ICS 33.030
CCS M21

团　　体　　标　　准

T/CAAAD 015—2024　T/CCSA 073-2024

基于人工智能的营销视频自动生成服务技术要求

Technical requirements for AI-based marketing video auto-generation service

2024－12－01 发布　　2025－01－01 实施

中国广告协会　　中国通信标准化协会　　发布

版权声明

邮箱：IPR@ccsa.org.cn digitalad@china-caa.org

电话：010-62302847 010-65924878

目　次

前　言 II
1 范围 1
2 规范性引用文件 1
3 术语和定义 1
3.1 1
4 缩略语 1
5 技术框架 2
6 脚本要求 2
6.1　基本要求 2
6.2　基础信息 2
6.3　内容结构信息 3
7 素材要求 5
7.1　基本要求 5
7.2　台词类创作视频素材 5
7.3　公共视频素材 6
7.4　其他视频素材 6
8 基于 AI 的自动生成功能要求 6
8.1　通用要求 6
8. 2 粗剪模块 7
8. 3 精剪模块 8
8.4　管理要求 10
9 反馈调整 10

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。本文件由电信终端产业协会提出并归口。

本文件起草单位：北京快手科技有限公司、中国信息通信研究院、利欧集团数字科技有限公司、北京风行在线技术有限公司、北京沃东天骏信息技术有限公司、北京回旋加速网络科技有限公司、上海微盟科技集团有限公司、OPPO广东移动通信有限公司、瓴羊智能科技有限公司、联通在线信息科技有限公 司

本文件主要起草人： 陈权、马也、王波、田凯斌、李长城、李晗、江鹏、盖坤，谷晨、落红卫、 杨正军、周崧弢、潘冲、张泽华、龚涛、贾晟、付艳艳、姚栋、曹珣。

基于人工智能的营销视频自动生成服务技术要求

1　范围

本文件规定了基于人工智能的营销视频自动生成服务服务的技术框架、脚本要求、素材要求、基于 AI 的自动生成功能要求、反馈调整。

本文件适用于对基于人工智能的营销视频自动生成服务进行设计和研发，也可为第三方评估机构 对基于人工智能的营销视频自动生成服务活动进行评估提供参考。

2　规范性引用文件

本文件没有规范性引用文件。

3　术语和定义

GB/T 25069 界定的以及下列术语和定义适用于本文件。

3.1

脚本 footscript

通过文本形式定义视频的故事情节和结构。

注：脚本由创作者上传。标准化的脚本定义是支撑视频自动生成的基础。

3.2

镜头 shot

不需要切换机位、在时间上连续、组成完整短视频的若干段时间不等的片段。

注：实际拍摄过程中，通常先将剧本撰写为按“镜头”为单元的“分镜头本”，拍摄完后再进行拼接。

3.3

镜号 shot number

镜头（3.2）的顺序编号。

4　缩略语

下列缩略语适用于本文件。

AI	人工智能	Artificial Intelligence
APP	自动语音识别	Automatic Speech Recognition
BGM	背景音乐	BackGround Music
JPG	联合图像专家组	Joint Photographic Experts Group
MPV	MPEG 视频	MPEG Video
PNG	便携式网络图形	Portable Network Graphics
TTS	文本到语音	Text to Speech
T2V	文本到视频	Text to Video

