The Platonic Representation Hypothesis

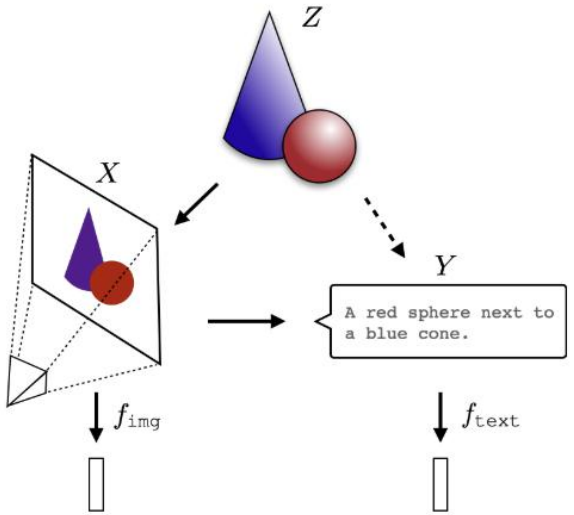
Minyoung Huh [*1] **Brian Cheung** [*1] **Tongzhou Wang** [*1] **Phillip Isola** [*1]

Platonic Representation Hypothesis

柏拉图表象假说（Platonic Representation Hypothesis）是一个理论概念：
随着模型规模的扩大和训练任务的多样化，不同的模型在表示数据的方式上越来越趋于一致。这种趋同指向一个共享的统计模型，这个模型能够捕捉到现实世界的基本结构，类似于古希腊哲学家柏拉图关于理想现实的概念。

The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.



世界（Z）可以用许多不同的方式来看待，如：图像（X），文本（Y）等。在每种模态上学习的表征将收敛到 Z 的类似表征。

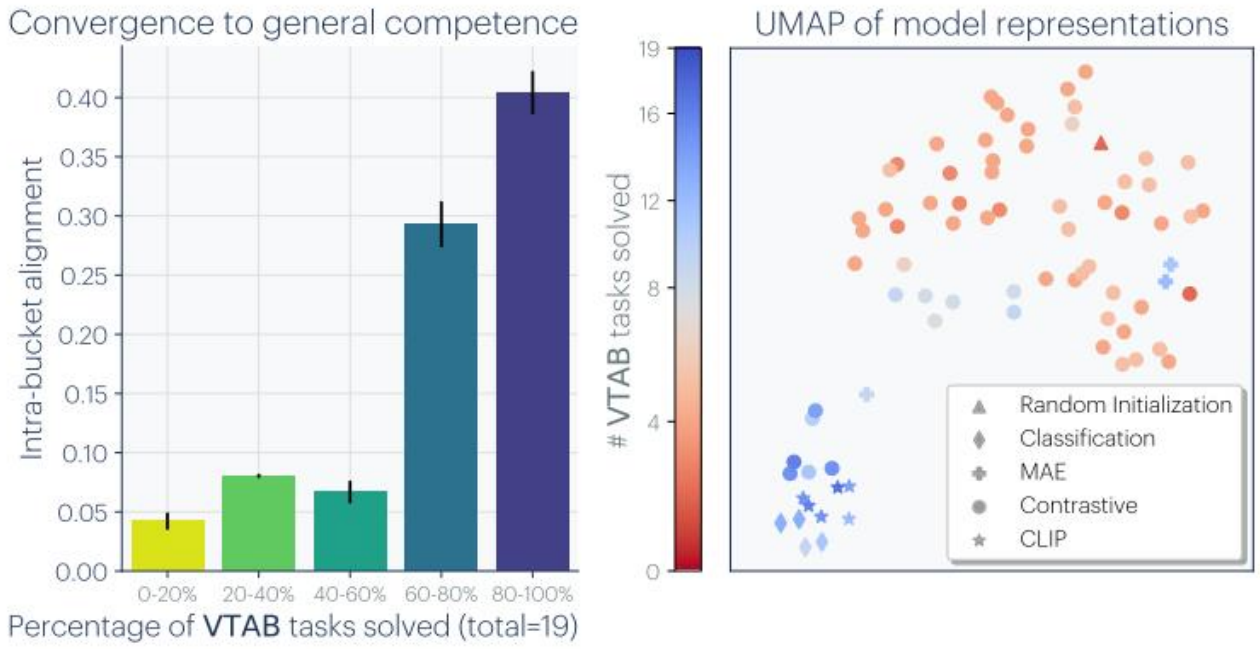
这个假说与柏拉图的洞穴寓言比较类似，其中真实世界被视为理想的形式，而我们通过感官体验到的是这些理想形式的影子或映射。

不同模态的训练数据是洞穴上不同的投影，我们假设，模型正在恢复对洞穴外实际世界的更好表示。

Different models, with different architectures and objectives, can have aligned representations

- 核心观点-表征趋同：不同的AI模型，不论其架构、训练目标或数据类型如何，其内部表示（表征）数据的方式正变得越来越相似。（model stitching ）

Alignment increases with scale and performance



强者大多相似，弱者各有不同:

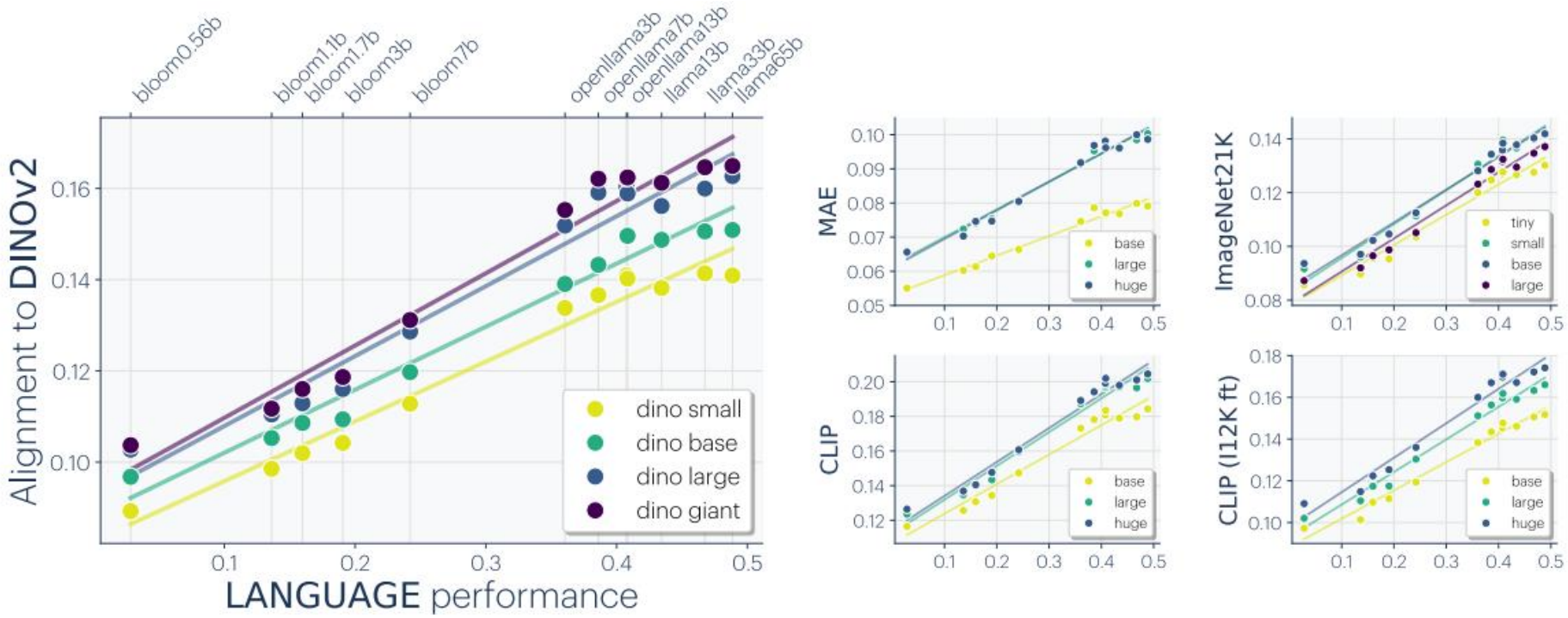
左图：解决更多VTAB（视觉任务自适应benchmark）任务的模型往往彼此更加一致。

