多模态神经数据库

Giovanni Trappolini
trappolini
@diag.uniroma1.it 罗马第一大学 罗马，拉齐奥，意大利

Andrea Santilli
santilli@di.uniroma1.it 罗马第一大学 罗马，拉齐奥，意大利

Emanuele Rodolà
rodola@di.uniroma1.it 罗马第一大学 罗马，拉齐奥，意大利

Alon Halevy
ayh@meta.com Meta AI 门洛帕克，加利福尼亚州，美国

Fabrizio Silvestri
fsilvestri@diag.uniroma1.it 罗马第一大学与 ISTI - CNR，意大利拉齐奥大区罗马

arXiv:2305.01447v1 [cs.MM] 2 May 2023

摘要

文本、图像及其他模态所包含的松散结构化数据日益增多，因此需要新的查询方式。多媒体信息检索填补了这一空白，并在近年来取得了令人瞩目的进展。搜索和检索大型多媒体档案等任务的性能大幅提升，这在很大程度上得益于多模态深度学习的最新发展。然而，该领域的方法仍受限于其支持的查询类型，尤其是无法回答类数据库查询。因此，受神经数据库领域近期研究的启发，我们提出了一种新框架，称为多模态神经数据库（MMNDB）。多模态神经数据库能够大规模回答涉及不同输入模态（如文本和图像）之间推理的复杂类数据库查询。本文介绍了首个能够满足这组要求的架构，并通过多个基线方法对其进行测试，展示了现有模型的局限性。结果表明，这些新技术有望处理来自不同模态的非结构化数据，为该领域未来的研究铺平道路。用于复现实验的代码将发布在 https://github.com/GiovanniTRA/MultimodalNeuralDatabases

CCS概念

• **信息系统** →**信息检索**；**多媒体与多模态检索**。

关键词

多媒体信息检索、数据库、神经网络

ACM参考格式：

Giovanni Trappolini、Andrea Santilli、EmanueleRodolà、Alon Halevy 和 Fabrizio Silvestri。2023年。《多模态神经数据库》。载于 *ACM Conference* 会议论文集（*Conference'17*）。ACM，美国纽约州纽约市，11页。https: //doi.org/10.1145/nnnnnnn.nnnnnnn

允许为个人或课堂使用而免费制作本作品的全部或部分数字版或纸质副本，但前提是不得出于营利或商业利益而制作或分发副本，并且副本必须保留本声明及首页的完整引文。对于本作品中由ACM以外的其他方拥有版权的组件，必须遵守其版权规定。在注明出处的情况下，允许摘编。若要以其他方式复制、重新出版、发布到服务器或重新分发到列表，则需要事先获得具体许可和/或支付费用。请通过 permissions@acm.org向权限部门申请许可。*Conference'17*，*2017*年7月，美国华盛顿特区© 2023年计算机协会。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$ 15.00 https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

近年来，可获取数据的数量和种类都急剧增加；随着智能眼镜等更多设备日益普及，这一趋势很可能会加速。当前设备生成的数据主要是文本和图像，而智能眼镜可能会增加个人创建的音频和视频数据量。随着生成式人工智能的兴起，我们很可能会看到有价值的生成数据呈爆炸式增长。多媒体信息检索（MMIR）一直受到信息检索领域科学家和从业者的关注。多媒体信息检索旨在应对在多媒体集合上处理查询所面临的挑战。由于数据可用性大幅提升，多媒体信息检索也日益受到关注。该领域已经探索了从大型图像档案库中检索、以图搜图，以及基于人脸或指纹 [5]的检索等主题。然而，本文提出了一个新颖且具有变革性的观点：鉴于人工智能领域对各个技术领域产生的巨大影响，我们认为，多媒体信息检索领域需要探索能够处理更具表达力的类数据库查询的系统，这些查询称为多模态神经数据库（*MMNDBs*）。

我们通过一个示例说明多模态神经数据库的潜力。考虑对图像档案库进行以下查询：包含乐器的图像有多少张？假设集合中的图像都标注了其中识别出的物体（例如小号、鳄梨、人）。因此，多媒体信息检索系统可能能够返回含有小号或其他乐器的图像。然而，要找出哪些物体属于管乐器（或更细致的类别），还需要额外执行一步推理，即与乐器数据库进行连接。此外，要统计满足条件的图像数量，还需要对答案集的大小进行推理；数据库系统经常执行这一操作，但多媒体信息检索系统并不支持。示例还可以更复杂，例如找出照片中最常出现的乐器，或只考虑在举办过奥运会的城市拍摄的照片。由上述示例可见，多模态神经数据库的一项关键需求是具备对集合进行推理的能力。

在这篇观点论文中，如图 1 所示，我们提出研究、设计并构建多模态神经数据库，将大型多模态模型、多媒体信息检索与数据库查询处理的能力结合起来。我们受到神经数据库相关研究的启发 [26, 28, 29]，这类研究已引起自然语言处理、数据库和信息检索领域的关注。然而，我们将自身定位为多媒体信息检索领域的演进，因此与上述研究有所不同

借助现代的多模态人工智能技术，尤其是近期提出的技术。

我们开发了首个基于原则的原型，以展示所提出任务的可行性。我们稍后会强调，这只是解决多模态神经数据库的一种可能架构，未来的研究将揭示新的策略。总体而言，我们的原型基于检索器-推理器-聚合器模型。给定一个查询，检索器会从数据库中返回与该查询相关的一小部分文档。然而，通常即使是这部分文档，其数量也过大，无法作为单个阅读器的输入，而阅读器本质上是一个 Transformer。因此，系统会并行运行多个推理器副本，每个副本都会为该查询生成一个部分结果。最后，多模态神经数据库的聚合器组件会根据这些中间结果生成查询结果。例如，如果查询要统计包含人物的图像数量，中间结果将为 1 或 0，具体取决于图像中是否包含人物。聚合器会将所有的 1 相加。

多模态神经数据库系统将设计用于处理范围广泛的多媒体数据，包括图像、视频、音频和文本。系统能够处理自然语言查询，让用户可以直观、轻松地表达查询。系统还能够从多媒体数据中提取特征，并利用这些特征提升检索任务的性能。

本文介绍了实现多模态神经数据库的初步探索，旨在构建足够灵活、能够支持未来模型的系统。我们研究针对图像集合的查询，并验证了所提出架构的若干方面，如图2所示。我们开展了大量实验，展示了所提出任务在一部分潜在查询类型上的可行性与潜力。最后，我们讨论了本文分析以及多模态神经数据库这一概念所引出的未来研究方向。

2 多模态神经数据库 S

我们将来自不同模态的文档语料库称为多模态数据库。本文对文档所作的定义有意保持宽泛。一般而言，文档可以是任何独立完整的数据片段。多模态数据库可以包含迥然不同的数据源。例如，其中可能包含自然语言形式的信息、图像、声音、地理标记信息、时间戳等诸多内容。与传统数据库不同，多模态数据库是非结构化的，也就是说，它不需要具有模式，甚至不需要任何特定的顺序，而可以只是由这些文档构成的无序、非结构化集合。

多模态数据库出现在多种情境中。如今已有的一种情境是在线社交媒体：用户会发布不同类型的内容（文本、图像、网络迷因、视频、音频）。在这种情况下，每条帖子都是多媒体数据库中的一个文档，此外，数据库还必须追踪用户之间的友谊关系图。智能眼镜在不久的将来普及后，还会出现另一种情境：记录用户一天的生活。只需进行一些简单的活动，比如在酒吧买杯咖啡，眼镜就会捕捉感官数据、拍摄在场人物及其所吃食物的照片（视频）（遵循届时采用的任何隐私规范），以及酒吧里播放的背景音轨，还有购物收据的照片。