url	统一资源定位	Uniform Resource Locator

5　技术框架

基于人工智能的营销视频自动生成服务系统的技术框架见图 1。

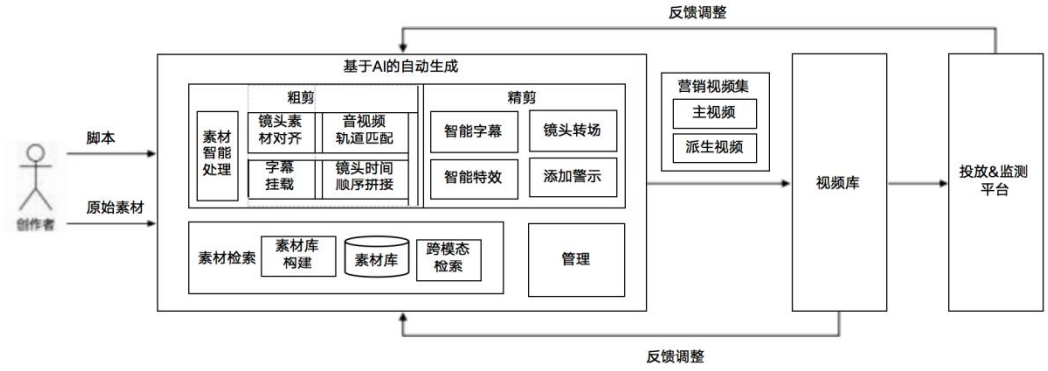

图 1 技术框架

基于人工智能的营销视频自动生成服务提供商（以下简称为“服务提供商”）向创作者提供服务， 通过创作者提供的结构化输入，基于人工智能技术自动生成营销视频集，在被创作者采纳后，进入营销 链路的投放和监测环节，根据营销效果对视频集的自动生成进行反馈优化。

结构化的输入包括脚本和素材。服务提供商提供基于 AI 的粗剪、精简、素材库构建及检索能力、以 及管理功能，输出包括主视频和派生视频的营销视频集，作为个性化营销投放的索引集。

注：在营销视频自动生成服务中，创作者通常是广告主，或受广告主委托的代理。

6　脚本要求

6.1　基本要求

脚本应包括基础信息和内容结构信息，并满足以下要求：

a)　基础信息要素应包括脚本名称、产品名称、关联行业、视频比例，宜支持基础信息要素的扩 展；

b)　内容结构信息应支持镜头信息和画中画信息；

c)　内容结构信息宜支持警示语信息；

d)　服务提供商宜向创作者提供标准化的脚本样例；

e)　服务提供商宜根据脚本基础信息，向创作者提供相应脚本内容结构信息的缺省值。

6.2　基础信息

6.2.1　脚本名称

应由创作者自定义，为字符串，宜有长度限制。

6.2.2　产品名称

应标识脚本描述的产品，或是与脚本关联的产品，为字符串，宜有长度限制。

6.2.3　关联行业

应描述脚本以及最终生成营销视频集所关联的行业，如服装配饰、教育培训等。

6.2.4　视频比例

应定义生成视频集的展现比例。

注：通常视频比例包括竖版视频 9:16，横版视频 16:9。

6. 3　内容结构信息

6.3.1　镜头信息

6.3.1.1　概述

应包括营销视频主体镜头，宜包括片头镜头和片尾镜头。营销视频主体镜头可包括 1 个或多个镜头。 镜头信息元素包括镜号、素材类型、台词、画面描述、音乐风格、人声朗读、标题、结构标签、参考时 长、景别、运镜技巧、分支主题、剧情描述等。

6.3.1.2　镜号

所有镜头都应包括镜号。镜号从 1 开始，应满足以下要求：

a)　每新增一个镜头，镜号增加 1；

b)　应支持镜号的切换，即将一个镜号向前移动或向后移动；

c)　应支持镜号的删除，删除后后续镜号依次减少 1。

6.3.1.3　视频素材类型

所有镜头都应包括视频素材类型。

视频素材类型包括台词类创作视频素材、非台词类创作视频素材、平台公共视频素材，宜支持扩展类型，并满足以下要求：

a)　当素材用作首帧镜头或尾帧镜头时，应明确标识 。

b)　不同的视频素材类型，可具有不同的镜头信息元素。

注 1：台词类创作素材，通常指创作者自主拍摄的真人出镜视频素材，其中有音频台词，脚本中有 对应的文本台词，且根据音频台词和文本台词的对应关系在粗剪阶段进行对齐处理。

