概念还是技能？重新思考多模态模型的指令选择

Andrew Bai, Justin Cui, Ruochen Wang, Cho-Jui Hsieh

Department of Computer Science
University of California, Los Angeles

Abstract

视觉-语言指令调优实现了两个主要目的：学习视觉概念和学习视觉技能。在本文中，我们发现视觉-语言基准测试主要通过训练具有类似技能或视觉概念的指令来获益。这一发现启发了我们设计了一种简单的目标训练数据选择方法，以优化给定基准的性能。我们首先从基准中提取概念/技能，确定该基准是主要受益于类似概念还是技能，最后选择概念/技能最匹配的指令。在对 10 多个基准进行的实验中，我们的方法验证了其有效性，与现有的最佳基线相比，平均在所有基准上提高了 0.9 %，在以技能为重点的子集上提高了 1.5 %。我们的发现强调了识别指令选择中固有权衡的重要性，需要在获得概念知识和视觉技能之间取得平衡。

1 介绍

近来，视觉语言模型的进展是由大型预训练语言模型（LLMs）与强大的视觉编码器的模块化组合推动的，通常通过一个转换视觉特征到与 LLMs 输入空间兼容的格式的模态适配器进行连接。这种架构允许模型在纳入来自预训练视觉主干的丰富视觉理解的同时，利用 LLMs 的语言能力。然而，仅仅连接这些组件对于实现强大的多模态表现是不够的；通常需要在配对的视觉语言数据上进行广泛的持续预训练以弥合模态差距。此外，视觉语言指令调优——在其中模型进一步在特定策划的提示上进行训练以教授特定的多模态能力——已成为将联合模型的行为与期望的视觉语言任务对齐的关键。

视觉-语言指令微调已成为一个强大的框架，用于使多模态模型与人类期望的行为对齐，特别是在需要视觉感知和语言推理的任务中。这个微调过程有两个主要目的：首先，它帮助模型学习将视觉表示与相应的文本概念关联起来；其次，它使得获取新的视觉能力成为可能，如计数对象、推理空间关系和推断物理属性。这双重角色对于在需要对已见或未见指令进行泛化的现实应用中部署视觉-语言模型至关重要。

然而，尽管指令微调被广泛采用，用于评估的视觉-语言基准测试在所测量的能力类型上往往有所不同。有些侧重于概念识别（例如，视觉问答），而另一些则强调推理或基于技能的任务（例如，计数、关系理解或常识推理）。这种差异引发了一个基本问题：指令选择应该优先考虑与任务所需技能的对齐，还是与视觉内容中存在的概念的对齐？更具体地说，是否可以通过在一个强调类似技能的数据集（与强调类似视觉概念的数据集相比）上进行微调来提高基准测试的性能？

为了研究这一点，我们对各种视觉-语言基准进行了系统分析，旨在了解不同类型的指令相似性如何影响下游性能。我们的研究表明，现有的基准分为两类：一种是从涉及相似技能的指令中受益更多（例如，不同领域的物体计数）；另一种是从涉及相似概念的指令中受益更多（例如，基准中常见的物体或场景类别）。值得注意的是，这种模式在不同的架构和训练方案中均能保持，表明这种划分是任务性质内在的，而不仅仅是模型设计的附带产物。

受到这些发现的启发，我们提出了一种简单而有效的针对性指令选择方法，以适应每个基准的特点。具体来说，我们首先使用自动指令解析和技能/概念分类对齐，从给定的评估集中提取主要概念和技能。然后，我们通过验证性能差异判断基准是概念主导还是技能主导，最后选择最符合基准主导类型的训练指令。这种针对性的策略使模型能够聚焦于最相关的归纳偏见，无论是概念性还是程序性。

我们在超过十个标准的视觉-语言基准上评估我们的方法，涵盖了各种任务类型和难度级别。我们的结果显示了持续的性能提升，相比最强基线使用非定向指令微调，平均提高了 0.9 %。这些发现强调了基准感知指令选择的重要性，并为任务自适应多模态学习开辟了新的方向。

