UniversalRAG：在具有多种模态和粒度的多种语料库中进行检索增强生成

Woongyeong Yeo[1*] **Kangsan Kim**[1*] **Soyeong Jeong**[1] **Jinheon Baek**[1] **Sung Ju Hwang**[1,2]
KAIST[1] DeepAuto.ai[2]
{ wgcyeo, kksan07, starsuzi, jinheon.baek, sungju.hwang } @kaist.ac.kr

Abstract

检索增强生成（RAG）通过利用与查询相关的外部知识来支持模型响应，已在提高事实准确性方面显示出显著的潜力。然而，大多数现有的 RAG 方法仅限于文本语料库，尽管最近的研究已将 RAG 扩展到其他模态如图像和视频，但这些方法通常只在单一模态特定的语料库上操作。相比之下，现实世界的查询在所需知识类型上差异很大，单一类型的知识源无法解决这些问题。为此，我们引入了 UniversalRAG，这是一种新颖的 RAG 框架，旨在从多种模态和粒度的异构来源中检索和整合知识。具体而言，我们观察到将所有模态强行融合到一个从单一组合语料库得出的统一表示空间中会导致模态差距，即检索过程中倾向于偏好与查询同模态的项目，为了解决这个问题，我们提出了一种模态感知路由机制，该机制能够动态识别最合适的模态特定语料库，并在其中进行目标检索。此外，超越模态的考虑，我们将每种模态组织成多个粒度级别，从而实现针对查询复杂性和范围的精细化检索。我们在跨越多种模态的 8 个基准测试上验证了 UniversalRAG，证明其优于模态特定和统一基准。我们的项目页面在 https://universalrag.github.io 。

1 介绍

近年来，我们见证了大型语言模型（LLMs）在各种任务中的卓越表现，例如问答 (**??**)，以及在各种服务中被广泛采用，例如 ChatGPT，来增强用户的日常生活。然而，LLMs 往往会生成事实错误或误导性的信息，尤其是在它们在训练过程中较少或未接触的话题上（例如近期事件）(**??**)。为了解决这个问题，检索增强生成（RAG）作为一种有前途的方法出现了，它允许模型的回答以从外部知识源检索的查询相关知识为基础，从而提高事实准确性 (**???**)。

然而，尽管其有效性，现有的 RAG 方法通常是为单一语料库和模态设计的，这限制了它们解决需要不同类型知识源的用户查询的能力。在实践中，如图 1 所示，用户查询在所需知识的类型上差异很大：有些最好通过文本回答（例如，表面层次的事实和定义），有些需要从图像中获取视觉理解（例如，对象的空间关系），还有些需要视频支持的时间推理（例如，包含动态场景的逐步说明）。相反，RAG 领域的起源主要集中于文本语料库 (**???**)，尽管最近的努力已将其扩展到文本以外的模态（如图像和视频）(**???**)，现有的 RAG 方法通常是针对特定模态和语料库的；因此，它们可能不适合用作一个灵活处理广泛查询的通用框架，其知识要求各不相同。

在这项工作中，我们提出了 UniversalRAG，这是一种新颖的 RAG 框架，它将分布于多个特定模态语料库的知识结合在一起，包括文本、图像和视频源，并利用它们在一个通用工作流程中生成基于基础的查询响应。为了实现这个目标，最直接的方法可能是汇总所有收集的异构知识语料库中的条目，并使用多模态编码器将其嵌入到统一的表示空间中（这种编码器通常被训练来对齐来自不同模态的输入，如果它们在语义上相似）。然而，尽管有这样的对齐努力，我们发现这种策略受到模态差距的影响，即根据模态而非其语义含义对输入进行聚类的倾向（在图 2 中可视化），这在不同设置下的先前工作中也有类似观察 (**??**)。因此，检索变得偏向于与查询共享相同模态的知识来源，忽视了来自其他模态的相关内容。

为了解决这一挑战，我们采取了不同的方法：引入一种感知模态的路由策略，而不是依赖将所有模态强制转换为共享表示的统一嵌入空间。具体来说，UniversalRAG 根据给定查询的模态需求动态确定检索最合适的知识来源，然后将检索过程路由到相应的模态特定语料库。值得注意的是，这一策略不仅通过避免直接的跨模态比较来绕过模态差异，还通过扩展路由逻辑而不修改现有模态特定的检索器，实现了新模态的无缝集成。

除了模态之外，另一个重要的维度是数据粒度（语料库中每个条目的大小或单位），这在检索精度和生成质量中扮演至关重要的角

[*] Equal contribution

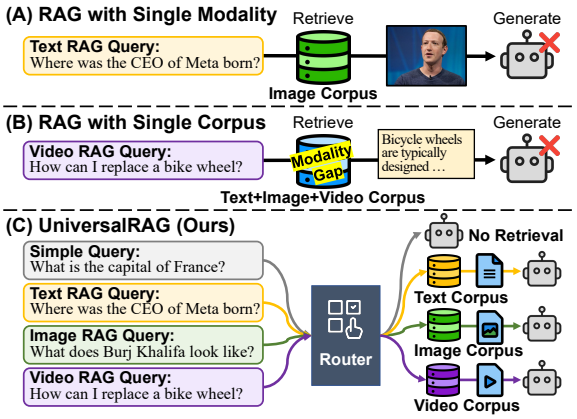

Figure 1: （a, b）现有 RAG 方法的局限性及（c）提出的 RAG 框架 UniversalRAG 的说明。

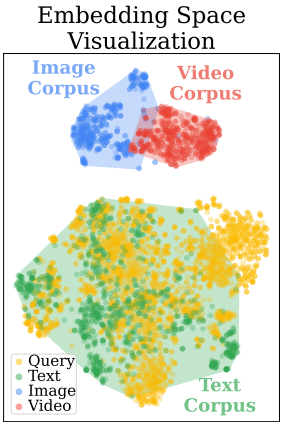

Figure 2: t-SNE 可视化统一嵌入空间。

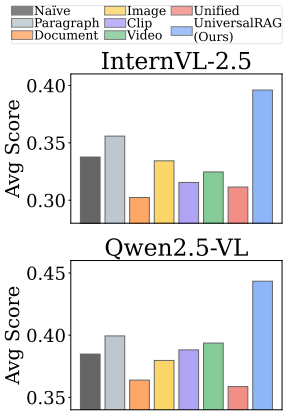

Figure 3: 基准方法和 UniversalRAG 的平均得分。

色（??），因为即使在同一模态中，不同的查询也受益于不同层次的粒度。这是因为过于细致的条目可能会稀释上下文，而过于粗略的条目可能会将不相关的信息捆绑在一起。例如，一个复杂的分析问题可能需要长篇文档或完整的视频来捕获足够的上下文，而一个简单的事实性问题可能最好以单个段落或短视频片段呈现。