右图：更有能力的模型间（蓝色）有更多类似的表示。更弱的模型则有更加分散的分布。

Representations are converging across modalities

- 核心观点-跨模态的趋同：不仅相同类型的模型之间存在趋同，不同数据模态（如视觉和语言）的模型在表示数据时也显示出趋同的趋势。（LLAVA）

研究目标是：随着模型规模的增大和性能的提升，不同模态的模型是否正在学习到一种通用的、模态无关的表示。以及这种通用表示是否适用于多种任务。



更强大的语言模型与更强大的视觉模型对齐得更好。这表明模型在学习过程中能够捕捉到更通用的语义信息，从而在多模态任务中表现更好。

为什么表征会趋同？--- 容量假设

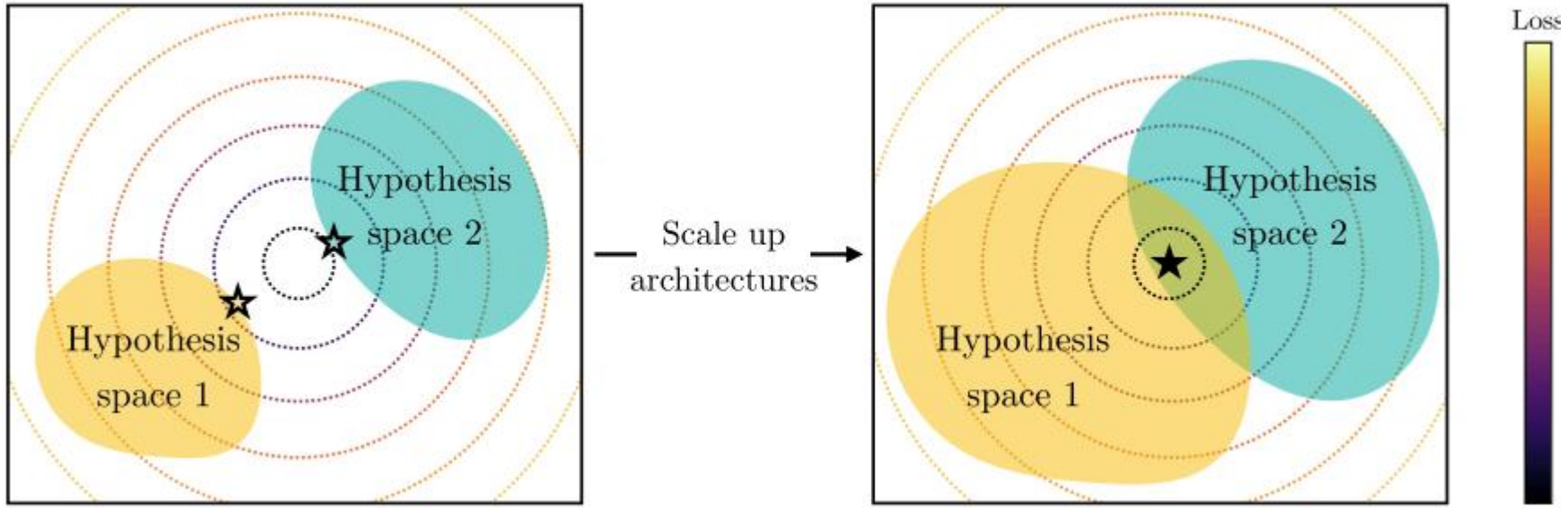
- 通过模型容量收敛（紫色）：缩放模型（即，使用更大的函数类F）以及改进的优化，应该更有效地找到对该最优表示的更好的近似。

能力假设：较大的模型比较小的模型更有可能收敛到最优表示。

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{x \sim \text{dataset}} [\mathcal{L}(f, x)] + \mathcal{R}(f)$$

trained model；training objective；function class；regularization

假设学习目标存在全局最优表示。然后，在足够的数据下，更大的模型以及改进的优化，应该更有效地找到对该最优的更好的近似，如图所示：



左图：两个小模型可能不覆盖最优，从而找到不同的解决方案。

右图：随着模型变得越来越大，它们覆盖了最优解并收敛到相同的解。

The Capacity Hypothesis

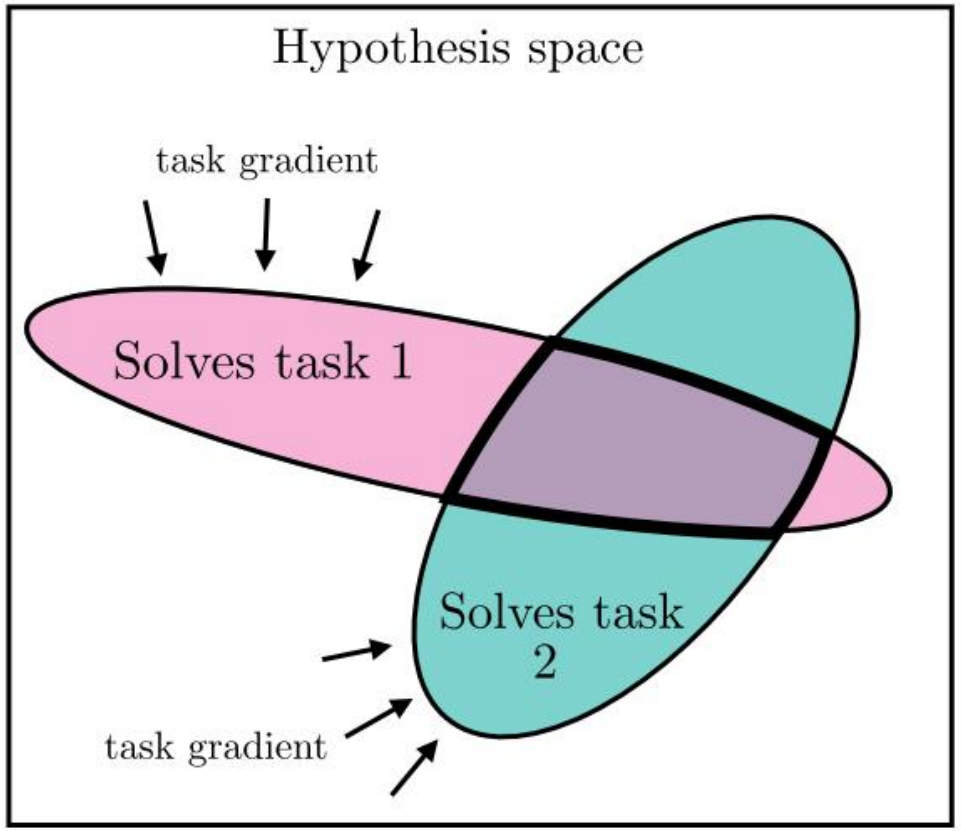
Bigger models are more likely to converge to a shared representation than smaller models.

为什么表征会趋同？--- 多任务假设

- 通过任务通用性进行收敛（绿色）：能够胜任N个任务的表征比能够胜任M < N个任务的表征要少。当我们用更多的数据训练更通用的模型来同时解决更多的任务时，我们应该期待更少的可能解决方案。

能够胜任N个任务的表征比能够胜任M < N个任务的表征要少。当我们训练更多的通用模型来同时解决更多的任务时，我们应该期待更少的可能解决方案。

$$\overbrace{f^*}^{\text{trained model}} = \arg\min_{f \in \underbrace{\mathcal{F}}_{\text{function class}}} \mathbb{E}_{x \sim \text{dataset}} [\overbrace{\mathcal{L}}^{\text{training objective}}(f, x)] + \underbrace{\mathcal{R}}_{\text{regularization}}(f)$$



每个训练数据点和目标（任务）都对模型施加了额外的约束。随着数据和任务规模的扩大，模型需要学习能够解决所有任务的表征，这样的交集是会变小的。多目标任务的解空间小于单目标任务的解空间。

The Multitask Scaling Hypothesis

There are fewer representations that are competent for N tasks than there are for $M < N$ tasks. As we train more general models that solve more tasks at once, we should expect fewer possible solutions.

