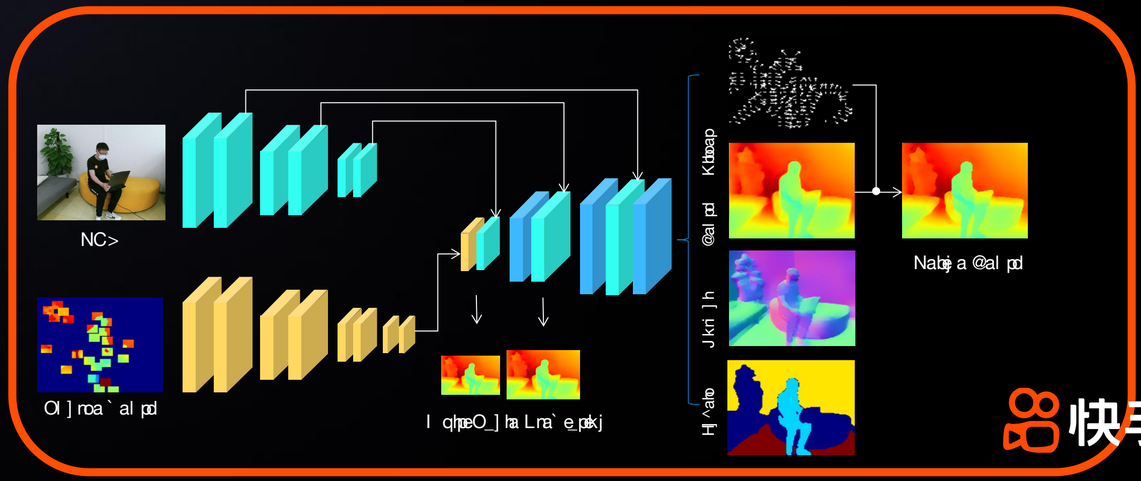
MR整体技术框图



位姿跟踪（RE0H=I ）流程图



深度估计网络框架图

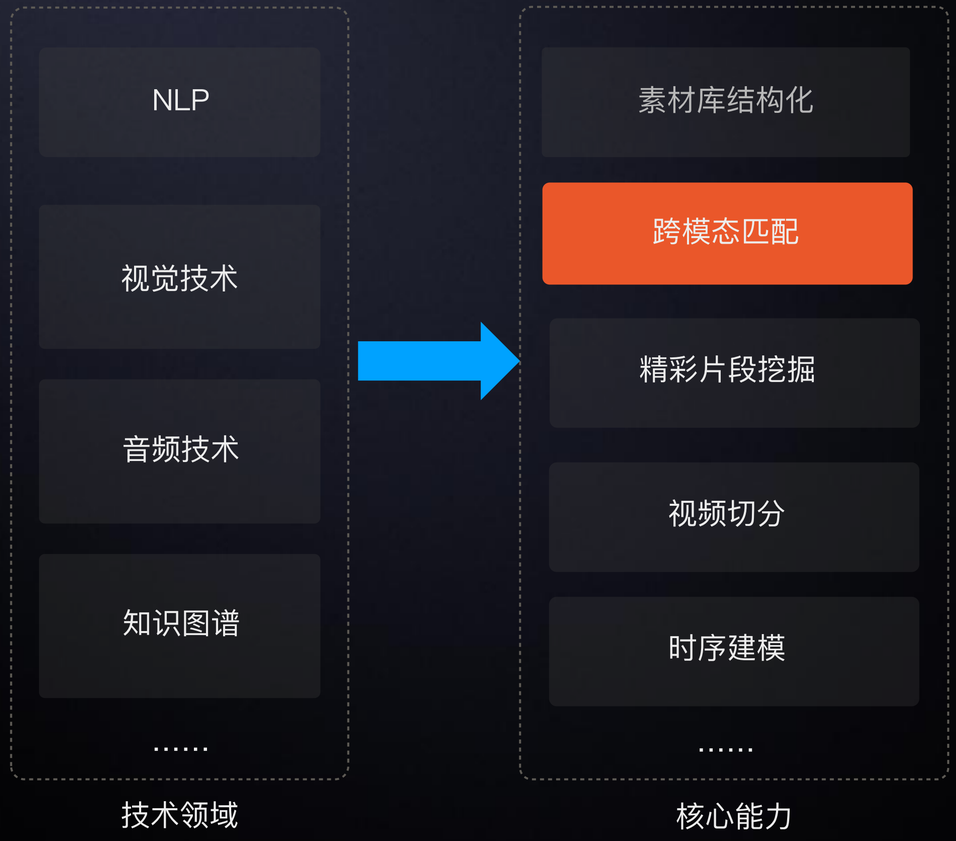


除了常规算法逻辑，我们针对于快手用户的设备分布和使用情况，做了.个方向的调整：
1、模块紧耦合设计，提升鲁棒性和尺度一致性
2、性能分级设计

最终实现MR系统覆盖80%的Android用户和99%的iOS用户。

智能创作

智能创作即基于素材的混剪，依托MMU的多项技术，构建智能短视频混剪技术，提升制作效率，补充供给。

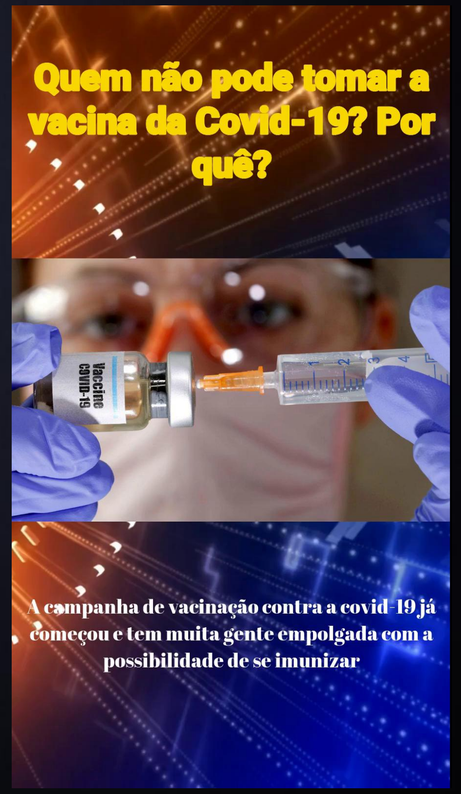


智能创作

在快手生态下，探索在用户增长、商业化广告、海外内容供给、新玩法等多个场景的落地。

国内热点新闻生产
（奥运）

广告生成



海外新闻自动生成

直播剪辑
（多场次）

快手

智能创作

原始版本

很多人不知道漠河冬天开水泼出能成雾霜。
很多人不知道乌苏里江大马哈鱼有多香。
很多人不知道新疆哈密瓜地里姑娘有多漂亮。
很多人不知道曾母暗沙海底有无数宝藏。
向更大的世界开始探索吧。
去体验
去感受
去交流
去求证
看看古老手艺如何惊艳时光。
看看翩翩少年如何奋发图强。
看看耄耋老人如何白头偕老。
看看芸芸众生如何逆风飞翔。

