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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large Language Models (LLMs), despite being trained on text alone, surprisingly develop rich visual priors. These priors allow latent visual capabilities to be unlocked for vision tasks with a relatively small amount of multimodal data, and in some cases, to perform visual tasks without ever having seen an image. This paper aims to demystify this phenomenon. Through systematic analysis, we reveal that these priors are not uniform but are composed of separable 'perception' and 'reasoning' priors with unique scaling trends and origins. We show that an LLM's latent visual reasoning ability is predominantly cultivated by pre-training on reasoning-centric data (*e.g.*, code, math, academia) and scales progressively. This reasoning prior acquired from language pre-training is transferable and universally applicable to visual reasoning. In contrast, the perception prior emerges more diffusely from broad corpora, and perception ability is more sensitive to the vision encoder and visual instruction tuning data. In parallel, text describing the visual world proves crucial, though its performance impact saturates rapidly. Leveraging these insights, we propose a data-centric recipe for pre-training vision-aware LLMs. The resulting 7B model trained on this recipe for 1T tokens, demonstrates stronger vision capabilities without compromising language proficiency. Our findings are grounded in over 100 controlled experiments consuming 500,000 GPU-hours, spanning the full MLLM construction pipeline—from LLM pre-training to visual alignment and supervised multimodal fine-tuning—across five model scales, a wide range of data categories and mixtures, and multiple adaptation setups. Together, this work provides a new way of deliberately cultivating visual priors from language pre-training, paving the way for the next generation of multimodal LLMs.

1 Introduction

A compelling phenomenon has emerged at the forefront of AI research: Large Language Models (LLMs), despite being trained exclusively on vast corpora of text, appear to develop profound priors about the visual world. This latent capability is paradoxical, suggesting that the statistical patterns within language might be rich enough to encode fundamental principles of vision, from object properties to spatial relationships, without ever observing a single image. This emergent visual prior presents in several surprising and powerful ways:

- **Programmatic visual knowledge.** LLMs possess a rich visual knowledge, enabling them to generate executable code that renders complex 2D and 3D scenes, from objects to spatial layouts (Sharma et al., 2024; Sun et al., 2025; Ge et al., 2025; Ashutosh et al., 2025). This demonstrates a grasp of visual concepts without ever seeing a single image. The resulting synthetic data is of sufficient quality to pre-train standard vision models for successful generalization to real-world images (Sharma et al., 2024).

Submitted to NeurIPS 2025 Workshop on Efficient Reasoning. Do not distribute.

- **Data-efficient visual adaptation.** LLMs are highly efficient for visual adaptation. With a vision encoder, high-level reasoning emerges from instruction tuning on a small scale of image-text pairs, bypassing the need for massive multimodal pretraining (Alayrac et al., 2022; Liu et al., 2023a; Li et al., 2023; Grattafiori et al., 2024; Tong et al., 2024a; Bai et al., 2025b). This data-efficient instruction tuning extends to unified models, where visual generation are unlocked with minimal data (Tong et al., 2024b). Furthermore, this efficiency enables adaptation to low-level visual tasks using vision-only data (Zheng et al., 2024; Du et al., 2025), proving that an LLM's reasoning framework can function independently of cross-modal alignment.

- **LLMs as strong vision encoders.** The learned representations of LLMs can directly benefit pure vision tasks without languages (Kumar et al., 2024; Pang et al., 2024; Lai et al., 2024; Bai et al., 2025a). When repurposed as visual encoders, the transformer layers of LLMs offer competitive performance on image classification, segmentation, and video understanding, even surpassing vision-specific backbones (Pang et al., 2024). These findings suggest that the hierarchical abstraction and long-range dependency modeling intrinsic to LLMs are not modality-specific, but rather capture general-purpose computational motifs that are well-suited to processing visual signals. This is also shown in neuron-level studies, which have identified multimodal neurons within LLMs that respond to the same abstract concept regardless of whether it is presented through text or vision (Schwettmann et al., 2023; Pan et al., 2023; Verma et al., 2024).

Collectively, these phenomena are not isolated curiosities; they point toward a deeper principle of representation learning. They lend strong empirical support to the Platonic Representation Hypothesis (Huh et al., 2024b; Jha et al., 2025), which posits that as models scale across diverse data and tasks, their latent representations—whether trained on text or images—converge toward a shared, underlying statistical model of reality. In this view, text and images are different "projections" or "shadows" of the world, and a powerful enough model can learn the structure of the world itself from any single projection. The visual priors in LLMs, therefore, may be a direct consequence of them recovering this unified internal world model from text alone.

These observations motivate a systematic investigation into the visual priors that LLMs acquire from text-only pre-training. We frame visual priors not as direct perceptual faculties, but as implicit knowledge or prior vision capabilities encoded in LLMs, whose primary effect is to grant both enhanced capability for vision tasks and greater ease of transfer to vision. We seek to determine their origins, dissect whether they form a uniform block of knowledge or are composed of distinct, separable abilities, and explore how they can be leveraged to build more capable MLLMs. Our methodology is centered on controlled ablation studies (Allen-Zhu, 2024), where we deconstruct the sources of different visual capabilities. By carefully manipulating pre-training model scale, data scale, data categories, data mixing ratios, vision-encoders components, and visual instruction tuning data, we reveal the underlying laws that govern them.

Our work presents the first systematic investigation into the nature and origins of visual priors in the pre-training of LLMs and shows three key contributions:

- **Structure of visual priors.** We establish that visual priors can be decomposed into perceptual and reasoning components.

- **Source of visual priors.** We identify that the model's latent visual reasoning is predominantly cultivated by and scales progressively with reasoning-centric data, whereas its perception ability emerges more diffusely from broad, diverse data.

- **Vision-aware language pre-training.** We propose a pre-training data-mixing strategy that strategically balances reasoning-centric and visually descriptive text to deliberately cultivate powerful visual priors for training LLMs that can result in stronger multimodal performance.

Beyond our primary findings, our work also introduces two resources valuable for the MLLM community: **(1): Blind visual instruction tuning:** A tric that serves as both a practical tool for improving visual adaptation and a probe to reveal how models can "hack" visual tasks with language. **(2): The Multi-Level Existence Bench (MLE-Bench):** A new benchmark specifically designed for the fine-grained evaluation of a model's perceptual abilities.

Ultimately, by demystifying the textual origins of these visual priors, this work contributes to a more fundamental understanding of how complex, seemingly modality-specific capabilities are encoded within language, thereby offering a clearer picture of the internal "world models" that foundation models learn from text alone and providing empirical support for the Platonic Representation Hypothesis.

2 Problem Formulation

In this section, we introduce our default training and evaluation settings.

2.1 Training protocol.

LLM pre-training setup. We follow standard practices and pre-train a suite of decoder-only Transformer models that closely adhere to the Llama-3 architecture (Grattafiori et al., 2024), spanning five model scales: 340M, 1B, 3B, 7B, and 13B parameters. These models are trained for varying numbers of tokens at 0B, 5B, 10B, 20B, 30B, 50B, 70B, 100B, and up to 1T tokens. We use a tokenizer with a vocabulary size of approximately 32000. Training is performed using the AdamW optimizer (Loshchilov and Hutter, 2017) with a peak learning rate of 3×10^{-4}, following a cosine decay schedule and a warm-up over the first 1024 steps. All models are trained with a context length of 2048 tokens and an effective global batch size of 1024. We fix the model size to 3B parameters and the total training data volume to 30B tokens as our default setting.

LLM pre-training data. Our training data is composed of a diverse mixture of 16 sources, including `academic`, `arts`, `biology`, `code`, `computer science`, `economics`, `encyclopedia`, `food`, `law`, `literature`, `mathematics`, `medicine`, `philosophy`, `politics`, `q-a forum`, and `web-crawl`. Each source contains at least 50B tokens.