为什么表征会趋同？--- 简单性偏置

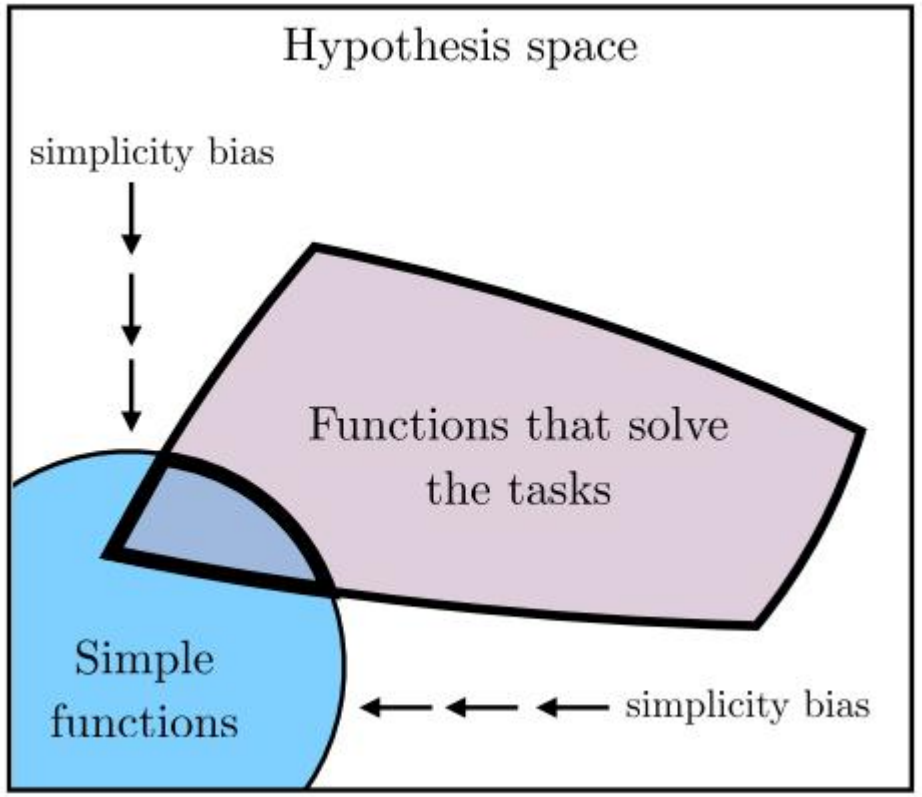
- 通过简单性偏置收敛（红色）：深度网络偏向于寻找数据的简单拟合，模型越大，这个bias就越大。因此，随着模型变大，我们应该期望收敛到更小的解空间。

对于拟合相同的数据，更大的模型对所有可能的方法有更大的覆盖范围。然而，深层网络的隐含的简单性偏向鼓励更大的模型找到这些解决方案中最简单的。

trained model / training objective

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{x \sim \text{dataset}} [\mathcal{L}(f, x)] + \mathcal{R}(f)$$

function class / regularization



深度网络偏向于寻找数据的简单拟合，这种简单性偏差可能来自深度学习中常用的显式正则化R（f）（例如，权重衰减和dropout）。然而，即使在没有外部影响的情况下，深度网络也会自然地遵循奥卡姆剃刀，隐含地倾向于适合数据的简单解决方案

The Simplicity Bias Hypothesis

Deep networks are biased toward finding simple fits to the data, and the bigger the model, the stronger the bias. Therefore, as models get bigger, we should expect convergence to a smaller solution space.

向什么样的表征收敛?

一系列对比学习模型收敛于一个关于 P(Z)（world model） 的表示：

底层现实 (Z)：世界由一系列离散事件构成，表示为：$\mathbf{Z} \triangleq [z_1, \ldots, z_T]$

观测 (X, Y)：我们通过不同方式来观测 Z。例如图像，文本，声音。假设这些映射都是双射 的，意味着每个真实事件都唯一对应一张图像和一段文本，信息没有损失 。

对比学习的目标函数：

$$\langle f_X(x_a), f_X(x_b) \rangle \approx \log \frac{\mathbb{P}(\text{pos} \mid x_a, x_b)}{\mathbb{P}(\text{neg} \mid x_a, x_b)} + \tilde{c}_X(x_a) \quad (3)$$

$$= \log \frac{P_{\text{coor}}(x_a \mid x_b)}{P_{\text{coor}}(x_a)} + c_X(x_a) \quad (4)$$

$$= K_{\text{PMI}}(x_a, x_b) + c_X(x_a), \quad (5)$$

将正负样本概率替换为共现概率，因为正样本对来自数据中的真实共现，而负样本对是随机组合的 。这个对数比率正是点互信息 (PMI) 的定义：即两个事件同时发生的概率与假设他们相互独立时发生的概率。

这证明了，对比学习的最终优化目标就是让模型内部的相似度度量（点积）去拟合真实世界数据中的 PMI。

将这个结论推广到所有模态：

因为假设观测函数是双射的，所以不同模态（图像 X, 文本 Y）保留了与底层现实 (Z) 完全相同的概率结构：

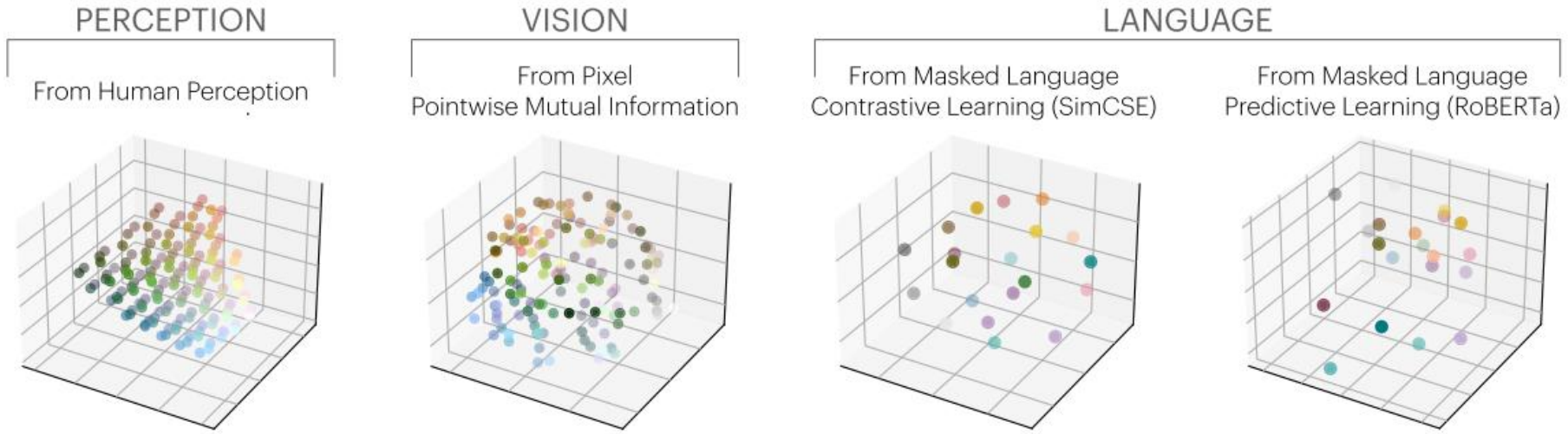
$$K_{\text{PMI}}(z_a, z_b) = \langle f_X(x_a), f_X(x_b) \rangle - c_X \quad (7)$$

$$= \langle f_Y(y_a), f_Y(y_b) \rangle - c_Y. \quad (8)$$

它表明，无论是在图像上训练的模型，还是在文本上训练的模型，只要学习方法是拟合共现统计规律，它们最终都会收敛到同一个表示内核 (representation kernel)，这个内核反映了共同的现实世界中的统计结构 (PMI) 。

向什么样的表征收敛?

当模型被训练来预测文本中共现的颜色时，学习到的颜色表示与人类对颜色距离的感知非常接近。这个实验试图在图像数据上重现这一现象，并比较文本和图像模态的学习结果。以验证在真实数据上确实发生了收敛。
感知基准：CIELAB颜色空间是一个三维空间，其中颜色之间的距离反映了人类对颜色差异的感知。



视觉共现：使用 CIFAR-10 图像数据集采样数据，计算统计点互信息 (PMI)。
语言共现：提取两个预训练好的大型语言模型（SimCSE 和 RoBERTa），对Lindsey & Brown数据集中20个核心颜色词的Embedding,计算欧氏距离，最后通过MDS转换为三维空间中的颜色点分布图 。

实验结果表明，无论是通过图像数据还是文本数据学习到的颜色表示，都与人类的感知（CIELAB 颜色空间）非常接近。这表明不同模态（视觉和语言）的模型能够学习到一致的颜色表示。

表征收敛的影响

1. Scaling 是重要的，但未必是高效的；
2. 训练数据可以跨模态共享：如果所有模型最终都趋向于一个独立于具体模态的“柏拉图式表征”，那么来自任何模态的数据都应该有助于模型学习这个表示 。这意味着，如果想训练出最好的视觉模型，不应该只用图像数据，还应该加入文本数据来训练 。事实上，这已经成为了普遍做法 。OpenAI的研究也表明，用图像数据训练模型可以提升其在文本任务上的表现 。
3. 扩大规模可能减少幻觉和偏见：如果模型随着规模的扩大，确实在不断收敛到一个更准确的现实世界模型，那么它们产生错误信息或“幻觉”的倾向就应该会随之减少 。当然，这取决于未来的训练数据是否足够好 。

