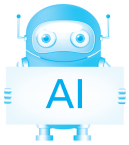
第 5 章

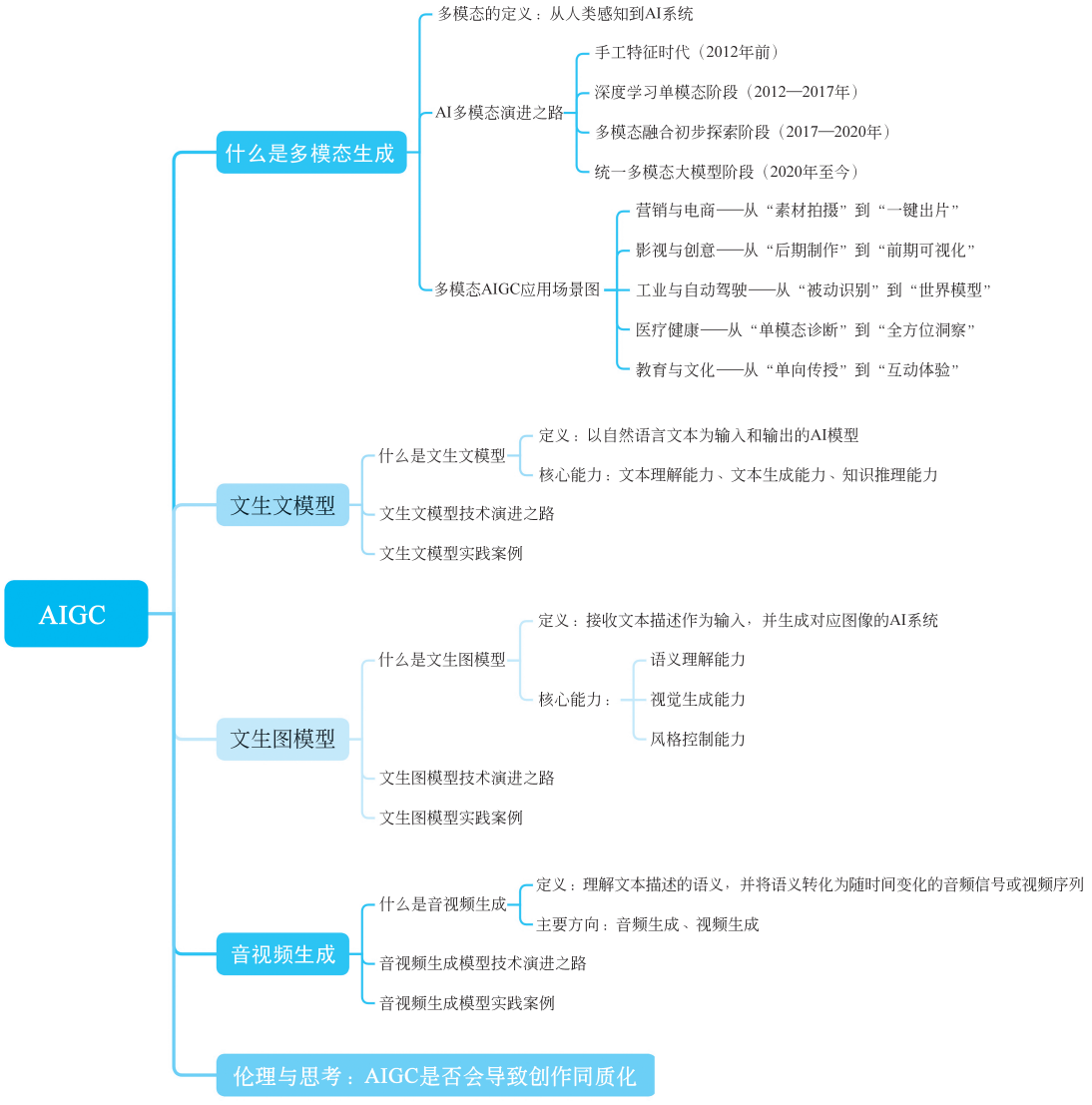


AIGC：多模态创作与跨学科实践

学习目标

- 理解多模态生成的基本概念与技术原理。
- 掌握文生文、文生图、AI 作曲、视频生成等核心 AIGC 工具的使用方法。
- 能够实施跨模态创作项目。
- 批判性思考 AIGC 对创作同质化的影响。
- 认识人类在 AI 时代的不可替代性。
- 应对 AIGC 时代的职业与发展挑战。