MLLM adaptation setting. We adopt a two-stage adaptation strategy following Cambrian-1 (Tong et al., 2024a) and Web-SSL (Fan et al., 2025), consisting of visual alignment and supervised fine-tuning. In the first stage, we train an MLP-based projector on top of a frozen vision encoder and language model to align visual features with the LLM. Unless otherwise specified, we use MetaCLIP-B/16 (Xu et al., 2023) as the default vision encoder. Extracted visual features are uniformly resized to a fixed length of 576 tokens. In the second stage, we perform supervised fine-tuning on a mixture of vision-language and language-only instruction data to enhance the model's multimodal instruction-following ability. Both the alignment and instruction tuning stages use the AdamW optimizer with a cosine learning rate schedule and linear warm-up, and models are trained for a single epoch. During alignment, we use a learning rate of 1×10^{-3} with a warm-up ratio of 6%. For instruction tuning, the learning rate is set to 4×10^{-5} with a 3% warm-up. Training is conducted with an effective global batch size of 512.

MLLM adaptation data. We adopt the Cambrian-1 and Web-SSL data pipeline, but with strategic data reductions to highlight the effect of vision priors. The initial alignment stage utilizes a 1M image-caption dataset, which is roughly 40% of the original dataset's size. This is followed by supervised fine-tuning on a curated 3.5M subset of the Cambrian-7M data. This subset is balanced with approximately 1.5M language-only and 2M vision-language paired instructions, resulting in a higher percentage of language-only instruction data than the original curation, as our models learn to follow language instructions during this phase.

For all experiments, we fix the random seed as 42. For MLLM adapatation, we use the same order for data loading in both alignment and instruction tuning stage to get stable results.

2.2 Evaluation protocol.

LLM evaluation. We conduct a comprehensive evaluation of our pre-trained models' language understanding and reasoning abilities. Following the benchmark suite used in Mamba (Gu and Dao, 2023) and GLA (Yang et al., 2023), we assess performance on two main fronts. For raw language modeling quality, we report the averaged perplexity (ppl) across Wikitext (Merity et al., 2016) and LAMBADA (Paperno et al., 2016). For reasoning, we evaluate zero-shot performance on a diverse suite of commonsense and question-answering tasks, including PIQA (Bisk et al., 2020),

HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), ARC (Clark et al., 2018), Copa (Reddy et al., 2019), SciQA (Auer et al., 2023), OpenbookQA (Mihaylov et al., 2018), and BoolQA (Clark et al., 2019). For a concise comparison, we report the unweighted averaged accuracy over all these benchmarks.

MLLM evaluation. To comprehensively assess the multimodal capabilities of our models, we follow Cambrian-1 and established a diverse evaluation suite comprising 16 public benchmarks. We group these benchmarks into four key categories to to isolate and probe the distinct components of the learned visual prior, ranging from fine-grained perception to abstract reasoning, and provide a holistic view of model performance:

- General: This category probes the model's ability to perform visual perception and connect it with commonsense knowledge, rather than complex, multi-step reasoning. It includes GQA (Hudson and Manning, 2019), MME (Fu et al., 2023), MMBench (Liu et al., 2024c), and SEED (Ge et al., 2023).
- Knowledge: This category evaluates the model's capacity to connect visual information with world and perform multi-step reasoning to solve complex scientific or mathematical problems. It covers ScienceQA (Lu et al., 2022b), MMMU (Yue et al., 2024), AI2D (Hiippala et al., 2021), and MathVista (Lu et al., 2023).
- OCR & Chart VQA: This category focuses on fine-grained perception, specifically the ability to accurately read and interpret dense textual and structured data within images. It comprises TextVQA (Sidorov et al., 2020), ChartQA (Masry et al., 2022), and OCRBench (Liu et al., 2023b).
- Vision-Centric: This category mainly probes abstract visual reasoning and rough perception skills, requiring the model to perform tasks such as spatial and 3D understanding, object counting, and IQ tests. It uses benchmarks including RealWorldQA (xAI, 2024), Blink (Fu et al., 2024), COCO, ADE, and Omni3D. COCO (Lin et al., 2014), ADE (Zhou et al., 2019), and Omni3D (Brazil et al., 2023) are proposed as CV-Bench from Cambrian-1 (Tong et al., 2024a).

The overall average result is based on all benchmarks. We also report the averaged multimodal evaluation accuracy for each category. To fairly assess a model's core vision ability, independent of its instruction-following capabilities, we address a challenge: models, especially smaller ones, often embed correct answers within conversational text rather than providing a direct response. Our evaluation uses a robust parsing strategy to extract the intended answer from this free-form text. This approach ensures a reliable assessment of all models, including those without language pre-training but only visual instruction tuning, making our results resilient to variations in response formatting. More details about this parsing strategy is presented in Appendix I.

LLM-vision alignment. To quantify the representational convergence and similarity between language and vision modalities, we measure the alignment between the feature spaces of LLMs and pretrained vision models, following the methodology of the Platonic Representation Hypothesis (Huh et al., 2024a). For this analysis, we use image-caption pairs from the Wikipedia-based Image Text (WIT) dataset (Srinivasan et al., 2021). Given an image x_i and its caption y_i, we compute vision and language kernels, K_{vision} and K_{lang}, from their respective model representations, f_{vision} and f_{lang}:

$$K_{\text{vision}}(i,j) = \langle f_{\text{vision}}(x_i), f_{\text{vision}}(x_j) \rangle$$
$$K_{\text{lang}}(i,j) = \langle f_{\text{lang}}(y_i), f_{\text{lang}}(y_j) \rangle$$

We then assess the alignment between these kernels using the mutual nearest-neighbor (m_{NN}) metric, which calculates the average overlap of the k-nearest neighbor sets (with $k = 20$) for each pair. The final alignment score for a given LLM is reported as the average of its m_{NN} scores against three strong vision backbones: ViT-Large (Dosovitskiy et al., 2021) (trained on ImageNet-21K (Deng et al., 2009)), DINOv2-Giant (Oquab et al., 2023), and CLIP-Huge (Radford et al., 2021).

For all evaluations, we fix the random seed as 42 and use temperature 0 for testing to get consistent results.

The rest of results are presented in Appendix.

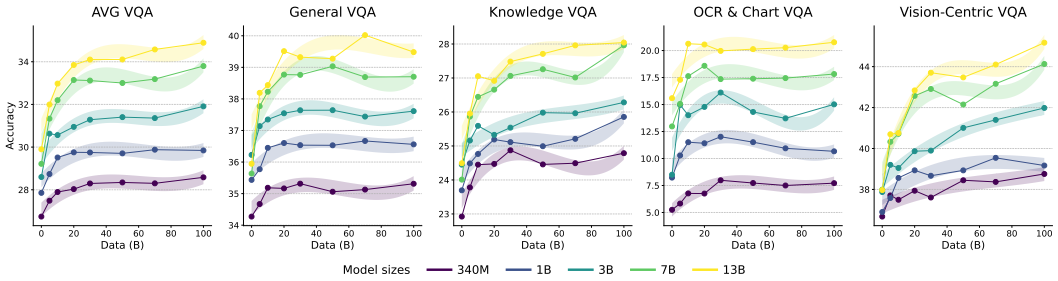

Figure 1: **Impact of model and data sizes.** The plots illustrate the performance of MLLMs, built upon LLMs of five different sizes (340M to 13B parameters), as a function of the amount of `web` pre-training data (0B to 100B tokens). The general trend shows that performance improves with both increasing model size and data volume, but the scaling behavior differs across task categories.

A Demystifying LLM Visual Priors: Studies and Findings

This section presents our main results and findings. We first conduct a series of controlled experiments to systematically deconstruct the origins of LLM visual priors. These studies investigate the impact of fundamental variables like model and data scale (Section A.1), data sources (Section A.2), and conceptual data mixtures (Section A.3), culminating in the derivation of a data mixture for more vision-aware LLMs (Section A.4). Building upon the rich data generated from these experiments, we then pivot to a broader analysis to uncover the internal structure and origin of the learned priors (Section A.5), and the ultimate source of these abilities within a multimodal system (Section A.6). Each subsection details its specific experimental setup or analytical approach, followed by the results and key findings.

A.1 Impact of model and data sizes.