《看见》文案节选

一键
成片

智能创作版本

智能创作

人工创作

一键成片

AI创作

智能创作

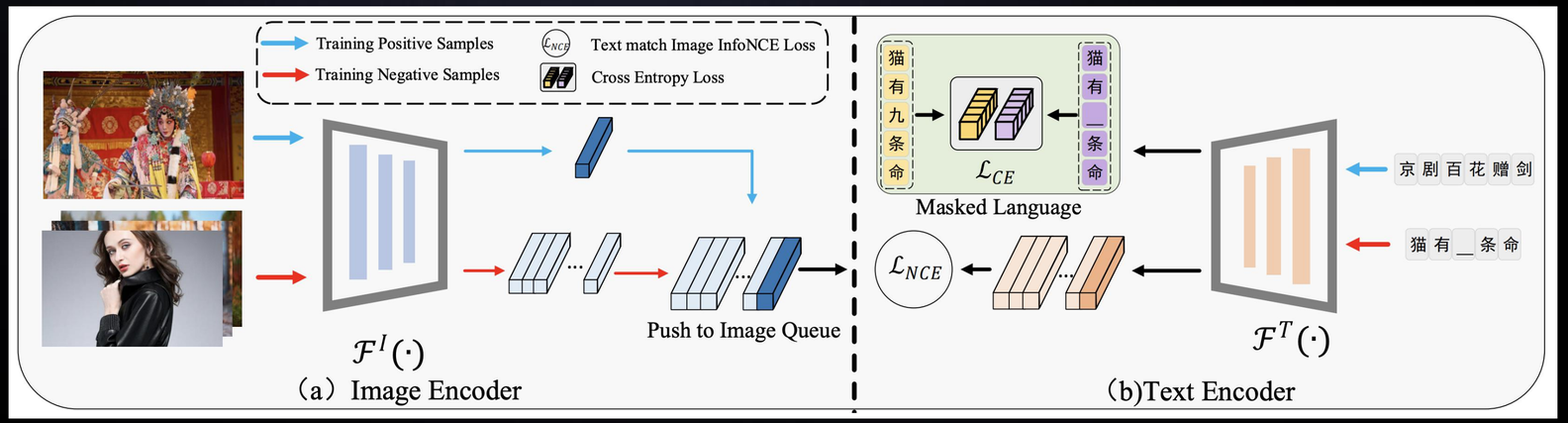
建立了百万级的高质量结构化素材库，通过基于互联网数亿图文数据的自研大规模预训练模型进行跨模态匹配，更好的克服了训练样本中的噪声，增强了对文本改写的鲁棒性。同时针对混剪创作需求，加入了素材序列优化和多模型融合排序等策略，最终融合智能TTS和配乐技术实现整片的合成。

百万级图片/视频素材库 → **跨模态语义匹配** → **整片合成**

- 违规素材清洗过滤
- 相似素材去重处理
- 标签结构化处理

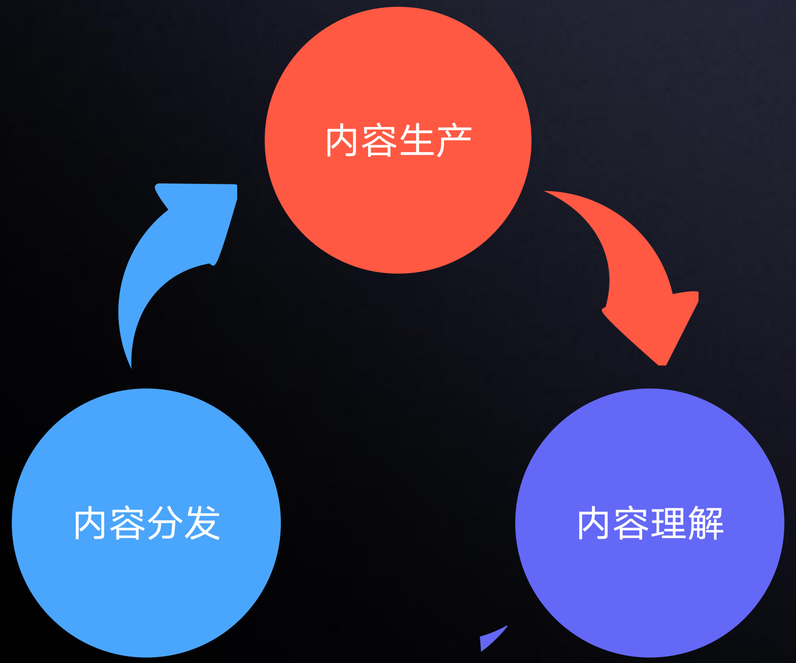
- 跨模态语义检索
- 时序一致性优化
- 粗排序+精排序召回策略

- 智能配乐
- 语音合成
- 转场特效



自研跨模态检索模型

AI技术在快手的应用



- **内容生产：**在APP中提供炫酷的视觉特效、魔法表情、一键出片、自动字幕等AI工具和玩法，依赖AR引擎、人脸&手势识别、语音转译、智能创作等自助研发技术。

- **内容理解：**基于对社区中视频、图像、音乐、语言语义、主播和创作者的理解，充分结构化解释快手的内容生态，实现了社区海量内容的分类管理、原创保护、安全审核、助力分发等诸多应用。

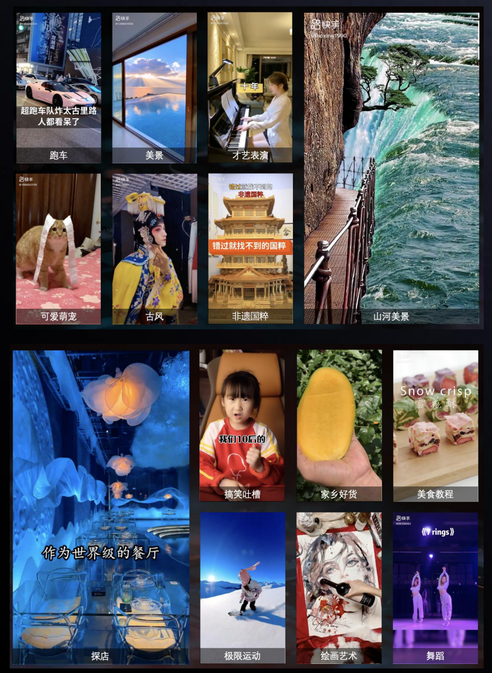
- **内容分发：**推荐是用户与视频的双向匹配，将百亿视频特征和亿万用户特征输入推荐系统，实现精准、个性化的推荐。

快手

AI在内容理解中的应用

终极目标：让机器像人类一样理解视频内容及用户生产的各种内容

通过计算机视觉、语音、自然语言处理、知识图谱、多模态等技术，准确高效地理解视频内容及用户生产的各种内容，并应用在推荐、搜索、广告、垂类运营、生态分析、内容安全等各种场景中。