思维导图



5.1 什么是多模态生成——文、图、音、视频的联动

5.1.1 多模态的定义：从人类感知到 AI 系统

人类感知世界从来不是单一感官的体验。当漫步在春日的花园，看到花朵的颜色、形状，蜜蜂的飞舞轨迹；听到鸟鸣、风吹树叶的沙沙声、远处的车流声；闻到花香、青草的气息；感受到阳光的温度、微风的轻拂；脑海中浮现“春暖花开”的诗句。这些感官信息在大脑中整合，形成对“春天”这个概念的完整理解。这种多感官整合能力，是人类智能的核心特征之一。这种多模态感知能力，使得人类能够：

（1）信息互补：当一种感官受到限制时，其他感官可以补偿（如黑暗中依靠听觉和

触觉）。

（2）信息验证：多种感官的信息相互印证，提高理解的可靠性（如闪电与雷声相互验证）。

（3）信息增强：多模态信息叠加，产生超越单一模态的理解（如电影中的视觉、听觉与字幕）。

理解了人类的多模态感知后，再来定义“模态”（modality）的概念。模态是指信息的某种特定表现形式或感知通道，如表 5.1 所示。每种模态都有其独特的数据结构（不同的表示方式和存储格式）、统计特性（不同的数据分布和统计规律）、语义密度（不同的信息承载方式和密度）。

表 5.1　常见模态类型及其数据结构特征

模 态 类 型	举　　例	数据结构特征
视觉模态	图像、视频	像素矩阵、张量
听觉模态	语音、音乐、环境音	时序波形、频谱
文本模态	书籍、文章	符号序列、词向量
触觉模态	纹理、温度、压力	触感数据、力反馈
其他模态	嗅觉、味觉、传感器数据	多维向量、时序数据

多模态（Multimodal）是指同时涉及多种模态的信息处理方式。例如：

（1）观看电影：视觉（画面）＋ 听觉（对白/音效）＋ 文本（字幕）。

（2）视频通话：视觉（对方形象）＋ 听觉（语音）＋ 文本（聊天记录）。

（3）自动驾驶：视觉（摄像头）＋ 激光雷达（距离）＋毫米波雷达（速度）＋ GPS（位置）。

这些不同模态的信息相互补充，相互增强，共同构成了对场景的完整理解。

AI 多模态（AI multimodal）是指能够同时理解、处理、融合和生成多种模态数据的人工智能系统，其核心目标是像人类一样通过多感官信息的整合来形成对世界的综合理解，让机器能够像人类一样同时处理多种类型的信息，在不同模态之间建立语义关联，并通过多模态信息的互补增强理解的鲁棒性。AI 多模态的核心能力可以分解为以下 4 个维度。

（1）多模态感知：同时接收来自多个异构模态的输入数据。

（2）多模态理解：学习并建立不同模态之间的语义关联，如图文匹配（判断图像和文本是否语义相关）、视觉问答（根据图像回答文本问题）、情感分析（结合面部表情和语音语调判断情感）。

（3）多模态融合：有效整合多模态信息，以增强理解和决策能力，如视频分类（同时使用视觉帧和音频流），情感识别（综合面部表情、语音语调和文本内容）。

（4）多模态生成：AI 模型能够跨越不同媒体形式，理解并生成文本、图像、音频、视频等多种模态的内容，如文生图（从文本描述生成对应图像）、图生文（为图像生成描述性文本）、音生视频（从音频生成匹配的视频内容）等，这种模态间的自由转换极大地降低了创作的门槛，使得创意可以更自由地流动。

如果考虑文本（text）、图像（image）、音频（audio）、视频（video）4 种主要模态，理论上存在着 16 种可能的单向转换关系，如图 5.1 所示。