反例和局限性

1. 不同模态可能包含不同的信息：不同模态（如视觉和语言）可能包含独特的信息，这些信息无法完全通过另一种模态来表达。例如，语言难以描述观看日全食的体验，而图像也无法传达“我相信言论自由”这样的抽象概念。
2. 双射投影假设：论文中的数学论证仅严格适用于Z的双射投影，这意味着所有投影中的信息等同于底层世界的信息。这并不适用于有损或随机的观测函数。
3. 社会学偏见：AI模型的发展轨迹受到研究者偏见和AI社区集体偏好的影响，这可能导致模型趋于模仿人类推理和表现，即使存在其他类型的智能。
4. 连续性和非有界世界：论文中的理想化世界模型假设了一个离散的事件序列，但在现实世界中，我们可能需要处理连续的和非有界的观测。
5. 如何测量对齐：论文中使用了特定的对齐度量方法（如PMI），但关于这些度量方法的优点和缺点存在争议。

Learning to See Before Seeing: Demystifying LLM Visual Priors from Language Pre-training

Junlin Han[1,2,†], **Shengbang Tong**[1], **David Fan**[1], **Yufan Ren**[1], **Koustuv Sinha**[1], **Philip Torr**[2], **Filippos Kokkinos**[1]

[1]Meta Superintelligence Labs, [2]University of Oxford
[†]Project lead

The Concept of a Visual Prior

作者发现，尽管LLMs仅通过文本训练，但它们却能涌现出关于视觉世界的深刻先验知识，这种能力表明语言中的统计模式可能足够丰富，能够编码视觉的基本原理，例如物体属性和空间关系，即使从未直接观察过图像。换句话说， AI 仅仅通过阅读description of the world就能学会看东西。

1. 程序化的视觉知识：LLMs能够生成可执行代码，渲染复杂的2D和3D场景，包括物体和空间布局。
2. 数据高效的视觉适应：LLMs通过在少量的图像-文本对上进行指令微调，可以实现高级推理，而无需进行大规模的多模态预训练。
3. LLMs作为强大的视觉编码器：LLMs学习到的表示可以直接用于纯视觉任务，无需语言。

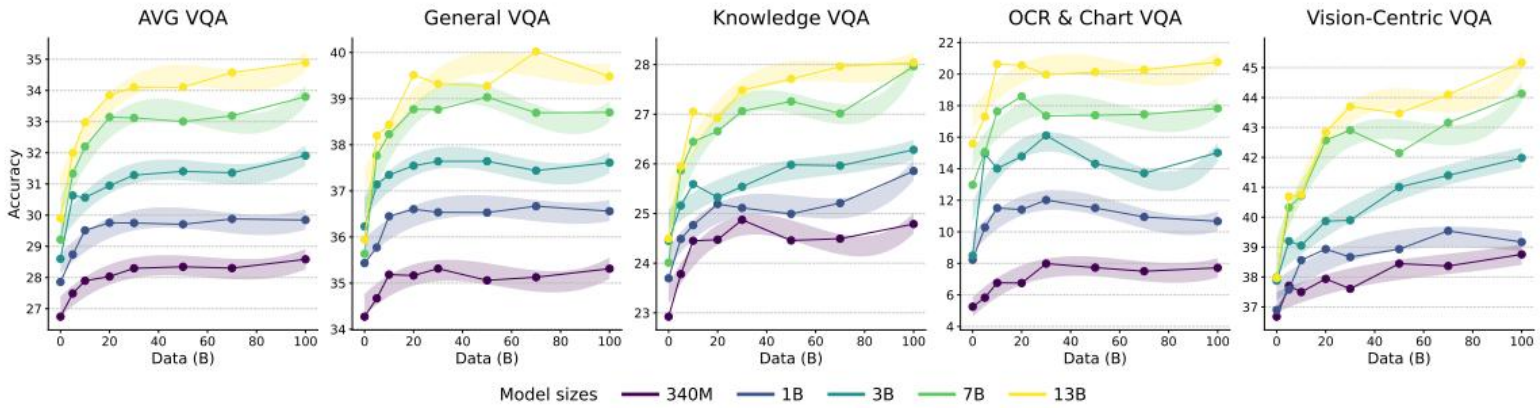
论文将这个现象拆解为三个可验证的子问题：

1. 结构问题：文本预训练所赋予的“视觉先验”是单一能力，还是可分离的复合能力？
2. 来源问题：若可分离，各个能力分别来源于哪种数据源？其 Scaling 规律与饱和点是什么？
3. 利用问题：能否基于上述规律，在纯文本预训练阶段主动配置数据配比，从而预训练出更适配视觉任务的LLM，减少后续多模态对齐成本？

Demystifying LLM Visual Priors: Studies and Findings

作者进行了一系列的受控实验，来系统地解构LLM视觉先验的起源。

首先是最基础的scale的影响，作者分析了模型和数据大小对下游多模态任务性能的影响。



在不同能力之间的这些不同的scaling模式表明，不同的视觉能力并不是scale uniformly，而是拥有不同的特性，这些特性决定了它们如何从增加的模型和数据大小中受益。

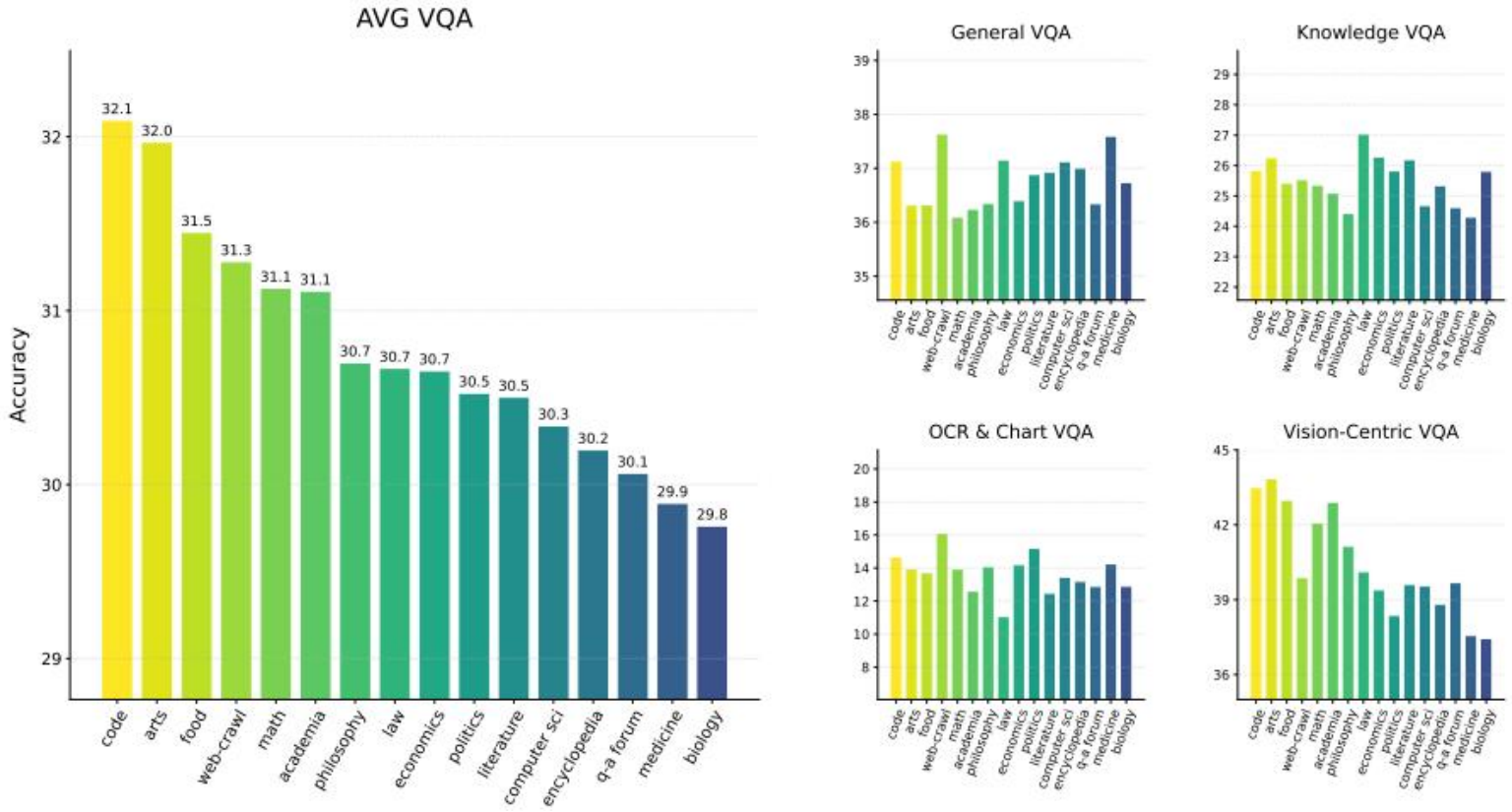
在General VQA和Knowledge VQA上的性能表现出类似的scale趋势，随着模型和数据大小的不断提高。而OCR&Chart VQA对模型大小的敏感度远远大于数据量。同时，vision-centric VQA则是更大的模型从更多的数据的scaling中提高的更快，而较小的模型则更早地停滞不前。

Finding 1: VQA performance scales positively with model and data size. However, this scaling is not uniform across all visual abilities.

