图文开源数据集:

常见的多模态任务常用的数据集: https://mp.weixin.qq.com/s/vzyOF4esJCkBZDMiNScE5A
多模态数据集汇总: https://blog.csdn.net/m0_59163425/article/details/142499141

0.coco，flickr8k，flickr30k

1.AI challenge: 数据集包含 30 万张图片，150 万句中文描述。训练集：210,000 张，验证集：30,000 张，测试集 A：30,000 张，测试集 B：30,000
https://blog.csdn.net/weixin_48344945/article/details/111859679

下载
百度网盘: https://pan.baidu.com/s/1g1XaPKzNvOurH9M44p1qrw 提取码: bag3

百度网盘: https://pan.baidu.com/s/1m-yFj6ST2KJlx7D77de6DQ 提取码: CPRK

谷歌网盘：https://drive.google.com/open?id=0ByB0MjjNghlyNkdhR3lIZGJneGM

2.Zero: 2300 万图文对、 230 万图文对
https://blog.csdn.net/c9Yv2cf9I06K2A9E/article/details/125240573

下载 https://zero.so.com/download.html

3.Wukong: 1 亿个图文对
https://zhuanlan.zhihu.com/p/472493389

下载 https://wukong-dataset.github.io/wukong-dataset/download.html

4.CapsFusion-120M: https://github.com/baaivision/CapsFusion huggingface ----------HTTP URL img ok

5.AnyWord-3M: https://github.com/tyxsspa/AnyText
下 载 https://modelscope.cn/datasets/iic/AnyWord-3M/summary modelscope ------xiazai

6.RLAIF-V-Dataset: 一个大规模多模态偏好数据集 huggingface ----------ok

7.LAION-5B:
80TB！ 58.5 亿！ 世界第一大规模公开图文数据集 LAION-5B 解读 https://zhuanlan.zhihu.com/p/571741834 Clip front
下载 https://laion.ai/projects/

8.LAION-400M:
史上最大多模态图文数据集发布！ https://mp.weixin.qq.com/s/vzyOF4esJCkBZDMiNScE5A
下载 https://laion.ai/blog/laion-400-open-dataset/

9.WuDaoMM: https://github.com/BAAI-WuDao/WuDaoMM 需申请

10.MINT-1T: : https://github.com/mlfoundations/MINT-1T huggingface

11.MMSci: MMSci 数据集是一个多模态、多学科的高质量学术文章和图表集合，涵盖 72 个科学领域。数据集包含 131,393 篇文章和 742,273 个图表，主要来源于 Nature Communications 期刊
下载 https://github.com/Leezekun/MMSci/blob/main/mmsci-data/README.md

12.Multimodal C4 (mmc4): Multimodal C4 的数据集，语料库包含 103M 文档，其中包含了 585M 张图片和 43B 个英文单词
https://github.com/allenai/mmc4?tab=readme-ov-file

13.ShareGPT4V: 数据集包含 120 万对图像-文本配对 https://sharegpt4v.github.io/