注 2：非台词类创作素材，通常指创作者自主提供的视频素材，其中没有音频台词或文本台词，如 无口播真人实拍、不含人物的视频等。

注 3: 平台公共视频素材指平台提供的素材库中的物料素材。

c)　应支持通过输入的描述、产品名称、行业等信息，筛选公共素材，并按照算法模型进行推 荐。应支持公共素材的预览，支持按时长筛选。

6.3.1.4　台词

台词用于挂载字幕，以及定位画中画的位置，应满足以下要求：

a)　台词宜按照人物角色、语义等进行拆分；

b)　对于台词类创作素材，应包括 1 个或多个台词信息。台词自动与视频素材中的音频匹配，实 现自动剪辑和对齐原始视频片段；

c)　对于其他类型的素材，可包括 0 个、1 个或多个台词信息。应基于台词自动合成语音，作为视频的配音音频。

6.3.1.5　画面描述

画面描述应满足以下要求：

a) 对于各种类型的素材，应支持 0 个或 1 个画面描述；

b) 当应用于公共类型的素材时，画面描述用于检索相应的视频片段，应满足 8.4 的要求。

6.3.1.6 音乐风格

对于各种类型的素材，应支持 0 个或 1 个 BGM 音乐风格，并满足以下要求：

a) 应提供音乐风格选项，如纯音乐、流行舞曲、民歌、经典等， 自动匹配背景音乐；

b) 应支持无背景音乐的选项；

c) 应支持当用户不选择音乐时，自动智能匹配音乐。

6.3.1.7 人声朗读

人声朗读基于 AI 技术，根据台词生成配音音频，应满足以下要求：

a) 人声朗读应区分性别、年龄、情绪、方言等；

b) 对于台词类创作素材，通常不包括人声朗读；

c) 对于其他类型的素材，应支持 0 个或 1 个人声朗读；

d) 应提供无人声朗读的选项。

6.3.1.8 标题

标题应满足以下要求：

a) 对于首帧镜头，可设置标题；

b) 当设置标题时，应支持不同的标题位置，如位于屏幕上方、下方。

c) 应支持以字符形式输入标题内容，应有字符串长度限制。

d) 宜支持主标题和副标题。

6.3.1.9 结构标签

对于各种类型的素材，应支持 0 个或 1 个结构标签。

注：结构标签通常以文本形式说明当前镜头的主旨，描述镜头在全局中的作用。

6.3.1.10 参考时长

对于各种类型的素材，应支持 0 个或 1 个参考时长。 注：参考时长指明镜头的时间长度。

6.3.1.11 景别

对于各种类型的素材，应支持 0 个或 1 个景别。

注：景别通常可包括大景、全景、中景、近景、特写、显微等。

6.3.1.12 运镜技巧

对于各种类型的素材，应支持 0 个、1 个或多个运镜技巧。

注：运镜技巧通常包括镜头的运用以及镜头的组合。镜头的运用如固定镜头、推、拉、摇、移、 跟 等。镜头的组合如淡入淡出、切换、叠化等。

6.3.1.13 剧情描述

对于各种类型的素材，应支持 0 个或 1 个剧情描述。

注：剧情描述通常以文本形式，详细说明镜头画面里的场景的变换和内容，简单的构图设计等。

6.3.2　画中画

6.3.2.1　基本要求

画中画在原有视频（通常称为“主视频”） 基础上，叠加新的视频画面，更好的呈现视频效果，应满足以下要求：

a)　应以台词为粒度添加画中画，支持跨镜号设置。

b)　当存在画中画时，音频应以主视频为主，必要时，画中画素材中的音频被去除。

c)　画中画元素应包括贴片方式、镜头信息。

d)　当镜号删除或切换时，相应的画中画内容应被清空。 6.3.2.2　贴片方式

应支持多种类型的贴片方式，如上下分屏、左右分屏、右上角、右下角、全覆盖-在前，全覆盖-在后。

6.3.2.3　镜头

镜头应满足以下要求：

a)　应满足 6.3.1 所述的镜头要求。

b)　当主视频为台词类创作素材时，应根据主视频的台词进行画中画匹配。

c)　当主视频为尾帧素材时，宜根据产品名称等信息进行画中画匹配。

d)　当主视频为其他类型素材时，宜根据画面描述等信息进行画中画匹配。