为了适应这一方面，我们进一步将每种模态分解为多个粒度层次，将它们组织成不同的语料库：文本文档被进一步分割成段落并存储在段落级语料库中，类似地，完整长度的视频被分割成短片段并存储，而图像由于本身就是碎片化的，保持不变。总体而言，有了这些模态感知和粒度感知的语料库（包括段落、文档、图像、片段和视频）以及一个附加的无检索选项以高效处理简单的查询（不需要外部知识），我们的 UniversalRAG 能够动态地将每个查询引导至最相关的知识来源，最终支持现实世界用户的多样化信息需求。我们在包含不同模态的 8 个基准测试上验证了 UniversalRAG（???????）。UniversalRAG 在平均得分上优于所有基线，表明在处理多样化查询时能表现出强大的性能。我们还通过实验结果研究了多模态和多粒度语料库的有效性。

2 方法

在本节中，我们介绍了 UniversalRAG，这是一种新的 RAG 框架，能够根据给定的查询，从涵盖多种模态和粒度的多样化语料库中检索知识。

2.1 预备知识

我们从预备知识开始，介绍 LVLMs 和 RAG 的正式描述。

大型视觉语言模型 为了扩展大型语言模型（LLMs）在文本之外的强大能力，并支持对视觉输入（如图像和视频）的理解，最近通过将视觉编码器整合到 LLMs 中引入了大型视觉-语言模型（LVLMs），使其能够处理包括图像和视频在内的文本和视觉输入。形式上，LVLM 的输入是一个序列 $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$，该序列可能包括文本和视觉标记，并生成一个输出标记序列 $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]$，表示为：$\boldsymbol{y} = \texttt{LVLM}(\boldsymbol{x})$。然而，尽管具有多模态能力，LVLMs 仍然局限于参数化知识，并常常在需要详细或基于具体信息的查询上遇到困难，这超出了预训练时所编码的内容。

为了应对仅限参数模型的前述限制，检索增强生成（RAG）从大型外部语料库中检索与查询相关的信息，并将其融入生成过程中。具体来说，在检索步骤中，$\texttt{Retriever}$ 从语料库 $\mathcal{C}$ 中选择相关上下文 $\boldsymbol{c}$，形式化为 $\boldsymbol{c} = \texttt{Retriever}(\boldsymbol{q}; \mathcal{C})$，其中 $\boldsymbol{c} \in \mathcal{C}$。在后续的生成步骤中，一个 LVLM 生成响应 $\boldsymbol{a}$，这以输入查询和检索到的上下文为条件，表示为 $\boldsymbol{a} = \texttt{LVLM}(\boldsymbol{q}, \boldsymbol{c})$。然而，大多数现有的 RAG 方法仅限于从单一模态（例如，仅限图像）内的单一语料库中检索，从而限制了它们处理多样化真实世界查询时对多模态信息的需求能力。

考虑到真实场景中的外部知识经常跨越多种模态——例如文本、图像和视频——我们定义了三个特定模态的语料库：文本语料库 $\mathcal{C}_{\texttt{text}} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n\}$、图像语料库 $\mathcal{C}_{\texttt{image}} = \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_m\}$ 和视频语料库 $\mathcal{C}_{\texttt{video}} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$。处理这种异构数据的一种常见方法是使用多模态编码器将所有项目统一到一个共享的嵌入空间中，从而形成一个统一的语料库 $\mathcal{C}_{\texttt{unified}} = \mathcal{C}_{\texttt{text}} \cup \mathcal{C}_{\texttt{image}} \cup \mathcal{C}_{\texttt{video}}$，在该空间中每个项目都被表示为向量（??），然后以 $\boldsymbol{c} = \texttt{Retriever}(\boldsymbol{q}; \mathcal{C}_{\texttt{unified}})$ 的形式进行检索。然而，我们的实验揭示了在这样统一的空间中存在明显的模态差距——如图 2 所示——其中查询本质上是文本的，往往更倾向于与文本语料库的项目对齐，而不考虑实际所需

的模态。因此，即使查询需要视觉或时间理解，检索器也会返回基于文本的内容，从而导致次优或不相关的响应。这个观察突显了统一检索策略的基本局限性，并激发了为不同模态维持独立特征空间的必要性。

现在我们开始介绍 UniversalRAG，这是一种新颖的框架，可以动态识别并将查询路由到最合适的检索知识的模式和粒度。

为了应对检索中的模态差异，我们为每种模态维持独立的嵌入空间，将整个语料库组织成三个不同的子语料库：$\mathcal{C}_{\texttt{text}}$ ，$\mathcal{C}_{\texttt{image}}$ 和 $\mathcal{C}_{\texttt{video}}$ ，其中每个都由模态特定的向量表示组成。然后我们引入一个路由模块 Router ，该模块动态选择每个查询最合适的模态。具体来说，给定一个查询 $\boldsymbol{q}$ ，Router 预测与查询相关的模态 $r \in$ { 'Text', 'Image', 'Video' }，形式化为 $r = \texttt{Router}(\boldsymbol{q})$ 。一旦确定了模态 r ，模态特定的 Retriever 从相应的语料库 $\mathcal{C}_r$ 中选择相关项 $\boldsymbol{c}$ ，并且 LVLM 基于该查询和检索的内容生成最终响应。然而，尽管这种设计减少了模态差异，仅根据模态区分语料库可能仍不够，因为不同的查询甚至在同一模态内也可能需要不同层次的细粒度。