理想情况下，我们希望能够像查询数据库一样，查询这些丰富、大规模且非结构化的数据集合。更进一步，与传统数据库不同，我们希望使用自然语言执行查询，而不是使用 SQL 这类僵化的语言。具体而言，给定一个多模态数据库 D 和一个查询 q，我们希望能够执行以下几种类型的查询：(a)**集合查询**；集合查询是抽取式查询，会从事实中返回文本片段列表，例如实体。(b) **布尔查询**；返回 True 或 False 作为答案。(c) **连接查询**；需要组合两个或更多文档来生成每个答案。

我们注意到，与传统数据库（甚至神经数据库）不同，多模态数据库可以生成由异构模态构成的答案。例如，集合查询可以无缝生成包含图像、音频和自然语言（及其组合）的答案。

设计多模态神经数据库面临着若干重大挑战。

首先，系统必须能够对输入的模态进行推理，这一点至关重要。例如，如果我要查找猫和狗打斗的图像，就需要识别出这两种动物都在场，而且 它们之间的互动确实是在打斗（背景里有一张迈克·泰森拳击的海报并不足以满足要求）。同样，如果查询提到有人在低声耳语或大声喊叫，系统就必须理解音频片段中的这些细微差别。近年来，深度学习技术，尤其是大型深度学习模型，展现出了出色的推理能力 [10]。视觉问答和多跳问答任务在自然语言处理方面已达到接近人类的表现 [30] ，在多模态场景中也展现出前景。

然而，这些模型通常规模极大，拥有数十亿个参数，因此带来了下一个挑战，即规模问题。面对大量文档构成的集合，对每个查询-文档对运行这类模型，甚至对每份文档都运行一次，都是不可行的。

开放域问答系统（ODQA）旨在回答自然语言文本中的查询，为扩展到更大的文档集合提供了一种方法。ODQA 首先从文档集合中检索相关文档，并将其与查询一同作为上下文输入 Transformer，从而回答查询。然而，Transformer 只能接受长度有限的上下文（目前为 512 至 1024 个词元）。尽管扩展这一长度是当前非常活跃的研究领域，但它很可能始终小于处理我们所追求的这类查询所需的长度。回答数据库查询所需处理的文档数量可能任意庞大，中间结果集也可能如此。相比之下，ODQA 系统通常处理答案较短、仅需将少量文档输入 Transformer 即可得出的查询。此外，多模态数据库是一个无序文档集合，因此我们无法利用任何局部性启发式来检索相关文档。

最后但绝不是最不重要的一点是，在多模态数据库中，还存在如何弥合不同模态之间差异的挑战。为了回答

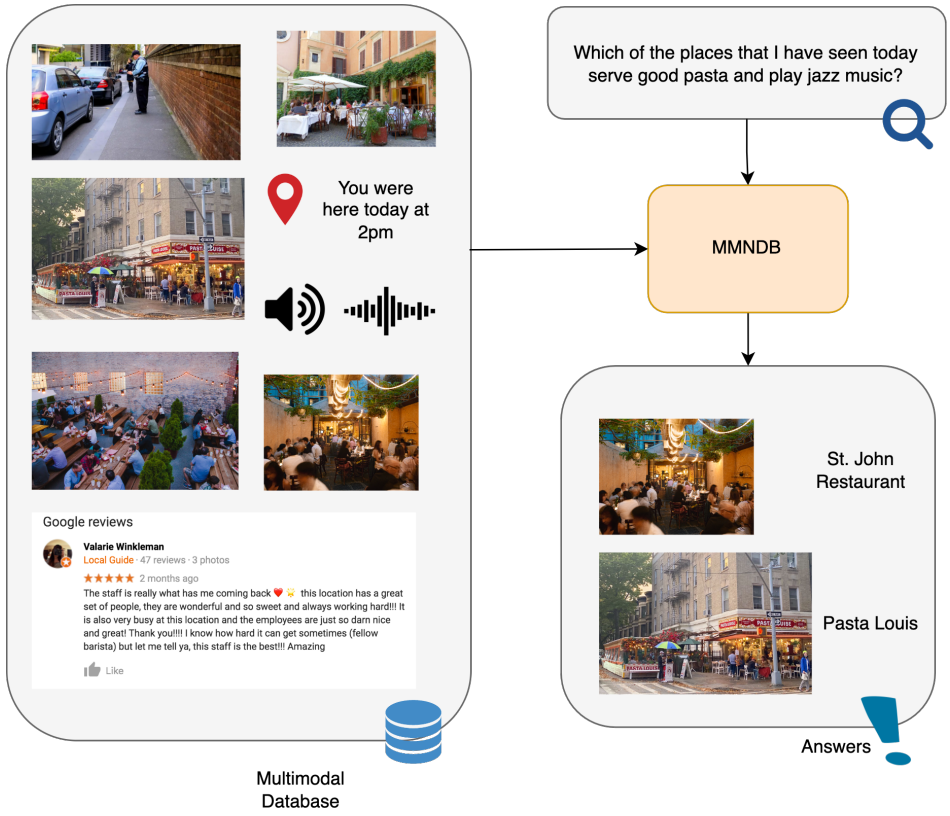


图1：多模态神经数据库的一种可能应用场景。想象一下，你戴着智能眼镜在城市中漫步，并将信息收集到多模态数据库中。到了晚上，你可能想知道哪些地方适合用餐并符合某些条件。多模态神经数据库可以通过回答以自然语言（或语音！）提出的类数据库查询，结合多个信息源和模态，帮助你做出决定。

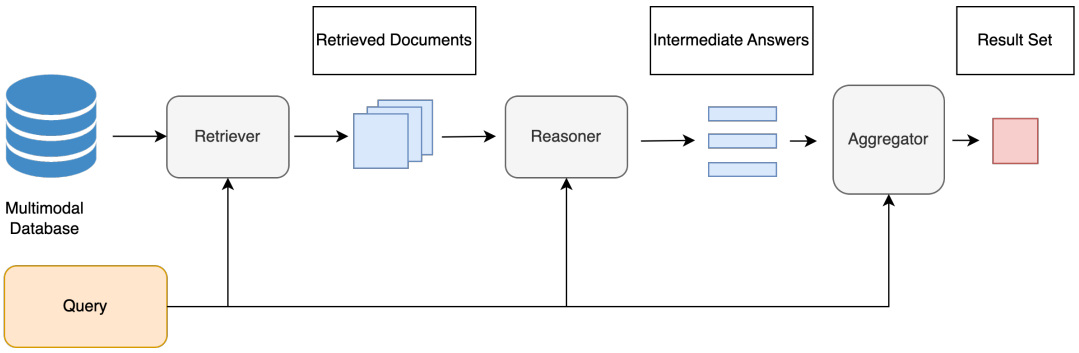


图2：我们提出的多模态神经数据库原型的模式。给定查询后，检索器模块会先对文档进行筛选。推理器会生成中间答案，然后由聚合器处理这些答案，生成最终答案。

关于多模态数据的查询，人们不仅需要处理、推理并整合来自不同文档的信息，还需要处理以不同模态呈现的文档中的信息。近年来，自然语言处理和计算机视觉领域的文献已经奠定了基础，并在该领域取得了卓越成果。随后出现了多模态模型，在文本到图像、图像到文本和文本到音乐任务中展现出优异表现。然而，目前大多数多模态模型要么处理文本与视觉任务，要么处理文本音频任务。尽管结合多种模态并非