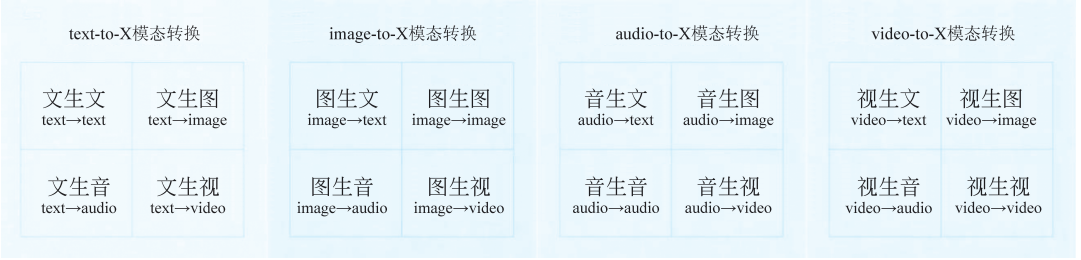


图 5.1　四种主要模态的转换关系

（1）文生类(text-to-X)：包含文生文、文生图、文生音(AI 作曲)、文生视。

（2）图生类(image-to-X)：包含图生文、图生图(风格迁移)、图生音(根据画面生成环境音)、图生视(静态图转动态视频)。

（3）音生类(audio-to-X)：包含音生文(语音转写)、音生图(音乐可视化)、音生音(变声器/声音克隆)、音生视。

（4）视生类(video-to-X) 包含视生文(视频摘要)、视生图(视频抽帧/关键帧提取)、视生音(视频自动配音)、视生视(视频风格化/动作捕捉迁移)。

5.1.2

5.1.2　AI 多模态演进之路

AI 多模态的发展经历了 4 个重要阶段，反映了 AI 从“单一感官”到“全知全能”的演进历程，如图 5.2 所示。

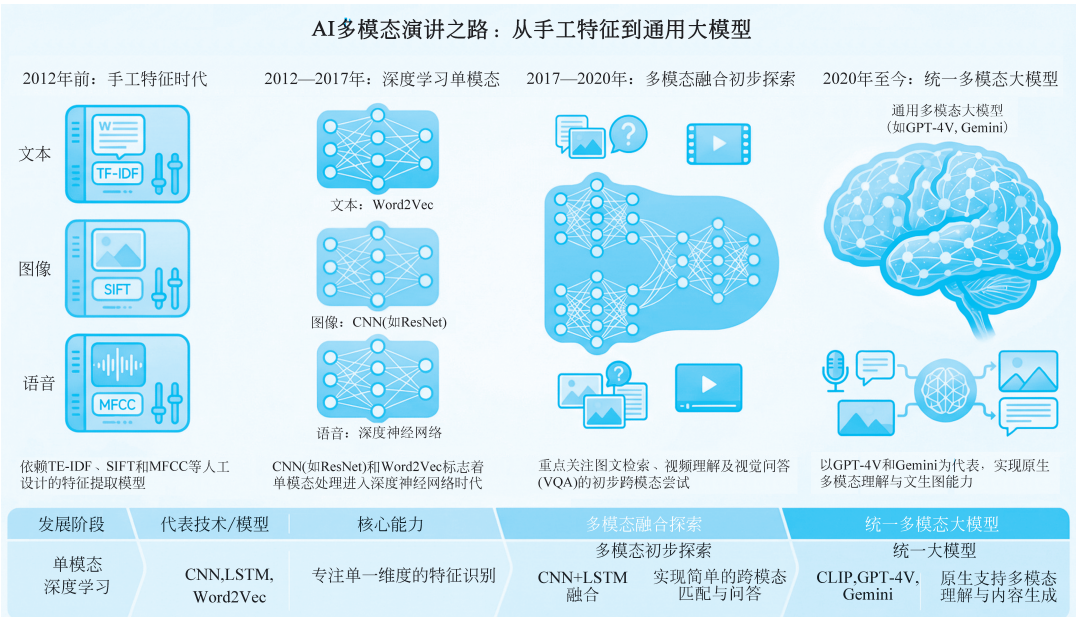


图 5.2　AI 多模态演进之路：从手工特征到通用大模型

阶段一：2012 年前，手工特征时代(单模态时代)。

2012 年以前，人工智能领域处于手工特征时代，亦可称为单模态孤立发展阶段。这一时期的核心特征是：机器处理不同模态数据的方式彼此割裂，文本被简化为词频统计，图像

被视为像素矩阵，语音则被抽象为声学参数，所有模态的表示都严重依赖人类专家设计的特征工程，缺乏对数据深层语义的自动理解能力。

由于各模态的特征空间相互独立，跨模态交互的尝试极为稀少且原始。例如，所谓的“基于文本的图像检索”，实际是通过人工为图片预先标注关键词，再将文本查询与这些标签进行字符串匹配，本质上仍是文本到文本的检索，并未实现视觉内容与语言语义关联。