为了灵活地满足不同查询的各种信息需求，我们扩展了 UniversalRAG，使其能够在每种模态内跨多个粒度级别运行，为文本和视频模态构建了两个语料库级别——细粒度和粗粒度。具体而言，虽然文本语料库最初是以段落级别组织的，每个项目通常包含关于单一实体的知识，但某些复杂查询需要跨多个段落进行推理。为了解决这个问题，我们构建了一个文档级的语料库 $\mathcal{C}_{\texttt{document}} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_l\}$ ，其中每个 $\boldsymbol{d}$ 是通过串联多个段落并编码结果文本获得的文档的向量表示。另一方面，原始视频语料库由完整视频组成，通常可能超过一小时长度，当某些问题仅需一个短片段即可回答时，检索整个视频效率不高。因此，我们将每个完整视频分段成多个固定时长的片段，构建一个片段级的语料库 $\mathcal{C}_{\texttt{clip}} = \{\boldsymbol{k}_1, \ldots, \boldsymbol{k}_p\}$ ，其中每个 $\boldsymbol{k}$ 表示从原始完整视频提取的剪辑视频片段的表示。注意，由于图像本质上是细粒度的，我们不会对图像语料库进行额外分割而保持原样。为此，Router 作出的路由决策分为六类：{ 'None'、'Paragraph'、'Document'、'Image'、'Clip'、'Video' }，并检索过程形式化如下：最后，LVLM 在检索内容 $\boldsymbol{c}$ 的条件下生成最终响应，该内容反映了为给定查询 $\boldsymbol{q}$ 确定的最合适的模态和粒度。此外，如果不需要检索（即 $\boldsymbol{c} =$ None ），LVLM 直接基于 $\boldsymbol{q}$ 生成响应，而无需任何额外上下文。

这里，我们探讨了两个路由器设计，它负责根据查询动态选择检索模式和粒度。

无需训练的路由器 无需训练的路由器利用预训练的 LLM 的固有知识和推理能力，将查询分类为适当的检索类型，而无需额外的训练。具体来说，对于一个查询 $\boldsymbol{q}$ ，LLM 会被给出描述路由任务的详细指令，并附有几个上下文示例，然后从六个预定义选项中预测最合适的检索类型。

我们进一步探讨了训练路由模块以实现更准确的路由决策。然而，这一策略的一个关键挑战是缺乏用于最佳语料选择的真实查询标签对。为了解决这个问题，我们利用现有基准的模态特定归纳偏差来构建路由器的训练数据集——也就是说，我们假设每个基准测试主要与特定的模态和检索粒度相关。具体来说，对于文本问答基准，来自仅基于模型参数化知识回答的数据集的查询被标记为“无”，来自单跳 RAG 基准的查询被标记为“段落”，而来自多跳 RAG 基准的那些被标记为“文档”。同样，来自基于图像的 RAG 基准的查询被标记为“图像”。对于视频问答基准，关注于视频中局部事件或特定时刻的查询——例如在特定时间戳识别动作——被标记为“剪辑”，而那些需要理解完整故事线或更广泛时间上下文的查询被标记为“视频”。利用这个构建的数据集，我们训练路由器在推理时预测给定查询的适当检索类型。

3 实验装置

在本节中，我们解释实验设置，包括数据集、模型、评估指标和实现细节。

为了评估我们的框架在不同模态下的性能，我们编制了一个全面的问答基准，涵盖六种不同的检索设置：无检索、段落、文档、图像、剪辑和视频。

对于无检索设置，我们使用 MMLU (**?**)，它评估模型的知识而不需要外部来源。对于文本检索设置，我们结合了三个基准：SQuAD (**?**) 和自然问题（NQ） (**?**) 作为单跳 RAG 基准，其中检索单元是段落，而 HotpotQA (**?**) 作为多跳 RAG 基准，其中检索单元是文档。对于图像检索设置，我们使用 WebQA (**?**) 的一个子集，由需要在外部图像中定位的查询组成。最后，对于视频检索设置，我们使用来自 LVBench (**?**) 、VideoRAG-Wiki (**?**) 和 VideoRAG-Synth (**?**) 的查询。其中，针对短片段或局部段落的查询被归类为剪辑级查询，而需要理解长视频或整个视频的查询被视为视频级查询。

检索语料库 为了支持跨模态和粒度的检索，我们为每种模态和粒度构建了专门的检索语料库。对于段落级检索，我们使用从 SQuAD 和自然问题中得出的维基百科段落语料库 (**?**) 。在

Table 1: 各种不同的 RAG 变体，包括 UniversalRAG 和基准方法，在特定模态基准上的结果。我们的方法论 UniversalRAG 由 彩色细胞 表示，包括分别为 DistilBERT 和 T5-Large 训练的方案，而 GPT-4o 以不需训练的方式运作。加粗字体表示每个指标的最佳性能；underline 表示 UniversalRAG 方法中的第二佳表现。R-L 和 BERT 分别指 ROUGE-L 和 BERTScore。