尚未探索的领域， [2], 但要达到适当水平、胜任当前任务，仍需进一步研究。具体而言，要妥善处理多模态神经数据库（MMNDB）任务，我们需要一种“真正的”多模态模型，能够根据输入的任何可能模态进行推理。关于这一问题及其他当前局限和未来研究方向的进一步讨论，请参阅第5节。

3 多模态神经数据库的首个原型

为了展示多模态神经数据库的可行性，本节针对一种受限情形，介绍此类系统的首个原型。我们考虑所有文档均为图像的数据库，以及能够表达计数、最大值和包含关系的查询；这些查询以自然语言提出。不过，正如下文所述，我们初步系统的架构也适用于更广泛的场景。我们希望这一架构能成为其他多模态神经数据库方法的基础。

我们的系统接收针对数据库 D 的输入查询 q。它包含三个组件。第一个组件是检索器，用于从 D 中选出与回答查询相关的部分文档。第二个组件是推理器，它会处理检索到的文档子集，并且可以并行处理。推理器会给出查询的部分答案。第三个组件是聚合算子，用于综合推理器提供的答案，从而计算出查询的最终答案。

我们的架构优势在于，在实现检索器和推理器时，能够利用多模态神经模型的最新进展。具体而言，这些模型能够将多种模态映射到同一嵌入空间，从而共同推理图像和文本的内容。例如，这些模型能够识别图像中的物体，并生成描述图像主要内容的文本描述。

在解释各个组件之前，我们先概述查询在系统中的完整处理流程。以查询“有多少人在海滩上穿着蓝色T恤弹吉他？”为例。推理器会考虑 D 中的单张图像，并使用最新的神经方法判断图像中是否有人在海滩上弹吉他。然而，对 D 中的每份文档都进行如此强大的推理并不可行，因此我们使用检索器，将范围缩小到D、 P（D， q）中的一小部分图像。随后，多个推理器实例并行处理检索到的 P（D， q）中的图像，以确定哪些图像符合查询。在我们的示例中，如果某张图像符合查询，推理器就返回1，否则返回0。然后，聚合器对所有的1进行计数，得出最终答案。下面我们将介绍各个组件。

检索器。检索器的目标是从 D 中返回与查询 q 相关的文档子集 $P(D, q)$。检索器的主要要求是具备可扩展性。虽然我们对检索器的推理要求不及推理器细致，但它应该能够筛除绝大多数不相关的图像。为了检索与查询相关的文档，我们将查询和文档编码到同一个潜在嵌入空间中。不过，如前所述，文档的嵌入不能 依赖于查询 q，否则我们就必须为每个给定查询重新计算 D 中每篇文档的嵌入。因此，正如我们在第 4.1 节中所述，我们考虑了几种以独立于查询的方式对 D 中的文档进行嵌入的方法。

推理器。推理器的一个实例 $P(D, q)$ 将 D 中的一个文档作为输入，并返回针对查询 A_p 的中间答案。在上面的示例中，推理器会根据图像是否满足查询中的条件，返回 1 或 0。不过，中间结果也可能有所不同。对于“图像中人数最多是多少？”这样的查询，推理器会针对每张图像返回其中的人数。再比如，对于“数据库中最常出现的乐器是什么？”这一查询，推理器会输出它在图像中识别出的每种乐器及其出现次数。

推理器的关键作用，恰恰在于推理图像与查询之间的关系。在我们的示例中，推理器需要判断是否有人穿着蓝色服装、弹吉他的人是否就是这个人，以及他们是否身处海滩。推理器利用神经模型的最新进展来完成此类推理：这些模型将图像和文本嵌入到同一个潜在空间中，并生成图像的文本描述。不过值得注意的是，这些模型会动态计算查询和图像的嵌入，而嵌入取决于二者本身，即 $F(I|T) \neq F(I)$，反之亦然；其中 I 是图像， T 是自然语言（也可以是任意两种模态）。这会带来重大的计算影响。事实上，要回答查询，就需要处理所有可能的 D， q 组合。此外，由于只有在推理时才能知道查询，因此无法预先计算嵌入。因此，要在整个数据库上运行推理器显然并不可行。为此，我们在流水线中引入了一个额外模块，即检索器。

聚合器。聚合器接收查询以及来自所有推理器实例的中间输出，并生成查询的答案。从概念上讲，这是系统中最简单的组件，因为只需根据查询的语义对中间结果进行聚合。在我们的示例中，聚合器会统计返回 1 的图像数量。对于统计人员总数的查询，聚合器会将推理器返回的中间结果相加。

4 实验

本节介绍我们为验证原型的潜力而开展的实验。我们首先描述实验设置。

4.1 实验设置

本节概述了用于验证我们方法有效性的实验设置。

数据集。 我们的实验使用 MS-COCO 数据集（Common Object in Context） [18], ，这是计算机视觉领域最受欢迎的基准数据集。我们使用作者发布的最新版本。COCO 数据集包含约 123K 张带标签图像。每张图像都配有 5 条图像描述，并标注了其中识别出的物体。这些物体

表1："混合策略"下不同检索器模型的比较。CLIP 中采用残差网络作为骨干网络的版本，其 F1值和精确率较高；而基于视觉变换器的模型召回率更高。我们选择后者，因为它能让推理器模块接收尽可能多的相关信息，最终降低流水线误差。

模型	μF1	μ召回率	μ精确率	F1	召回率	精确率
RN50	0.315 ± 0.002	0.819 ± 0.003	0.195 ± 0.002	0.320 ± 0.018	0.731 ± 0.035	0.302 ± 0.026
RN50x4	0.424 ± 0.002	0.794 ± 0.003	0.290 ± 0.002	0.447 ± 0.022	0.717 ± 0.031	0.419 ± 0.027
RN50x16	**0.440 ± 0.002**	0.791 ± 0.003	**0.305 ± 0.002**	**0.478 ± 0.023**	0.710 ± 0.029	**0.457 ± 0.028**
RN50x64	0.331 ± 0.002	0.837 ± 0.003	0.206 ± 0.002	0.384 ± 0.019	0.759 ± 0.034	0.343 ± 0.025
RN101	0.344 ± 0.002	0.873 ± 0.003	0.214 ± 0.002	0.388 ± 0.021	0.809 ± 0.028	0.317 ± 0.024
ViT-B/32	0.378 ± 0.002	0.876 ± 0.003	0.241 ± 0.002	0.395 ± 0.018	0.813 ± 0.022	0.298 ± 0.019
ViT-L/14	0.324 ± 0.002	0.931 ± 0.002	0.196 ± 0.001	0.329 ± 0.015	0.894 ± 0.018	0.219 ± 0.013
ViT-L/14@336px	0.337 ± 0.002	**0.932 ± 0.002**	0.205 ± 0.002	0.347 ± 0.016	**0.905 ± 0.015**	0.228 ± 0.014

来自一个包含 80 个物体类别、共 1.5M 个物体实例的集合。该数据集分为训练子集和评估子集，分别包含 118K 张和 5K 张图像。我们使用训练集训练或微调我们的方法，并在评估集上报告结果。