2 相关工作

2.1 视觉语言模型

大型语言模型（LLMs）(Dubey et al., 2024; OpenAI; Team et al., 2024; Chiang et al., 2023) 在各种任务中表现出了惊人的性能。这一成功主要归功于在万亿级的标记上进行的预训练，随后是如人类反馈强化学习（RLHF）(Christiano et al., 2017) 等后训练技术。在其文本领域能力的基础上，最近的研究已将 LLMs 扩展到处理诸如图像等额外的模态。MiniGPT-4 (Zhu et al., 2023) 集成了一个预训练的视觉变换器（ViT）骨干与一个 Q-Former 和一个单一线性投影层，将其与 Vicuna 语言模型结合，以在多模态任务中实现强大的性能。同时，InstructBLIP (Dai et al., 2023) 采用类似方法生成基于图像和文本提示的指令跟随响应。LLaVA (Liu et al., 2023b, 2024) 通过首先使用 CLIP (Radford et al., 2021) 编码器对图像进行编码，然后通过线性 MLP 将视觉特征映射到文本嵌入空间，从而将大型语言模型转换为多模态模型，这形成了一种简单但有效的集成策略。

选择

2.2 视觉指令数据选择

数据集 (Har-Peled and Mazumdar, 2004; Roux et al., 2012; Campbell and Broderick, 2018; Mirzasoleiman et al., 2020) 已被广泛研究，以提高模型训练效率。最近，这一研究方向已扩展到视觉-语言模型，旨在在保持性能的同时降低训练成本。Coincide (Lee et al., 2024) 将数据集划分为多个子集，并根据聚类的可转移性保留样本。ICONS (Wu et al., 2024) 通过测量与验证集的梯度相似性来量化各个数据点的影响，从而识别重要样本。Prism (Bi et al., 2025) 引入了一种无训练方法，利用皮尔逊相关分析来衡量 MLLMs 的内在视觉编码能力。相比之下，PreSel (Safaei et al., 2025) 以不同的方式解决问题：它首先应用过滤机制识别高质量图像，然后仅针对选中的样本生成指令。尽管在减少数据的情况下取得了强性能方面的显著进展，但关于不同任务如何受到基础概念或技能影响的研究仍然有限。

我们明确区分了一个二分法：概念，指的是图像中存在的视觉实体、属性和对象，即外观；技能，指的是正确解释或回答有关这些实体的问题所需的推理操作或判断策略，即如何分析它们。这一区分反映了 Whitehead et al. (2021) 在视觉问答（VQA）中提出的组合性观念，他们明确地建模了一种技能-概念的分解，以实现对技能（例如“颜色”判断）和概念（例如“汽车”）的新组合的泛化。虽然概念使模型根植于视觉内容中，但技能捕捉了推进下游任务成功的潜在推理模式——如计数、空间推理或趋势分析。Coincide (Lee et al., 2024) 整合了概念和技能，共同定义了一种相似性的概念，然后用来最大化所选样本的多样性。在我们的案例中，解耦这两个轴线允许选择方法基于模型需要识别的对象与需要推理的方面来选择训练样本，从而更精确地与各种视觉语言基准的认知需求对齐。

3 方法论

3.1 问题表述

设 $\mathcal{I} = \{(x_i, y_i)\}$ 表示一个多模态指令示例池，其中每个 x_i 是一个自然语言指令与一个或多个图像配对，y_i 是预期的响应。给定一个目标下游任务集 $\mathcal{T} = \{T_1, ..., T_K\}$，我们的目标是选择一个子集 $\mathcal{I}^* \subset \mathcal{I}$，在指令调整后最大化在下游任务上的性能。每个任务 T_k 对应一个基准数据集，其特征是特定的输入输出格式和评估指标。我们假设，最有益的指令子集因任务而异，并且指令的视觉内容与任务需求之间的对齐在下游泛化中起着至关重要的作用。

4 方法论

本研究引入了一种基于检索的框架，用于比较以概念优先和技能优先的选择策略以策划视觉-语言指令调优数据。该方法明确地将视觉概念与视觉技能的概念分开。对于给定的指令，从候选池中检索出相似的例子，这个过程通过在概念空间或技能空间中进行最近邻搜索来实现。