		Text							Image		Video					Avg.
		MMLU	SQuAD		NQ		HotpotQA		WebQA		LVBench	VideoRAG-Wiki		VideoRAG-Synth		
	Models	Acc	EM	F1	EM	F1	EM	F1	R-L	BERT	Acc	R-L	BERT	R-L	BERT	
InternVL-2.5-8B	Naïve	64.50	7.82	16.86	24.71	38.11	12.92	20.87	40.63	90.30	28.60	15.74	84.20	14.93	85.73	33.76
	Paragraph	64.50	20.62	30.97	35.14	47.89	14.45	23.05	35.72	89.13	29.19	14.82	84.08	19.15	86.53	35.59
	Document	51.50	6.33	13.72	23.57	32.66	19.71	28.49	28.92	87.45	28.80	13.28	83.75	18.51	86.12	30.24
	Image	54.50	7.41	15.74	23.57	32.96	13.11	20.18	46.50	91.32	31.64	17.26	83.79	20.72	87.02	33.43
	Clip	53.50	4.58	12.52	13.86	21.82	9.38	16.51	39.53	90.27	35.36	18.76	86.38	27.37	89.34	31.55
	Video	59.50	3.77	11.55	14.43	22.98	9.95	16.95	40.08	90.51	33.59	19.23	86.35	28.23	89.45	32.47
	Unified	59.00	4.72	12.81	17.00	27.87	9.67	17.08	41.71	90.27	27.23	15.87	83.96	19.03	86.46	31.15
	Random	55.50	7.68	16.20	22.71	32.79	12.82	20.37	38.37	89.71	31.15	16.55	84.79	21.02	87.37	32.27
	GPT-4o	61.50	18.33	28.09	33.43	46.28	17.80	26.10	45.39	91.10	33.01	14.65	84.11	19.68	86.83	37.56
	DistilBERT	62.00	19.14	29.71	33.57	46.45	19.43	28.35	46.40	91.29	35.16	19.23	86.35	28.15	89.44	39.60
	T5-Large	63.00	20.49	30.87	35.00	47.78	18.09	26.90	45.47	91.09	34.28	19.18	86.32	27.71	89.33	39.36
	Oracle	64.50	20.62	30.97	35.14	47.89	19.71	28.49	46.50	91.32	35.65	18.79	86.38	27.45	89.35	40.34
Qwen2.5-VL-7B	Naïve	73.00	10.78	19.85	17.29	25.71	18.47	25.47	61.26	94.39	29.38	14.26	83.04	10.52	84.34	38.48
	Paragraph	72.00	23.58	34.25	38.43	49.37	19.04	26.54	53.42	92.65	27.13	14.88	83.30	12.62	84.93	39.94
	Document	66.50	8.76	15.13	23.14	31.02	20.96	28.78	54.37	92.71	27.23	14.78	83.33	11.39	84.50	36.39
	Image	68.50	11.19	18.30	16.14	23.14	16.94	23.01	64.39	94.73	30.17	16.17	83.62	13.35	85.10	37.97
	Clip	68.50	10.65	17.66	15.14	22.69	16.46	22.86	62.78	94.38	33.50	18.39	85.04	20.53	87.75	38.81
	Video	70.00	11.05	18.07	14.00	21.42	17.42	23.74	63.89	94.54	32.81	19.34	85.64	23.31	88.52	39.36
	Unified	71.50	7.95	15.06	12.29	19.81	14.35	21.11	55.64	93.07	30.14	15.00	82.74	11.38	84.16	35.87
	Random	72.00	12.67	19.49	20.86	29.23	18.47	25.09	59.67	93.85	28.80	15.96	83.91	15.63	86.01	38.50
	GPT-4o	71.50	21.70	30.62	36.57	48.11	20.19	28.00	63.58	94.58	32.42	14.87	83.29	12.69	85.01	42.61
	DistilBERT	73.50	21.83	32.52	37.57	48.27	20.96	28.87	64.20	94.70	33.01	19.34	85.64	23.18	88.46	44.34
	T5-Large	72.50	23.58	34.12	38.29	49.22	19.52	27.32	63.53	94.55	33.01	19.34	85.62	22.85	88.38	44.01
	Oracle	73.00	23.58	34.25	38.43	49.37	20.96	28.78	64.39	94.73	33.20	18.43	85.05	20.70	87.80	44.35
Phi-3.5-Vision-Instruct	Naïve	61.00	9.30	18.32	10.43	18.49	14.26	21.01	54.01	93.01	29.58	15.94	83.64	34.58	90.66	35.16
	Paragraph	58.50	22.24	33.38	34.86	46.07	17.03	24.82	59.90	93.65	28.21	17.31	85.02	32.11	89.94	39.61
	Document	52.50	6.47	12.95	16.43	24.80	17.80	25.86	57.46	93.18	29.09	14.05	84.18	33.27	90.18	34.95
	Image	55.50	8.36	15.20	9.86	15.73	13.68	18.70	63.25	94.13	31.15	15.16	85.02	34.18	90.32	35.24
	Clip	54.00	7.68	13.38	11.43	16.48	13.40	18.73	60.22	93.60	35.06	19.50	86.04	36.34	90.97	35.62
	Video	53.00	8.09	14.05	9.29	15.09	13.11	17.91	59.90	93.50	32.13	19.33	86.14	36.71	90.95	34.56
	Unified	55.00	6.47	14.63	5.86	13.48	11.87	18.46	51.05	92.67	28.50	18.09	84.76	35.78	90.82	32.47
	Random	55.50	9.57	16.67	14.00	21.72	15.12	21.15	58.84	93.48	29.77	16.94	85.02	33.88	90.31	35.28
	GPT-4o	57.50	20.35	30.18	32.86	44.19	16.84	25.09	62.88	94.11	32.62	16.79	84.95	32.01	89.93	40.48
	DistilBERT	57.00	20.62	31.90	33.71	44.87	18.18	26.30	63.39	94.14	34.87	19.33	86.14	36.48	90.91	41.78
	T5-Large	58.50	22.37	33.36	34.71	45.94	17.61	25.98	62.69	94.04	34.97	19.33	86.10	36.31	90.87	42.08
	Oracle	61.00	22.24	33.38	34.86	46.07	17.80	25.86	63.25	94.13	34.57	19.53	86.04	36.20	90.97	42.46

文档级检索的情况下，我们遵循 LongRAG (**?**) 的构建方法来建立综合维基百科文章的语料库。关于图像检索，我们使用由 WebQA 数据集中图像组成的检索语料库。对于与视频相关的检索，我们定义了两个独立的语料库：视频检索语料库由来自 LVBench 和 VideoRAG 的全长 YouTube 视频组成，而片段级检索语料库则由从这些视频中提取的剪辑片段组成。数据集构建的进一步细节见附录 A 。

Table 2: 路由器在域内和域外数据集上的检索方法中的准确性和生成性能。

	In-Domain		Out-Domain	
Models	Router Acc	Avg Score	Router Acc	Avg Score
Random	16.67	32.27	16.67	29.99
Unified	16.67	31.15	16.67	28.92
GPT-4o	57.23	37.56	69.49	36.85
DistilBERT	66.42	39.60	39.62	32.58
T5-Large	59.99	39.36	47.47	35.27
Ensemble	63.99	39.43	61.55	35.22

3.1 模型

我们将 UniversalRAG 与以下八个不同的基线进行比较：1）Naïve 不检索外部知识直接回答查询。2）Paragraph，3）Document，4）Image，5）Clip，和 6）Video 仅从其各自的特定模态语料库中检索信息。7）Unified 在多模态编码器 InternVideo2 (**?**) 的单一统一嵌入空间上检索不同语料库中的所有数据，类似于 (**??**) 。8）Random 随机选择一个特定模态语料库进行检索。我们还实现了 UniversalRAG 的三个变体，它们的检索组件不同。9）GPT-4o 采用 GPT-4o (**?**) 作为无需训练的路由器。10）DistilBERT 和 11）T5-Large 分别使用在路由数据集上训练的 DistilBERT (**?**) 和 T5-Large (**?**) 。12）Oracle 是我们的理想设置，其中每个查询都被路由到最合适的特定模态语料库，模拟完美的路由。