查询。 我们使用 MS-COCO 数据集构建查询。对于计数查询类型，我们可以提出如下查询："数据库中有多少个 {object}？"，其中对象可以是 COCO 数据集中包含的任意对象类别。同样，对于最大值查询类型，我们可能会关注数据集中某种特定类型标注最频繁出现的图像。最后，对于包含查询，我们关注标注满足特定条件的图像。

模型。 下面介绍我们在整个实验中使用的神经模型。

对于推理器，我们采用 OFA [32]。OFA 是一种深度学习模型，经过广泛的多模态（文本和图像）任务训练，任务范围从图像描述生成到图像生成，在未见过的任务上也表现出色。OFA 是开源的（代码和权重），目前是表现最佳的多模态模型之一。我们测试了 OFA 的四个不同版本，即 medium、base、large 和 huge，其中最大的版本参数量接近 10 亿。OFA 是一种基于 Transformer 的模型，它为输入（即文本和视觉内容）构建联合表示，并据此生成文本响应。我们再次强调，只要有充足的计算资源，流水线中的这一模块就具有高度并行化能力，因此能够在几秒内生成中间答案。

对于检索器，我们采用了 CLIP 模型 [22]。这些模型通过匹配图像描述和图像，以无监督的对比式方式进行训练。它们接收文本或图像，并将其映射到共享潜在空间中，以便后续进行推理并衡量它们之间的距离；相似的图像—描述对在该空间中彼此接近。我们测试了 8 个不同版本的 CLIP，分别为 RN50、RN101、RN50x4、RN50x16、RN50x64、ViT-B/32、ViT-L/14 和 ViT-L/14@366px。CLIP 的显著特点是生成的嵌入是静态的，也就是说，它们不依赖于查询。这使我们能够预先计算所有图像的嵌入，这意味着推理时只需计算查询的嵌入。计算出嵌入后，还需要一种策略来选择哪些文档被视为相关（并传递给推理器），哪些则不相关。为此，我们设计了三种策略：

(i) *TopK*：在这种情况下，我们计算文档嵌入与查询嵌入之间的点积，对结果进行排序，然后选出 TopK 个文档。

(ii) 阈值：我们计算文本嵌入与图像嵌入之间的余弦相似度，并返回余弦相似度高于特定阈值 τ的所有文档。该阈值取决于所使用的特定 CLIP 模型，范围为 0.15 到 0.4。

(iii) 神经选择器：在这里，我们训练一个小型神经网络，该网络根据 q 和 D 嵌入，返回一个二元结果，以指示文档是否与查询相关，以及是否应将其返回。

实际参数数量取决于所采用的 CLIP 版本，但始终在数千量级。值得注意的是，尽管与推理器所用的拥有 10 亿参数的大型模型相比，这种方法仍然具有更好的可扩展性，但它需要进行"动态"处理；也就是说，选择哪些文档取决于一个神经模型对所有 q、D 对进行评估。

在实际系统中，可以借鉴在线聚合文献中的技术来规避上述一些问题 [12]。实践中，我们可以根据图像嵌入与查询的点积对图像进行排序，然后按预定大小分批处理 w。当满足特定的容差标准时，我们便停止处理；具体而言，当模型预测为相关的文档不超过 c 个时，即停止处理。

这引出了我们的第四种策略，我们称之为混合策略。顾名思义，我们将前面介绍的两种策略——神经选择器和 TopK——结合起来。具体而言，我们取 TopK（K 较小）检索到的文档集合与神经选择器检索到的文档集合的集合并集，并将其传递给推理器。

表 2：查询类型计数的性能结果。PerfectIR 设置作为理想的上限，展现了 MMNDB 框架的全部潜力。另一方面，完整流程（完整）展现出出色的准确率和 Δ 误差，但总误差虽表现良好，却仍未达到 PerfectIR 的水平。我们通过实证表明，这并非由检索器模块引入的噪声所致，因为 NoisyIR 设置取得了出色的结果。相反，这是由于检索器选取了有害文档，这些文档会误导推理器，导致较大的假阳性误差，并最终造成较大的总误差。

	总误差 ↓				Δ误差 ↓			准确率 ↑		
股票	误差	真正例误差	假阳性误差	假阴性误差	Δ 误差	Δ 真阳性误差	Δ 假阳性误差	准确率	准确率 TP	准确率 FP
完美 IR	**0.46 ± 0.07**	0.46 ± 0.07	N/A	N/A	4.64 ± 1.94	4.64 ± 1.91	N/A	0.60 ± 0.02	0.60 ± 0.02	N/A
含噪 IR	0.77 ± 0.16	0.46 ± 0.07	**0.31 ± 0.15**	N/A	**2.66 ± 1.04**	4.64 ± 1.91	**0.31 ± 0.02**	**0.81 ± 0.01**	0.60 ± 0.02	0.92 ± 0.01
受损红外光谱	1.24 ± 0.32	0.46 ± 0.07	0.78 ± 0.32	N/A	4.22 ± 1.41	4.64 ± 1.91	2.31 ± 1.15	0.70 ± 0.02	0.60 ± 0.02	0.76 ± 0.02
Full	1.27 ± 0.17	**0.42 ± 0.07**	0.76 ± 0.13	0.09 ± 0.02	3.33 ± 1.16	4.83 ± 2.03	1.96 ± 0.80	0.73 ± 0.02	**0.61 ± 0.02**	0.75 ± 0.02
FT模型										
完美 IR	**0.14 ± 0.01**	0.14 ± 0.01	N/A	N/A	1.46 ± 0.10	1.46 ± 0.10	N/A	0.67 ± 0.02	0.67 ± 0.02	N/A
含噪 IR	0.22 ± 0.01	0.14 ± 0.01	0.08 ± 0.01	N/A	**0.90 ± 0.06**	1.46 ± 0.10	0.43 ± 0.05	**0.86 ± 0.01**	0.67 ± 0.02	0.93 ± 0.01
受损红外光谱	0.54 ± 0.05	0.14 ± 0.01	0.40 ± 0.05	N/A	1.25 ± 0.08	1.46 ± 0.10	1.04 ± 0.09	0.73 ± 0.01	0.67 ± 0.02	0.73 ± 0.02
Full	0.99 ± 0.06	**0.11 ± 0.01**	0.79 ± 0.06	0.09 ± 0.02	1.10 ± 0.07	**1.42 ± 0.10**	0.99 ± 0.07	0.72 ± 0.01	**0.69 ± 0.02**	0.72 ± 0.02

4.2 结果

本节将介绍支持本文所述思想的实验证据。首先，我们将展示用于测试单个架构组件性能的结果。随后，我们将评估整个流水线。所有指标的结果均与其标准误差一并报告。

我们首先评估检索器策略。我们认为，对于我们的流水线而言，优秀的检索器应具有较高的召回率，因为每个未能检索到的相关文档都会产生错误，而该错误会传播到后续组件，最终影响响应。因此，我们明确倾向于选择召回率较高的模型和策略。我们在 4 种精心设计的策略上，分别测试了 8 个 CLIP 模型版本。为节省篇幅，我们仅展示最终选定设置——混合策略——下各个模型的结果，以及使用最佳模型 ViT-L/14@366px 时不同策略之间的比较。在表1中，可以看到混合策略设置下各个模型的性能。首先可以注意到，尽管 μ 与宏观指标的数值尺度有所不同，至少对于精确率和召回率而言，不同模型之间的排名并未发生明显变化。此外，虽然 ViT-L/14@366px 在 F1值和精确率方面都不是最佳模型，但按召回率衡量时，它表现最佳。事实上，在这方面它始终优于其他模型，唯有其孪生模型 ViT-L/14 例外；二者的性能差异微乎其微。由于参数数量和整体复杂度的差异也几乎可以忽略，我们认为没有理由不采用前者。在表3中，我们报告了所测试的 4 种检索策略的结果。同样，尽管阈值策略的精确率最高，神经选择器（尤其是混合策略）总体表现最佳，其 F1值相当，召回率则高得多。在表4中，我们展示了所测试的各个 OFA 版本之间的性能差异。为避免重复，我们仅展示 COUNT查询类型的结果，因为这些模型之间的差异在各项任务中都大致相同。在此情况下，与检索器不同，我们发现不同模型版本的结果存在显著差异