4.1 概念表示

概念表示来源于每条指令关联的图像。图像通过一个预训练的视觉编码器传递，结果的嵌入用于表征视觉内容。这种表示在捕捉图像之间的语义相似性时，对语言信息保持不敏感。

4.2 技能表示

本工作的主要技术贡献在于构建技能表示。与概念不同，视觉技能在数据集中没有直接标注，必须推断出来。为了解决这个问题，我们开发了一条自动化流程来提取和编码每个指令的技能信息：

1. 通过大型语言模型进行技能分离对于每个指令-图像-答案三元组，利用大型语言模型提出问题：“正确回答该指令需要哪些视觉技能？”回答包括一个简洁的技能列表，例如物体计数、空间推理或细粒度属性识别。

2. 技能嵌入提取由语言模型生成的技能描述通过预训练的句子嵌入模型被转换为固定维度的向量。这些嵌入定义了一个技能空间，其中具有相似认知要求的指令彼此接近。

这个过程产生了一种表示法，该表示法明确地将图像描绘的内容与解释所需的推理技能分开。

给定一个查询指令，最近邻检索可以应用于概念空间或技能空间。概念优先选择检索出概念嵌入最接近的示例。技能优先选择检索出技能嵌入最接近的示例。由此产生两个不同的数据子集，它们仅在选择标准上有所不同。

4.3 下游评估

接下来对整理后的数据集进行指令调优，以实现概念驱动与技能驱动数据选择对下游多模态性能影响的直接比较。具体而言，对于给定基准，使用其概念邻居训练的模型记为“概念 ↑ ”，技能邻居训练的模型记为“技能 ↑ ”。

5 实验

我们在十二个多样的视觉语言基准测试和两个指令数据集上评估我们提出的指令选择策略。我们的目标是在数据预算限制下，与几个无目标基线相比，衡量概念和技能针对的指令选择的有效性。

5.1 实验设置

数据集 我们在两个数据集上进行了训练预算限制实验：LLaVA-1.5 (Liu et al., 2023a) 包含 665k 个示例，和 ALLaVA-4V (Chen et al., 2024) 包含 1.2M 个示例。每个示例包括一个或多个指令-响应对和一个相关的图像。我们对 LLaVA-1.5 指令池进行了 5 % 和 10 % 的采样预算实验，对 ALLaVA-4V 指令池进行了 2.5 % 的采样预算实验。

我们将以概念为优先（`Concept ↑ `）和以技能为优先（`Skill ↑ `）的目标选择策略与三个无目标的基线进行比较：随机采样、Coincide (Lee et al., 2024) 选择样本以最大化多样性，以及 PreSel (Safaei et al., 2025) 仅依赖于图像进行指令选择。值得注意的是，虽然 ICONS (Wu et al., 2024) 也是目标选择的相关基线，但由于缓存每个训练实例的 LoRA 梯度所需的高计算成本，我们无法再现他们的结果。

我们使用 LLaVA-1.5 框架实现指令微调，该框架整合了 CLIP-ViT 图像编码器和 Vicuna-7B 语言模型。微调通过 LoRA 适配器进行，以提高参数效率。所有模型使用 AdamW 优化器在一个周期内进行训练，学习率为 2×10^{-5}，批量大小为 128，图像分辨率为 224×224 像素。训练分布在四个 NVIDIA A6000 GPU 上。对于具有报告标准差的方法，性能在三个随机种子上取平均值。

我们预先计算了所有指令示例和基准样本的嵌入。我们使用 FAISS 库进行高效的最近邻搜索，以识别每个基准中最相似的前 k 个指令样本。概念嵌入是通过在图像上使用零样本 CLIP 图像特征提取器获得的。技能嵌入是通过首先使用 GPT-4o 提示指令中的问题，以获取回答问题所需的相关视觉技能，接着使用开源句子转换器 MiniLM-L6-v2 对技能描述进行编码，以提取 384 维的密集嵌入。

5.2 评价协议

我们在零样本设置中对每个模型在下游基准上进行评估，报告任务特定的指标，如准确率（GQA, ScienceQA）和精确匹配（TextVQA, OCR-VQA），如果适用的话。随机基线用三个随机种子重复，以考虑训练的变异性，我们报告平均性能及其标准差。由于计算限制，对于其他基线，我们报告一次实验运行的结果。随机基线的标准差用于确定其他基线是否在统计上显著优于随机基线。