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

We begin our analysis by investigating the fundamental impact of scale. To study how model size and pre-training data volume influence downstream multimodal capabilities, we perform a set of experiments to pre-train five LLMs of varying sizes (340M, 1B, 3B, 7B, and 13B parameters). Each model size was trained on eight different scales of data, ranging from 0B to 100B tokens. The training dataset is `web-crawl` for all experiments.

As illustrated in Figure 1, a clear and consistent trend emerges: larger models and more pre-training data generally lead to stronger downstream multimodal performance. This holds true for the overall average VQA. However, a closer look at the different VQA categories reveals significant nuances. Performance on General VQA and Knowledge VQA demonstrates a similar scaling trend, consistently improving with both model and data size. In sharp contrast, OCR & Chart VQA is far more sensitive to model size than data volume; the performance gap between models is significantly wider, and gains from additional data saturate very quickly. Meanwhile, Vision-Centric VQA also presents a unique pattern where the largest models benefit disproportionately from more data, while smaller models plateau much earlier. These divergent scaling patterns across different abilities demonstrate different visual abilities do not scale uniformly, but instead possess different properties that govern how they benefit from increased model and data size.

A.2 Impact of pre-training data sources.

> ***Finding 2:*** Different visual abilities are decoupled. Specific categories of text pre-training data can enhance certain visual capabilities in the resulting MLLM; in particular, data related to reasoning and the visual world significantly improves performance on vision-centric tasks.

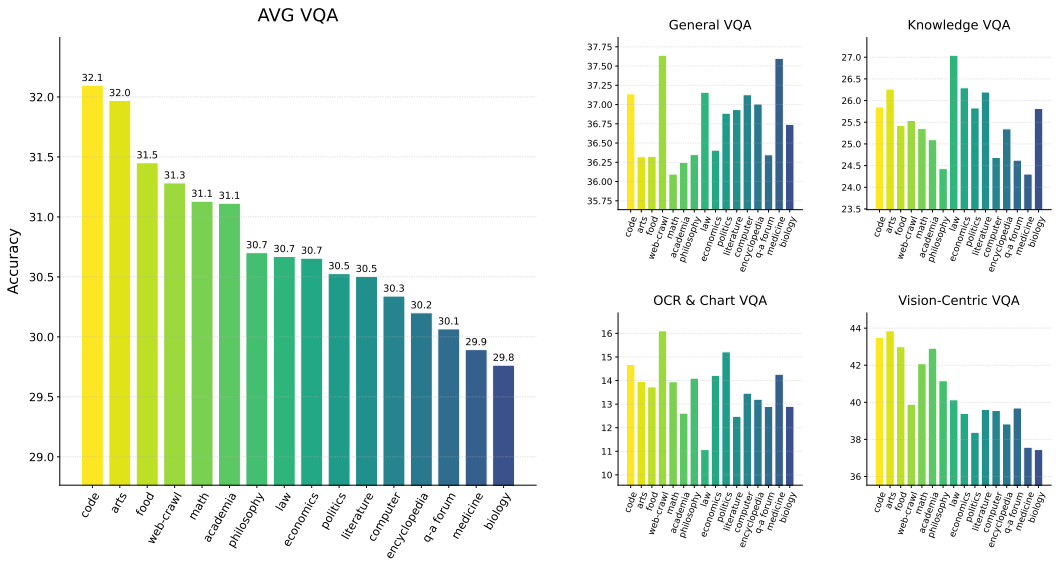

Figure 2: **Impact of pre-training data sources.** The bar charts illustrate the downstream VQA performance of MLLMs built upon a 3B parameter LLM, where each LLM was pre-trained on 30B tokens from a single, specific data source. The plots show that performance varies significantly depending on the pre-training sources.

To study the impact of different pre-training data sources, we conduct a set of controlled experiments. We fix the model size to 3B parameters and the total training data volume to 30B tokens. We then pre-train 16 distinct models, each trained exclusively on data from one of the 16 sources outlined in our pre-training sources (*e.g.*, `academia`, `biology`, `code`, etc.). This setup allows us to attribute performance variations directly to the specific data source used for pre-training.

As illustrated in Figure 2, the results reveal a significant variance in downstream multimodal performance depending on the pre-training data source. This divergence suggests that different categories of text data contribute to distinct and non-uniform visual priors. Notably, strong performance on Vision-Centric VQA tasks is highly correlated with two types of data: reasoning-centric (e.g., `code`, `mathematics`, `academia`) and corpora rich in visual-world descriptions (e.g., `arts`, `food`). The top-performing models in Vision-Centric VQA, all scoring above 42%, are trained on these specific sources.

The key takeaway from this is that the visual prior acquired from text-only pre-training is likely not a single entity. Rather, the variance in performance across data categories indicates that it may be composed of separable components or abilities. We hypothesize that this occurs because different pre-training data sources emphasize distinct underlying principles, thereby cultivating different facets of the overall visual prior. This also raises the possibility that these abilities can be selectively nurtured by curating the pre-training data, which will be studied in Section A.4.

A.3 Impact of conceptual data categories and proportions.

> ***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 data in the pre-training mix progressively enhances visual abilities, with performance gains observed up to a 75% ratio.

To move beyond the broad, pre-defined data sources analyzed in Section A.2, we investigate the impact of more fundamental conceptual categories within the text. Our findings in the previous section show that reasoning-centric categories and categories related to the visual world were the most potent drivers of downstream visual capabilities. To dissect this phenomenon further, we focus our next set of experiments specifically on these high-impact domains.

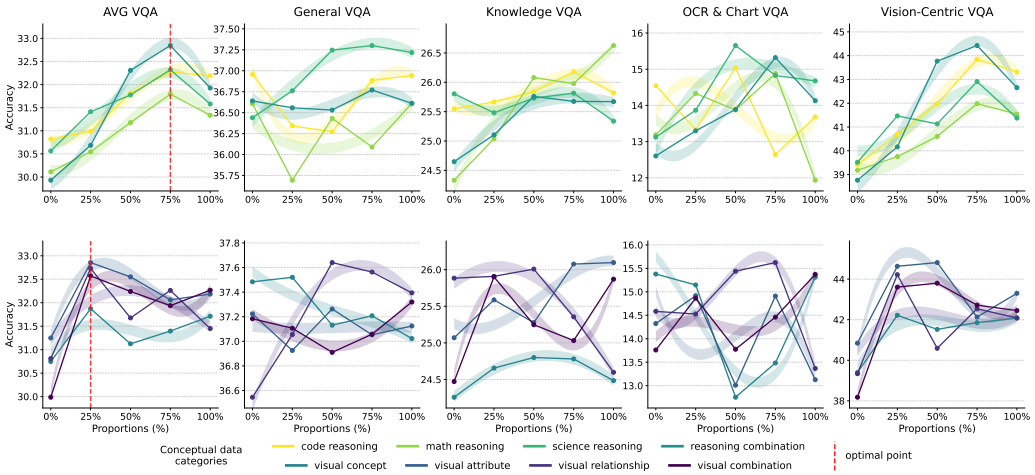

Figure 3: **Impact of conceptual data categories and proportions.** The plots illustrate how varying the proportion of specific conceptual data categories in the pre-training mix affects downstream MLLM performance across VQA benchmarks. Top Plots (Reasoning-Centric Data): The upper set of lines shows that increasing the share of reasoning-centric data leads to progressive and significant performance gains, with benefits scaling up to a 75% proportion before plateauing. This indicates that a strong reasoning foundation is critical for enhancing visual abilities. Bottom Plots (Visual World Data): In contrast, the lower set of lines, representing data that explicitly describes the visual world demonstrates rapidly diminishing returns. While a small amount of this data is crucial to establish a baseline,

We begin by creating a data pool of approximately 300B tokens, comprising all sources used in Section A.2. To partition this corpus conceptually, we employ a 32-B LLM (Yang et al., 2025a) to classify the text into finer-grained categories. The classification is performed on 1024-token segments, and is treated as a multi-label task, allowing each segment to be assigned to multiple conceptual categories. Detailed classification settings and results for each data source are presented in the Appendix H.

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.