在所测试的模型中，较大的模型明显大幅优于较小的模型。此外，OFA-huge 的总误差和 Δ 误差均低于 OFA-large，但后者的准确率更高。我们选择 OFA-large 有两个原因：（i）相比另外两项指标，我们更重视准确率；（ii）它的参数量只有 huge 版本的一半（0.5B 对 1B）。我们还报告了 OFA-large 的微调版本（OFA-large FT）的结果。该版本是在相同任务上，使用 $5e-5$ 的学习率，在训练集上对 OFA-large 微调 10 个周期得到的。微调 OFA 模型显著提升了其在 MMNDB任务上的表现。

所跟踪的指标虽然有所调整，但与表 2 所报告结果的测试相同。在这里，我们测试推理器的能力以及完整流程。我们在 4 种不同情境下进行测试，分别考察使用原始模型或微调模型的情况，并报告 10 项不同指标。我们使用 PerfectIR 设置作为基线。在此设置中，检索到的文档集合 D_r 就是实际相关的文档集合，直接取自真实值。当然，这是一个理想设置，我们假设检索器完美无缺；它相当于我们方法的一种上限。完整流程则指我们的实际设置，其中混合策略检索器会传入检索到的文档集合。我们收集的指标分为两类：以"total"为前缀的指标针对整个流水线；其他不含"total"的指标则用于测试推理器 A_p 生成的中间答案。具体而言，准确率指中间答案 A_{p_i}与其真实值完全相同的百分比，然后对所有查询取平均值。随后，我们将这项计算进一步划分为两个互不重叠的集合：真正例（TP）的准确率，即对 D_r 中实际相关文档的准确率；以及假正例（FP）的准确率，即对 D_r 中本不应被检索到的文档的准确率。请注意，根据定义，在 PerfectIR 中，FP 文档集合为空。由于当前任务的查询类型是计数，我们还关心中间答案与真实值的接近程度。我们通过指标 Δ 误差来跟踪这一点。这里，与……类似，

表3：不同检索策略的比较。阈值策略的 F1值和精确率得分较高，而"混合策略"的召回率较高。我们再次选择召回率较高的策略，即混合策略，因为这样推理器模块能够接收尽可能多的相关信息，最终降低流水线误差。

选择策略	μF1	μ召回率	μ精确率	F1	召回率	精确率
Top-K	0.211 ± 0.001	0.683 ± 0.004	0.125 ± 0.001	0.201 ± 0.009	0.852 ± 0.018	0.125 ± 0.009
阈值	**0.351 ± 0.003**	0.226 ± 0.003	**0.791 ± 0.006**	**0.445 ± 0.029**	0.377 ± 0.030	**0.776 ± 0.022**
神经	0.337 ± 0.002	**0.932 ± 0.002**	0.205 ± 0.002	0.343 ± 0.016	0.898 ± 0.018	0.235 ± 0.016
混合策略	0.337 ± 0.002	**0.932 ± 0.002**	0.205 ± 0.002	0.347 ± 0.016	**0.905 ± 0.015**	0.228 ± 0.014

表4：我们在 PerfectIR 设置下测试了不同的神经模型，作为推理器的构建模块。较小的模型显然无法与较大的模型竞争。OFA-huge 的总误差较小，而 OFA-large 的准确率较高。我们选择后者，因为我们更看重准确率而非其他指标，而且它的参数量只有前者的一半。我们还报告了一个微调版本，其表现显著优于原始版本。

模型	总误差↓	Δ 误差↓	准确率 ↑
OFA-base	0.831 ± 0.024	2.876 ± 0.217	0.094 ± 0.014
OFA-medium	0.871 ± 0.013	2.869 ± 0.180	0.074 ± 0.005
OFA-large	0.460 ± 0.073	4.645 ± 1.944	**0.597 ± 0.022**
OFA-huge	**0.392 ± 0.025**	**2.363 ± 0.179**	0.533 ± 0.023
OFA-large FT	**0.138 ± 0.011**	**1.455 ± 0.100**	**0.668 ± 0.018**

准确率指标方面，我们记录中间答案 A_{p_i} 与真实值之间的平均绝对偏差，并对所有查询取平均值。再次，我们将其拆分为两个组成部分，即真正例和假正例。

从这两个指标来看，完整流程的结果与 PerfectIR 版本相比具有竞争力，甚至更好。进一步检查后，我们也能推断出原因。事实上，在完整流程中，假阳性文档会被加入计算中。这些文档中有许多实际上更容易处理，因为它们不包含感兴趣对象，并且可以产生值为 0 的中间答案，从而同时提高完整流程版本的准确率并降低其 Δ 误差。在实验中，我们还注意到，原始模型在某些实例中难以生成有用的中间结果。例如，模型会给出"很多"或"很少"这类模棱两可的答案。通过一些提示工程，明确要求模型"用数字作答"，可以缓解这个问题，但无法彻底解决。因此，如前所述，我们生成了推理器的微调版本，该版本提高了准确率分数，并大幅降低了 Δ 误差。

最后，我们报告总误差指标的结果。在该指标下，我们考察流水线 o的最终结果，计算其与真实值之间的绝对偏差，并对所有查询取平均，再按基数进行归一化。PerfectIR版本在这项任务中表现出色，充分证明了本文所提出任务的可行性。完整流程虽然

表5：MAX查询类型的结果。可以立即看出，微调过程显著提升了MAX查询类型的性能。特别是，我们注意到，微调后的模型较少给出"许多"和"大量"这类不确定的中间答案，而这些答案与该查询密切相关。我们还注意到，与其他查询相比，完整流程设置的结果与PerfectIR非常接近。我们认为，这是因为有害文档的影响较小；也就是说，有害文档不太可能成为MAX查询的候选对象。

股票	总误差↓	Δ 错误↓	准确率 ↑
完美 IR	**2.845 ± 1.759**	**29.263 ± 17.598**	0.188 ± 0.044
含噪 IR	4.576 ± 2.486	41.438 ± 21.343	0.200 ± 0.045
受损红外光谱	4.258 ± 2.035	53.325 ± 23.933	0.188 ± 0.044
Full	4.280 ± 2.014	53.063 ± 24.027	**0.213 ± 0.046**
FT模型			
完美 IR	**0.229 ± 0.035**	1.813 ± 0.271	**0.575 ± 0.056**
含噪 IR	**0.229 ± 0.035**	**1.800 ± 0.273**	0.550 ± 0.055
受损红外光谱	0.303 ± 0.060	2.100 ± 0.320	0.525 ± 0.056
Full	0.317 ± 0.056	2.263 ± 0.342	0.563 ± 0.055