6.3.3　警示语

警示语应满足以下要求：

a)　警示语元素应包括位置及内容。宜支持警示语库。

b)　应支持不同的警示语位置，如右上方、右下方等。

c)　应支持以字符形式输入警示语内容，应有字符串长度限制。

d)　警示语库应支持服务商提供和创作者自行录入，典型的警示语如未成年人禁止饮酒、本广告仅供参考、加盟有风险投资须谨慎等。

7　素材要求

7.1　基本要求

创作者上传素材，用于生成营销视频。素材应满足以下要求：

a)　一个镜头应对应一个或多个素材；

b)　素材应包括视频素材、应用 logo、背景图素材、背景视频素材等；

c)　应支持多种素材上传方式，包括但不限于：用户自行上传素材，服务商提供素材库等。

d)　上传形成的素材库应支持按照产品名称、素材类型、创建时间等进行筛选。

应支持上传的素材应满足以下要求：

a)　应支持 JPG、PNG、MP4、MPV 格式的素材格式；

b)　应限制单个素材不应过大；

注：实际中通常不超过 1G。

c)　应限制单次上传素材的数量。

注：实际中通常不超过 50 个。

7. 2　台词类创作视频素材

创作者上传台词类创作视频素材时无需考虑实际顺序。上传的台词类创作视频素材应满足以下要求：

a)　有明显的、可被系统识别的开始标记；

注：常用的开始标记方式，如在视频素材内容正式开始时，清晰喊出开始标志词“开始！”；

b)　有明显的、可被系统识别的结束标记。

注：常用的结束标记方式，如在视频素材内容全部结束时，清晰喊出结束标志词“结束！”；

c)　多人拍摄时，收音应尽量适中和均衡，如尽量避免一个过远一个过近；

d)　应支持在有口误的情况下，短暂停顿后，从这句话的开头重新口播。

7. 3　公共视频素材

根据镜头信息中的画面描述检索公共视频素材，应满足以下要求：

a)　支持 T2V 跨模态检索；

b)　支持时长筛选；

c)　支持基于不同属性的检索；

注：如人体属性（如长发、短发、男、女）、衣服属性（衣服类型、衣服颜色）、人物姿态（走 路、跳舞、坐着等）、场景（湖边、路上）

d)　支持不同属性的组合检索；

注：如黑色连衣裙女生在湖边；短发女生走路；运动鞋开箱；湖边小镇

7.4　其他视频素材

当非台词类创作视频素材的镜头中含有台词时，应通过 TTS 根据台词产生音频，并支持选择是否去除上传视频素材的原声音频。

8　基于 A I 的自动生成功能要求

8.1　通用要求

8.1.1　数据

对人工智能算法处理的数据，应满足以下要求：

a)　应对训练数据进行预处理，包括识别并删除恶意样本、识别并修复或过滤被污染数据；

b)　宜对训练数据集采取必要的保护措施，确保数据的保密性、完整性；

c)　宜构建不同于训练数据集的标准测试数据集，模型上线前应通过标准测试数据集的测试；

d)　应构建合理的训练数据集。

注：通常考虑数据集的规模、均衡性、准确性等。

8.1.2　模型

应用于营销视频自动生成服务的人工智能模型应满足以下要求：

a)　应定期针对模型机制机理、应用结果进行审核、评估，并对模型训练和模型推理进行动态改 进和调整；

b)　应支持模型训练优化，如模型组合、参数调整、实验调优等。

8.1.3　素材库构建

应考虑不同行业的特点，构建匹配不同行业特征的素材库。

素材可包括但不限于：视频片段、台词文案、警示文案、背景音乐、特效音乐、特效贴纸、特效视 频、LOGO 标签、图片、插画、字幕库等。

注：特效贴纸通常支持静态 png 格式，或动态 gif 格式。

素材库应以结构化的方式记录素材的信息，包括但不限于：

a) 素材的类型，必要时可对类型进行细分，记录二级类型。如：文本/关键词、图片/贴纸、饮 品/特效音等

b) 素材的来源，如创作者上传、广告库等；

c) 素材的标签，如某带特效音的贴纸特效素材，其标签可能为“类型为特效，行业为社交通 讯，贴纸地址为某 url，音效地址为某 url”；