With this conceptually partitioned dataset, we conduct a series of controlled mixing experiments to study how varying the proportion (mixing ratio) of these data types affects the final MLLM's performance. For each conceptual category, we train five separate models, systematically varying its proportion in the data mixture to 0%, 25%, 50%, 75%, and 100%. The remainder of the data for each run is sampled proportionally from the other categories within the 300B token pool to maintain a constant total training volume.

As shown in Figure 3, the results reveal a critical divergence in how different conceptual data categories contribute to visual priors. The impact of reasoning-centric data is profound and progressive, with performance scaling steadily up to a 75% proportion, especially obvious with vision-centric abilities. The contribution from data explicitly describing the visual world saturates quickly; a small initial amount seems to be crucial, but further increases yield diminishing returns.

A.4 Deriving a data mixture for more vision-aware LLMs.

> ***Finding 4:*** Maximizing MLLM VQA performance is best achieved by pre-training on a data mixture heavily skewed towards reasoning-centric content but with necessary vision world knowledge. The balance point between language and vision proficiency is reached via a calibrated data mixture between language-favorable and vision-favorable pre-training data mixture.

Building on our findings, our goal is to derive a single, practical data mixture that excels not just in language tasks but also serves as a powerful foundation for MLLMs. We narrow our focus for this analysis to six primary data categories: `web-crawl`, `encyclopedia`, `academia`, `literature`, `math`, and `code`. Our approach proceeds in three stages: first, we determine a vision-favorable conceptual blend; second, we identify a language-favorable mixture; and finally, we try to derive a balanced mixture by interpolating between these two optima.

Vision-favorable mixture. To define a target for strong MLLM visual performance, we conduct a search at the conceptual category level. Specifically, we perform a grid search across 24 data blends constructed by sampling from a space where the `reasoning combination` ranges from 50% to 85% and the `visual combination` ranges from 5% to 30%, following the conclusions drawn from Section A.3. The comprehensive results of this search are presented in Table 1.

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

Table 1: **Grid Search for an optimal-vision data mixture.** Results from pre-training a 3B parameter LLM on 30 distinct data blends, each totaling 30B tokens. The table explores how varying the proportions of reasoning-centric and visual-world text affects various capabilities, measured by Avg VQA. The data highlights a performance peak for vision tasks at a mixture of approximately 60% reasoning and 15% visual content.

From this search, we find that the best-performing models for downstream MLLM tasks emerge from a mixture containing approximately 60% reasoning and 15% visual content. This result show a powerful visual foundation is not built by simply maximizing exposure to visual descriptions, but by establishing a strong reasoning faculty, which is then grounded by a smaller amount of visual-world knowledge.

Language-favorable mixture. We begin by establishing a language-favorable mixture that achieves the best performance on our language task suite. Guided by recent literature and empirical testing, we identify this as a mix of 50% `web-crawl`, 2.5% `encyclopedia`, 2.5% `academia`, 20% `literature`, 5% `math`, and 20% `code`. This blend, designated as mix0 in Table 2, serves as our baseline for strong language proficiency, achieving the highest text accuracy (53.0%) and the best perplexity (13.46) in our experiments.

Balanced mixture. To reconcile these two objectives, we seek a single, balanced mixture that offers strong performance across both modalities. We achieve this by performing a series of interpolation experiments, detailed as mix0 through mix10. We shift the data composition from our language-favorable baseline (mix0) towards a conceptual endpoint representing the vision-favorable blend (approximated by mix9 and mix10). To ensure stabilized results, each model in this series is trained for 50B tokens.

The performance metrics in Table 2 reveal the expected trade-off: as the mixture becomes more reasoning-centric, vision accuracy (v-acc) generally improve, while language proficiency (t-acc and ppl) shows a slight decline. Our analysis identifies mix6 as the balanced mixture, achieving the highest overall rank. Mixtures in its vicinity (e.g., mix5, mix7, mix8) also achieve high rankings. This demonstrates that a carefully calibrated data mixture can cultivate powerful visual priors without substantially compromising core language abilities.

Recipe	Data Source Mixture (%)								Performance Metrics			Overall Rank
	web-crawl	encyclopedia	academic	literature	math	code	reasoning	visual	t-acc (%)	ppl (↓)	v-acc (%)	
mix0	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	26.0	5.5	12.5	2.0	13.0	41.0	61.3	13.3	52.1	13.88	**33.4**	11

Table 2: **Deriving a data mixture for more vision-aware LLMs.** This table details a series of 11 data mixtures, from mix0 (language-favorable blend) to mix10 (approximating the vision-favorable blend), all trained on a 3B-parameter LLM with 50B tokens. The experiment systematically shifts the data composition towards a higher proportion of reasoning-centric content (math, code, academia). The results highlight a trade-off, with mix6 emerging as the most balanced mixture, achieving top-ranked overall performance by improving visual capabilities without a significant drop in language proficiency.

A.5 The structure and origin of learned visual priors.

> ***Finding 5:*** The learned visual prior is not a single entity but decomposes into at least two priors with different origins. We reveals a perception prior (linking General and OCR) and a reasoning prior (linking Knowledge and Vision-Centric tasks). These priors also scale independently.

We now synthesize our previous results to investigate the internal structure of the visual prior. Is it a single, uniform ability, or a composite of different, separable visual skills? To answer this, we conceptualize the visual prior as a collection of distinct abilities, each measured by one of our four VQA categories.

Internal structure of visual priors. We aggregate the performance data across all 105 3B models from our previous experiments, encompassing variations in data sources, mixing ratios, and training scales. We then compute the Spearman correlation matrix across the four VQA performance categories to identify which abilities scale together and which diverge.

The results in Figure 4 suggest a potential internal structure within the visual prior, hinting at a separation into at least two distinct types of abilities. We observe a moderate correlation (0.37) between General and OCR performance. This connection seems to point towards a perception prior, as success in both categories relies heavily on the model's perceptual acuity—the ability to accurately process raw visual input—rather than complex, multi-step reasoning.

In contrast, we find another moderate correlation (0.33) between the Knowledge and Vision-Centric tasks. This link appears to emerge because both categories often require abstract inference that goes beyond simple perception. For instance, the Knowledge category demands multi-step reasoning to solve complex scientific or mathematical problems, while Vision-Centric tasks include challenges like

visual IQ puzzles, object counting, and correspondence matching, necessitate a blend of perception and reasoning, often with a heavier reliance on the latter.

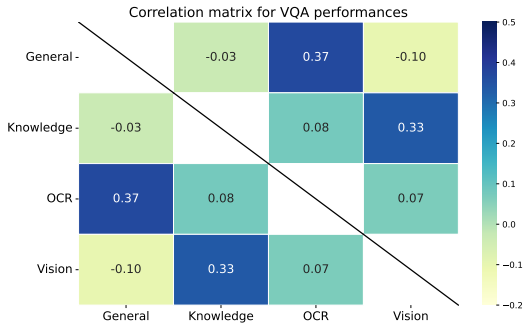

Figure 4: **Correlation matrix for VQA performances.** The matrix reveals two loosely-coupled skill clusters: one axis for perception (General/OCR) and another for reasoning (Knowledge/Vision-Centric).

The correlation matrix also reveals very weak, or even slightly negative, correlations between these two groups (perception-heavy vs. reasoning-heavy) [1]. This lack of a strong positive correlation raises the possibility that these are largely independent abilities, potentially stemming from loosely-coupled priors within the LLM's representation. Our observations on the separability of these visual priors in MLLMs align with and extend the findings of recent research Chen et al. (2025), which identified a similar dissociation through test-time parameter merging.

Different origins of priors. The statistical independence of these two priors implies they are cultivated through different mechanisms. As our analysis in Section A.2 and Section A.3 demonstrated, the reasoning prior is from reasoning-centric data and can be predictably enhanced by increasing the proportion of reasoning-centric data.

In contrast, the origins of the perception prior appear more diffuse. A signal comes from our single-source experiments (Section A.2), where `web-crawl` data yields the best performance on General and OCR tasks. However, `web-crawl` is an extremely general category, and no other, more specific data category consistently boosts perceptual abilities. This 'mixed effect' suggests the perception prior may be a general byproduct of large-scale language modeling, emerging from the sheer diversity of language rather than a specific category. To further investigate this emergent prior and characterize its properties more directly, we introduce a multi-level existence benchmark (MLE-Bench) designed to assess a more pure perception abilities (with less reasoning required) across multiple levels. The detailed study using this benchmark is presented in Section B.1.