我们在十二个基准上评估我们的方法：VQAv2, GQA, VizWiz, ScienceQA (SQA-I), TextVQA, POPE, MME, MMBench (en), LLaVA-Bench, AI2D, OK-VQA 和 ST-VQA。这些基准涵盖各种任务，包括通用 VQA、OCR 和科学推理。

5.3 实验结果

表 1 , 2 和 3 显示了基准测试中基线方法的比较。基准测试分为三个部分，以水平线分隔：顶部部分对应于概念导向任务，而底部部分对应于技能导向任务。中间部分由基准测试组成，最佳表现的方法并未超过随机均匀基线平均值加其标准差。最佳表现的方法用粗体突出显示，第二好的方法用下划线标出。

LLaVA 5 % 表 1 展示了在 5 % LLaVA 预算下的结果。在这种低资源环境下，指令选择尤其具有影响力。针对性的策略在大多数基准测试中取得了显著的改进，特别是在具有特定领域特征或推理要求的技能为重点的任务中。

LLaVA 10 % 表格 2 显示了使用 10 % 预算的结果。虽然增加的数据量减少了方法之间的性能差距，但有针对性的选择仍然可以在具有强领域或技能匹配的基准测试中提供显著的改进。值得注意的是，更多的基准测试受益于

Table 1: 关于 LLaVA-1.5 的 5 个% 数据选择的实验结果。

Benchmark	Untargeted			Targeted (Ours)		
	Random	Coincide	PreSel	Concept ↑	Skill ↑	C-S
VizWiz	28.4 ± 0.7	<u>28.7</u>	28.4	30.2	28.6	+1.6
LlaVa-Bench	66.7 ± 1.4	<u>68.0</u>	68.4	<u>68.0</u>	67.4	+0.6
VQAV2	72.2 ± 0.2	73.3	<u>72.5</u>	72.1	71.6	+0.5
TextVQA	52.0 ± 0.3	52.1	<u>51.1</u>	54.8	54.4	+0.4
GQA	52.7 ± 0.5	53.6	52.0	54.1	<u>54.0</u>	+0.1
MME	1259.3 ± 12.8	1343.0	<u>1326.3</u>	1302.4	1248.2	+54.2
MMBench(en)	55.8 ± 1.4	55.2	<u>57.0</u>	57.1	<u>57.0</u>	+0.1
POPE	84.3 ± 0.6	83.9	<u>84.4</u>	83.5	84.7	-1.2
STVQA	44.7 ± 0.1	46.5	45.1	<u>46.8</u>	47.5	-0.7
SQA-I	65.9 ± 0.5	66.3	<u>66.7</u>	65.8	68.7	-2.9
AI2D	50.8 ± 0.6	50.5	<u>51.1</u>	49.1	53.8	-4.7
OK-VQA	45.2 ± 0.7	51.7	42.4	43.2	<u>50.8</u>	-7.6

技能集中的选择，这表明信号在概念集中的选择中更容易达到饱和。

ALLaVA 2.5 % Table 3 报告了 2.5 % ALLaVA 设置的结果，其中指令更加多样且噪声更大。在这种设置中，目标基线显著优于随机基线，突出了在数据较少的情况下，尤其是对于更大且更多样化的指令数据集，指令质量和对齐的重要性。

6 讨论

6.1 不同的优先级化方法对不同的下游任务有益处

我们的研究结果表明，数据选择策略的有效性在很大程度上取决于基准测试的性质。概念优先选择策略往往对那些成功主要依赖于识别和定位图像内物体的基准测试有利。这类任务，比如 VQAv2、VizWiz 和 MME，主要是一些相对简单的是非题或简答题，只要正确识别相关的视觉元素即可作答。相比之下，技能优先选择在需要超出物体识别的推理或专业视觉能力的基准测试中显示出明显的优势。数据集如 SQA-I、OK-VQA 和 AI2D 需要具备常识推理、细微属性区分、读取嵌入文本（OCR）、或基于视觉证据的多步骤推理等能力。性能上的差异表明，概念驱动和技能驱动的选择是互补的：前者增强了模型基于视觉内容给出答案的能力，而后者提升了其在该内容上执行更复杂推理的能力。

