**深度学习技术贯穿多模态图像检索的各环节：**

## （一）多模态特征提取技术

### CNN 基础架构

* + 核心作用：作为图像特征提取的基石，通过卷积层、池化层逐层抽象视觉特征（边缘→纹理→语义），典型代表如 AlexNet（2012 年）、VGGNet（2014 年），解决了传统手工特征（SIFT、 HOG）泛化能力弱的问题。
  + 技术局限：深层网络易出现梯度消失，且难以建模全局语义关联。

### 残差与多尺度网络

* + ResNet（2015 年）：引入跳跃连接与残差学习机制，突破网络深度限制，使深层特征提取成为可能，成为后续多模态模型的主流视觉骨干。
  + U-Net（2015 年）：通过编码器 - 解码器结构 + 跳跃连接实现多尺度特征融合，虽起源于分割任务

，但因其细粒度特征保留能力，被广泛用于医疗影像等精准检索场景的特征提取。

### Transformer 视觉适配

* + ViT（Vision Transformer，2020 年）：将图像切割为 Patch 序列输入 Transformer，打破 CNN 局部感受野局限，实现全局特征建模，成为多模态大模型的核心视觉编码器。

## （二）跨模态对齐技术

### 注意力机制演进

* + 早期共注意力（如微软动态共注意力模型）：通过联合学习聚焦图文关键区域，初步实现模态关联建模，使跨模态检索准确率提升至 60%-70%。
  + Transformer 自注意力：CLIP（2021 年）采用双编码器结构，通过对比学习让图文特征在统一语义空间对齐，解决了传统方法 “语义鸿沟” 问题。

### 动态交互模块

* + BLIP-2（2023 年）的 Q-Former 模块：通过可学习查询向量实现图文特征实时对话，提升细粒度语义对齐精度，支撑复杂推理任务。

## （三）多模态融合机制

### 阶段式融合

* + 早期融合：特征提取阶段拼接图文特征（如 Google Multi-modal Fusion Network），但易丢失单模态细节。
  + 晚期融合：特征提取后再融合，保留细节但泛化性差，代表模型如早期图文检索系统。

### 深度融合模型

* + 自编码器（AE）/ 变分自编码器（VAE）：通过共享隐空间强制不同模态特征对齐，适用于小规模数据场景。
  + 生成对抗网络（GAN）：通过生成器与判别器博弈，增强跨模态特征的一致性，提升检索鲁棒性

。

# 二、框架演进与发展趋势（2014 年至今）

## （一）奠基阶段（2014-2018 年）：CNN 主导

* **核心框架**：ResNet（解决深度问题）、U-Net（多尺度特征）、基础注意力网络。

### 技术突破：

1. 以 CNN 为视觉骨干，RNN/LSTM 处理文本，实现初步跨模态关联（如商品图文匹配）。
2. 引入注意力机制，使模型可动态关注关键信息（如微软共注意力模型）。

* **局限**：依赖大量标注数据，对齐精度有限，未形成统一语义空间。

## （二）预训练突破阶段（2019-2022 年）：Transformer 驱动

* **核心框架**：Transformer 统一架构、对比学习范式。

### 技术突破：

1. CLIP（2021 年）：通过 4 亿图文对预训练，采用 “图像编码器（ResNet/ViT）+ 文本编码器（- Transformer）” 双架构，实现零样本跨模态检索，ImageNet 零样本分类准确率达 75%。
2. Flamingo（2022 年）：冻结视觉编码器，仅训练跨模态注意力层，降低 75% 计算成本，支撑视频问答等复杂任务。

* **标志**：从 “任务特定训练” 转向 “预训练 - 微调” 模式，突破数据依赖瓶颈。

## （三）多模态大模型阶段（2023 年至今）

* **核心框架**：纯 Transformer 架构、动态交互模块。

### 技术突破：

1. 统一 token 化：GPT-4 将图像离散为网格 token，与文本统一处理，支持任意模态输入输出。
2. 无编码器架构：Fuyu-8b 直接输入图像像素至 Transformer，简化流程并降低延迟，适配实时检索场景。
3. 轻量化优化：知识蒸馏压缩百亿参数模型，平衡性能与算力需求。