输入：多模态内容

输出：理解结果

构建海量视频图书馆–河图体系

海量视频结构化管理和应用，自动化完成视频分类、精细化标签

视频质量自动解析

自动化挖掘优质视频，过滤劣质视频

内容安全体系

全面建设AI智能审核，保安全，提效率，省成本

评论分析和理解

挖掘神评热评，促进社区活力，提升创作率

音乐理解

识别音乐特征，AI歌手与AI音乐

知识图谱

做视频百科，挖掘有用的内容

商业内容理解

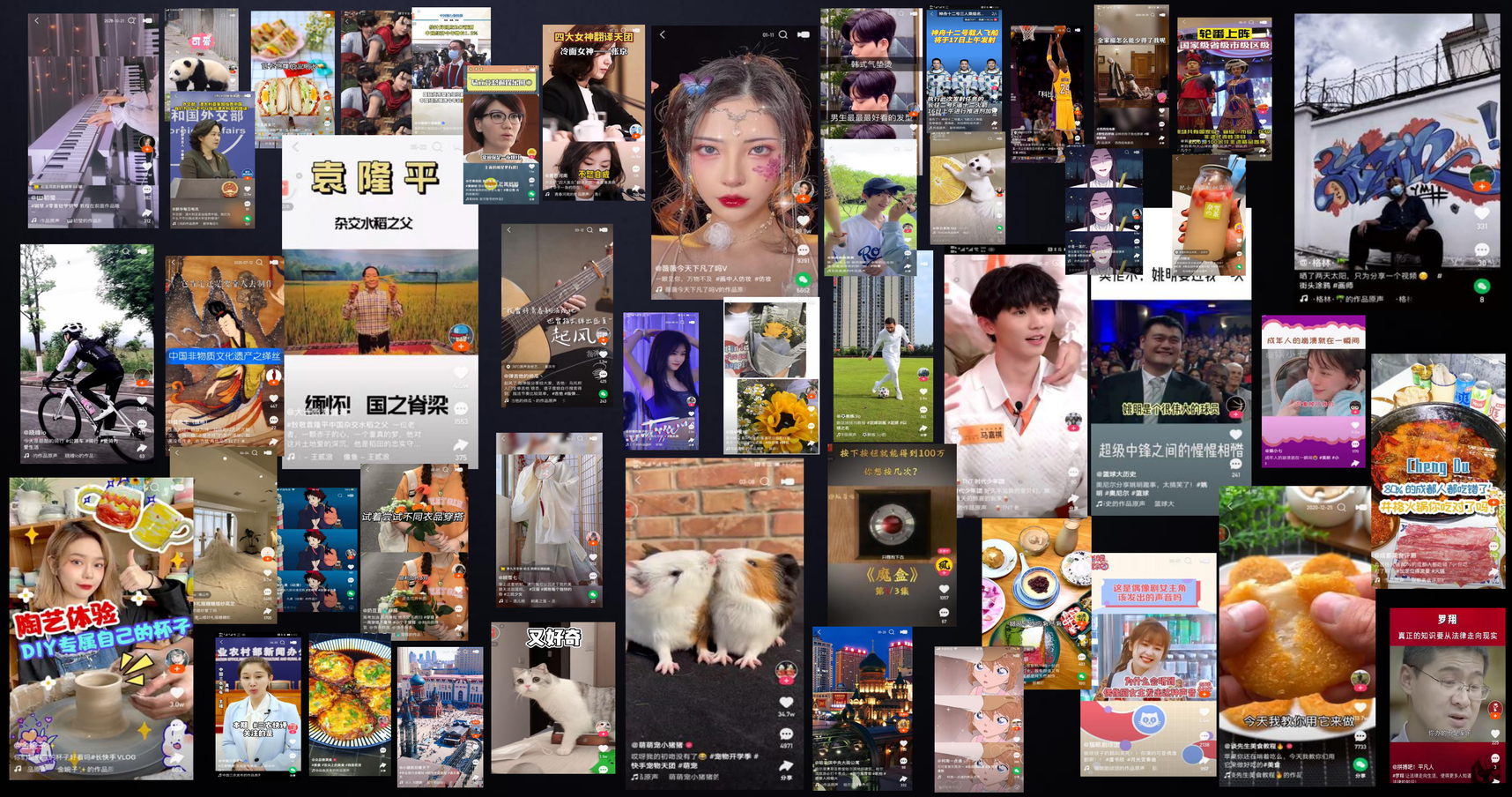
挖掘视频电商意图，提升卖货效率

其他AI能力

快手

智能视频分类与标签

河图体系的核心是：通过结构化的方式解释快手的视频，在结构化之前，视频的内容时纷乱复杂的。



快手

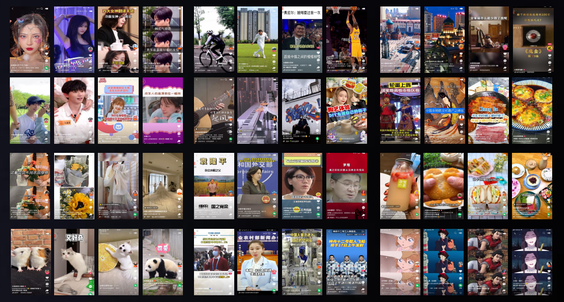
智能视频分类与标签

河图体系可通过三种方式表达视频内容，视频类目、视频标签、多模态向量化表征。

图书馆

- **分门别类**：哲学、文学、历史、计算机
- 主题摘要：题材、人物、事件...
- 隐式特征：相似的图书在位置上比较近

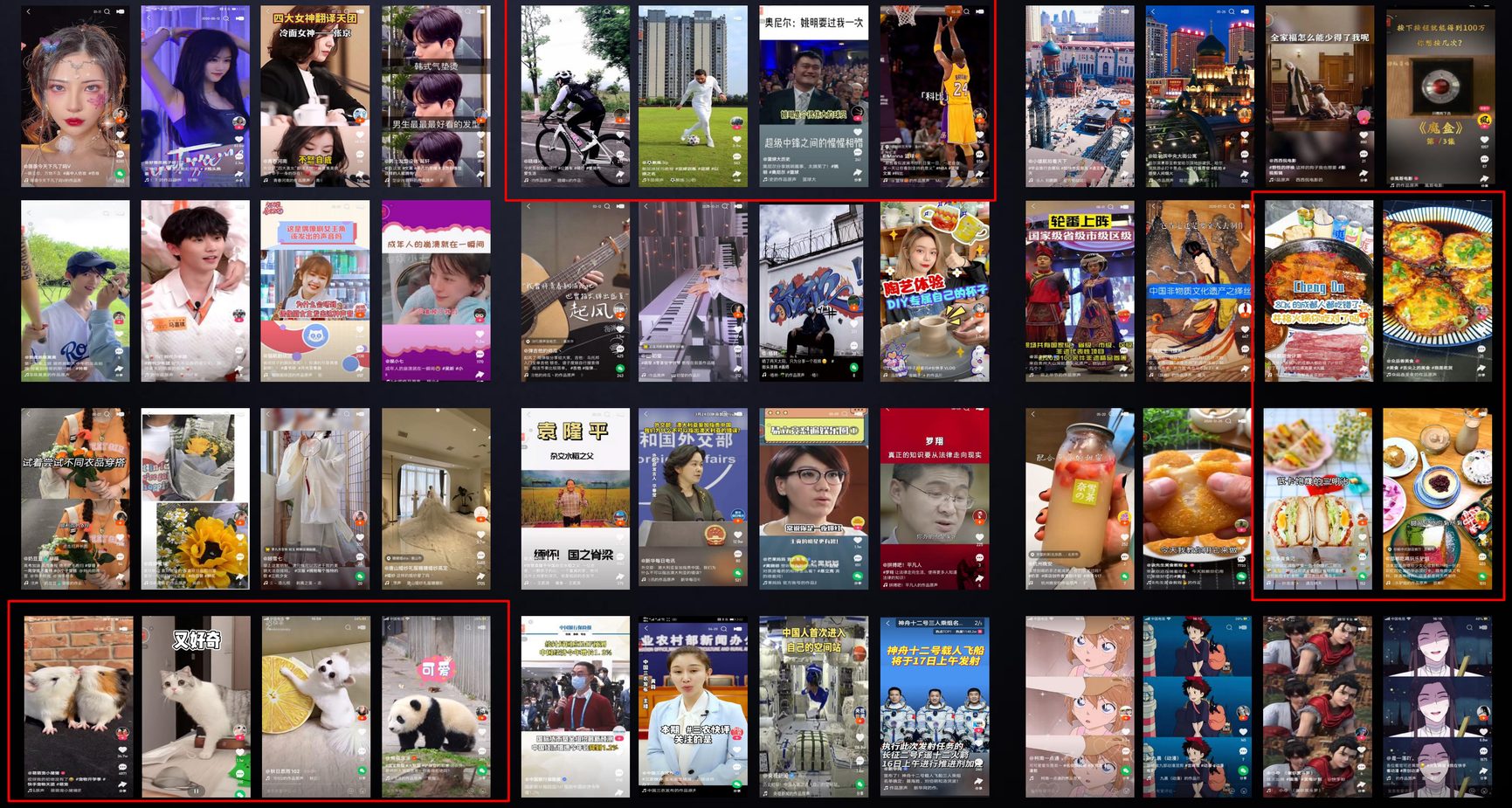
视频库



- **视频类目**：影视、颜值、资讯、搞笑
- 主题标签：形式、IP、场景、POI
- Embedding：相似视频在特征空间上比较近

智能视频分类与标签

通过河图结构化后，可以将视频按照多种维度归类和汇总，如萌宠、体育、美食等ź

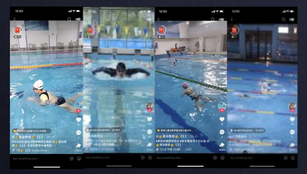
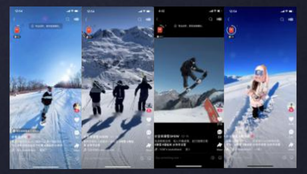
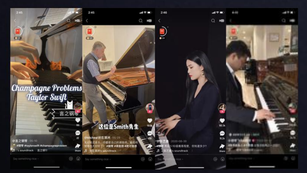