6.2 非针对性基准隐含地权衡性能

虽然 Coincide 和 PreSel 等不针对特定目标的基线方法平均表现良好，但它们往往在各个任务之间做出隐含的权衡。这些方法偏向于全局频繁出现的指令模式，这可能有利于像 VQAv2 或 GQA 这样的通用基准，但会降低在特定领域或技能密集型任务上的表现。在原始基线研究中未报告的多个以概念为中心的基准中，我们观察到，当应用这些启发式方法时，这些未报告任务的表现往往较差。这突显了理解不针对特定目标选择所带来的隐藏偏见和权衡的重要性。这种全局平均方法的另一个后果是，像 TextVQA 或 SQA 这样的特殊案例（需要更专业的技能，例如 OCR、组合推理）容易被忽视。我们的研究表明，在通常不针对特定目标的选择策略中加入更加强调技能多样性的焦点，可能有助于缓解这些限制，降低以平均表现为目标的优化成本，同时支持那些处于主流分布之外的任务。

6.3 混合选择策略

我们还研究了结合概念导向和技能导向策略是否能提供两全其美的效果。对于每个基准，我们根据视觉概念和技能为每个指令计算了相关性得分，然后探索了多种方法以将这些得分组合成一个统一的选择标准。具体而言，我们尝试了 (1) 将概念和技能相关性得分相加，(2) 取两个得分的最大值，以及 (3) 均分选择预算，以便一半的指令基于概念相关性选择，另一半基于技能相关性选择。

表 4 展示了在一个代表性概念和技能目标的子任务上的三种混合方法。令人惊讶的是，这些混合策略中没有一种能够始终优于单一目标的方法。在几乎所有的基准测试中，表现最差的混合变体都比不上两个目标策略中较好的那一个。这表明不加区分地结合这两个信号会

Table 2: 关于 LLaVA-1.5 的 10 个% 数据选择的实验结果。

Benchmark	Untargeted			Targeted (Ours)		
	Random	Coincide	PreSel	Concept ↑	Skill ↑	C-S
VizWiz	28.8 ± 1.1	29.7	29.8	29.2	30.8	-1.6
LlaVa-Bench	67.0 ± 3.2	68.6	66.9	68.4	68.2	+0.2
VQAV2	74.0 ± 0.2	75.0	74.0	74.5	73.6	+0.9
TextVQA	53.5 ± 0.7	53.4	53.3	55.0	55.7	-0.7
GQA	56.0 ± 0.2	56.6	56.0	56.6	57.2	-0.6
MME	1349.6 ± 34.1	1382.4	1387.3	1368.6	1302.7	+65.9
MMBench(en)	58.1 ± 0.8	60.8	57.7	57.7	59.5	-1.8
POPE	84.0 ± 1.0	84.3	84.3	84.9	84.3	-0.6
STVQA	47.1 ± 0.6	48.2	47.9	47.7	48.7	-1.0
SQA-I	67.7 ± 0.5	66.9	66.0	67.1	70.4	-3.3
AI2D	52.2 ± 0.5	53.3	51.5	51.9	54.2	-2.3
OK-VQA	49.5 ± 1.3	53.7	50.4	50.7	51.9	-1.2

Table 3: ALLaVA 2.5 % 数据选择的结果。

Benchmark	Untargeted	Targeted (Ours)		
	Random	Concept ↑	Skill ↑	C-S
VizWiz	21.0 ± 0.4	31.1	30.5	+0.6
LlaVa-Bench	63.2 ± 2.5	76.6	70.7	+5.9
VQAV2	52.5 ± 2.8	72.3	70.9	+1.4
TextVQA	36.6 ± 3.2	52.2	52.8	-0.6
GQA	34.8 ± 1.9	52.7	51.7	+1.0
MME	854.4 ± 97.8	1208.5	1222.9	-14.4
MMBench(en)	29.9 ± 1.8	52.0	54.8	-2.8
POPE	76.1 ± 1.6	83.3	82.3	+1.0
STVQA	29.8 ± 3.5	46.8	47.5	-0.7
SQA-I	44.9 ± 5.4	62.1	66.5	-4.4
AI2D	42.4 ± 2.6	50.3	54.0	-3.7
OK-VQA	0.4 ± 0.3	38.4	24.7	+13.7