Demystifying LLM Visual Priors: Studies and Findings

然后是探究了预训练数据源的影响。

将模型大小固定为3B参数，将总训练数据量固定为30B（单源数据）。



下游多模态任务的性能因数据源的不同而有显著差异。这种差异表明，不同类别的文本数据导致了不同的视觉先验。其中，vision centric VQA任务上的出色表现与两类数据高度相关：以推理为中心的(例如，代码、数学、学术)和丰富的视觉世界描述的语料库(例如，艺术、食物)。

特定类别的语言预训练数据可以增强结果MLLM中的某些视觉能力；特别是reasoning-centric和related to the visual world的数据显著提高了以视觉为中心的任务的表现。

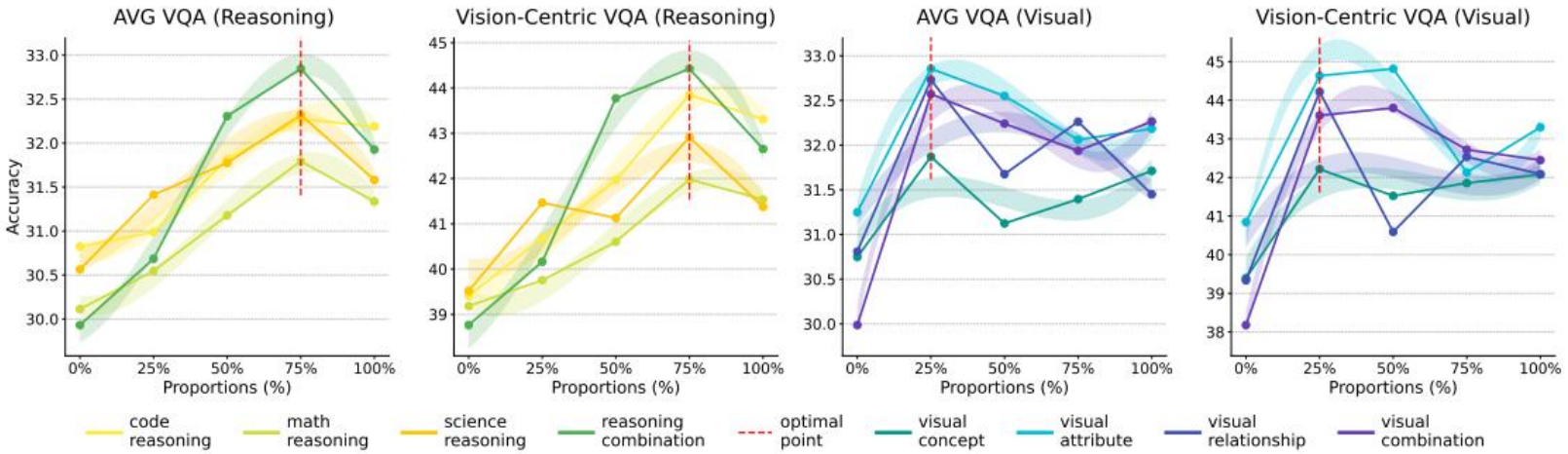
Finding 2: Specific categories of language pre-training data can enhance certain visual capabilities in the resulting MLLM; in particular, data related to reasoning and the visual world significantly improve performance on vision-centric tasks.

Demystifying LLM Visual Priors: Studies and Findings

Impact of reasoning and visual data categories and proportions.

The reasoning-centric data was partitioned into **code reasoning**, **math reasoning**, **science reasoning**, and a **reasoning combination** category, which aggregates the three aforementioned categories. Concurrently, we define four categories for data related to the visual world:

- **visual concept**: Text naming visual entities like objects, people, places, and scenes.
- **visual attribute**: Descriptions of visual properties such as color, shape, texture, and style.
- **visual relationship**: Language detailing spatial arrangements or part-whole connections.
- **visual combination**: A combination of all three visual categories.



reasoning-centric data带来渐进且显著的性能提升，并且比例大概在75%的时候达到平台期。而Visual world data 大概在25%左右就饱和了。

少量的Visual world data是重要的，但它的贡献很快就会饱和；相比之下，增加reasoning-centric data在预训练混合中的比例会逐渐增强视觉能力，大概在**75%**达到平台期。

Finding 3: A small amount of data about the visual world is crucial, but its contribution saturates quickly; in contrast, increasing the proportion of reasoning-centric data in the pre-training mix progressively enhances visual abilities, with performance gains observed up to a 75% ratio.

Demystifying LLM Visual Priors: Studies and Findings

接下来作者想找到一个最佳的数据混合方案，在语言任务上表现优异，作为MLLM的base也能有比较好的表现。

1. identify a data mixture that excels at visual tasks.

强大的视觉基础能力不是通过简单地最大限度地增加视觉描述来建立的，而是通过建立强大的推理能力来建立的。

2. Language-favorable mixture.
确定Mix0，作为text的基线，在实验中获得了最高的text acc(53.0%)和最低的困惑度。

3. Balanced mixture.
进行一系列内插实验，寻找平衡两种任务的性能的数据配比。

Data Ratio		Avg VQA	Data Ratio		Avg VQA
reasoning	visual		reasoning	visual	
50	5	30.7	55	5	30.9
	10	31.3		10	31.7
	15	31.8		15	32.2
60	5	31.9	65	5	32.0
	10	32.4		10	32.2
	15	**32.7**		15	32.5
	20	32.5		20	32.1
	25	32.4		25	31.9
	30	31.6		30	31.4
70	5	31.9	75	5	31.6
	10	32.3		10	31.5
	15	32.6		15	32.4
80	5	31.5	85	5	31.2
	10	32.4		10	31.6
	15	32.2		15	31.8

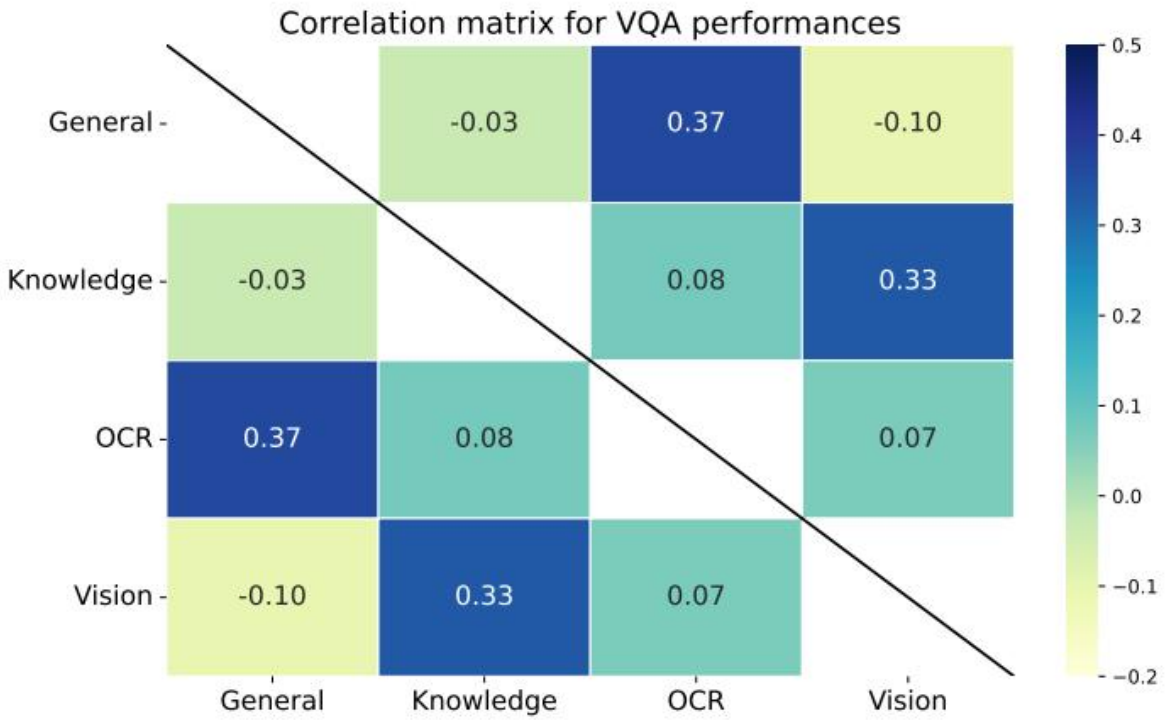
Recipe		Data Source Mixture (%)								Performance Metrics			Overall Rank
		web-crawl	encyclopedia	academic	literature	math	code	reasoning	visual	t-acc (%)	ppl (↓)	v-acc (%)	
mix0	language	50.0	2.5	2.5	20.0	5.0	20.0	33.1	21.7	**53.0**	**13.46**	32.4	5
mix1		48.3	3.4	2.9	17.0	5.8	22.5	36.2	20.6	52.8	13.48	32.4	4
mix2		46.7	4.3	3.3	14.0	6.7	25.0	39.4	19.4	52.6	13.51	32.6	8
mix3		45.0	5.2	3.8	11.0	7.5	27.5	42.6	18.2	52.5	13.56	32.9	9
mix4	↑	43.3	6.1	4.2	8.0	8.3	30.0	45.7	17.1	52.4	13.62	32.7	10
mix5	↕	41.7	7.1	4.6	5.0	9.2	32.5	48.9	16.0	52.6	13.57	33.0	6
mix6	↓	40.0	8.0	5.0	2.0	10.0	35.0	52.0	14.8	52.7	13.52	33.3	**1**
mix7		36.5	7.0	7.5	2.0	11.5	35.5	55.5	14.4	52.5	13.56	33.1	3
mix8		33.0	6.5	9.5	2.0	12.0	37.0	57.2	14.0	52.7	13.52	33.2	2
mix9		29.5	6.0	11.5	2.0	12.5	38.5	59.0	13.6	52.3	13.71	33.2	7
mix10	vision	26.0	5.5	12.5	2.0	13.0	41.0	61.3	13.3	52.1	13.88	**33.4**	11