我们使用以下指标评估 UniversalRAG 和基准模型的性能。对于有选择题的基准，我们使用 Top-1 准确率（Acc），该指标显示有多少问题得到正确答案。对于答案少于几个词的基准，我们使用精确匹配（EM），它检查预测的答案是否完全匹配真实答案，以及 F1 分数（F1），它测量答案和参考答案之间的词级重叠。最后，对于答案长于一个句子的基准，我们使用 ROUGE-L，它抓取预测答案和真实答案之间最长的匹配序列，和 BERTScore，它使用上下文嵌入测量答案和标注之间的语义相似性。

为了有效地从不同模态中检索信息，我们利用了模态特定的编码器：bge-large-en-v1.5 作为文本编码器，InternVideo2 作为视觉编码器。对于响应生成，我们使用了多种 LVLMs，包括 InternVL2.5-8B、Qwen2.5-VL-7B-Instruct 和 Phi-3.5-Vision-Instruct。对于路由器模块，可训练路由器以学习率 2e-5 进行训练 5 个周期，并基于验证准确性选择最佳的检查点。在无训练设置中，通过如图所示的提示实例化 GPT-4o。更多详细信息见附录。

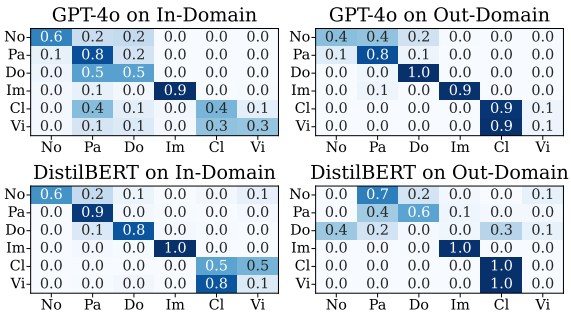

Figure 4: 使用不同模型对域内和域外查询进行路由器预测的混淆矩阵 。

Table 3: 不同粒度对三个模型在两个基准上的性能影响。Gn 表示粒度。

Models	Gn	HotpotQA EM	HotpotQA F1	LVBench Acc
GPT-4o	✗	14.26	22.95	32.32
	✓	17.80	26.10	33.01
DistilBERT	✗	14.55	23.08	33.20
	✓	19.43	28.35	35.16
T5-Large	✗	14.35	23.03	33.20
	✓	18.09	26.90	34.28

4 实验结果与分析

我们现在展示我们的结果和深入分析。

在这里，我们展示了跨越多种模态和粒度级别的多样化检索场景的整体结果。

首先，图 3 展示了在八个多模态基准中 UniversalRAG 和基线模型的平均得分，结果的详细分解在表 1 中提供。UniversalRAG 始终在平均得分上超越所有基线，展示了通过自适应语料库选择利用多种模态的有效性。与提供有限信息的单一模态语料不同，UniversalRAG 为每个查询动态选择最相关的模态，从而实现更准确的检索和生成。

有趣的是，UniversalRAG 显著优于 Unified 基线，突显了我们路由策略在现实的多模态环境中的有效性。具体来说，由于 Unified 基线在其统一嵌入空间中存在模态差距，通常只检索文本数据，因而性能受损。UniversalRAG 通过使用路由器来选择单一模态特定的语料库进行检索，从而有效地解决了模态差距问题。鉴于在没有模态差距的情况下构建跨模态的统一嵌入空间具有内在挑战性，我们基于路由器的策略为解决此问题提供了一个有前途的方向。

在 UniversalRAG 模型中，经过训练的路由模型在所有使用不同 LVLMs 的实验中表现出优于无训练路由模型的结果。这一改进是由于训练过的路由器在训练过程中被明确优化用于路由任务，从而导致更优异的路由性能。因此，带有训练路由器的 UniversalRAG 模型更能识别出最优的数据来源并生成更精确的答案。尽管如此，无训练路由器仍然优于其他基线方法，包括随机路由器，表明在我们的框架内零样本路由仍然有效。为了进一步了解路由对整体系统性能的影响，我们分析了每种路由模型的准确性和相应的整体得分。图 4 展示了零样本和训练路由模型的混淆矩阵。虽然两种路由器总体上都能成功地将输入引导到适当的模态，但训练路由器显示出比无训练模型更高的准确性。值得注意的是，对于剪辑和视频模态，存在一些错误路由的查询，这主要是由于在区分两种不同的细粒度时的模糊性。然而，输入仍然被正确地路由到视频模态，突出了路由机制的鲁棒性。如表 2 所示，我们的路由方法在路由准确性方面明显优于随机和统一基线。这种准确性的提高直接转化为更好的整体性能，展示了准确路由与端到端有效性之间的强相关性。这些结果强调了将查询正确路由到适当模态语料库的重要性，展示了在多模态 RAG 场景中可靠路由器的必要性。

Table 4: 路由器模型大小变化时的路由器准确性。

Models	# params	Router Acc
T5-Small	60M	51.16
T5-Base	220M	63.65
T5-Large	770M	59.99
T5-XL	3B	67.50

为了进一步研究纳入多层次粒度的有效性，我们在粗粒和细粒检索设置下评估了 UniversalRAG。在无粒度（粗粒度）设置中，路由器将查询分类为四种广泛的模式：无、文本、图像或视频。在粒度化（细粒度）设置中，我们进一步细分模式以更精确地进行检索：文本被分为段落和文档级别，而视频被分为片段和完整视频。为了进行基准测试，我们使用 HotpotQA 评估跨多个实体的文档级推理，使用 LVBench 进行片段级任务，因为其问题通常可以使用短视频片段进行回答。如表 3 所示，具有粒度的 UniversalRAG 在所有路由器模型的两个基准上均一致优于没有粒度的模型。这强调了支持文本和视频语料库中不同粒度水平可通过使模型能够检索针对每个查询量身定制的适当信息来提高 UniversalRAG 的性能。相比之下，没有粒度控制的模型对所有查询应用相同的粒度水平，这可能导致信息检索不足或过量。因此，支持多层次粒度对于适应性处理各种用户查询至关重要。

在这里，我们对性能提升进行详细分析。

域外数据集的结果 为了调查我们方法的泛化能力，我们在五个未见过的数据集上评估了 UniversalRAG，每个基准的详细描述见附录 A.2 。如表 2 所示，GPT-4o 达到了最高的路由准确率，甚至超过了其在领域内的表现，展示了强大的泛化能力。然而，训练的路由器在域外数据上的表现不佳，表明路由器过拟合于训练数据，主要是由于训练数据中查询的多