This may also suggest that perception ability aligns more with the platonic representation hypothesis; that is, it requires greater data diversity for effective multi-modality alignment. We find that the Spearman correlation (from our 105 experiments) between the LLM-vision alignment score and OCR is 0.42, while being only 0.22 for reasoning tasks. The stronger correlation for OCR suggests that the perception prior is more intricately linked to cross-modal alignment, which thrives on the rich and diverse language found in broad corpora. In contrast, the weaker correlation for reasoning tasks shows that reasoning is thus less dependent on the specific signals used for vision-language alignment. We present a more detailed study of this phenomenon in Section B.3.

A.6 Deconstructing multimodal abilities: vision or language.

> ***Finding 6:*** Visual reasoning ability is primarily shaped by reasoning prior acquired from language pre-training; perception is more dependent on vision encoders and on downstream visual instruction tuning data.

Here, we conduct further analysis to first verify the universality of learned visual priors and then deconstruct the source of different multimodal abilities, distinguishing between those inherited more from the LLM and those acquired more from the visual instruction tuning.

Universality of the learned visual priors. To test the general influence of the visual prior, we apply two more vision encoders (DINOv2-L (Oquab et al., 2023) and MAE-H (He et al., 2022)) other than

[1] Note that this categorization into perception-heavy and reasoning-heavy tasks is a conceptual simplification intended to facilitate our analysis. In practice, the boundary between perception and reasoning is not always clear.

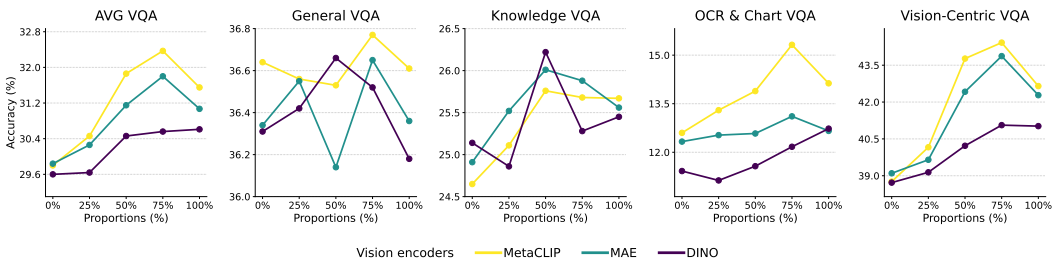

Figure 5: **Universality of the learned visual priors.** The plots show the VQA performance of MLLMs built using three distinct vision encoders based on the proportion of reasoning-centric data used in the LLM's pre-training mix. Despite differences in their absolute performance, all three configurations show a consistent improvement on reasoning-heavy tasks as the LLM's reasoning pre-training proportion increases, similar to trends observed before, demonstrating the universality of the reasoning prior.

our default MetaCLIP-B/16. We pair these with LLMs pre-trained on varying proportions of our `reasoning combination` conceptual data category, from 0% to 100%.

As illustrated in Figure 5, the results reveal a dual-faceted pattern. Firstly, they confirm the universality of the reasoning prior. For reasoning-heavy tasks, all three vision encoder configurations exhibit a nearly identical, strong upward trend in performance as the proportion of reasoning data in the LLM's pre-training increases. This demonstrates that the visual reasoning prior cultivated in the LLM is a foundational, modality-agnostic prior that benefits the multimodal system regardless of the specific vision encoder used.

In contrast, the perception prior lacks this universality. The performance trends for perception-oriented tasks are inconsistent across the different vision encoders. Instead of following a unified pattern, the performance curves for different vision encoders vary significantly from one another. This suggests that perceptual abilities are more sensitive to the specific characteristics of the vision encoder (Liang et al., 2025; Tong et al., 2024c) and are not systematically influenced by the LLM's perception prior, further showing the decoupling of the two priors.

Source of abilities from visual priors and visual instruction tuning. Second, we conduct targeted studies to determine whether key skills—namely, perception and reasoning—originate primarily from the LLM's visual priors or the subsequent visual instruction tuning. stage. We use an MLLM to classify our Cambrian-7M dataset that contains 5M text-image pairs into these two categories, resulting in 1.8M perception and 0.6M reasoning data, and the remaining 2.6M data as others.

We partition our instruction-tuning data into perception, reasoning, and other categories and trained five tuning configurations that ablate perception and reasoning data in stages (100% $\rightarrow$ 50% $\rightarrow$ 0%) while leaving other data unchanged. Our model with full perception and reasoning data achieves a baseline performance of 37.98% on General VQA, 25.75% on Knowledge VQ, 17.74% on OCR & Chart VQA, and 43.48% on Vision-Centric VQA. The results, presented in Figure 6, show two observations: (1) reducing perception-targeted tuning produces the largest performance drops on perception-heavy benchmarks (OCR & Chart and General) and modest drops on reasoning tasks (Vision-Centric and Knowledge); (2) removing reasoning-targeted tuning causes only small incremental drops on perception tasks and modest drops on reasoning tasks.

Together, results in this section show two mechanisms. First, the LLM encodes a robust, transferable visual reasoning prior primarily via language pre-training; this prior benefits reasoning-centered VQA across different vision encoders. Second, perception performance depends more on vision-encoder characteristics and on subsequent supervised visual instruction tuning: perception performance gains require more encoder- and vision-supervision-specific interventions.

B Discussion and Hypotheses

This section transitions from empirical findings to a more speculative exploration of the underlying mechanisms of visual priors. The following subsections present three key hypotheses about the

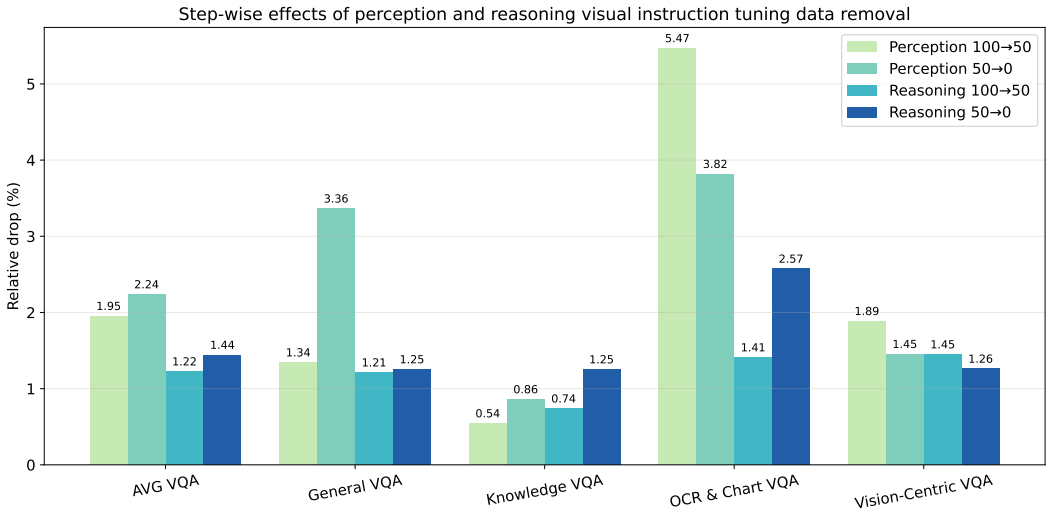

Figure 6: **Step-wise effects of perception and reasoning visual-instruction tuning data removal.** Relative incremental performance drop (%) on AVG VQA and per-category VQA (x-axis). Each bar shows the step-wise change when ablating perception or reasoning instruction data in stages (100% $\rightarrow$ 50% $\rightarrow$ 0%); each reported percent is computed relative to the immediately preceding configuration. Removing perception-tuning data produces the largest incremental decline on OCR & Chart VQA and General VQA (showing perception's stronger dependence on supervised vision-side tuning), while removing reasoning-tuning data yields only small incremental declines on perception tasks and modest declines on Vision-Centric and Knowledge VQA.

structure of the perception prior, the universal nature of reasoning, and the role of data structure in cross-modal alignment. These hypotheses are not presented as definitive conclusions but as frameworks for interpreting the results and for future research.