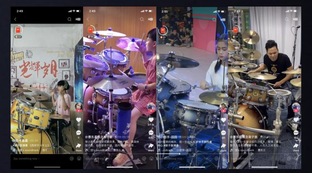
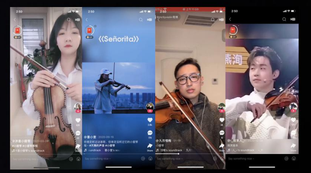
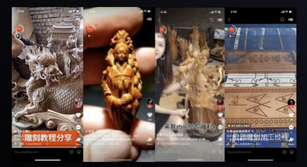
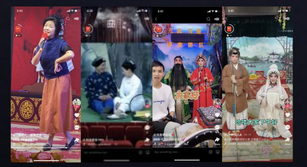
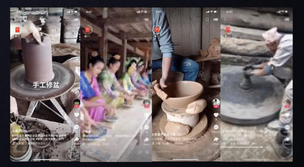
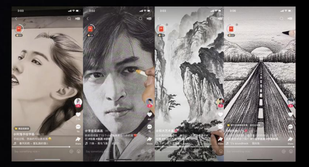
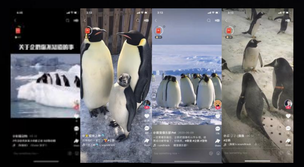
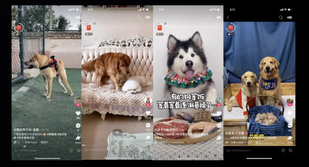


快手

智能视频分类与标签

一级

二级

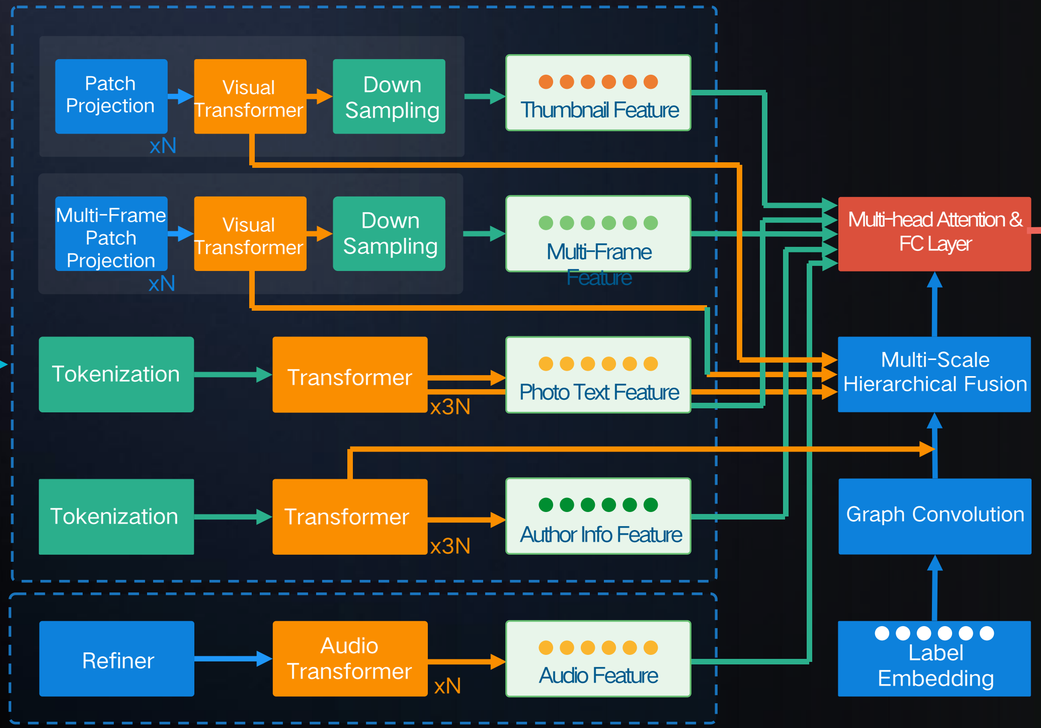
	游泳	滑雪	健身	…	篮球
运动				…	
	钢琴	吉他	架子鼓	…	小提琴
音乐				…	
	木雕	戏曲	制陶	…	绘画
才艺				…	
…	…	…	…		…
	企鹅	兔子	猫	…	狗
动物				…	

快手

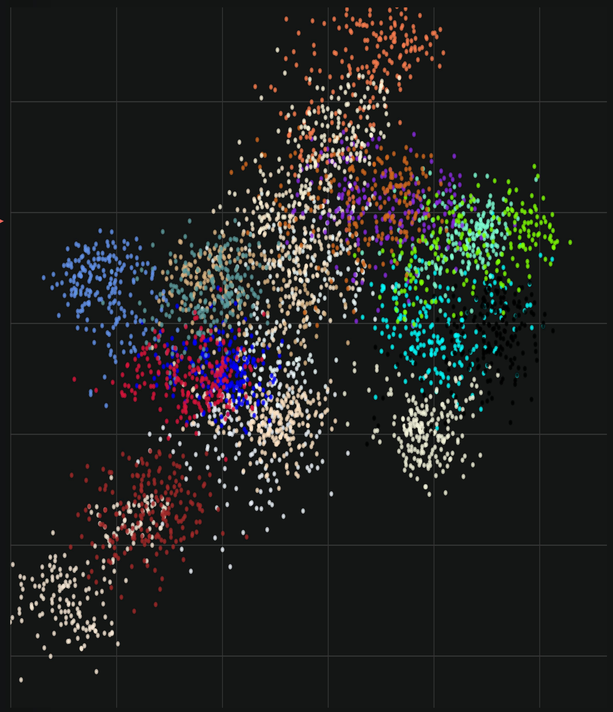
多模态Embedding

将数百亿量级的快手作品，浓缩到一个指定维度的特征空间。

丰富多彩的快手作品



多模态表征模型



Embedding分布的可视化结果

快手

多模态Embedding检索示例

快手

视频万物检索与识别

视频级检索->元素级检索：不同粒度的视频相似检索

- 视频级：视频主题或高层语义相似，为用户推荐更多内容相似的视频，提升短视频的推荐体验
- 元素级：视频实体元素相似，用户对视频细粒度实体感兴趣兴趣，如明星、商品、品牌、地标、IP等