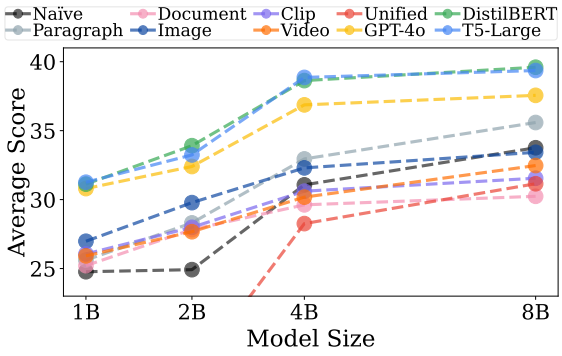

Figure 5: 生成性能与不同生成模型 (InternVL2.5) 大小的变化。

样性不足。图 4 进一步突出显示了领域内和领域外数据集之间的性能权衡。得益于其强大的路由能力，GPT-4o 也实现了最高的平均 QA 得分，优于训练的路由器和基线模型。

作为解决两种设置之间性能取舍的方案，我们引入了一个使用训练过的路由器和免训练路由器的集成路由器。具体来说，如果训练过的路由器的置信度得分足够高，则选择其路由结果；否则，利用免训练路由器的响应。这种策略使得我们可以针对与域内数据集特征相似的查询利用训练过的路由器，而对于不熟悉或域外的查询则依靠免训练路由器的通用路由能力。正如表格 2 所示，具有集成路由器的 UniversalRAG 在域内和域外基准测试中都表现出了更好的性能。

为了评估路由器大小对路由准确性的影响，我们使用不同模型大小的训练路由器来评估 UniversalRAG。具体来说，我们训练了四种不同参数数量的 T5 模型变体，并使用 InternVL2.5 作为生成器来测量路由器准确性。如表 4 所示，路由器准确性随着模型大小的变化而显著变化，这表明较大的模型在跨模态和粒度的路由决策中更有效。

为了观察 UniversalRAG 的性能如何随 LVLM 的规模变化而变化，我们用不同规模的 InternVL2.5 模型对我们的模型和基线进行评估，结果如图 5 所示。在所有模型规模中，UniversalRAG 的得分持续增加，并且优于其他基线。这表明 UniversalRAG 具有可扩展性，并暗示其性能可以通过使用更大的 LVLM 来提高。

我们在附录 C 中展示了 UniversalRAG 的案例研究。

5 相关工作

在大型语言模型（LLM）强大性能的基础上，研究人员努力使 LLM 能够理解视觉信息。**?** 通过采用基于 CLIP 的 (**?**) 图像编码器，率先开发了大型视觉语言模型（LVLM），这一编码器使语言模型能够在其文本特征空间内理解输入图像。随后，各种图像理解语言模型被引入，每种模型在 LLM 上使用不同的视觉编码器 (**???**) 。随着图像理解性能的提升，一些研究将这些方法扩展到了视频数据，这些视频数据可以视为一系列的图像帧 (**???**) 。借助更大的训练数据集和改进的模型结构，当前的 LVLM 在多项基准测试评估中表现出了强大的图像和视频理解能力 (**????**) 。然而，独立的 LVLM 通常由于其基础语言模型所固有的有限知识边界而容易出现幻觉。

检索增强生成 检索增强生成（RAG）可以通过在生成答案时结合外部知识来解决上述挑战；然而，传统的 RAG 方法仅依赖于文本数据，而最近的研究已经开始探索在多样的多模态语料库上实现 RAG，突出了其在超越仅文本设置方面的显著潜力。具体来说，基于图像的 RAG (**??**) 是多模态 RAG 的首次尝试，检索并使用视觉信息来回答查询。此外，**?** 最近将 RAG 扩展到视频，捕捉过程相关问题的视觉和时间元素。尽管有这些进展，大多数现有方法只考虑单一模态语料库，而这在现实世界中是不切实际的，因为真实查询可能需要来自任何模态的信息。因此，利用所有可用数据来生成最佳答案至关重要，而不是将模型限制在有限的模态上。最新的方法 (**??**) 支持从多模态语料库中进行检索，但通常从所有可用模态中检索，然后在检索后或甚至是在生成后再决定使用哪个，这种做法效率低下，并且未能使检索适应查询的特定需求。

处理多样化的查询需要一种 RAG 方法，该方法根据具体的上下文和查询进行调整，而不是使用单一固定的方法。一种有前途的方法是根据预定义的复杂级别 (**???**) 路由查询，将它们分类为不需要检索、单步检索或多步检索，以平衡性能和延迟。另一种策略是利用模型信心 (**??**) ，仅在模型信心低时检索外部信息，从而有效地将资源分配给具有挑战性的查询。尽管自适应检索已成为 RAG 的核心，现有的基准 (**??**) 主要评估纯文本系统，尚未解决如何在多种模态间调整检索的问题。在实际场景中，查询受益于不同的数据类型，因此在混合模态语料库中识别最合适的检索模态是至关重要的。

检索的粒度，即索引语料库的大小，是检索中的一个关键设计选择，因为它显著影响 RAG 的性能和效率。**?** 发现，从用命题索引的语料库中检索性能优于句子级或段落级检索的表现。最近的研究 (**??**) 也显示，考虑多种粒度可以实现更好的检索性能。同样，粒度感知的文本到视频检索被研究以查找不仅仅是整个视频，而是从视频语料库中找出与查询相关的特

定片段（**?**）。因此，在多模态语料库中，仅仅选择合适的模态是不够的；系统还应该识别出适合检索的最优粒度水平。

6 结论

在本文中，我们提出了 UniversalRAG，这是一种新颖的 RAG 框架，旨在从具有多样化模态和粒度的语料库中检索。通过一种模态和粒度感知的路由机制，UniversalRAG 动态选择最适合每个查询的知识来源，有效解决了模态差距和固定粒度检索所带来的限制。在 8 个基准上的广泛评估表明，UniversalRAG 在多种模态中持续优于模态特定和统一的基线，展示了强大的性能。此外，我们的分析突出强调了细粒度检索的重要性以及无训练路由器和有训练路由器的互补优势。这些发现展示了 UniversalRAG 作为一项可以适应的解决方案的潜力，以异质外部知识为 LVLMs 提供支持，为更可靠的多模态推理和模态感知信息整合开辟了新方向。

A 数据集的附加细节

表 5 概述了我们实验中使用的所有数据集及其对应的数据语料库，包括目标模态类型以及查询和语料库的大小。我们将每个数据集按 3:7 的比例分为训练和测试。下面提供了每个数据集的详细说明。