虽然取得了不错的分数，但其表现仍落后于 PerfectIR 设置。为进一步探究这种性能差异，我们将总误差分解为各个组成部分。再次说明，TP 指正确检索到的文档，FP 指错误检索到的文档，而假负例（FN）指本应检索到却未能检索到的文档（根据定义，在 PerfectIR 的情况下，后两项均为零）。我们注意到，两个版本的总误差 TP 实际上相当；在完整流程的情况下略低，因为有少数较难处理的文档未被检索到。进一步检查后，我们发现总误差 FN 几乎可以忽略不计，这意味着总误差差距并非由文档未被检索到造成。实验证据清楚表明，这一差距实际上是由假正例造成的，也就是那些本不应被检索到、却仍然被检索到的文档。

为了进一步研究这一现象，我们设计了一个名为 NoisyIR 的附加设置。在该设置中，我们假设 D_r 与 PerfectIR 中一样，由相关文档集合构成；不过，我们还随机加入了一些不相关文档（300 篇）。我们发现，NoisyIR 设置的表现仅略逊于

表 6：IN 查询类型的结果。我们再次观察到微调后的模型在性能上存在差距。具体而言，微调版本生成的答案对噪声的鲁棒性要强得多。此外，尽管总体结果令人满意，我们观察到完整流水线的错误率有所上升。我们将此归因于有害文档误导推理器，使其错误预测对象是否存在；DamagingIR 设置下的高损失也印证了这一点。

原始模型	总误差 ↓	准确率 ↑
完美 IR	**0.131 ± 0.014**	0.869 ± 0.014
含噪 IR	0.404 ± 0.176	**0.906 ± 0.007**
破坏性 IR	0.829 ± 0.357	0.811 ± 0.013
Full	0.793 ± 0.150	0.672 ± 0.018
FT模型		
完美 IR	**0.060 ± 0.007**	0.940 ± 0.007
含噪 IR	0.085 ± 0.007	**0.946 ± 0.004**
有害信息检索	0.436 ± 0.054	0.838 ± 0.008
Full	0.330 ± 0.015	0.877 ± 0.007

PerfectIR设置，表明我们的模型实际上能够抵御噪声。

在这项实验之后，我们设计了一种新设置，与NoisyIR相同，但负样本文档不再是随机选取的。事实上，我们选择与查询的CLIP嵌入相似度最高的不相关文档。我们将这一设置称为DamagingIR。结果清楚地表明，这些文档会“欺骗”推理器，使其生成错误的中间答案，从而造成较大的假阳性误差，并最终导致总误差显著增大，使PerfectIR版本与完整流程版本之间出现性能差异。

此前 [26] 已经观察到DamagingIR；据作者所知，这一问题尚未得到充分解决。本节末尾，我们将对此问题作更全面的讨论。

在表 6 中，我们展示了包含查询类型的结果。这类查询回答的是“有多少张图片中有 {object}？”之类的问题。在这种情境下，我们考察两个指标，与 COUNT 设置中定义的指标相对应。首先是准确率，即中间结果 A_{p_i} 与各自真实值完全相同的比例。总误差表示满足条件的文档总数与其真实值之间的绝对偏差，之后对所有查询取平均值，并按基数进行归一化。我们可以立即发现，推理器的微调版本通常优于其原始版本。我们还注意到，完整流程取得了不错的结果，尽管表现不及近乎完美的 PerfectIR。再次印证这一点，而且比之前更加明显的是，我们可以将性能下降归因于 DamagingIR，也就是说，假阳性文档成功“欺骗”了模型，让它认为图像中有某个对象而实际上并没有；这一机制下观察到的性能下降也证明了这一点。

最后，我们报告MAX查询类型的结果。这种查询会返回集合中某个特定对象实例数量最多的文档。我们在相同的4种场景下进行测试，并报告三个指标。Δ 和总误差与前述设置中的指标互为镜像，而总准确率则表示找到正确文档的查询所占百分比。这是股票推理器与其微调版本之间差异最显著的场景。我们认为，这一差距源于前文提到的问题：对于某个特定对象实例数量较多的图片，模型会给出“很多”之类不确定的答案，而微调模型不存在这一问题。此外，我们注意到，PerfectIR 版本与完整流程版本之间的差异相当小。这是因为，与另外两种场景不同，假正例文档不太可能成为MAX类型查询的有力候选项，因此不会影响最终结果。我们还发现，即使模型未能检索到实例数量最多的正确文档，找到的图片中该对象的实例数量也相近，这一点可由总误差看出。

总体而言，结果非常令人鼓舞，充分展现了多模态神经数据库的潜力。我们成功构建了一个有效且高效的检索系统，召回率很高。推理模块以及整个流水线都表现出良好的性能和抗噪能力，错误率低、准确率高，同时也能抵御噪声。不过，与信息检索领域的其他系统一样，该系统容易受到 DamagingIR 的影响，导致误报增加。我们认为，解决这一问题能够进一步提升多模态神经数据库的性能，使其接近最优水平。

5 未来研究方向

多模态神经数据库的引入为新颖且令人振奋的研究方向铺平了道路；本节将探讨其中一些较有意思的方向。

本文展示了所提出任务的可行性，但仍有许多尚待探索的开放问题。

首先，任何数据库系统的一项关键特性都是能够更新其中的信息。在典型的数据库系统中，人们理应能够按需删除、添加或修改信息。在我们当前的范式下，这并非易事，仍需进一步研究。

在这一方面，考虑时间对数据库的重要性至关重要。我可以提出这样一个数据库查询：“今年下午 1 点到 3 点之间，我去得最多的地方是哪里？”

此外，我们将范围限定为仅两种模态，具体来说，是一个纯图像数据库。扩展可用模态显然大有裨益。此外，我们不仅可以考虑文档，还可以考虑文档及其元数据。举例来说，每当我们用智能手机拍照时，都会收集各种信息，例如位置和时间，这些信息显然有助于构建此类数据库。

继续聚焦于智能手机领域，最近，视频片段分享在社交网络用户中变得非常流行。在视频上执行类数据库查询是一个开放性问题，带来了

诸多挑战。其中，至关重要的是能够识别跨帧实体，从而有效回答查询。虽然对于文本而言，识别实体（如人物）通常是可行的，但在考虑不同模态时，这一任务要复杂得多。解决这一问题将对多模态神经数据库的发展至关重要。

在我们的演讲中，我们强调，所提出的架构并非解决这一问题的唯一途径。事实上，近期我们见证了大型基础模型解决各种任务的强大能力，其中 ChatGPT 和 GPT-x 模型总体上引领着这一趋势 [23]。我们认为，这些大型基础模型也能推动该领域的发展。然而，这并非易事，还有一些问题需要解决。这些模型需要大量数据进行预训练；由此便引出一个问题：如何将训练期间使用的记忆与实际的多模态数据库隔离开来，以避免知识污染。所谓知识污染，是指预训练所用的数据在生成完全无关情境下的回答时泄露出来这一众所周知的现象。知识污染已被证明给许多应用带来麻烦，据称有些系统甚至泄露了私钥或个人电话号码。此外，这些大型模型能否实现真正的多模态能力，仍是一个开放的研究方向，也是构建对话式多模态系统的一大障碍。