问题核心：

- 如何判断视频主体元素、建立细粒度实体元素Embedding

技术挑战：

- 大规模弱监督模型训练：如何使用海量弱标注视频数据
- 视频细粒度表征学习：如何分辨细粒度实体差异
- 视频多模态语义表征：如何融合视觉、语音、文本信息

不同用户会产生不同兴趣点

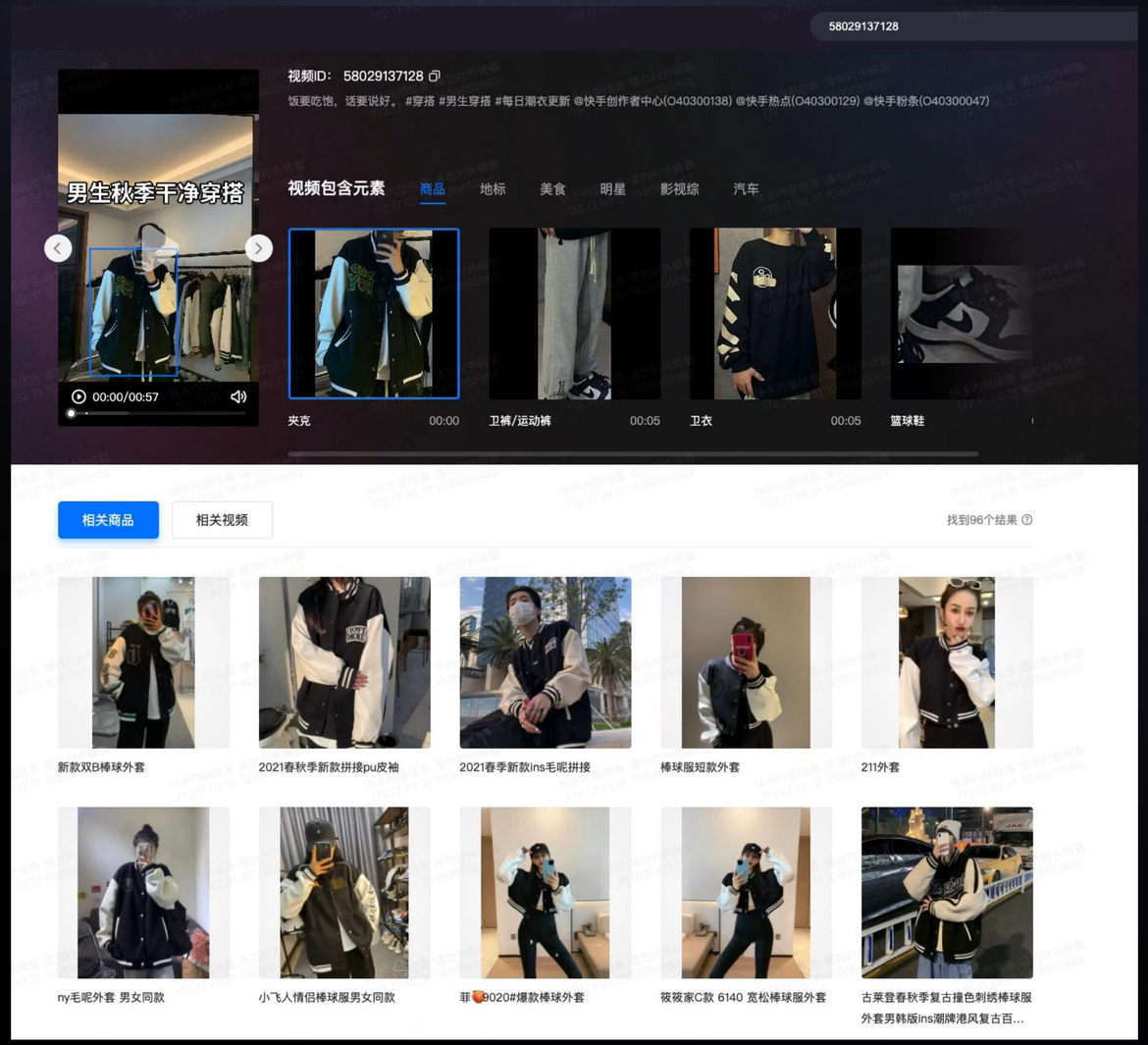
视频主题：
长袖蕾丝拼接连衣裙黑色

- 哪一款画作？
- 连衣裙款式？
- 沙发款式？
- 手包品牌？
- 是否为明星？
- 装修风格？
- 口红色号？

快手

视频万物检索与识别Demo

- 视频中海量实体元素的实时识别与检索，包括人物、商品、品牌、IP、运动、建筑、宠物、汽车等



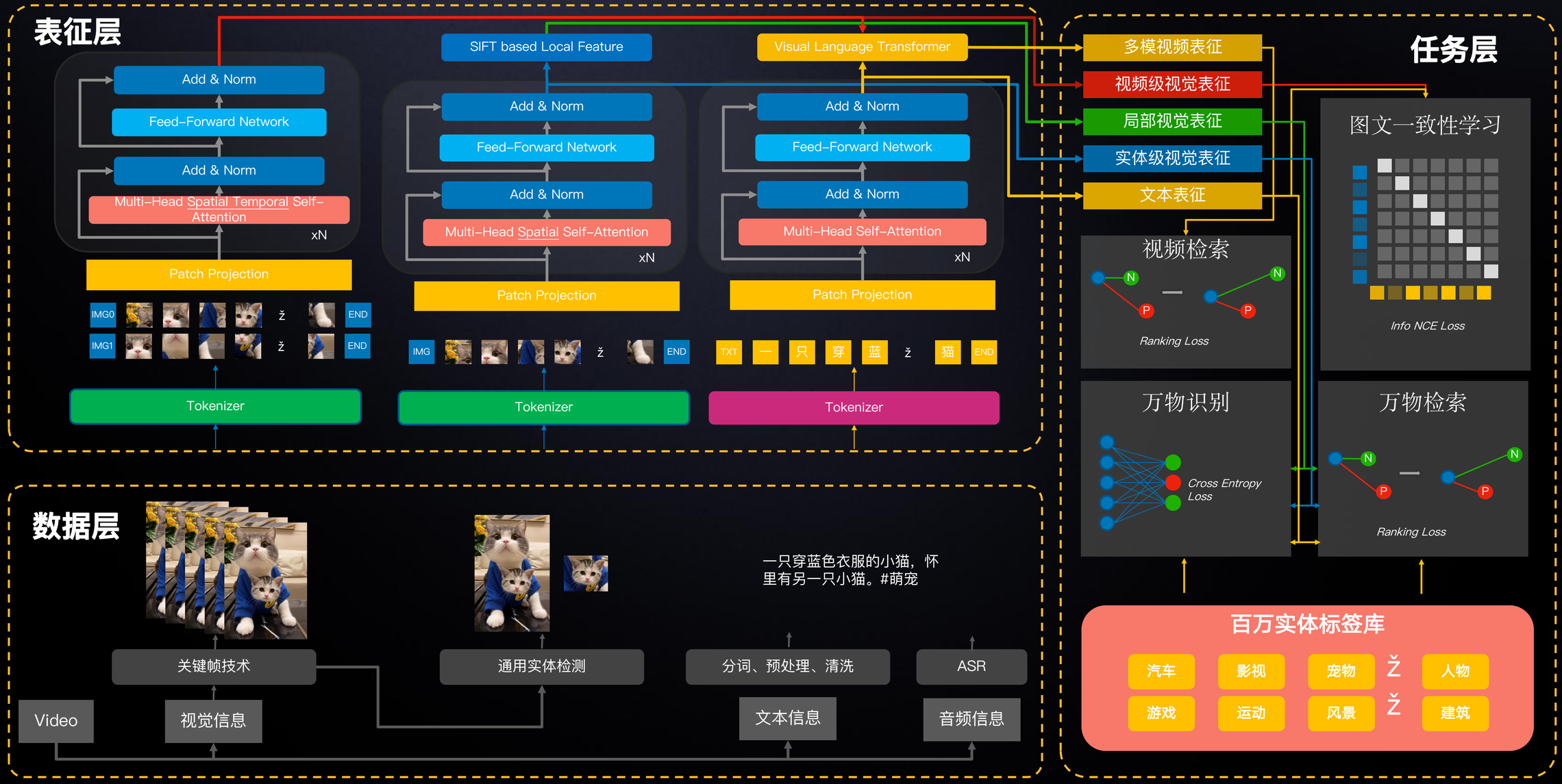