d) 素材的文本内容，如“**神器 ”；

e) 素材的状态，如“有效、失效 ”；

f) 素材的上传者联系方式，如邮箱。

8. 2 粗剪模块

8.2.1 概述

粗剪模块包括镜号对齐、智能去口误、素材智能增强、视频混剪、以及画中画处理等。

镜号对齐指将素材与脚本镜号进行匹配对齐。将用户输入的脚本和素材，转化为结构化信息，确定脚本中每个镜号的台词，对应哪一个素材的哪个时间段，以及需要的操作等。

8.2.2 台词类创作视频素材镜号对齐

8.2.2.1 单机位

在单机位场景下，应按照如下流程完成镜号对齐：

a) 对视频素材的音频台词进行 ASR；

b) 对齐镜头中的台词与 ASR 识别结果，应支持模糊匹配；

c) 依据对齐的结果，获取镜头中台词对应的素材视频时间戳；

d) 串联组合成片。

8.2.2.2 多机位交叉

对于多机位交叉场景，同一个镜头对应多个不同机位角度的视频素材，应按照如下流程完成镜号对齐：

a) 对每个视频素材（片段）的音频台词进行 ASR；

b) 对齐镜头中的台词，与每个视频素材的 ASR 结果，识别出台词的每句话分别对应哪些视频素材，作为对应的后哦选素材；

c) 从候选素材中选择合适角度的机位素材；

d) 注：常见的选择依据如是否有人正对着镜头说话、镜头的远近、说话的时长等

e) 依据选择的结果，获取镜头中台词对应的机位素材视频时间戳；

f) 串联组合成片。

8.2.3 其他类型视频素材镜号对齐

其他类型视频素材镜号对齐应满足以下要求：

a) 当脚本存在台词，或存在画面描述时，宜基于跨模态检索模型识别哪些素材的视觉或音频匹配程度较高，作为对应的候选素材；宜支持分析语义信息进行优化匹配。

b) 可基于素材时长进行镜号对齐。

c) 当存在多种可能性时，可支持以随机方式或流量优选方式进行选择。

8.2.4 智能去口误

对于台词类创作视频素材，应支持根据台词内容，智能去除实拍中的口误片段，包括但不限于支持忘词、重复、口误等情况的去除。

8.2.5　素材智能增强

素材智能增强应满足以下要求：

a)　包括色彩增强、画面去抖、音频降噪等；

b)　色彩增强应支持对用户上传的原始素材进行对比度、亮度、饱和度等的自动调整，提升视觉 观感的统一；

注：调色时宜遵循画面原本色彩倾向的色系，如室外调色偏向冷色调，室内调色偏向暖色调。

c)　音频降噪应支持对用户上传的原始素材进行背景杂音过滤等，提升有效人声的清晰程度；

d)　素材智能增强的处理不应改变原始素材的时长。

8.2.6　视频混剪

根据创作者输入的画面描述等文本信息，从公共素材库中检索出合适的候选素材。视频混剪应满足以下要求：

a)　基于素材的视觉表征、文本表征、结构化标签等信息，构建索引库；

b)　基于创作者输入的画面描述等文本信息，以及索引库中的视频素材信息，进行文本-视频的 跨模态检索；

c)　对检索结果进行过滤、处理，形成粗剪视频集中的镜头。

8.2.7　画中画

通过自动智能化的镜头组合，将画中画视频与主视频在时间轨道上对齐。画中画应满足以下要求：

a)　应支持内容匹配和时间匹配；

b)　内容匹配应能适配不同的画中画素材类型，满足以下要求：

1）　对于公共类型的画中画素材，根据画面描述检索素材；

2）　对于实拍类型的画中画素材，根据台词抽取画中画素材的对应片段；

3）　对于录屏类型的画中画素材，根据主视频画面内容做自动匹配。

c)　时间轨道匹配应根据画中画对应的镜头信息中的台词，计算主视频轨道上的时间戳，插入对应的画中画副视频。

8. 3　精剪模块

8.3.1　概述

基于粗剪模块的时间轴，进行精细化的处理，包括智能字幕、智能特效、警示语等。通过精剪模块 进行视频的组合派生，生成个性化投放的候选集合空间。

8.3.2　智能字幕

8.3.2.1　基本要求

应支持将台词以字幕的形式正确挂载到素材视频、高显关键词、敏感词替换、智能样式匹配。

8.3.2.2　字幕挂载