阶段二：2012—2017 年，深度学习单模态阶段。

在 2012—2017 年的深度学习单模态阶段，人工智能各子领域分别借助深度神经网络实现了感知能力的巨大飞跃，但模态间的交互仍处于早期探索，缺乏统一的语义表征空间。

这一时期的标志性突破首先体现在计算机视觉领域。2012 年，AlexNet 在 ImageNet 竞赛中以压倒性优势夺冠，证明了深度学习自动学习视觉特征的能力远超 SIFT、HOG 等手工特征，由此开启了数据驱动的视觉时代。随后，VGG 通过堆叠小卷积核加深网络，ResNet 引入残差连接解决了深层网络退化问题，使得视觉特征的提取越发精准和丰富。与此同时，自然语言处理领域也迎来了关键进展：2013 年，Word2Vec 提出了“语义由分布决定”的分布式表示思想，将单词映射到连续的向量空间，为后续跨模态对齐奠定了概念基础；而 RNN 及其变体 LSTM 则显著增强了对文本序列的建模能力，虽然仍面临长程依赖的挑战，但已经能够捕捉一定程度的上下文信息。

在各自模态能力提升的基础上，研究人员开始尝试跨模态任务的初步探索，典型代表是图像描述生成和图文检索。图像描述生成通常采用编码器-解码器架构：用 CNN 提取图像特征作为“视觉编码”，再将其输入 LSTM，逐词生成描述文本。这类方法虽然实现了模态间的协作，但 CNN 和 LSTM 往往分开预训练后简单拼接，需要大量人工标注的图文对进行监督学习，且缺乏对视觉与语言深层对齐的建模。图文检索则更直接地计算图像特征与文本特征之间的相似度，常采用晚期融合（如分别提取特征后计算余弦距离）或简单的特征拼接，本质上仍是独立模态表示的后期匹配。

总体而言，该阶段的核心局限性在于：不同模态仍由异构的神经网络架构处理（CNN 处理图像，RNN 处理文本），它们各自在独立的特征空间中进行表示，缺少一个能够对齐视觉与语义的联合嵌入空间。模态间的融合停留在浅层或任务特定的拼接层面，远未实现真正的语义交互。这些局限性恰恰催生了下一阶段对多模态融合与统一表征的深入研究。

阶段三：2017—2020 年，多模态融合初步探索阶段。

在深度学习单模态能力趋于成熟后，研究者开始将目光转向如何让不同模态的信息在模型中实现真正的“对话”，而非简单的特征拼接。这一阶段（2017—2020 年）可称为多模态融合的初步探索期，其核心矛盾在于：如何设计有效的交互机制，使视觉、语言等模态能够在深层语义上相互对齐与协同推理。

2017 年 Transformer 架构的提出为多模态融合带来了革命性工具。其核心的注意力机制能够动态计算输入元素之间的相关性，天然适配多模态任务——例如，在图像描述生成中，模型可以学习在生成每个单词时关注图像中对应的区域；在视觉问答中，注意力能够定位问题所指的图像区域，实现精准的跨模态对齐。虽然注意力思想此前已有应用（如 Show、Attend 和 Tell 在 2015 年已将注意力用于图像描述），但 Transformer 提供了统一、高效的实现框架，使得注意力成为多模态建模的标配。

这一时期涌现了大量探索多模态融合的典型任务与模型架构。

视觉问答(visual question answering，VQA)：模型需要同时理解图像内容和自然语言问题，并给出正确答案。这一任务迫使模型进行深度的跨模态推理，例如判断问题关注的对象、属性或关系，进而定位图像中的相应证据。早期的 VQA 模型常采用双塔结构(CNN 编码图像，LSTM/GRU 编码文本)，再通过注意力机制融合特征；或采用单塔结构将图像区域特征与词特征拼接后送入共同网络。

视频理解：视频本质上是图像序列与音频的复合模态。研究者尝试融合时空特征，例如使用 3D 卷积提取短期运动特征，结合 CNN 提取空间语义，再用 LSTM/GRU 建模长期时序依赖，以完成动作识别、视频描述等任务。这类融合往往涉及多级特征拼接或注意力加权。