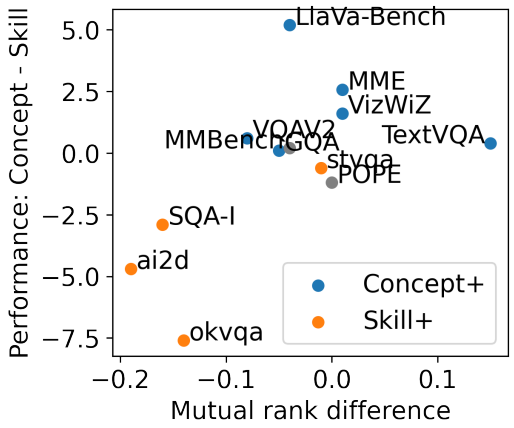

Figure 1: 在优先考虑概念与技能邻居时，互相排名差异和性能差异的分散情况。

削弱主导对齐因子的效果，无论是概念还是技能，并且没有直接结合概念和技能信号的简单方法。我们的研究结果暗示，基准测试往往从一种类型的对齐中受益更大，而不是两者的混合。

Table 4: 在 LLaVA-1.5 上进行 5 个% 数据选择的实验结果，用于比较结合概念和技能信号的混合策略。

Benchmark	Concept-Skill Hybrids				
	Max	Sum	Split	Concept ↑	Skill ↑
GQA	53.9	54.9	53.7	54.1	54.0
MME	1312.3	1312.5	1337.2	1302.4	1248.2
SQA-I	70.1	69.3	68.1	65.8	68.7
OK-VQA	41.9	47.4	47.4	46.8	47.5

6.4 通过互相关排名预测基准对齐

未能实现两全其美的尝试强调了能够预测哪个维度——概念或技能——对于某个给定基准更为相关的重要性。与其尝试结合这两种选择启发式，不如学习自动选择它们之间更为有效前进的路径。我们探讨能否通过一个简单的相互排名分析预测哪个对齐因素在某个基准上占主导地位。其目标是在不进行两个单独的微调实验的情况下，推断某个基准更可能从概念目标选择或技能目标选择中获益更多。

对于给定的基准，我们首先在整个指令池上构建两个相关性排名。概念排名根据样本与基准图像的视觉概念相似度（使用概念嵌入空间相似度）对样本进行排序，而技能排名则根据它们的视觉技能相似度（使用技能嵌入空间相似度）进行排序。

为了捕捉这两个排名之间的关系，我们检查每个列表中的第一个最近邻，并测量该样本在对方排名中的出现位置。具体来说，对于技能排名第一的样本，我们记录其在概念排序中的排名，对于概念排名第一的样本，我们记录其在技能排序中的排名。我们将这两个交叉排名分别称为 $R_{c|s}$ 和 $R_{s|c}$ 。

解释交叉排名 交叉排名揭示了在给定的基准中概念和技能共现的强度：

- 如果 $R_{c|s}$ 较低（即，技能邻居的概念等级较低），这表明技能相似的样本在视觉上是多样化的，这意味着该基准强调的是一般推理能力而不是特定的视觉领域。
- 如果 $R_{c|s}$ 很高，这表明技能相似的样本也共享相似的视觉概念，这意味着所需的技能与特定领域相关。
- 相反，高 $R_{s|c}$ 表示概念相似的样本往往涉及相似的技能，而低 $R_{s|c}$ 表示概念邻居在技能需求方面是异质的。

通过比较这两个交叉等级，我们可以推断出基准测试的主要对齐因素。一般技能基准往往具有较低的 $R_{c|s}$ 但较高的 $R_{s|c}$ ，因为一般技能广泛应用于许多概念，但概念邻居主要对应于需要类似低级技能的简单问题。另一方面，特定技能基准往往具有较高的 $R_{c|s}$ 但较低的 $R_{s|c}$ ，因为在技能上相似的样本通常共享重叠的视觉结构，而概念邻居可能不会表现出所需的专门技能。这种不对称的模式使我们可以将基准测试归类为主要由技能驱动还是概念驱动，而无需明确运行两种目标选择策略。