万物识别

万物检索

点击播放

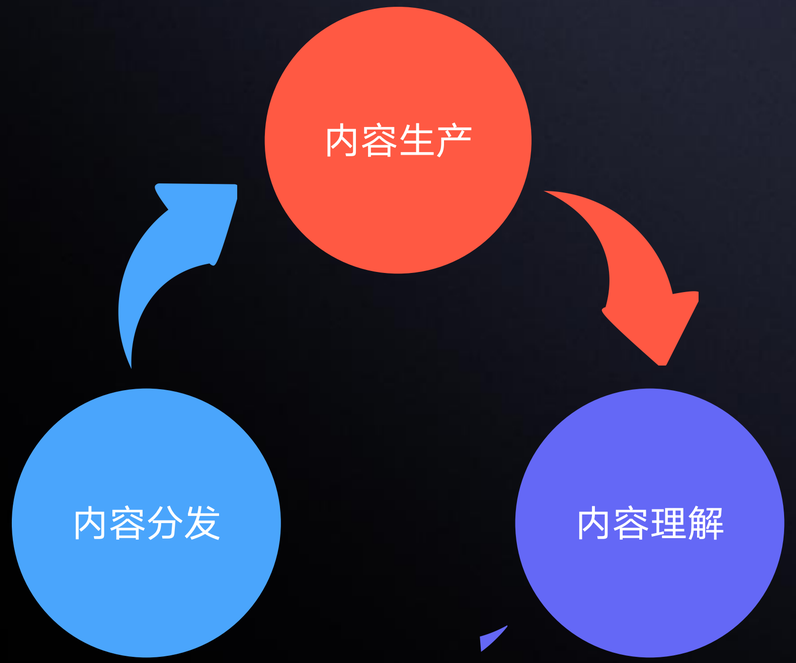
视频万物检索算法框架

基于视频多模态方法，构建视频细粒度元素检索系统，对视频中的主体元素进行检测、识别、检索。



快手

AI技术在快手的应用



- **内容生产：**在APP中提供炫酷的视觉特效、魔法表情、一键出片、自动字幕等AI工具和玩法，依赖AR引擎、人脸&手势识别、语音转译、智能创作等自助研发技术。

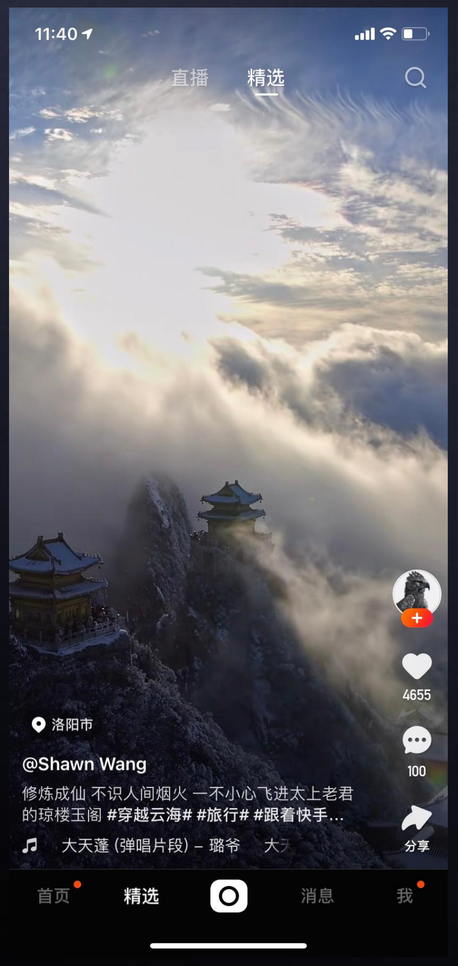
- **内容理解：**基于对社区中视频、图像、音乐、语言语义、主播和创作者的理解，充分结构化解释快手的内容生态，实现了社区海量内容的分类管理、原创保护、安全审核、助力分发等诸多应用。