A.1 域内数据集

MMLU 我们使用 MMLU (**?**) 作为一个数据集，其中包括可以在不需要检索情况下回答的查询，这是一个涵盖广泛任务的基准测试，包括问题解决能力（例如，小学数学、计算机科学）和世界知识（例如，法律、世界宗教）。具体来说，我们使用开发集划分中的所有任务的问题。

SQuAD SQuAD v1.1 (**?**) 是一个基准数据集，由众包工作者根据一组维基百科文章生成的问题组成。每个问题在给定适当的上下文段落时都可以回答。从数据集的 100,000 多个问答对中，我们随机抽取开发集的 1,060 对。对于上下文检索，我们利用完整提供的维基百科语料库，将每篇文章分段为最多 100 个字的段落。

自然问题 (NQ) 我们还使用 Natural Questions (**?**)，这是一个包含真实用户对谷歌搜索引擎提出的查询的问题回答数据集，其答案基于支持的维基百科文章进行标注。我们随机抽取开发集的 1,000 个问答对，并在与 SQuAD 相同的设置中制定文本语料库，将维基百科语料库分段为最多 100 个字的段落。

HotpotQA HotpotQA (**?**) 是一个基于维基百科的问答基准，但包含被注释为需要在多个文章上进行推理的复杂查询。我们利用测试集的 1,492 个随机抽取的问答对。由于需要在多个文件上进行多跳推理，我们根据 LongRAG (**?**) 的方式分组多个相关文件来制定文本语料库，这些文件可以超过 4K 个标识符。

WebQA WebQA (**?**) 是一个基准测试，旨在评估大型语言模型在开放域环境中通过多种信息源（包括文本和图像）进行推理的能力。由于该数据集最初是用特定问题的检索源构建的，这些检索源结合了文本和图像，我们提取了一组仅需一张图像即可检索的问题子集。然后，我们使用 GPT-4o 和图 **??** 中的提示进一步过滤这些问题，以确保问题不依赖于某张特定图像，最终得到一个包含 2,000 个问答对的集合。

LVBench LVBench (**?**) 是一个为长视频理解而开发的基准，特点是基于 YouTube 视频由标注者生成的问题，这些视频的平均时长超过一个小时。由于该基准最初是为非 RAG 任务设计的，我们将原本的文本-视频交错查询重新表述为仅文本格式，以便使用 GPT-4o 与我们实验设置对齐，该设置包括视频元数据和一个提示（图 **??**）。每个查询都与特定的视频和相应的时间范围相关联。值得注意的是，大多数查询附有不到五分钟的时间戳，从而专注于较长视频中的短片段。在训练中，我们使用这些短时间戳查询作为片段级数据集。

视频 RAG 我们还利用在 VideoRAG (**?**) 中介绍的视频级 RAG 测试的 VideoRAG-Wiki 和 VideoRAG-Synth 基准。这些基准构建于 HowTo100M (**?**) 语料库之上，该语料库是一个大规模的 YouTube 教学视频集合，查询来自 WikiHowQA (**?**) 和基于视频合成生成的问答对。由于它们缺乏时间戳标注，我们使用 GPT-4o 确定那些更适合通过全视频检索而不是从真实视频中截取短片段来解答的视频级查询，然后将其用作训练路由器的视频级数据集。

A.2 域外数据集

与领域内数据集不同，领域外数据集仅用于评估，以评估我们的路由方法的泛化能力，并且只由测试拆分组成。

TruthfulQA TruthfulQA (**?**) 包含常识性问题，旨在测试大型语言模型（LLMs）是否能避免各种类别（包括健康、法律和政治）中的常见错误信仰或误解。我们使用数据集的多项选择版本，每个问题只有一个正确答案。

TriviaQA TriviaQA (**?**) 是一个阅读理解数据集，由来自维基百科和网络的证据文本和琐事问题配对组成。为了区分那些需要文本检索的查询与不需要的查询，我们基于 GPT-4o 是否能在不访问外部文本的情况下生成精确匹配答案来对每个查询进行分类。我们从开发拆分中随机抽取问答（QA）对。借鉴了在 SQuAD 和 NQ 中使用的预处理策略，所有支持证据文档被分段成不超过 100 个词的段落。

我们还使用 LaRA (**?**)，该工具设计用于理解长篇文献如学术论文和小说。在我们的用例中，我们专注于这些文档的一个子集，特别是排除了“比较”任务的查询，因为我们的目标是检索辅助生成（RAG），而不是阅读理解。此外，我们稍微重新格式化剩余的查询以与一般 QA 格式对齐。考虑到源材料的长度，每个文档在文档级语料库中被视为一个独立的条目。

视觉-RAG Visual-RAG (**?**) 是一个针对视觉知识密集型问题设计的问题回答基准，专为文本到图像检索任务定制。我们利用提供的查询

Table 5: 数据集摘要。平均语料长度是指文本语料库的平均标记数和视频语料库的平均持续时间。

Dataset	Gold Retrieval	# Queries	Corpus Size	Avg. Corpus Length
In-Domain Datasets				
MMLU	None	285	-	-
SQuAD	Paragraph	1,060	1.19M	100 tokens
Natural Questions	Paragraph	1,000	850k	100 tokens
HotpotQA	Document	1,492	509k	693 tokens
WebQA	Image	2,000	20k	-
LVBench	Clip/Video	1,376	94	3,941s
VideoRAG-Wiki	Clip/Video	374	9k	378s
VideoRAG-Synth	Clip/Video	374		
Out-of-Domain Datasets				
TruthfulQA	None	790	-	-
TriviaQA	Paragraph	661	661k	100 tokens
LaRA	Document	112	34	28k tokens
Visual-RAG	Image	374	2k	-
CinePile	Clip/Video	1,440	144	158s

全套，但每个类别抽样五张图像来构建图像检索池，确保高效的文本到图像检索。

CinePile CinePile (**?**) 是一个长视频问答基准，其问题基于来自 YouTube 的电影片段。由于该基准最初是为视频理解任务设计的，而不是为 RAG 设计的，我们使用与 LVBench 相同的程序重新制定每个查询。对于 144 个可用视频中的每一个，我们从测试集随机选择 10 个问题。由于 CinePile 没有提供细粒度注释，我们使用 GPT-4o 将问题分类为两类——片段级和完整视频级粒度，遵循在 VideoRAG 中使用的相同方法。