多模态机器翻译：在文本翻译的基础上，引入图像作为辅助信息。例如，在翻译含有歧义的词汇时，模型通过注意力机制参考对应图像区域来消除歧义，从而提升翻译质量。这证明视觉模态能够为语言任务提供补充语义。

这一阶段的多模态融合表现出以下特点：①架构多样，但尚未统一：不同任务往往设计专用的融合策略，如早期融合(输入级拼接)、中期融合(特征级交互)或晚期融合(决策级投票)，且常采用异构网络(CNN 处理视觉，RNN/Transformer 处理语言)；②预训练范式缺失：与单模态领域(如 ImageNet 预训练、BERT 预训练)不同，多模态模型大多从零开始在特定任务数据上训练，缺乏大规模通用预训练的支持，融合的有效性高度依赖任务数据的规模和人工标注质量；③交互层次较浅：尽管引入了注意力，但许多模型仍停留在特征对齐或加权求和层面，未能实现模态间深层次的语义推理与知识迁移。

尽管存在上述局限，这一时期的探索为后续突破奠定了重要基础：它验证了注意力机制在多模态融合中的有效性，并催生了构建联合语义空间的迫切需求，直接推动了下一阶段(2020 年至今)基于对比学习和大规模预训练的统一多模态大模型时代的到来。

阶段四：2020 年至今，统一多模态大模型阶段。

进入 21 世纪 20 年代，多模态技术迎来了爆发式发展。研究者不再满足于为特定任务设计复杂的融合架构，而是探索能够统一建模文本、图像、音频、视频的大模型。这一阶段的核心突破在于大规模对比学习与生成式预训练的结合，使模型首次拥有了接近人类的跨模态理解和创造能力。从 CLIP 构建共享语义空间，到 DALL·E 实现语义到图像的生成，再到 GPT-4V 和 Gemini 的原生统一架构，多模态模型逐步实现了理解→生成→原生交互的跨越。2024—2025 年，技术演进进入架构统一与推理增强期；至 2026 年年初，则进一步向基础设施化、无编码器原生架构和具身智能方向深化，如表 5.2 和表 5.3 所示。

表 5.2　统一多模态大模型阶段国外多模态大模型发展全景(2020—2026 年)

时　间	公司/机构	模　型	核心贡献	模态能力
2020—2021 年：奠基期				
2021.01	OpenAI	**CLIP**	图文对比学习，奠定统一语义空间基础	图文对齐
2021.01	OpenAI	**DALL·E**	首个文生图大模型，开启 AI 绘画时代	文本→图像
2022 年：扩散模型爆发期				
2022.04	OpenAI	**DALL·E 2**	扩散模型＋CLIP，大幅提升生成质量	文本→图像

续表

时　间	公司/机构	模　型	核心贡献	模态能力
2022.04	DeepMind	**Imagen**	基于扩散模型的高质量文生图	文本→图像
2022.08	Stability AI	**Stable Diffusion**	开源文生图，引爆 AI 绘画大众化	文本→图像
2022.04	DeepMind	**Flamingo**	800 亿参数，小样本学习能力强	图文理解
2022.09	Meta	**FLAVA**	统一视觉—语言预训练框架	图文理解
2023 年：原生多模态开启期				
2023.03	OpenAI	**GPT-4V**	首个原生多模态商用模型	图文理解、对话
2023.07	Google DeepMind	**RT-2**	机器人视觉—语言—动作模型	具身智能
2023.08	Google	**PaLM-E**	具身多模态模型	视觉、语言、动作
2023.09	OpenAI	**DALL · E 3**	与 ChatGPT 深度集成	文本→图像
2023.11	xAI	**Grok（初代）**	3140 亿参数 MoE 架构，搭载实时联网功能，直接访问 X 平台数据，标志性幽默应答风格	文本为主，实时数据
2023.12	Google DeepMind	**Gemini 1.0**	原生全模态架构	文本、图像、音频、视频
2024 年：架构统一与全模态萌芽期				
2024.02	Google DeepMind	**Gemini 1.5 Pro**	百万级上下文，长视频理解	全模态
2024.03	Anthropic	**Claude 3 Haiku/ Sonnet/Opus**	首次引入视觉理解能力	图文理解、文档分析
2024.03	xAI	**Grok-1.5**	在 MATH 基准测试中得分为 50.6%，显著提升了推理能力；Grok-1 开源（Apache 2.0 协议）	文本、推理
2024.04	xAI	**Grok-1.5V**	引入视觉理解能力	图文理解
2024.05	OpenAI	**GPT-4o**	全模态实时交互，320ms 延迟	文本、视觉、音频实时
2024.05	Meta	**Chameleon**	统一 Transformer 架构	图文统一处理
2024.05	Anthropic	**Claude 3.5 Sonnet**	视觉理解大幅增强；引入 Artifacts 实时预览	图文理解、代码生成
2024.08	xAI	**Grok 2**	在聊天、编码和推理方面能力强大，提供文本和视觉理解先进功能	图文理解、代码
2024.09	Mistral	**Pixtral 12B**	开源多模态	图文理解
2024.10	智源研究院	**Emu3**	原生多模态世界模型（*Nature* 发表）	理解、生成、预测统一
2024.11	Mistral	**Pixtral Large**	124B 参数，复杂数学推理	图文理解、推理