最大化MLLM在VQA任务上的性能，需要大量reasoning-centric data（50%），以及少量必要的Visual world data（15%）。通过在语言友好的数据和视觉友好的数据之间进行校准的数据混合，可以达到语言能力和视觉能力之间的平衡点。

Demystifying LLM Visual Priors: Studies and Findings

The structure and origin of learned visual priors.

现在综合之前的findings来研究视觉先验的内部结构，他是单一可分离的还是统一的视觉能力?
首先将视觉先验概念化为不同能力的集合，每种能力都由四个VQA类别中的一个来衡量。



综合前面的105个不同版本的3B模型的性能数据，计算得到Spearman相关性矩阵。结果表明，视觉先验可能包含至少两种不同的能力类型。

1. 感知能力：包括General和OCR；
2. 推理能力：包括Knowledge和Vision-Centric任务；

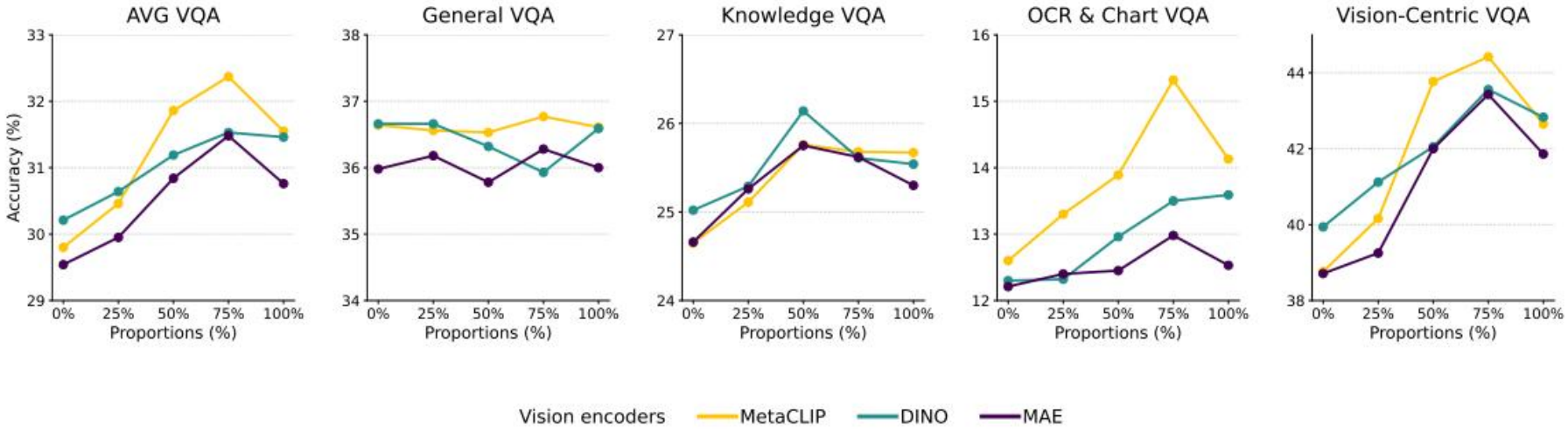
并且感知能力和推理能力之间的相关性非常弱，甚至略为负相关，说明他们是相对独立的。
推理能力主要来源于reasoning-centric data。感知能力的起源则更加分散，在前面实验中，研究者发现，web-crawl数据训练的模型在General和OCR上表现最佳。这说明，感知能力可能源于llm处理大量多样化语言数据时的自然涌现，而不是某个特定数据类别的直接贡献。

Finding 5: The learned visual prior is not a single entity but decomposes into at least a perception prior and a reasoning prior with different origins.

Demystifying LLM Visual Priors: Studies and Findings

Deconstructing multimodal abilities: vision or language.

进行了进一步的分析，首先验证所学到的视觉先验的普遍性，然后剖析不同多模态能力的来源，区分那些更多地继承自语言模型（LLM）的能力和那些更多地通过视觉指令微调获得的能力。



横轴表示与训练中与推理相关的数据比例。首先，确认了推理先验的普遍性。对于推理密集型任务，随着LLM预训练中推理数据比例的增加，所有三种视觉编码器配置都表现出几乎相同的、明显的性能上升趋势。这表明，在LLM中培养的视觉推理先验是一个基础性的、与模态无关的先验。

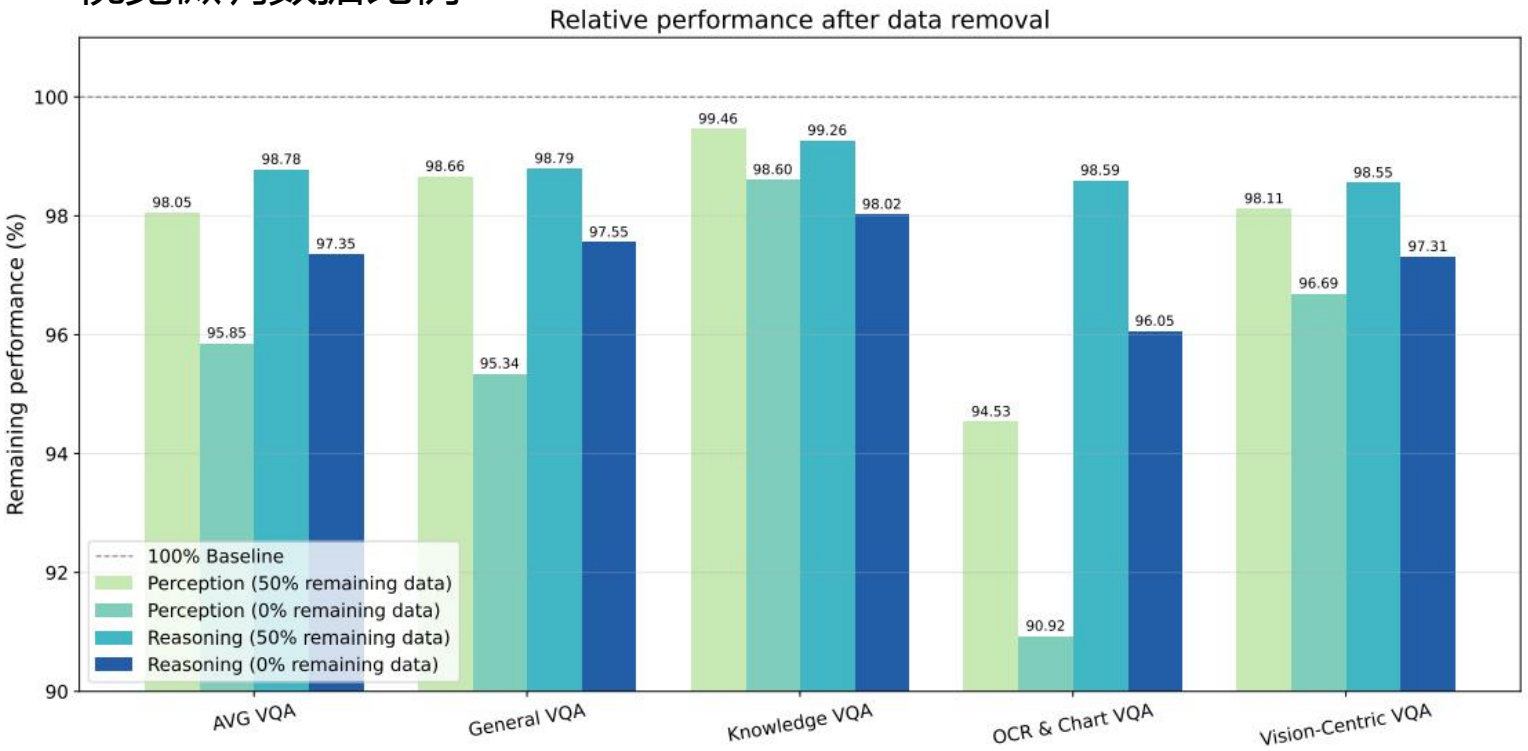
相比之下，感知先验缺乏这种普遍性。对于感知导向的任务，不同视觉编码器的性能趋势更加不一致。它们并没有遵循一个统一的模式，而是不同视觉编码器的性能曲线彼此不同。这表明感知能力对视觉编码器的具体特性更加敏感，而不是仅仅依赖于预训练数据。

Finding 6: Visual reasoning ability is primarily shaped by reasoning prior acquired from language pre-training; perception ability is more dependent on post-training (visual instruction tuning).