结果 图 1 显示了互相关等级差（$R_{s|c} - R_{c|s}$ ）在 x 轴上的散点图，以及模型在优先考虑概念与技能时的性能差异在 y 轴上的散点图。靠近左下角的数据点对应于以技能为目标的基准测试，而右上角的数据点对应于以概念为目标的。简单的交叉排名启发法成功预测了首选的对齐类型，从而实现了自动化且轻量级的基准感知指令选择策略。我们发现这种预测方法比简单组合两种策略（如前一小节所述）更为一致，因为它明确识别了哪个因素——概念或技能——在基准测试的数据分布中占主导地位。

为了验证所提出的技能提取流程是否捕捉到有意义且不平凡的信息，我们检查了 SQA-I 基准的技能描述，并将其与 LLaVA-1.5 训练混合中的最近邻进行了比较（见表 5 ）。检索到的邻居展示了基于技能的表示与解决每个问题所需的推理步骤紧密一致，而不仅仅是简单的物体计数或识别。

例如，“解读跨月的图表趋势”和“从视觉线索中理解植物的生长要求”等技能突出了从图像本身无法显现的一些微妙、依赖于情境的细节。这一合理性检验表明，技能嵌入在根据所需的推理类型进行指令分组时是有效的，而不是按表层的视觉相似性。

7 结论

在这项工作中，我们研究了视觉-语言指令选择的问题，重点考察了选择策略如何影响后续的性能。我们引入了基于视觉概念和技能的目标指令选择方法，并证明了不同的基准测试在概念和技能的倾向上表现出明显的偏好。我们的实验揭示了在与其对应重点对齐的任务中，概念和技能目标的选择可以优于无目标的基线，特别是在指令预算有限的情况下。此外，对概念和技能邻居进行简单的互相排名分析显示出清晰的不对称模式，这表明一个基准测试是概念驱动还是技能驱动。这一见解指向了一种轻量且自动的选择正确选择策略的方法，而无需代价高昂的尝试与错误过程。

8

局限性 虽然我们的研究结果突出了有针对性地选择指令的价值，但仍存在若干限制。首先，我们的视觉概念和技能分类是通过启发式和半自动标注结合而成的，这可能无法充分捕捉多模态能力的细微差别或相互依赖性。其次，我们的方法假设在选择过程中可以访问基准级别的信息，而这在涉及新颖或私有任务的部署场景中可能并不一定成立。此外，我们的方法目前将每个基准单独对待，并未考虑在为多个任务选择的指令上联合调整模型时的多任务泛化或干扰效应。最后，我们策略的有效性可能因模型大小和预训练质量而异，并且需要进一步研究以评估其在不同架构间的可迁移性。

在未来的工作中，我们的目标是开发自动化的方法，在没有监督的情况下推断基准对齐偏好，实现对未见任务的动态零样本选择。我们还计划探索更细粒度的选择标准，这些标准考虑了概念和技能属性之间的交互效应，并扩展我们的框架到多任务和持续指令调整设置。

References

Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2025. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. arXiv preprint arXiv:2502.12119 .

Trevor Campbell and Tamara Broderick. 2018. Bayesian coreset construction via greedy

Table 5: SQA 样本的技能描述与其在 LLaVA-1.5 训练混合中的最近邻的比较。

Samples from benchmark	Nearest neighbors in training set
To interpret the graph accurately and identify temperature trends across the months.	The ability to interpret and analyze text and numerical data related to temperature ranges.
One must analyze the beak shape of the birds to determine adaptations for cracking hard seeds.	One needs to observe details of the bird's beak shape, feeding behavior, and surrounding environment.
Identifying and comparing the number of pink balls in each solution.	One must identify and differentiate between various types of balls in the image.
The ability to compare the number of seedlings in different pots and analyze growth differences is required.	Observation, interpretation of visuals, and understanding of growth requirements for plants.
Identifying organisms based on scientific names and recognizing their taxonomic relationships within a specific context.	Identifying the animal's species and recognizing its characteristics for accurate scientific naming.

iterative geodesic ascent. In International Conference on Machine Learning , pages 698–706. PMLR.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. Preprint , arXiv:2402.11684.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90 % * chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) , 2(3):6.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. Advances in neural information processing systems , 30.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems , 36:49250–49267.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. arXiv e-prints , pages arXiv–2407.

