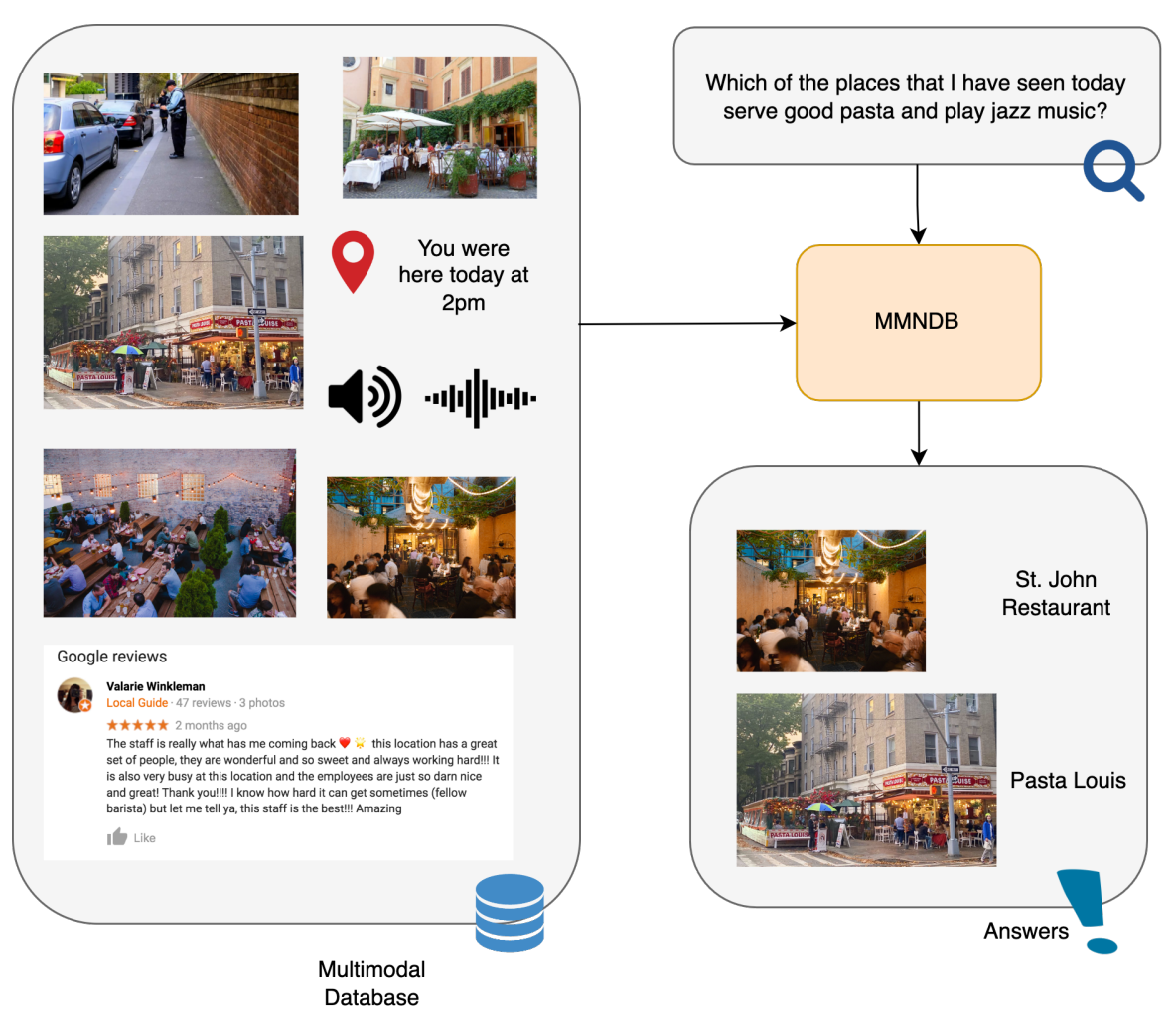
**1 Multimodal Databases**

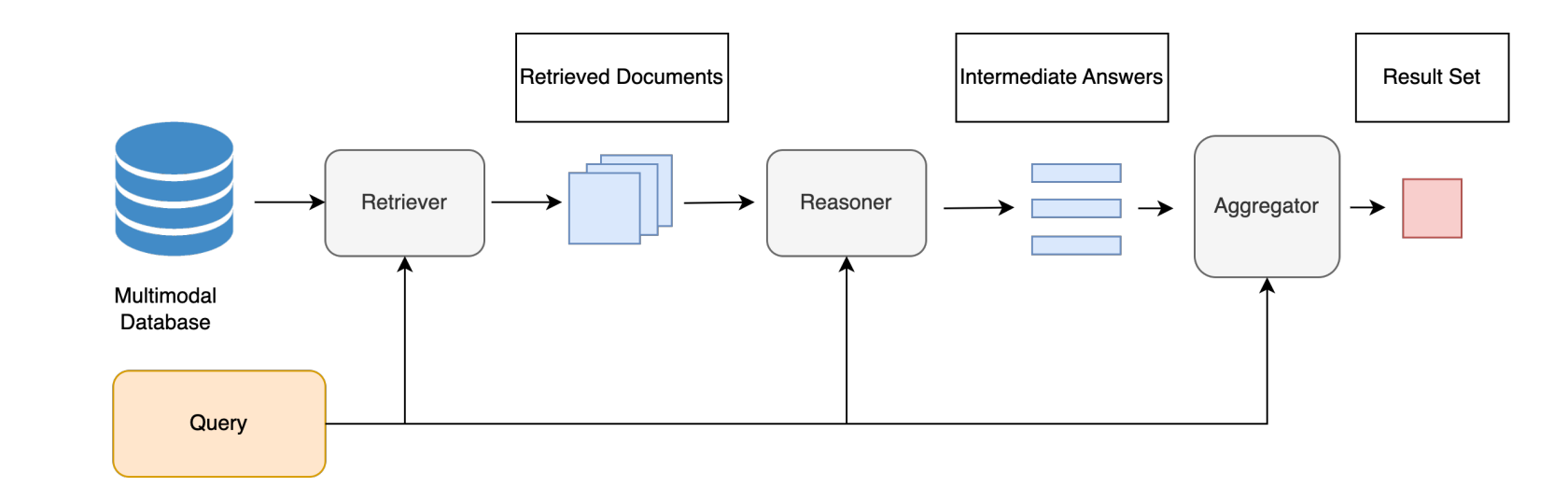
**1.1 Design && Innovation**

## 1.1.1 Multimodal Neural Databases（SIGIR-2023）

（1）尽管多媒体信息检索取得了较大进展，但仍然受限于支持的查询类型，于是文章根据神经数据库工作提出了新的框架：多模态神经数据库（mmndb）；mmndb可以回答复杂的类似数据库的查询，针对不同输入模式（如文本和图像）进行大规模推理；



（2）文章提出：检索器——推理器——聚合器模型；给定一个查询检索器从数据库中返回与该查询相关的一小部分文档、系统并行运行推理器的多个副本，每个推理器为查询生成部分结果；最后聚合器组件将从中间查询结果创建查询结果；



（3）理想的数据库：数据是多媒体的，例如社交圈的帖子、智能眼镜带来的一天；查询利用自然语言而非SQL语言；

（4）文章提出的系统的第一个原型：所有文档都是图像的数据库；自然语言提出的查询；数据集：MS-COCO dataset (Common Object in Context)（大约包含123K个带标签的图像，每张图片和五个标题关联，分为训练子集和测试子集，分别包含118k和5k个图像）；采用训练/微调的方法；推理器使用OFA（在各种多模态任务上进行训练，目前最好的开源多模态模型之一）；检索器使用CLIP模型，能通过匹配字幕和图像以一种无监督的对比方式进行训练，CLIP创建的嵌入是静态的不依赖于查询，我们可以预先计算图像的嵌入，文中使用了八个版本的模型：RN50, RN101, RN50x4, RN50x16, RN50x64, vitb /32, ViTL/14, ViTL/14@366px；