Demystifying LLM Visual Priors: Studies and Findings

Deconstructing multimodal abilities: vision or language.

感知和推理--是否主要来自LLM的视觉先验或视觉微调。
视觉微调数据比例：



减少感知数据的影响：
感知密集型任务：在OCR & Chart和General任务上，减少感知数据导致性能大幅下降。
推理任务：在Vision-Centric和Knowledge任务上，减少感知数据导致性能适度下降。

减少推理数据的影响：
感知任务：在OCR & Chart和General任务上，减少推理数据导致性能小幅下降。
推理任务：在Vision-Centric和Knowledge任务上，减少推理数据导致性能适度下降。

•推理能力的普遍性：推理能力的普遍性表明，通过增加推理类数据的比例，可以显著提升模型在推理任务上的表现，而不依赖于具体的视觉编码器。
•感知能力的依赖性：感知能力的依赖性表明，为了提升感知能力，可能需要针对不同的视觉编码器进行优化，并进行更多的视觉监督微调。

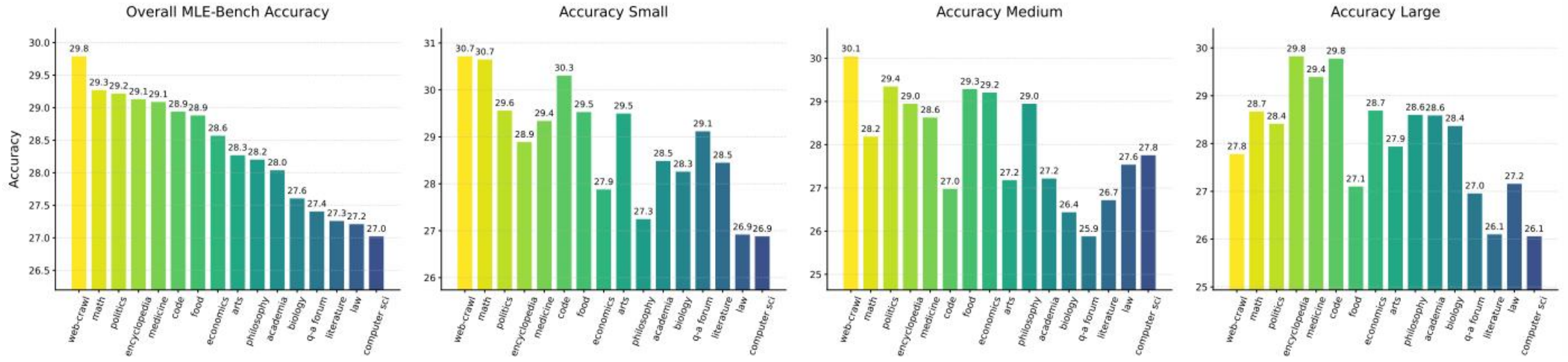
Discussion and Hypotheses

The structure and origin of learned visual priors.

之前的分析表明，感知先验的起源是分散的，主要从多样化的数据中涌现出来。这意味着感知能力并不是由某个特定的数据类别直接训练出来的，而是通过处理大量多样化的数据自然形成的。这引发了对感知先验内部结构的进一步思考：感知先验是一个统一的能力，还是具有更细致的特征？

为了进一步研究这种涌现的感知能力，并更直接地描述其特性，研究者引入了一个多级存在基准（MLE-Bench）。

MLE-Bench包含关于图像中物体或场景存在性的四选一问题。这些问题根据目标物体的相对大小进行分类，大小通过物体占据的像素百分比来衡量。

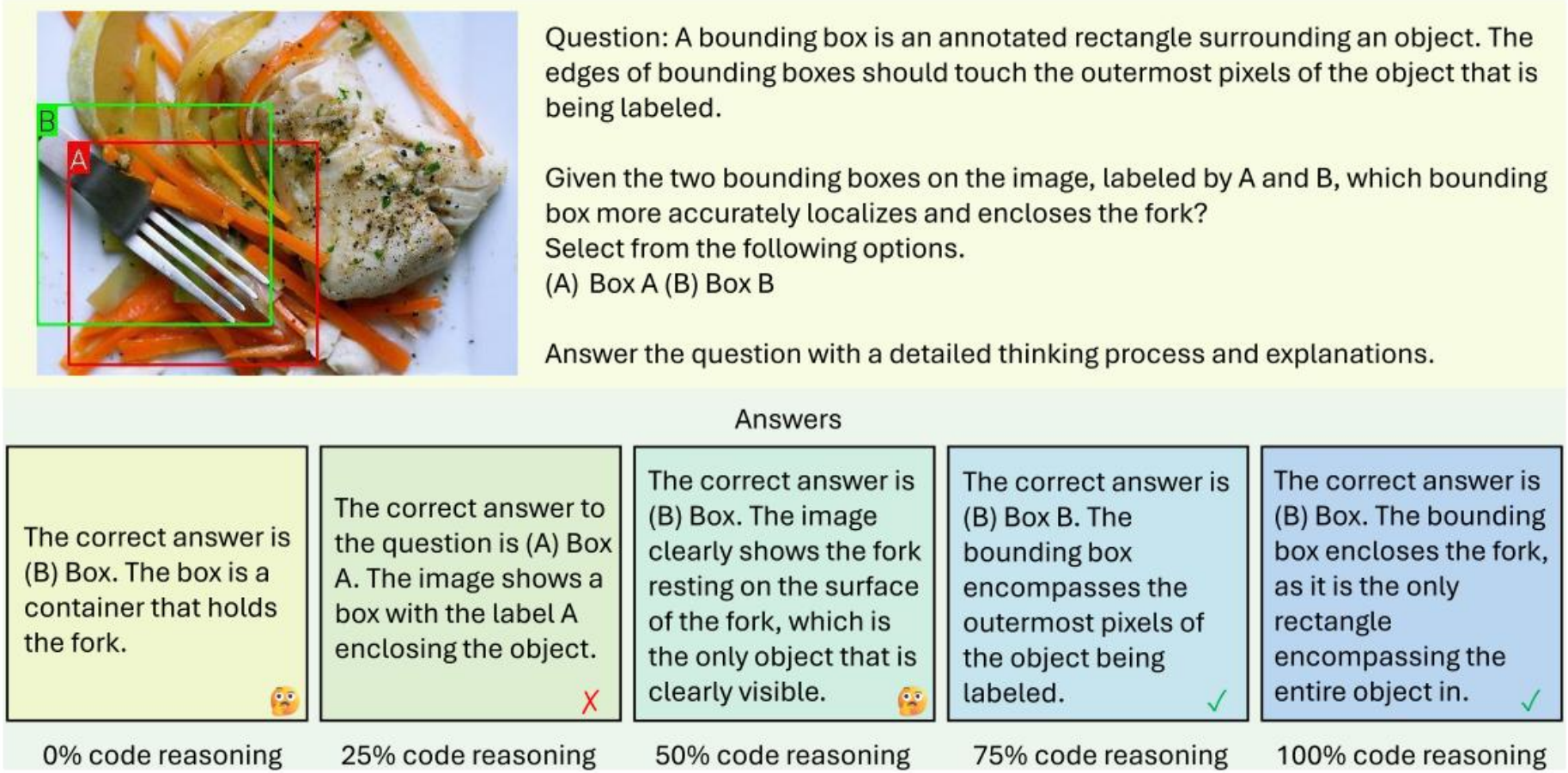


网页抓取数据的模型表现最佳（证实了数据多样性是先验感知的关键，其中可能包括了更多更大的场景下的更小的东西，提高了模型对更小的实体的敏感性），在小到中等大小物体的优势明显，在大物体上的表现差距缩小。这些结果表明感知先验确实是尺度依赖的。从不同数据获得的感知先验表现出尺度依赖性，其作用对于感知中小型物体最为明显。

Discussion and Hypotheses

Is reasoning a universal, cross-modal skill already acquired during pre-training?