Sariel Har-Peled and Soham Mazumdar. 2004. On coresets for k-means and k-median clustering. In Proceedings of the thirty-sixth annual ACM symposium on Theory of computing , pages 291–300.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. 2024. Concept-skill transferability-based data selection for large vision-language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing , pages 5060–5080.

Bo Li*, Peiyuan Zhang*, Kaichen Zhang*, Fanyi Pu*, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. 2024. Lmms-eval: Accelerating the development of large multimoal models.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition , pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. Advances in neural information processing systems , 36:34892–34916.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. 2020. Coresets for data-efficient training of machine learning models. In International Conference on Machine Learning , pages 6950–6960. PMLR.

OpenAI. OpenAI Platform — platform.openai.com. https://platform.openai.com/docs/models. [Accessed 27-07-2025].

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning , pages 8748–8763. PmLR.

Nicolas Roux, Mark Schmidt, and Francis Bach. 2012. A stochastic gradient method with an exponential convergence _rate for finite training sets. Advances in neural information processing systems , 25.

Bardia Safaei, Faizan Siddiqui, Jiacong Xu, Vishal M Patel, and Shao-Yuan Lo. 2025. Filter images first, generate instructions later: Pre-instruction data selection for visual instruction tuning. In Proceedings of the Computer Vision and Pattern Recognition Conference , pages 14247–14256.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 .

Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. 2021. Separating skills and concepts for novel visual question answering. Preprint , arXiv:2107.09106.

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. 2024. Icons: Influence consensus for vision-language data selection. arXiv preprint arXiv:2501.00654 .

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 .

A 提取技能描述的提示

```
Here is a list of questions about an
image:
[QUESTION_1]
[QUESTION_2]
[QUESTION_3]
Don't answer the above questions
directly. What visual skills are
required to answer these questions?
Answer in one short sentence with
less than 20 words without any extra
reasoning.
```

我们遵循 COINCIDE (Lee et al., 2024) 的评估方法来评估重叠的基准测试。唯一的小差别在于，我们采用了 `gpt-4o-mini` 来判断 LLaVA-Bench 的结果。我们还通过排除不可回答的子集来计算 VizWiZ 的准确性。我们发现，对于性能较差的模型来说，不可回答子集的表现更好，这可能是因为模型对每个问题都输出“不可回答”作为响应。对于新基准测试，我们直接使用了 LMMS-Eval 库提供的实现 (Li* et al., 2024) 。

我们重新实现了 Coincide (Lee et al., 2024) 和 PreSel (Safaei et al., 2025) 的结果。重新实现 Coincide 非常简单，因为他们完全开源了代码库，并且他们的选择方法支持选择任意数量的样本。PreSel 需要先在数据集的一个随机 5 % 子集中训练一个参考模型，然后依赖于参考模型提供的信号进行选择。直接与 PreSel 在 5 % 的结果进行比较是不公平的，因为那仅仅是 5 % 的随机样本。因此，PreSel 的 5 % 基线是通过使用参考模型额外选择 5 % 的数据来实现的，有效地利用了 10 % 的数据。然而，即使有这个优势，PreSel 仍然无法超越其他方法。

B 关于技能描述的额外定性研究

Table 6: OK-VQA 样本与其在 LLaVA-1.5 训练混合物中的最近邻样本之间的技能描述比较。

Samples from benchmark	Nearest neighbors in training set
One must recognize the vehicle type and its design characteristics to determine its historical context and invention date.	Identifying the vehicle type and recognizing its historical context.
One needs to identify the meal' s ingredients, presentation style, and cultural context to suggest a suitable side dish.	One must identify cultural elements in the dish's presentation, ingredients, and style.
Identifying the pastry type, size, and any visible toppings or fillings.	One must identify colors, textures, and shapes of the filling in the pastry.
One must identify size, shape, and features typical of buses versus vans to answer the question.	You need to identify and differentiate types of buses based on visual characteristics.
Identifying logos, labels, and packaging design details in the image to determine the origin of the beverage.	The ability to identify brand logos and labels on beverage packaging.