B 额外的实现细节

为了有效利用视觉和文本信息进行视觉元素检索，我们采用了集成方法，将视觉和文本相似度分数结合，其中视觉信息的权重比例为 0.8。文本信息包括图像的描述和视频的脚本。在生成阶段，我们仅使用基于相应嵌入向量的余弦相似度选出的最优结果。此外，对于检索和生成阶段，我们统一每个视频采样 32 帧。可训练路由器在 5 个 epoch 内训练，学习率为 2e-5，选择基于验证准确性的最佳状态。

我们在表 6 中展示了三个路由器针对每个数据集的路由结果。在域内数据集上，GPT-4o 常常难以区分段落和文档 RAG 查询，并错误地将 VideoRAG 查询导向文本库。与此同时，两个训练的路由器在所有域内数据集上表现出强大的分类性能。在域外数据集上，GPT-4o 对大多数数据集泛化良好，除了基于图像的 RAG 查询。相比之下，训练的路由器则无法针对每个查询分类出适当的粒度。这主要是由于训练数据多样性有限，导致对已见示例的过拟合。

不同领域数据集上 UniversalRAG 模型和基线的

B.1 域外数据集的详细结果

QA 评估结果如表 7 所示。总体而言，UniversalRAG 模型优于基线。GPT-4o 在所有数据集上表现出色，主要得益于路由器在未见过的查询上卓越的泛化能力，如第 **??** 节所讨论的。然而，与领域内数据集的结果相比，训练后的路由器表现有所下降，因为它们的路由器在处理未见过的查询时经常误判。

C 定性结果

我们展示了一些案例研究，以证明 UniversalRAG 的有效性。表格 **??** 比较了各种 RAG 方法的结果，包括传统的单模态方法和 UniversalRAG 在 WebQA 数据集查询上的表现。传统方法如 TextRAG 和 VideoRAG 未能生成准确的答案——TextRAG 检索的段落缺乏相关的视觉细节，而 VideoRAG 更适用于时间推理任务。相比之下，UniversalRAG 正确地将查询路由到图像模态，识别出需要颜色的视觉信息，并成功生成了正确的回答。这凸显了模式感知路由在利用正确模态语料库中适当数据的优势，展示了 UniversalRAG 在准确答案生成方面自适应选择最具信息性模态的能力。

除了模态路由外，我们观察到 UniversalRAG 还受益于在适当的粒度检索信息。表 **??** 显示了来自 HotpotQA 的结果，其中查询需要对多个文本来源进行复杂推理。虽然段落级粒度未能提供足够的推理背景，UniversalRAG 将查询路由到文档级语料库，以检索准确推理所需的所有文本信息。同样，对于视频查询，表 **??** 显示了来自 LVBench 的结果，其中查询只需要回答完整长视频的一小段片段。虽然全视频级别的检索包含了不相关的内容，并且均匀采样的 32 帧未能捕捉必要的信息，片段级别的检索则专注于更小、更相关的视频片段，以确保

Table 6: 路由结果在域内和域外数据集的表现。

		In-domain Dataset								Out-domain Dataset				
		Text				Image	Video			Text			Image	Video
		MMLU	SQuAD	NQ	HotpotQA	WebQA	LVBench	VRAG-Wiki	VRAG-Synth	TruthfulQA	TriviaQA	LaRA	Vis-RAG	CinePile
	Models	200	742	700	1045	1392	829	374	374	790	661	112	374	1440
gpt-4o	None	117	44	57	38	25	3	8	27	374	121	0	0	0
	Paragraph	39	512	588	505	102	17	304	271	205	509	4	18	0
	Document	44	185	39	502	44	34	52	53	206	27	108	0	6
	Image	0	1	5	0	1210	44	1	9	2	4	0	356	1
	Clip	0	0	6	0	0	622	0	0	3	0	0	0	1354
	Video	0	0	5	0	11	109	9	14	0	0	0	0	79
DistilBERT	None	120	2	1	1	0	0	0	0	2	1	42	0	0
	Paragraph	49	679	669	150	30	1	0	6	629	274	21	0	1
	Document	11	32	12	866	12	3	0	0	53	338	2	2	1
	Image	0	6	11	17	1351	7	0	0	12	32	4	371	1
	Clip	5	0	1	8	5	818	0	2	3	7	34	0	1436
	Video	15	23	6	3	2	0	374	366	91	9	9	1	1
T5-Large	None	110	4	1	0	1	0	1	1	21	4	30	0	0
	Paragraph	79	731	698	461	145	13	5	15	709	558	74	0	13
	Document	2	3	0	571	6	15	0	0	35	94	0	0	2
	Image	0	1	0	9	1234	15	0	0	1	2	0	374	4
	Clip	0	0	1	2	13	784	0	3	2	1	1	0	1420
	Video	9	3	0	2	1	2	368	355	22	2	7	0	1

Table 7: UniversalRAG 和基线在域外数据集上的详细结果。

	Text					Image		Video	Avg.
	TruthfulQA	TriviaQA		LaRA		Vis-RAG		Cinepile	
Models	Acc	EM	F1	R-L	BERT	R-L	BERT	Acc	
Naïve	64.68	49.47	57.92	23.15	87.62	6.24	80.98	30.76	33.88
Paragraph	58.73	54.61	65.14	20.23	86.48	4.74	80.77	30.07	33.88
Document	28.73	39.94	44.73	25.18	86.83	4.34	81.14	32.64	26.68
Image	57.85	45.23	52.50	21.40	87.09	7.31	82.32	34.03	33.35
Clip	51.01	31.62	42.40	19.64	87.50	6.92	81.32	35.63	29.59
Video	47.34	33.59	43.82	19.89	87.19	70.4	81.42	37.43	29.47
Unified	52.15	35.70	45.01	21.28	86.83	4.31	80.47	30.76	28.92
Random	51.27	42.66	51.51	21.81	87.27	5.81	81.09	32.43	29.99
GPT-4o	55.19	54.01	64.05	24.93	88.42	7.17	82.26	35.56	36.85
DistilBERT	57.85	42.51	51.01	20.96	87.26	7.34	82.32	35.63	32.58
T5-Large	57.85	50.08	60.16	20.63	86.73	7.31	82.32	35.49	35.27
Oracle	64.68	55.52	64.85	25.18	86.83	7.31	82.32	37.71	38.26

只考虑最相关的视觉细节，从而导致更准确的答案。