B.1 Is the perception prior multi-level? An evaluation using the MLE-Bench

> ***Hypothesis 1:*** The perception prior derived from diverse data exhibits scale-dependency, with its benefits being most pronounced for the perception of small and medium-sized objects.

Our previous analyses show that the perception prior is diffuse in origin, emerging most strongly from diverse data. This leads to a question about its internal structure: is this prior a uniform ability, or does it possess finer-grained characteristics?

To study this question, we introduce the Multi-Level Existence Bench (MLE-Bench), a benchmark designed to probe perception with greater precision. MLE-Bench consists of 4-choice questions about the existence of objects or scenes within an image. We categorize questions based on the target object's relative size, measured by the percentage of pixels it occupies. In total, MLE-Bench comprises 1,861 images, with a distribution of 732 questions for small objects (0-30%), 698 for medium objects (30-60%), 397 for large objects (60-90%), and 34 for very large objects/scenes (90-100%). This structure allows us to deconstruct "perception" into distinct, scale-dependent components. Further details on the benchmark's construction are presented in Appendix J. We also present a holistic evaluation of common MLLMs on MLE-Bench, with detailed results available in Appendix K.

We evaluate our 16 single-source pre-trained models (from Section A.2) on MLE-Bench, with the results presented in Figure 7. The 3B LLM model trained on web-crawl remains the top performer overall, confirming that data diversity is key for perception prior. Its advantage is most pronounced for small-to-medium objects (0-60% pixel range), where it establishes a clear lead over models trained on other data sources. In contrast, for large objects that dominate the visual scene, this performance gap diminishes significantly.

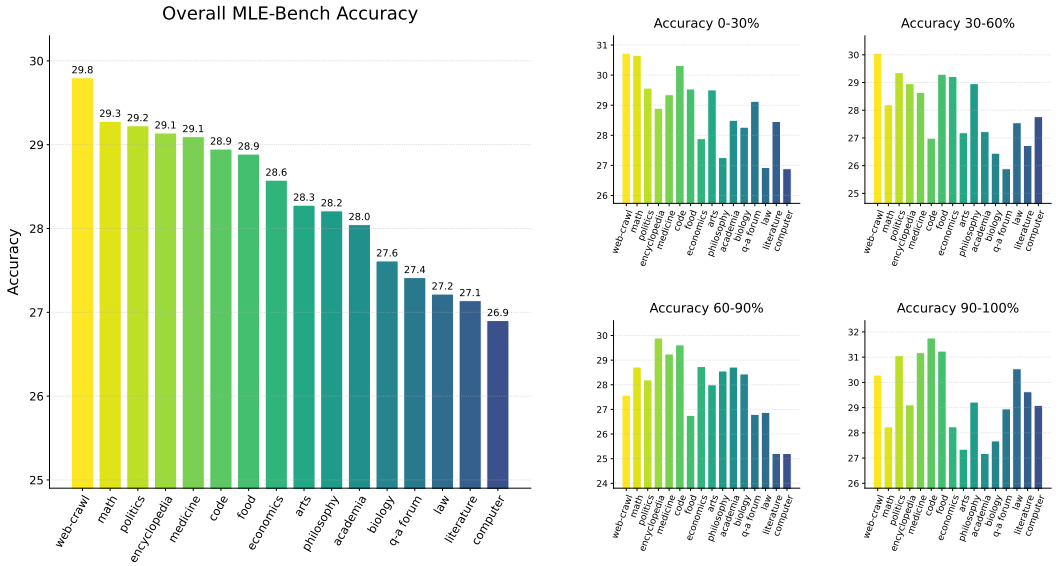

Figure 7: **Performance of MLLMs on the Multi-Level Existence Bench (MLE-Bench).** The left plot shows the overall accuracy for models pre-trained on 16 different single-source data types. The four smaller plots (right) detail performance on objects of varying relative sizes, from small (0-30% of image pixels) to very large (90-100%). The results demonstrate that pre-training on the broad and diverse `web-crawl` corpus is most effective in gaining perception prior, with its advantage being particularly pronounced for perceiving smaller objects.

These results indicate that the perception prior is indeed scale-dependent. A possible explanation is that diverse, unstructured text like web-crawl contains a vast vocabulary describing a wide array of entities, including smaller, often overlooked details within a larger scene. This textual richness forces the model to learn representations sensitive to fine-grained visual concepts, a capability less critical when identifying large, obvious objects. This finding refines our understanding of the perception prior, revealing that it is not a uniform faculty.

B.2 Is reasoning a universal, cross-modal skill already acquired during pre-training?

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

Our findings suggest a hypothesis: the reasoning capabilities an LLM acquires from text are not bound to the linguistic domain. We posit that by pre-training on reasoning-centric data, a model learns abstract, generalizable principles of logic, structure, and compositionality. This foundation is largely modality-agnostic, allowing the model to apply this faculty to other domains, including vision, since the reasoning process likely occurs within the language domain.

To verify this, we propose an experiment that directly probes the quality of the models' visual reasoning processes. For reasoning-focused VQA tasks (Knowledge and Vision-Centric), we switch the evaluation from prompting for a direct answer to answer with detailed explanations. This will require each model to produce a detailed explanation of its reasoning. We then use a separate LLM as a judge to evaluate the quality of these reasoning traces based on a clear rubric, assessing criteria such as: (1) Logical Soundness: The percentage of reasoning traces that are coherent and reasonable; and (2) Reasoning Depth: the average length of the reasoning trace measured by text count taken to reach the conclusion.

The results, presented in Figure 8 (a), strongly support our hypothesis that the reasoning capabilities an LLM acquires from text are transferable to vision. We observe a clear trend: as the proportion of reasoning-centric data increases, the models generate visual reasoning that is both more logically sound and significantly longer. For instance, increasing the proportion of `code reasoning` data from

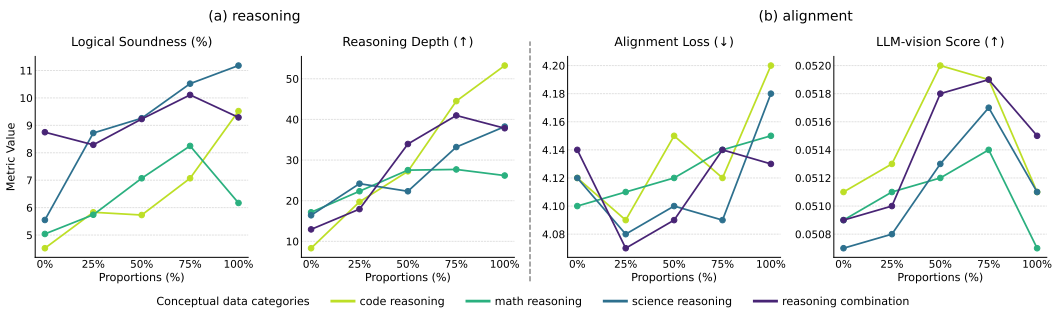

Figure 8: **Qualitative impact of reasoning-centric data on visual reasoning (a) and representation alignment (b).** The plots show how varying the proportion of different reasoning-centric data categories in the pre-training mix impacts metrics of visual reasoning quality and cross-modal alignment. The left two plots indicate that more reasoning data leads to more coherent and detailed visual reasoning. The right two plots reveal a more complex relationship, with the LLM-vision alignment score showing a generally positive but non-monotonic trend and the alignment loss exhibiting more noisy but generally reverse behavior.