本文前面的findings支持了这个假设：LLM从文本中获得的推理能力是模态无关的，可以扩展到视觉领域。
语言模型（LLM）通常通过大量的文本数据进行预训练，这些文本数据中包含逻辑、结构和组合性的信息。通过预训练，模型能够学习到这些抽象和概括的原则。



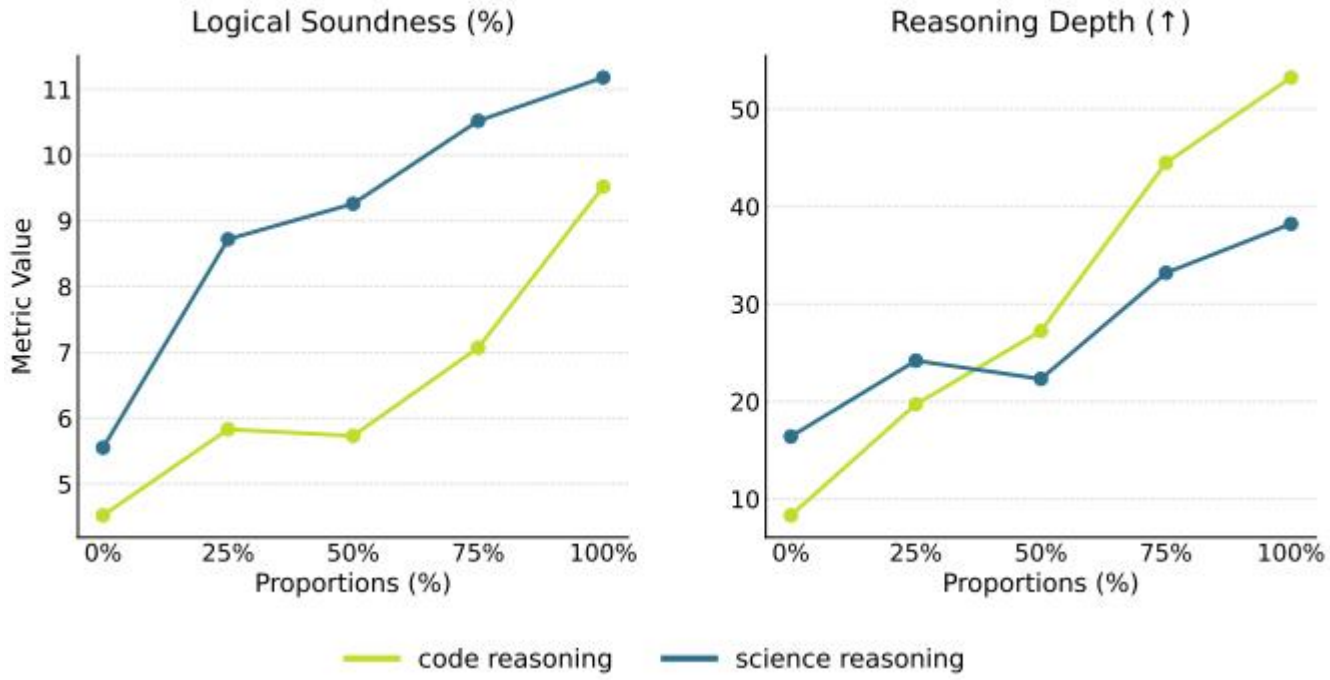
Qualitative impact of reasoning-centric data on visual reasoning.

Discussion and Hypotheses

Is reasoning a universal, cross-modal skill already acquired during pre-training?

使用一个单独的LLM as a judge;
评估标准：
(1)逻辑可靠性：连贯和合理的推理痕迹的百分比；
(2)推理深度：通过文本计数衡量得出结论的推理痕迹的平均长度。

我们观察到一个明显的趋势：随着以推理为中心的数据比例的增加，模型生成的视觉推理在逻辑上更可靠，而且明显更长。这表明，该模型正在应用从文本中学习的通用、抽象的推理框架来解决视觉问题。



Qualitative impact of reasoning-centric data on visual reasoning tasks.

Hypothesis 2: The reasoning capabilities an LLM acquires from text are fundamentally modality-agnostic. Language reasoning skills can be directly transferred to solve visual problems.

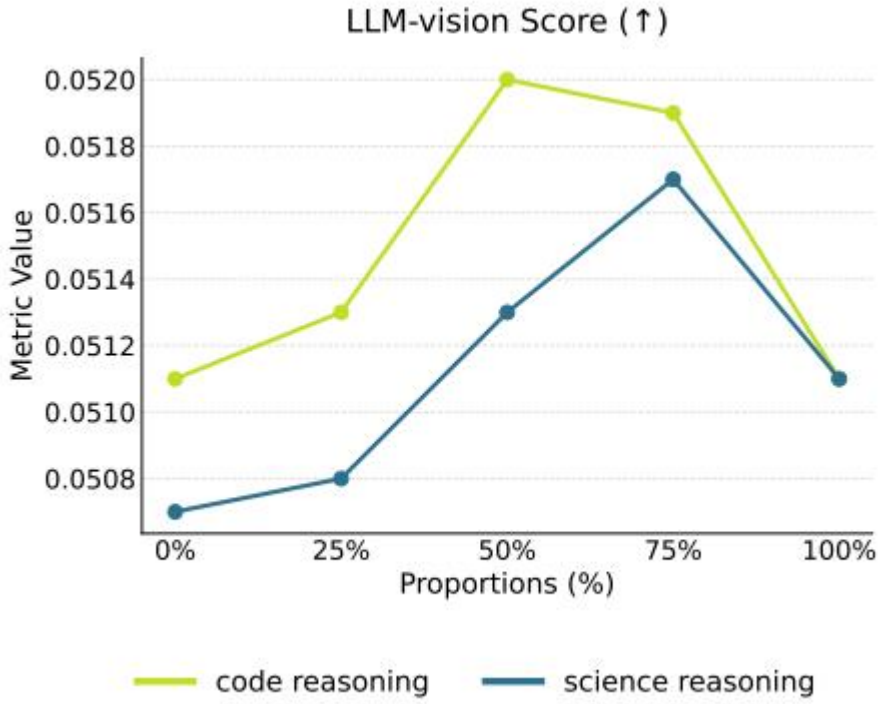
Discussion and Hypotheses

Does language data structure drive representational alignment with vision?

代码和数学等领域的数据本质上是高度结构化的。它受严格的语法、逻辑依赖和层次结构的控制。同样，视觉数据远不是像素的随机集合。它有自己丰富的结构：比如空间关系和层次结构。

因此，本文假设，这种共享的结构基础意味着从结构化文本中学习的表征本质上更类似于视觉领域，因此更容易转移到视觉领域。

随着结构化推理数据比例的增加，对齐分数通常会提高，这表明从抽象结构中会学习到更一致的latent space。然而，在75%达到峰值，然后下降。这可能是因为模型完全在推理数据中学习抽象结构，但缺乏来自其他文本类型的必要词汇来有效地将其映射到不同的视觉概念中。



Hypothesis 2: The reasoning capabilities an LLM acquires from text are fundamentally modality-agnostic. Language reasoning skills can be directly transferred to solve visual problems.

Scaling Up and Training a Vision-Aware LLM

Building upon our findings, we scale up our approach to validate our findings and develop a vision-aware LLM on a larger-scale. The goal is to test whether the principles identified in our controlled, smaller-scale studies hold true when applied to larger training runs. To this end, we pre-train two 7B parameter LLMs, each on 1T tokens, based on the two data mixtures identified previously:

- **Language-favorable model**: Following the mix0 mixture, which is the best-performing blend for pure language tasks.
- **Balanced model**: Based on the mix6 recipe, our proposed balanced mixture is designed to deliberately cultivate strong visual priors without compromising language proficiency.

Model	Language		Vision				
	ppl	avg acc	General	Knowledge	OCR&Chart QA	Vision-Centric	Overall
Language-Favorable	8.72	0.647	46.92	28.35	21.49	46.31	37.32
Balanced	7.49	0.655	49.59	29.02	23.63	46.59	38.64

Balanced model在language任务上表现是旗鼓相当的，并且在所有视觉任务上比Language-Favorable model 表现好。在预训练时观察到的一个有趣的现象是，Balanced model的语言表现最初落后，在大约600Btoken后开始反超。这可能表明，当预训练token数足够大时，当以大量的世界知识为基础时，与推理相关的token数量的好处可以更有效地释放出来，最终导致语言方面的强劲表现。

Hypothesis 2: The reasoning capabilities an LLM acquires from text are fundamentally modality-agnostic. Language reasoning skills can be directly transferred to solve visual problems.