- **内容分发：**推荐是用户与视频的双向匹配，将百亿视频特征和亿万用户特征输入推荐系统，实现精准、个性化的推荐。

快手推荐场景

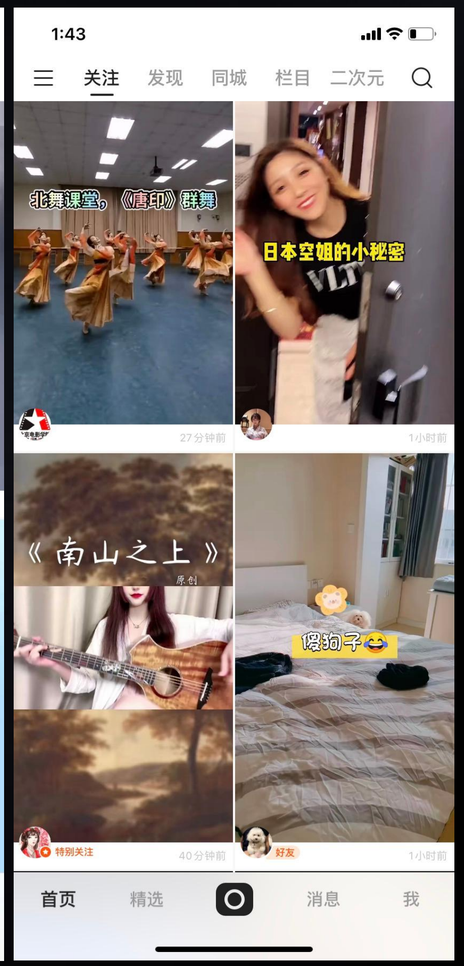
无处不在的推荐场景

- 单列：沉浸式
- 双列：选择权
- 关注页：半熟人半陌生人社区，私域流量
- 同城页：身边触手可及的生活

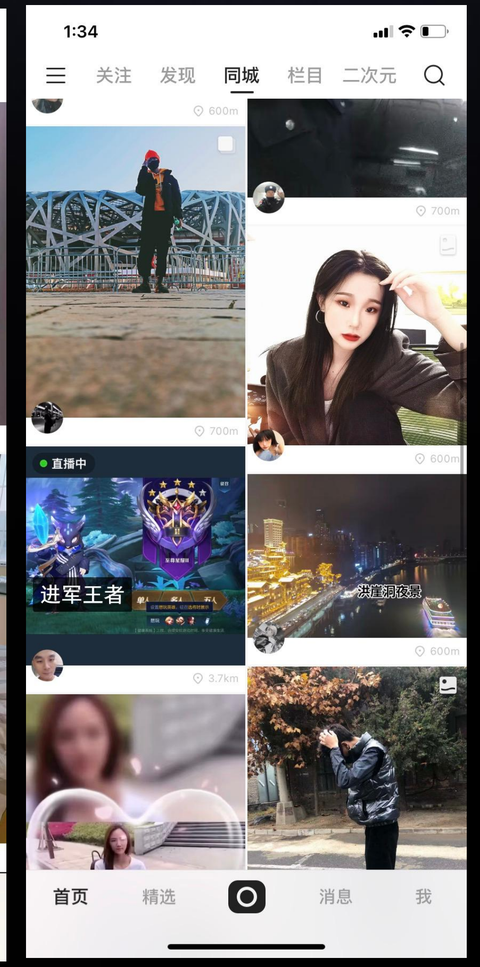


单列精选页

双列发现页



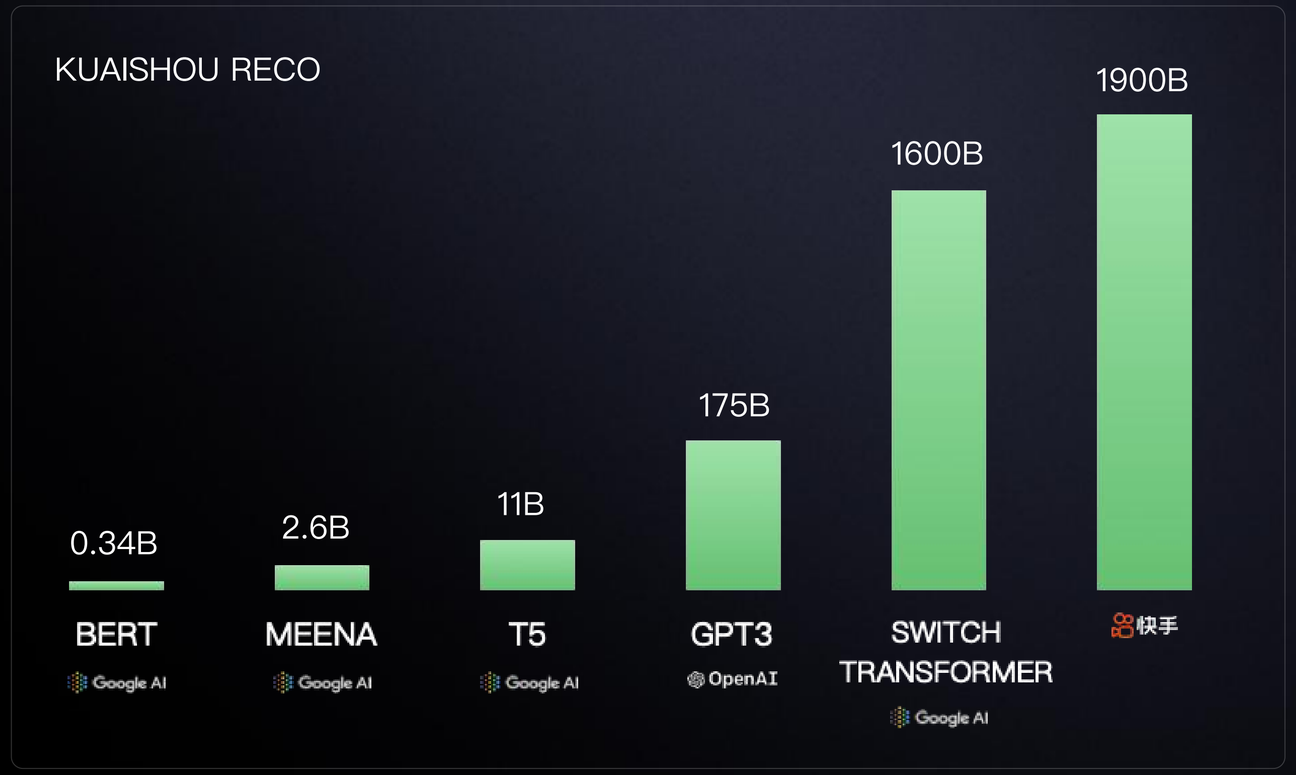
关注页



同城页

快手

快手的推荐系统规模



业界主流模型参数量(十亿)

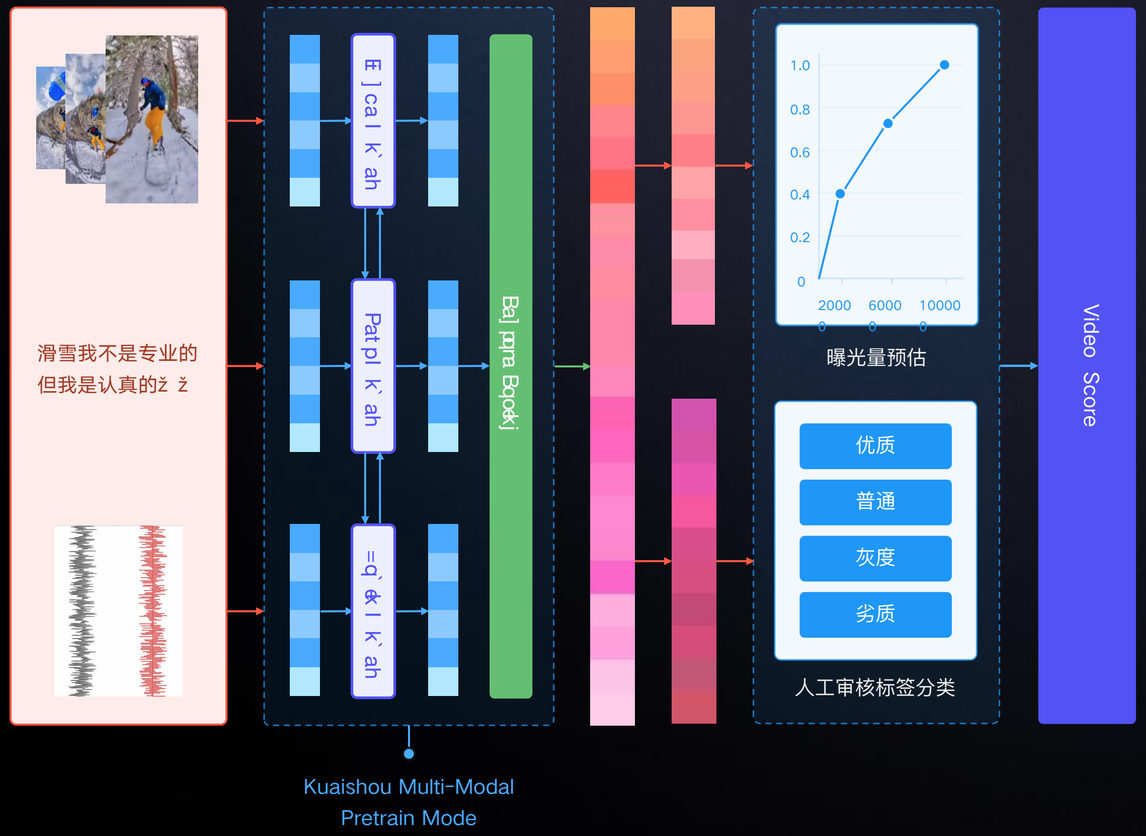
短视频行业有其独特的挑战，诸如用户量多，视频上传量大，作品生命周期短，用户兴趣变化快等等。

快手的推荐团队基于Transformer和MMoE模型，落地了业内首个万亿参数精排模型，对用户的长短期兴趣进行了精确的建模。快手精排模型的参数规模达到了1.9万亿，处理了超过万级的用户历史序列以及千亿的模型特征量。