最后，我们考察了最一般情形下的多模态神经数据库。不过，人们可能会对具有更明确准则和目标的特定场景感兴趣。例如，在某些情况下，人们可以准确预知将会出现哪些类型的查询。在这种情况下，可以制定相应策略来优化系统。比如，在传统数据库系统中，为常见查询建立索引或创建视图是普遍做法。为多模态神经数据库实现等效流程仍有待探索。

6 相关工作

多媒体信息检索（MMIR）弥合多模态非结构化数据与结构化数据库系统之间的鸿沟，一直是信息检索领域的核心工作之一 [11]。前者在网络上极为丰富，但与后者相比，难以处理和查询。研究尤其关注基于内容的图像检索 [15, 25, 27]，近年来也关注跨模态检索 [13, 31], ；深度学习 [14] 的最新进展使这些方法成为可能。具体而言，图文检索 [9, 21, 24, 35–37] 领域涌现了大量此类方法。然而，这些系统主要致力于根据给定查询（例如文本）检索相关文档（例如图像）。相比之下，多模态神经数据库专注于针对大型数据集合回答类数据库查询，而当前的跨模态检索方法尚无法做到这一点。

多模态神经模型近年来，能够处理不同形式数据（主要是图像和文本）的多模态神经模型得到了迅速发展，并被应用于各种场景。通常，这通过单个神经多模态编码器[3, 8, 17, 32, 33] ，或通过每种模态各自的编码器来实现，这些编码器经由共享空间共同对齐 [22, 34]。在多模态神经数据库中，我们利用这一特性，将独立的编码器系统用作检索器，预先计算并索引视觉标记，从而减少运行时的计算量和耗时，因为此时只需使用文本编码器来计算查询的文本嵌入。不过，由于计算成本高昂，直接将这些神经模型应用于 MMNDB任务并不具备可扩展性。我们将它们作为架构中的组件，借助它们在其他视觉语言任务中取得的成功。

视觉问答（VQA）大多数此类多模态视觉文本模型都在视觉问答任务上进行评估 [10], ，其目标是根据关于图像的问题生成准确且语义连贯的回答。通常，这些任务需要运用推理能力及其他能力，即使对于当前的神经网络架构而言也并非易事。与多模态神经数据库的任务相比，视觉问答定义在单个图像—问题对上，需要对图像进行推理以回答问题。与多模态神经数据库任务更为接近的是开放域问答（OpenQA） [39] 及其多模态变体 WebQA[7] ，它们旨在针对大规模非结构化文本文档回答自然语言问题。与多模态神经数据库任务相比，它们的范围有所不同，涉及多模态开放域问答；而我们的重点是高效回答针对不同格式（例如图像）的文档集合所提出的类数据库查询。

数据库查询回答人们已投入大量精力，将自然语言表达的查询转换为适用于结构已知数据库的 SQL 查询 [1, 16, 38], ，并且在将这种方法用于半结构化数据和知识库方面也取得了进展 [4, 20]。

近期，Thorne 等人 [28, 29] 提出了 NeuralDB，旨在对一组文本文档执行数据库查询，而无需将数据或查询转换为预定义的数据库模式，而是采用并行神经技术。他们的方法非常有效，但存在以下局限：（i）聚合算子需要预处理和分析；（ii）仅适用于简单查询；（iii）只能处理文本格式的数据。本文以这项研究方法为基础，针对第三项局限展开工作，将原始架构扩展至多模态文档处理。

检索增强模型 近期，关于检索增强神经模型的研究受到了广泛关注 [6]。当前大多数模型着重于通过外部记忆或检索机制增强现有语言模型的能力；该机制会根据输入查询检索相关文档，从而减少参数数量和非事实性错误 [19]。

7 结论

本文提出通过引入多模态神经数据库，拓展多媒体信息检索领域。多模态神经数据库有望大规模回答复杂的类数据库查询，这类查询涉及跨多种模态的推理。我们提出了一种解决这一问题的首个系统化方法，其架构由三个模块组成：检索器、推理器和聚合器，并通过一系列丰富的实验展示了该系统的可行性与潜力。我们还探讨了本文所介绍系统可能催生的未来研究方向。多模态神经数据库确立了新的研究

议程，力求同时充当信息检索与数据库系统之间的桥梁，并缩小二者之间的差距。我们相信，多模态神经数据库不仅有望显著推动多媒体信息检索领域的发展，也有望推动整个信息检索领域的发展。

8 致谢

本研究部分由欧盟资助的意大利大学与研究部国家复苏与韧性计划下的 FAIR（PE0000013）和 SERICS（PE00000014）项目，以及欧洲研究委员会启动资助 No. 802554（SPECGEO）和 PRIN 2020项目 n.2020TA3K9N“LEGO.AI”提供支持。