续表

时　间	公司/机构	模　型	核心贡献	模态能力
2024.11	**Anthropic**	**Claude 3.5 Sonnet（Computer Use Beta）**	全球首个支持“计算机使用”的模型	视觉—动作、Agent
2025 年：全模态爆发期				
2025.01	Google	**Veo 2**	4K 分辨率视频生成	文本→视频
2025.02	xAI	**Grok 3**	免费向公众开放	图文理解
2025.04	xAI	**Grok 3.5**	早期测试版向 SuperGrok 订阅者发布	图文理解
2025.07	xAI	**Grok 4**	256K 长上下文窗口，每秒 344 token 的推理速度；推出虚拟伴侣 Companion Mode（Ani、Valentine 等角色）	长上下文、多模态交互
2025.08	xAI	**Grok Imagine**	文本转视频生成器，马斯克称其为“AI 版的 Vine”	文本→视频
2025.09	xAI	**Grok 4 Fast**	接近 Grok 4 的推理表现，平均减少 40% 推理 token	文本、推理
2025.10	xAI	**Grok Companion Mika**	推出新的虚拟伴侣角色 Mika	多模态交互
2025.11	xAI	**Grok 4.1**	性能优化版本	图文理解
2025.02	Anthropic	**Claude 3.7 Sonnet**	全球首个“混合推理架构”（双脑 AI），SWE-bench Verified 70.3%	图文理解、深度推理
2025.02	Google DeepMind	**Gemini 2.0**	原生工具调用	全模态智能体
2025.05	OpenAI	**GPT-5**	推理与多模态深度融合	全模态理解与推理
2025.09	OpenAI	**Sora 2**	视频生成 GPT-3.5 时刻	文本→视频
2025.10	Google	**Veo 3.1**	音频＋视频统一	音画同步生成
2025.10	智源研究院	**Emu3.5**	Next-State Prediction 范式	世界模型升级
2025.11	Google DeepMind	**Gemini 3**	推理与代理能力	全模态智能体
2025.12	Meta	**Seamless M4T v2**	全向翻译	语音、文本互译
2025.12	MiniMax	**M2**	2300 亿 MoE 架构	Agent 工作流优化
2025.12	Amazon	**Nova 系列**	全模态家族（Micro、Lite、Pro、Premier）	文本、图像、视频
2026 年：基础设施化与具身智能期				
2026.01	xAI	**Grok 4.20**	2026 年 2 月 15 日正式发布	图文理解

续表

时　间	公司/机构	模　型	核心贡献	模态能力
2026.01	**NVIDIA**	**Alpamayo**	开源自动驾驶 VLA	视觉-语言-动作
2026.02	xAI	**Grok Imagine 1.0**	10 秒视频生成，720p 分辨率，支持情感语音和音画同步，过去 30 天生成 12.45 亿个视频；推出 Imagine API 支持开发者集成	文本→视频、图像→视频、视频编辑
2026.02	Anthropic	**Claude Opus 4.6**	最强智能体模型，百万上下文，支持“智能体团队”	全模态理解、Agent 集群
2026.03	Google DeepMind	**Gemini Embedding 2**	原生多模态嵌入模型	文本、图像、视频、音频、PDF 统一嵌入
2026.03	OpenAI	**GPT-5.3 系列**	涵盖 Nano/Mini/Pro/Codex 等多版本	全模态家族
2026. Q1（即将发布）	xAI	**Grok 5**	6 万亿参数（前代 3 万亿的两倍），原生多模态架构，支持实时视频理解；计划挑战《英雄联盟》顶级战队（限制人类视觉和反应速度），验证 AGI 在复杂游戏中的适应性	全模态、视频理解、实时决策