在这个复杂的推荐系统落地过程中，内容理解能力发挥了不可或缺的作用。

快手

视频冷启动推荐



◀ 基于内容理解的冷启动模型

基于行为的冷启动：

给每个视频分配一定流量，通过线上用户行为反馈，预估视频内容质量；

短视频作品冷启阶段，行为稀疏，推荐模型预估结果不置信，导致流量浪费；

基于内容理解的视频质量估计：

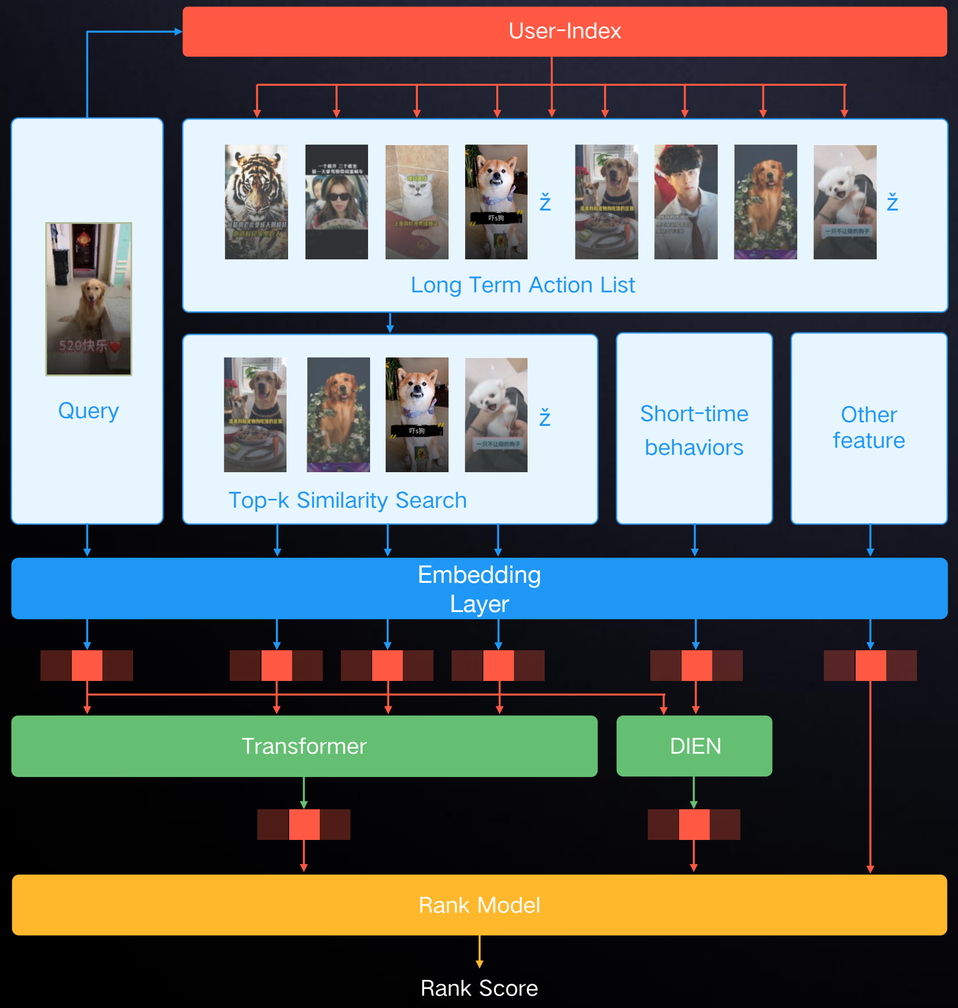
利用多模态预训练模型进行Fine-tuning；

预估视频在推荐系统中的成长潜力；

与视频内容质量正相关，帮助优质作品获得更多的曝光。

快手

推荐长期兴趣建模



相比电商、新闻等领域，短视频的内容更丰富，玩法更多样。因此用户对于短视频内容的兴趣也更广泛。 如果高效地利用好这些用户行为历史，以提升推荐效果是推荐模型长期以来难以解决的问题。

高质量的多模态Embedding为基于内容理解的推荐提供了新的可能性。

在基于内容理解的推荐场景中：
Embedding被用于从用户消费历史中精准检索与当前要预估作品最相关的作品；
进而得出用户对要预估作品的【感兴趣】程度；
通过长期兴趣的引入，结合短期行为特征，大幅提升了人均视频观看时长等核心指标。

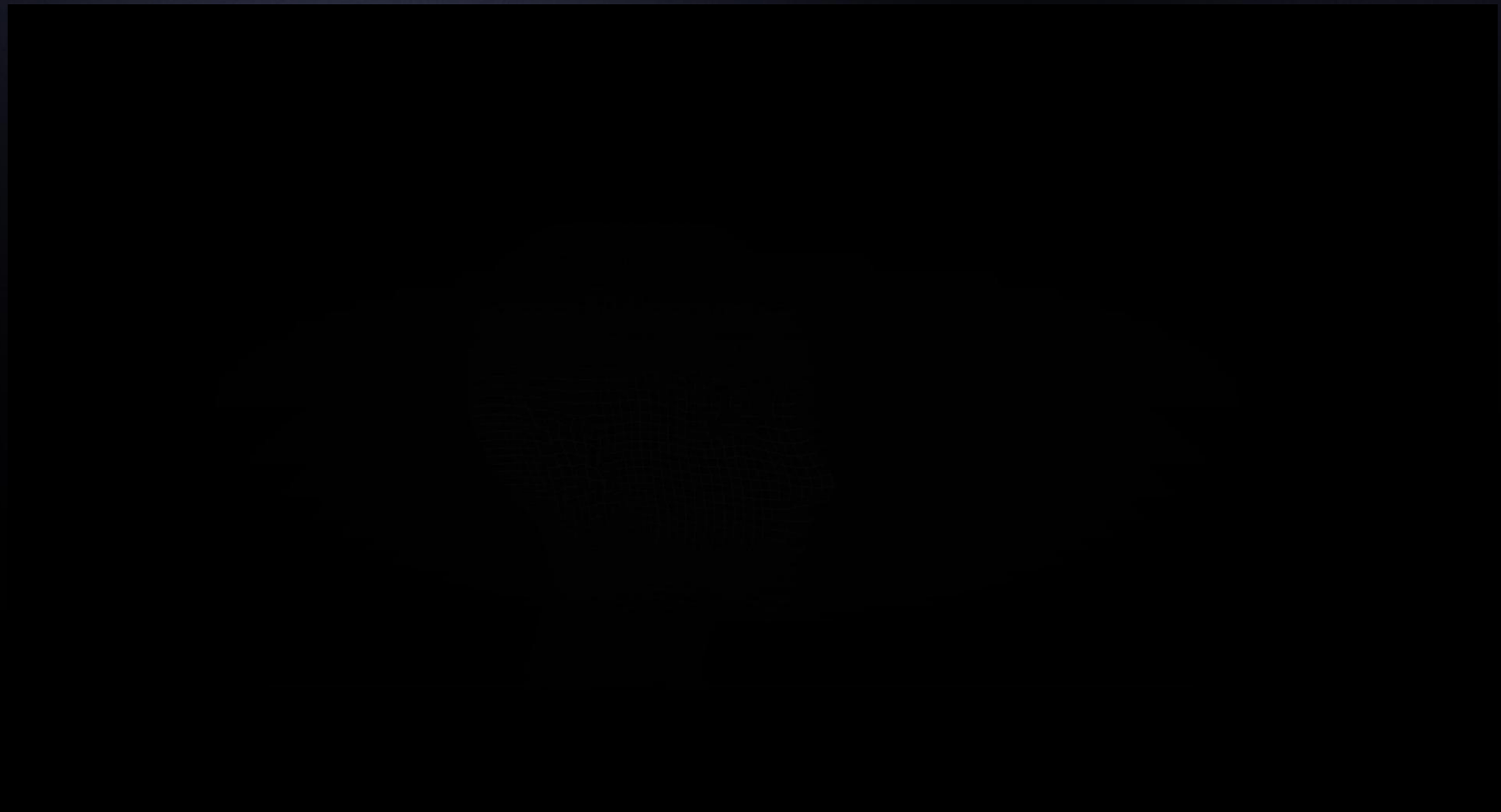
快手

快手AI开放平台：https://ai.kuaishou.com/

联系我们：ai-service@kuaishou.com

快手

彩蛋：快手虚拟人（敬请期待）



为一线互联网公司核心技术人员提供优质内容

- ☑ TGO专访
- ☑ 技术干货
- ☑ 每周精要
- ☑ 行业趋势

关注 InfoQ 公众号

InfoQ

促进软件开发及相关领域知识与创新的传播

THANKS