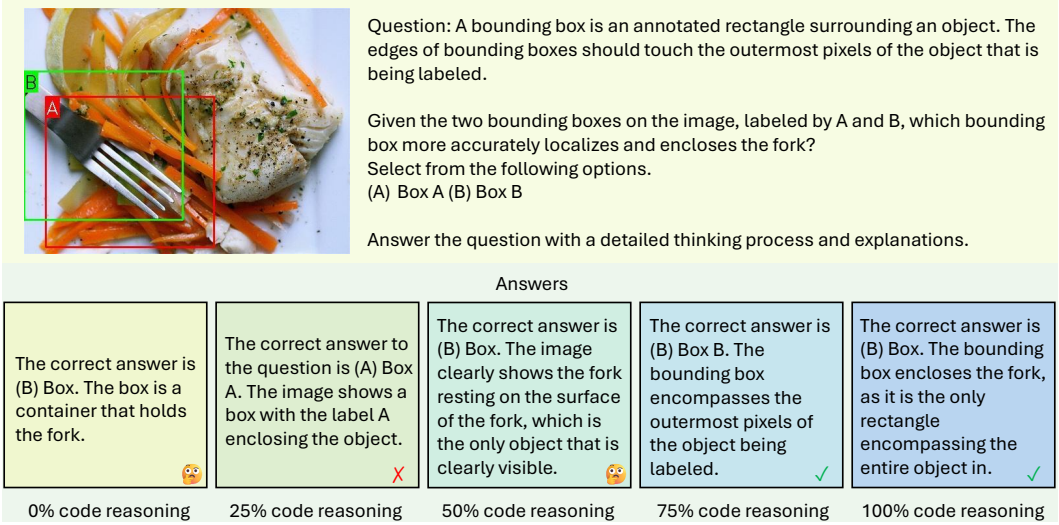

Figure 9: **Qualitative impact of reasoning-centric data on visual reasoning.** The figure displays the answers from five models—pre-trained with 0% to 100% `code reasoning` data—to a visual question requiring the application of a specific rule. Answers show a clear improvement in reasoning quality: the model with 0% `code reasoning` provides a simplistic justification, while the models with 75% and 100% `code reasoning` produce more detailed reasoning that correctly applies the definition from the prompt.

0% to 100% boosts Logical Soundness from 4.52% to 9.52% and more than sextuples the Reasoning Depth from 8.31 to 53.25. This demonstrates that the model is is applying a general, abstract reasoning framework, learned from text, to solve visual problems. The particularly dramatic increase in Reasoning Depth for code-trained models may also reflect a stylistic transfer; pre-training on `code reasoning`, which is often structured in long, logically coherent sequences, likely predisposes the model to generate longer, more structured step-by-step explanations.

Figure 9 provides a qualitative example of this phenomenon. It showcases how models trained with more `code reasoning` data produce increasingly sophisticated and reasonable reasoning for a visual task. While the model with 0% `code reasoning` offers a simplistic justification, the model trained on 100% `code reasoning` provides a detailed, step-by-step explanation that correctly applies the abstract rule given in the prompt. This demonstrates that the model is applying a general, abstract reasoning framework, learned from text, to solve visual problems. Our conclusions here also reflect the results shown in very recent studies that reasoning abilities can transfer between languages at

test time (Yong et al., 2025), and that post-training such as language reinforcement learning can enhance or transfer to multi-modal reasoning (Rastogi et al., 2025; Wei et al., 2025; Yang et al., 2025b; Liu et al., 2025; Chen et al., 2025). We further demonstrate that reasoning abilities are highly modality-agnostic, to the extent that training solely on `code` can yield strong multimodal reasoning. Moreover, we show this transferability is not confined to post-training phases, but originates from pre-training itself.

B.3 Does language data structure drive representational alignment with vision?

> ***Hypothesis 3:*** The structural properties of language data do not exclusively drive representational alignment with vision.

An alternative, or perhaps complementary, hypothesis centers on the structural similarities between the data modalities themselves. Data from domains like `code` and `mathematics` is inherently highly structured. It is governed by strict syntax, logical dependencies, and hierarchical compositions. Similarly, visual data is far from being a random collection of pixels. It is rich with its own structure: spatial relationships between objects, part-whole hierarchies, and the implicit rules of physics and geometry. We thus hypothesize that this shared structural foundation means that representations learned from structured text are intrinsically more similar to, and thus more readily transferable to, the visual domain.

To test this hypothesis, we analyze two metrics across models trained with varying proportions of structured reasoning data: the LLM-vision alignment score and the average alignment loss during the MLLM alignment stage. Our analysis, presented in Figure 8 (b), yields mixed results. The LLM-vision alignment score partially supports our hypothesis. As we increase the proportion of structured reasoning data, the alignment score generally improves, indicating a more congruent latent space. However, this trend is not monotonic; the score often peaks at a 75% ratio and then decreases at a 100% proportion. This might be due to a model trained purely on reasoning data learning abstract structure but lacking the necessary vocabulary from other text types to effectively map it onto diverse visual concepts, thus hindering the final alignment.

The alignment loss metric provides a less clear picture, generally exhibiting a decrease from 0% to 25% reasoning data and a noticeable increase from 75% to 100%, despite showing more inconsistent fluctuations between 25% and 75%. This suggests a nuanced relationship. Qualitatively, we observe a general inverse trend between the LLM-vision alignment score and the alignment loss. This relationship may indicate that both metrics are capturing related aspects of cross-modal alignment, potentially lending partial support to the structural similarity hypothesis. However, given these mixed signals and the lack of a clear, consistent trend, this analysis does not provide definitive evidence to confirm the hypothesis. It thus remains a compelling direction for future research to untangle the precise interplay between abstract structure and semantic grounding in forming cross-modal representations.

C Scaling Up and Training a Vision-Aware LLM

Building on the principles identified in the controlled, smaller-scale studies, this section details the process of scaling up the approach to validate the findings.

C.1 Settings and models

Building upon our findings, we scale up our approach to validate our findings and develop a vision-aware LLM in a larger-scale. The goal is to test whether the principles identified in our controlled, smaller-scale studies hold true when applied to a large-scale training run. 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.

We conduct the pre-training for each model on 128 A100 GPUs for approximately 32 days. We use an processing approximately 4.2M tokens per step and the model is trained for 250000 steps. Following pre-training, we adapt both 7B LLMs into MLLMs. For this stage, we utilize the complete Cambrian data suites. Specifically, we use the full 2.5M image-caption dataset for the visual alignment stage, followed by visual supervised fine-tuning on the full 7M vision-language instruction dataset.

C.2 Blind visual instruction tuning

We also introduce a "blind visual instruction tuning" trick that provides a more effective starting point for visual adaptation. This trick involves an initial instruction tuning phase using only the textual data while withholding the corresponding images. This initial "blind" stage allows the model to first focus on learning the instruction-following format of the task. Consequently, the subsequent standard tuning phase with images can be more dedicated to learning the core vision capabilities, rather than simultaneously learning how to follow instructions. Furthermore, this process enables the model to effectively leverage its pre-existing language priors to solve VQA questions that may not strictly require visual input, a known phenomenon and potential "short-cut" on some benchmarks (Tong et al., 2024a). This trick can lead to broad performance improvements in most of the tasks.

Note that this trick should not be a standard in practice, since when no images are provided, models should identify their absence [2] rather than encouraging more hallucination. We leave this as an optional trick and introduce this phenomenon to the community for future investigation. This also shows that "vision" can be "learned" from the language side through an unconventional, shortcut-based mechanism.

Model	Language		Vision				
	ppl	avg acc	General	Knowledge	OCR&Chart QA	Vision-Centric	Overall
Lang-Favorable	8.72	0.647	46.92	28.35	21.49	46.31	37.32
Lang-Favorable (+Blind)			48.16	30.30	20.77	47.01	38.20
Balanced	7.49	0.655	49.59	29.02	23.63	46.59	38.64
Balanced (+Blind)			50.90	31.25	22.60	47.32	39.56

Table 3: **Performance comparisons of the Lang-Favorable and Balanced models across both language and vision-language benchmarks.** The table summarizes key language metrics (perplexity and accuracy) and provides average scores for a suite of vision tasks, categorized as General, Knowledge, OCR & Chart QA, and Vision-Centric. It also shows the impact of applying our blind visual instruction tuning trick (+Blind). The results demonstrate that the Balanced model, pre-trained with our vision-aware data mixture, exhibits competitive language proficiency while consistently outperforming the Lang-Favorable model on all visual tasks, with the blind tuning method providing an additional performance boost.