参考文献

[1] Ion Androutsopoulos、Graeme D Ritchie 和 Peter Thanisch。1995年。数据库自然语言接口——简介。自然语言工程1, 1（1995年），29–81。[2] Alexei Baevski、Wei-Ning Hsu、Qiantong Xu、Arun Babu、Jiatao Gu 和 Michael Auli。2022年。Data2vec：一种用于语音、视觉和语言自监督学习的通用框架。载于国际机器学习大会。PMLR，1298–1312。[3] Hangbo Bao、Wenhui Wang、Li Dong、Qiang Liu、Owais Khan Mohammed、Kriti Aggarwal、Subhojit Som 和 Furu Wei。2021年。Vlmo：采用模态专家混合的统一视觉-语言预训练。*arXiv* 预印本 *arXiv:2111.02358*（2021年）。[4] Jonathan Berant、Andrew Chou、Roy Frostig 和 Percy Liang。2013年。基于问答对对 Freebase 进行语义解析。载于*2013*年自然语言处理经验方法会议论文集。1533–1544。[5] Alessandro Bozzon 和 Piero Fraternali。2010年。多媒体与多模态信息检索。载于*Search Computing*。Springer，135–155。[6] Deng Cai、Yan Wang、Lemao Liu 和 Shuming Shi。2022年。检索增强文本生成的最新进展。载于第*45*届*ACM SIGIR*信息检索研究与发展国际会议论文集（西班牙马德里）（*SIGIR '22*）。计算机协会，美国纽约州纽约市，3417–3419。https://doi.org/10.1145/3477495.3532682[7] Yingshan Chang、Mridu Narang、Hisami Suzuki、Guihong Cao、Jianfeng Gao 和 Yonatan Bisk。2021年。WebQA：多跳与多模态问答。*CoRR* abs/2109.00590（2021年）。arXiv:2109.00590 https://arxiv.org/abs/2109.00590[8] Xi Chen、Xiao Wang、Soravit Changpinyo、AJ Piergiovanni、Piotr Padlewski、Daniel Salz、Sebastian Goodman、Adam Grycner、Basil Mustafa、Lucas Beyer 等。2022年。Pali：联合缩放的多语言图像-语言模型。*arXiv* 预印本 *arXiv:2209.06794*（2022年）。[9] Mengjun Cheng、Yipeng Sun、Longchao Wang、Xiongwei Zhu、Kun Yao、Jie Chen、Guoli Song、Junyu Han、Jingtuo Liu、Errui Ding 等。2022年。ViSTA：用于跨模态检索的视觉与场景文本聚合。载于*IEEE/CVF*计算机视觉与模式识别会议论文集。5184–5193。[10] Yash Goyal、Tejas Khot、Douglas Summers-Stay、Dhruv Batra 和 Devi Parikh。2017年。让 VQA 中的 V 发挥作用：提升图像理解在视觉问答中的作用。载于*IEEE*计算机视觉与模式识别会议论文集。6904–6913。[11] Alon Y Halevy、Oren Etzioni、AnHai Doan、Zachary G Ives、Jayant Madhavan、Luke K McDowell 和 Igor Tatarinov。2003年。跨越结构鸿沟。载于*CIDR*。[12] Joseph M. Hellerstein、Peter J. Haas 和 Helen J. Wang。1997年。在线聚合。*SIGMOD Rec.* 26, 2 1997 171 182 10 1145 253262 253291（6月），–。[13] 2019 https://doi.org/。/。Peng Hu、Liangli Zhen、Dezhong Peng 和 Pei Liu。可扩展的跨模态检索深度多模态学习。载于第*42*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集。635–644。[14] Yann LeCun、Yoshua Bengio 和 Geoffrey Hinton。2015年。深度学习。《自然》521, 7553（2015年），436–444。[15] Michael S Lew、Nicu Sebe、Chabane Djeraba 和 Ramesh Jain。2006年。基于内容的多媒体信息检索：现状与挑战。*ACM*多媒体计算、通信与应用汇刊（*TOMM*）2, 1（2006年），1–19。[16] Fei Li 和 Hosagrahar V Jagadish。2014年。为关系数据库构建交互式自然语言接口。*VLDB*基金会论文集8, 1（2014年），73–84。[17] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。2023年。BLIP-2：利用冻结图像编码器和大语言模型引导语言-图像预训练。*arXiv* 预印本 *arXiv:2301.12597*（2023年）。[18] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。2014年。Microsoft COCO：情境中的常见物体。载于计算机视觉—*ECCV 2014*：第*13*届欧洲会议，瑞士苏黎世，*2014*年*9*月*6*日至*12*日，论文集，第*V*部分第*13*卷。Springer，740–755。[19] Grégoire Mialon、Roberto Dessì、Maria Lomeli、Christoforos Nalmpantis、Ram Pasunuru、Roberta Raileanu、Baptiste Rozière、Timo Schick、Jane Dwivedi-Yu、Asli Celikyilmaz、Edouard Grave、Yann LeCun 和 Thomas Scialom。2023年。增强语言模型：综述。https://doi.org/ 10 48550 2302 07842 。/ARXIV.。[20] 2015 Panupong Pasupat 和 Percy Liang。组合语义解析半结构化表格。*arXiv* 预印本 *arXiv:1508.00305*（2015年）。[21] Leigang Qu、Meng Liu、Jianlong Wu、Zan Gao 和 Liqiang Nie。2021年。面向图文检索的动态模态交互建模。载于第*44*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集。1104–1113。[22] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021年。通过自然语言监督学习可迁移视觉模型。载于国际机器学习大会。PMLR，8748–8763。[23] Alec Radford、Jeff Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。2019年。《语言模型是无监督的多任务学习者》。OpenAI 博客。https://openai.com/blog/better-language-models/[24] Jun Rao、Fei Wang、Liang Ding、Shuhan Qi、Yibing Zhan、Weifeng Liu 和 Dacheng Tao。2022年。性能提升来自何处？——关于图文检索可复现性的疑问。载于第*45*届*ACM SIGIR*信息检索研究与发展国际会议。2727–2737。[25] Yong Rui、Thomas S. Huang 和 Shih-Fu Chang。1999年。图像检索：当前技术、有前景的方向和未解决的问题。《视觉通信与图像表示期刊》10，1（1999），39–62。https://doi.org/10.1006/jvci. 1999.0413[26] Artsiom Sauchuk、James Thorne、Alon Halevy、Nicola Tonellotto 和 Fabrizio Silvestri。2022年。相关性在自然语言处理任务中的作用。载于第*45*届*ACM SIGIR*信息检索研究与发展国际会议。1785–1789。[27] A.W.M. Smeulders、M. Worring、S. Santini、A. Gupta 和 R. Jain。2000年。早期发展阶段结束时的基于内容的图像检索。《*IEEE* 模式分析与机器智能汇刊》22，12（2000），1349–1380。https://doi.org/10.1109/34.895972[28] James Thorne、Majid Yazdani、Marzieh Saeidi、Fabrizio Silvestri、Sebastian Riedel 和 Alon Halevy。2021年。对文本进行数据库推理。载于计算语言学协会第*59*届年会暨自然语言处理国际联合会议第*11*届会议论文集（第*1*卷：长篇论文）。3091–3104。[29] James Thorne、Majid Yazdani、Marzieh Saeidi、Fabrizio Silvestri、Sebastian Riedel 和 Alon Halevy。2021年。从自然语言处理到神经数据库。载于*VLDB*基金会论文集，第14卷。VLDB 基金会，1033–1039。[30] Alex Wang、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel R Bowman。2018年。GLUE：自然语言理解多任务基准与分析平台。*arXiv* 预印本 *arXiv:1804.07461*（2018）。[31] Kaiye Wang、Qiyue Yin、Wei Wang、Shu Wu 和 Liang Wang。2016年。跨模态检索综合综述。*arXiv* 预印本 *arXiv:1607.06215*（2016）。[32] Peng Wang、An Yang、Rui Men、Junyang Lin、Shuai Bai、Zhikang Li、Jianxin Ma、Chang Zhou、Jingren Zhou 和 Hongxia Yang。2022年。通过简单的序列到序列学习框架统一架构、任务和模态。*arXiv* 预印本 *arXiv:2202.03052*（2022）。[33] Wenhui Wang、Hangbo Bao、Li Dong、Johan Bjorck、Zhiliang Peng、Qiang Liu、Kriti Aggarwal、Owais Khan Mohammed、Saksham Singhal、Subhojit Som 等。2022年。图像作为外语：面向所有视觉及视觉语言任务的 BEiT 预训练。*arXiv* 预印本 *arXiv:2208.10442*（2022）。[34] Jiahui Yu、Zirui Wang、Vijay Vasudevan、Legg Yeung、Mojtaba Seyedhosseini 和 Yonghui Wu。2022年。CoCa：对比式图像描述模型是图文基础模型。*arXiv* 预印本 *arXiv:2205.01917*（2022）。[35] Tan Yu、Hongliang Fei 和 Ping Li。2022年。用于快速跨模态搜索的 Cross-Probe BERT。载于第*45*届*ACM SIGIR*信息检索研究与发展国际会议。2178–2183。[36] Tan Yu、Hongliang Fei 和 Ping Li。2022年。用于快速、可扩展文本图像检索的 U-BERT。载于*2022*年 *ACM SIGIR* 信息检索理论国际会议论文集。193–203。[37] Tan Yu、Yi Yang、Yi Li、Lin Liu、Hongliang Fei 和 Ping Li。2021年。用于高效跨模态检索的异构注意力网络。载于第*44*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集。1146–1156。[38] Jichuan Zeng、Xi Victoria Lin、Caiming Xiong、Richard Socher、Michael R Lyu、Irwin King 和 Steven CH Hoi。2020年。Photon：稳健的跨领域文本到 SQL 系统。*arXiv* 预印本 *arXiv:2007.15280*（2020）。[39] Fengbin Zhu、Wenqiang Lei、Chao Wang、Jianming Zheng、Soujanya Poria 和 Tat-Seng Chua。2021年。检索与阅读：开放领域问答综合综述。*CoRR* abs/2101.00774（2021）。arXiv:2101.00774 https://arxiv.org/abs/2101.00774