表 5.3　统一多模态大模型阶段国内多模态大模型发展全景（2020—2026 年）

时　间	公司/机构	模型/成果	核心特点	模态能力
2020—2021 年：萌芽探索期				
2021.03	北京智源研究院	**悟道 1.0**	发布四大模型：文源（26 亿参数中文语言）、文澜（10 亿参数图文多模态）、文汇（113 亿参数认知模型）、文溯（蛋白质预测），对外开放 API	中文语言、图文多模态、认知、蛋白质
2021.07	中国科学院自动化所	**紫东太初 1.0**	全球首个千亿参数的图、文、音三模态大模型，实现图像、文本、语音三类数据的相互生成，获世界人工智能大会 SAIL 奖	图文音三模态
2022—2023 年：先行探索期				
2022.03	百度	**文心一格**	国产文生图先行者	文本→图像
2022.08	阿里	**通义万相**	电商场景优化	文本→图像、图像编辑
2023.02	复旦大学	**MOSS**	国内首个对话式大型语言模型，200 亿参数	文本为主
2023.03	百度	**文心一言**	知识增强大模型	图文对话
2023.04	阿里	**通义千问**	多尺寸模型家族	图文理解
2023.04	知乎	**知海图 AI**	千亿级参数，在“知乎”业务中人工标注量降低 90%以上	文本理解

续表

时　间	公司/机构	模型/成果	核心特点	模态能力
2023.05	科大讯飞	讯飞星火认知大模型	文本生成、知识问答、数学能力早期领先	文本为主
2023.06	中国科学院	紫东太初 2.0	新增视频、3D 点云、传感信号，实现全模态扩展，应用于神经外科导航、工业焊接	全模态（图文音视频＋3D＋传感）
2023.07	华为	盘古大模型 3.0	5＋N＋X 三层架构，面向行业应用的系列大模型	多模态行业应用
2023.07	京东	言犀大模型	70%通用数据＋30%供应链原生数据，产业导向	多模态产业应用
2023.07	网易有道	子曰	国内首个教育领域垂直大模型	教育垂直
2023.08	字节跳动	豆包（云雀模型）	AI 对话产品，支持网页、iOS、安卓平台	对话、写作助手
2023.08	百川智能	百川 4	中文优化	图文理解、文生图
2023.09	腾讯	混元大语言模型	超千亿参数，预训练语料超 2 万亿 token	文本理解与创作
2023.09	阿里	通义千问	通过备案并向公众开放，同时宣布开源	图文理解
2023.09	金山办公	WPS AI	接入金山办公全线产品，提供 AIGC、Copilot、Insight 能力	办公智能
2023.10	智谱 AI	ChatGLM-4	开源友好，国内首个具备代码交互能力的大模型产品	视觉理解、文生图
2023.11	昆仑万维	天工 SkyAgents	Agent 框架	多模态交互
2024 年：密集爆发期				
2024.01	MiniMax	abab6	长文本＋多模态	文生图、语音交互
2024.03	阶跃星辰	Step-1V	千亿参数多模态大模型，在 GDC 大会首发	图文理解、视频理解
2024.03	阶跃星辰	Step-2（预览版）	万亿参数 MoE 语言大模型	文本为主
2024.05	快手	可灵 Kling	视频生成领先	文本→视频
2024.06	商汤	日日新 SenseNova 5.0	全栈能力	文生图、文生视频、数字人
2024.07	商汤	Vimi	首个可控人物类视频生成模型，面向 C 端用户	视频生成
2024.07	字节	豆包 Seed-MM	轻量化多模态	端侧部署、图文理解
2024.07	科大讯飞	讯飞星火 V4.0	音视频、图片解读等多领域复杂问题处理	多模态理解
2024.08	腾讯	混元 hunyuan-vision	SuperCLUE-V 银牌（国内第一、全球第二）	中文多模态理解