（5）由检索器提前计算了嵌入我们就可以选择哪些文档被认为是相关的，哪些不是，在选择上，文中提出了三种策略：

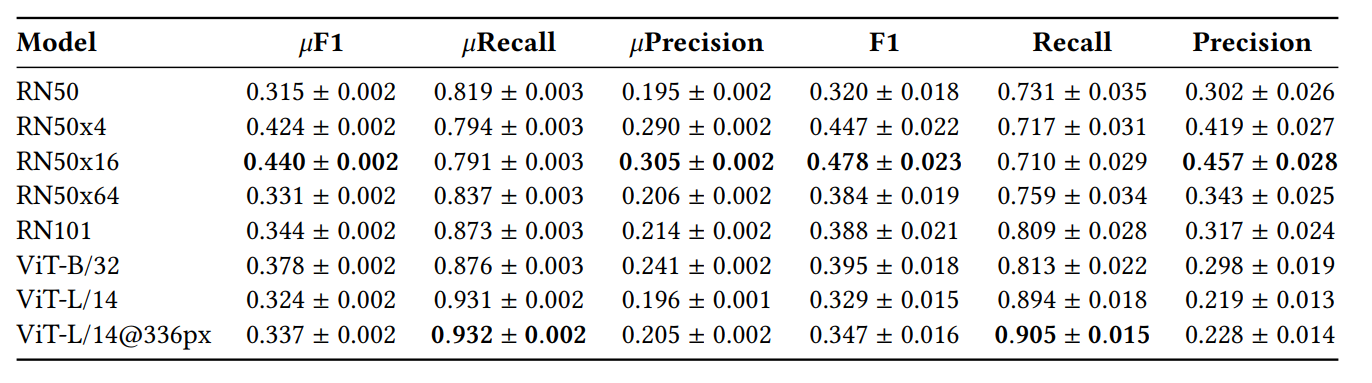
1、top k，计算文档和查询之间的点积，对他们进行排序选择top K；

2、阈值，计算文本和图像嵌入之间的余弦相似度，返回余弦相似度大于某个阈值的所有文档，该阈值取决于正在使用的特定CLIP模型，范围在0.15到0.4之间；

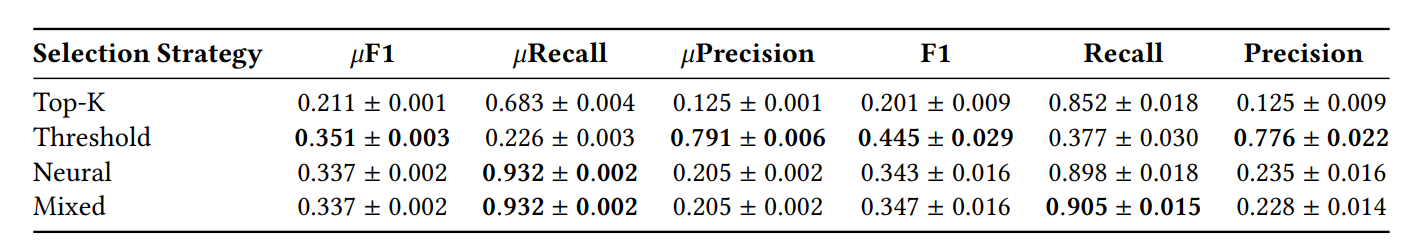
3、神经选择器，训练一个小的神经网络，给定𝑞和𝐷嵌入，返回一个二进制结果，表明文档是否与查询相关，以及是否应该返回；

混合策略：topK和神经选择器的并集；

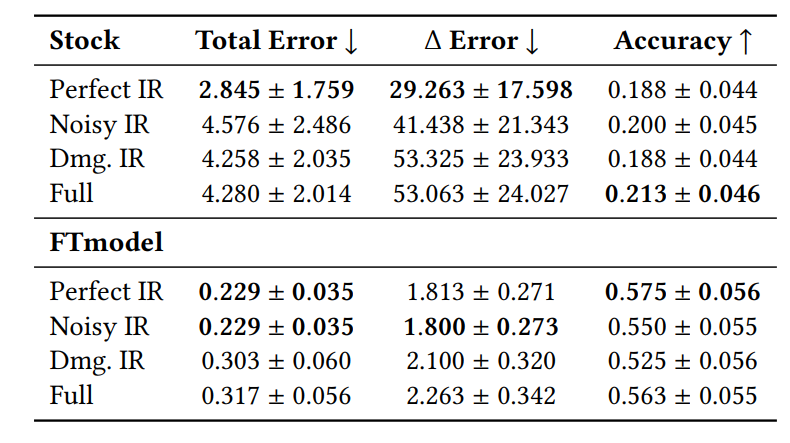
（6）结果1：“混合”检索策略下不同推理器模型的结果比较：



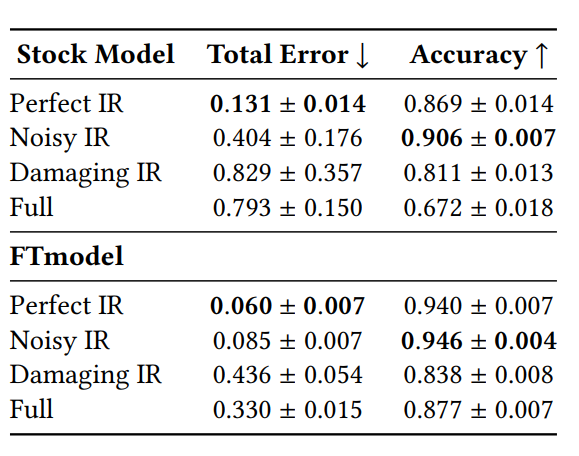
结果2：不同检索策略的比较：混合策略有更高的召回率：



结果3：查询max的结果：



结果4：查询in的结果



PerfectIR版本表现较好，证明了本文任务的可行性。Full Pipeline稍落后于PerfectIR设置。TP正确检索的文档，FP是错误检索的文档，假阴性(FN)指的是应该检索但没有检索的文档，在总误差上FN几乎可以忽略不计，意味着总误差的差距不是由于未检索文档造成的。实际上是由误报造成的，于是增加NoisyIR的附加设置：随机选取一些不相关的文档(300个)；同时增加底片文件不是随机取的DamagingIR设置，选取方式为与查询的CLIP嵌入最高的非相关文档。

（7）结果表明：多模态神经网络是非常有希望的，文章设法建立的高效的检索系统有很高的召回率，表明了巨大的潜力；