字幕挂载根据对视频素材的 ASR 识别结果的时间戳，将对应台词挂载到对应位置，应满足以下要求：

a)　对台词的标点符号进行处理；

注：如一般可将顿号变为空格，将顿号之外的标点符号去除。

b) 应设置合理的屏占比，确保字幕不应超过屏幕的合理区域；

c) 必要时对台词进行语义分词，在不拆分完整词语的前提下，将字幕拆成多行或放在多个画面中显示；

d) 应调整字幕在画面中的位置，保证字幕中心位置对齐和避免重叠；

e) 对于片头字幕，宜支持人脸规避。

8.3.2.3 高显关键词

高显关键词应满足以下要求：

a) 根据不同行业识别关键词；

b) 关键词应高亮显示，如使用更大的字号、使用花边等特殊字体样式、使用特效字体等；

c) 关键词词库宜支持更新和扩展。

8.3.2.4 敏感词处理

敏感词处理应根据法律法规及平台规则，识别出台词中的敏感词，并进行相应处理，如敏感词屏蔽、替换等。

8.3.2.5 智能样式匹配

智能样式匹配应满足以下要求：

a) 应支持不同样式的字幕；

b) 宜支持根据不同画面和场景，适配不同样式的字幕；

c) 应支持对于同一个脚本派生的不同视频，适配不同的样式。

8.3.3 智能特效

8.3.3.1 概述

智能特性可包括针对镜头的视频特效、针对镜头的音频特效、插入引导信息（如引导视频、尾帧视 频）、自动插入转场特效、镜头融合、贴纸特效、产品 logo 等。

8.3.3.2 贴纸特效

贴纸特效根据视频内容、脚本信息等识别出需要添加特效贴纸的时间和位置，检索相应的特效贴纸并完成添加，应满足以下要求：

a) 支持贴纸特效应与字幕关键词绑定，对应弹出特效贴纸和特效音；

b) 保证合理性，通常根据视频画面作为判断依据； 注：如爱心类贴纸通常匹配男女爱情。

c) 控制贴纸弹出的频率。

8.3.3.3 转场特效

转场特效应在不同镜头和视频片段切换的时候，根据前后画面的内容，自动识别需要添加视频转场 的位置，选择转场类型，添加对应转场视频特效。

8.3.3.4 logo 添加

应支持创作者配置是否添加 logo。

当创作者选择添加 logo 时，检索相应 logo 图片， 自动调整大小并添加到合适的位置。

8.3.3.5 引导视频添加

宜根据实时事件或节假日，检索匹配具有时效性的文案或者视频，作为头部帧。

注：通常用于实效性节假日等。

8.3.4　警示语添加

应支持创作者配置是否添加警示语。

当创作者选择添加警示语时，应满足以下要求：

a)　应支持创作者输入警示语，　自动从中选择合适的警示语并添加到合适的位置；

b)　宜支持自动从警示语库选择合适的警示语病添加到合适的位置。

8.4　管理要求

8.4.1　合成记录管理

合成记录管理应满足以下内容：

a)　应支持将合成视频关联到相应的脚本；

b)　应支持以下不同方式的筛选：

1）　基于脚本名称筛选：支持按照脚本名称和产品名称搜索生成记录

2）　基于产品名称筛选

3）　基于合成状态筛选：合成中、合成成功、合成失败

4）　基于时间筛选：按照合成的时间范围筛选

c)　应支持合成视频相关信息的展示，包括主视频和派生视频的视频集合，用于个性化曝光；

d)　对于合成失败的视频，应展示失败原因；

e)　对于派生视频，应支持替换音频、替换公共素材；

f)　应支持合成视频的推送；

g)　宜支持合成视频的导出。

8.4.2　脚本管理

脚本管理应满足以下要求：

a)　应支持将创建过的脚本关联到相应的合成记录。

b)　应支持按照脚本名称、产品名、创建时间进行历史脚本的查询。

9　反馈调整

应设计合理的评价体系，观测评价指标，反馈改进人工智能模型，满足以下要求：

a)　应以用户对生成的营销视频集的采纳率，以及投放后的效率为指标，进行反馈和优化；

b)　应在模型上线部署前进行评估，并在上线后定期进行评估反馈，必要时触发模型的优化；

c)　宜支持在线训练和优化。