C.3 Results

As shown in Table 3, the Balanced model, pre-trained with balanced recipe, exhibits competitive language proficiency. Notably, it achieves a lower (better) average perplexity of 7.49 compared to the Language-favorable model's 8.72, while also maintaining a slightly higher average accuracy (0.655 vs. 0.647). An interesting dynamic observed during pre-training was that the Balanced model's language performance initially lagged behind the Language-favorable model, beginning to surpass it after approximately 600B tokens. This may suggests that when the pre-training token volume is sufficiently large, the benefits from reasoning-related tokens can be more effectively unleashed when grounded in a substantial amount of world knowledge, ultimately resulting in strong performance also on the language side.

On VQA benchmarks detailed, the Balanced model consistently outperforms the Language-Favorable model in most of the benchmarks, achieving a higher overall VQA average (38.64 vs. 37.32). This

[2]This type of hallucination may persists even in strong models including GPT-5 thinking, Gemini 2.5, and Claude Opus 4.1. Models will exhibit such hallucinatory behaviors when answering VQA questions without actual visual context given. Some examples are presented in Appendix L. This highlights a systemic issue that warrants more future investigation.

confirms that the deliberate pre-training on a data mixture rich in reasoning and visual-world text successfully imbues the LLM with stronger visual priors in a large-scale.

Furthermore, the application of our blind visual instruction tuning trick yields additional overall performance gains for both models by clear margins. The performance gains are most pronounced in the Knowledge categories, while tasks in OCR & Chart VQA, conversely, suffer a performance drop. This pattern suggests that while the blind tuning phase provides a more effective initialization for the subsequent visual instruction tuning, it gains more from strengthening the model's ability to leverage its internal knowledge and reasoning priors to "short-cut" the problem. This is highly beneficial for knowledge-intensive tasks where answers can be inferred or reasoned from textual context and the LLM's pre-existing knowledge.

D Related Work

D.1 From LLMs to MLLMs.

With the rapid development of LLMs (Radford et al., 2021; Google, 2023; Touvron et al., 2023), a direction of work extend LLMs to Multimodal LLMs. Pioneering works like Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023) connected pre-trained vision encoders to LLMs using connectors like cross-attention modules. Later models such as LLaVA (Liu et al., 2023a) demonstrate that even with a projection layer, LLMs can be extended MLLM with visual instruction tuning. This adapter-style architecture has been widely explored in numerous subsequent works (Liu et al., 2024a; Tong et al., 2024a; Laurençon et al., 2024; Liu et al., 2024b; AI@Meta, 2024; Bai et al., 2025b; Zhu et al., 2025; Team et al., 2025; Lin et al., 2024).

The success of visual instruction tuning has enabled open-source multimodal models to achieve performance even comparable to proprietary counterparts (Google, 2023; xAI, 2024; OpenAI, 2024). This success underscores that multimodal capabilities in adapted LLMs largely emerge through instruction tuning, effectively unlocking knowledge already embedded in pretrained language models (Zhou et al., 2024). Furthermore, recent studies (Tong et al., 2024a; Laurençon et al., 2024) highlight that improvements in the underlying language model remain the most impactful means to improve multimodal performance. Inspired by these insights, our work investigates the visual priors and inherent multimodal potential embedded within pretrained LLMs.

Though there are different ways of connecting vision to LLMs like the use of discrete tokenization (Wang et al., 2024; Deng et al., 2025; Team, 2024), we focus on adapter-style architectures, which are most widely used and permit clean analysis of visual priors from language pre-training.

D.2 The role of data in shaping foundation model capabilities.

Pretrained LLMs encode rich latent knowledge—even across modalities—depending heavily on the nature of their training data (Kaplan et al., 2020; Grattafiori et al., 2024; Han et al., 2022; Rae et al., 2021; Penedo et al., 2023; Lu et al., 2022a; Mayilvahanan et al., 2025). This has shifted the focus of research from simply scaling data to understanding the role of data and then strategically curating it to unlock specific, powerful abilities (Allen-Zhu, 2024; Aryabumi et al., 2024; Ye et al., 2024; Shinnick et al., 2025).

A prominent example, and one highly relevant to our findings, is the strategic inclusion of reasoning-centric data like `code`. Research has consistently shown that pretraining on a mix of text and `code` does more than just improve coding skills; it significantly enhances a model's foundational reasoning and ability to understand abstract, structural patterns (Muennighoff et al., 2023; Aryabumi et al., 2024; Ma et al., 2023; Zhang et al., 2025). This suggests that the pre-training data mixture may endows the model with latent, generalizable structures that can be activated for tasks beyond their original domain.

This has established a central challenge in the field: determining the optimal data mixture to cultivate these desired foundational abilities (Chen et al., 2024; Albalak et al., 2023; Ma et al., 2023; Xie et al., 2023; Touvron et al., 2023; Grattafiori et al., 2024; Zhang et al., 2024; Held et al., 2025; Bai et al., 2024; Albalak et al., 2023; Shukor et al., 2025a). This has spurred a move beyond simple heuristics toward quantitative frameworks which aim to predict a model's performance based on different data blends, thereby guiding the search for an optimal mixture.

However, much of this prior work has focused on optimizing data mixtures for core language proficiency. As we transition from LLMs to MLLMs, a critical question emerges: how do these text-only pre-training choices influence the model's visual priors and the potential for multimodal capabilities? Our work directly addresses this gap. We extend the investigation of language data composition's impact from the purely linguistic to the visual domain, systematically analyzing how different text sources contribute to the emergent visual priors in LLMs and seeking data mixture to help them "learn to see" more effectively from text pre-training.

E Limitations and Future Research Directions

While this work provides a systematic analysis of visual priors in LLMs, it is subject to several limitations that open avenues for future research.

First, our investigation primarily centers on adapter-style MLLM architectures. While this is a prevalent and effective paradigm, our findings may not fully generalize to other approaches, such as those that employ discrete visual tokenization (Team, 2024; Wang et al., 2024; Deng et al., 2025; Wu et al., 2024) or involve end-to-end joint training of vision and language components (Diao et al., 2024; Tao et al., 2025; Diao et al., 2025; Shukor et al., 2025b). In these latter cases, language and vision data are co-trained, making it hard to identify the priors originating solely from language. The dynamics of how visual priors are formed and utilized could differ in these models, which leaves a promising future direction.

Second, a significant area our study does not address is the safety and ethical implications of these learned visual priors. Language corpora are known to contain societal biases, stereotypes, and potentially harmful content (Bengio et al., 2024; Qu et al., 2023). Our analysis focused on capability, but did not investigate whether these text-based priors encode biased visual associations (*e.g.*, linking certain objects or roles to specific genders or races) that could manifest as harmful generation or classification behavior in a downstream MLLM. A thorough audit of the fairness and safety of these emergent priors is a critical next step.

Finally, our study is confined to the domain of static images, leaving the exploration of visual priors for dynamic modalities, such as video understanding, as an open question. For example, the temporal knowledge important for video understanding might be learned more from story-related data like `literature`. Investigating how different textual sources contribute to priors for temporal reasoning, action recognition, and causality in video is a rich area for future work.

F Conclusion

This work has undertaken a systematic deconstruction of the visual priors that LLMs acquire from text-only pre-training. Through a series of controlled experiments manipulating data composition, we moved beyond observing the phenomenon of vision priors to interrogating its fundamental drivers.

Our investigation provides a data-centric roadmap for developing multimodal systems, shifting the paradigm from serendipitous emergence to the deliberate cultivation of visual capabilities. By showing that core reasoning abilities are a transferable, modality-agnostic foundation, our work offers more empirical supports for the idea that models can learn a unified representation of the world from even a single modality.

Looking forward, we hope this research encourages a paradigm where LLM development is more considerate of vision and multimodality, prompting the cultivation of visual priors from the earliest stages of pre-training. We also hope it inspires a deeper investigation into the fundamental correlations between cross-modal representations, contributing to a more unified understanding of how knowledge is structured across modalities.

G Broader Impact

Our research provides a systematic analysis of how prior visual capabilities emerge in LLMs from language-only pre-training, shifting the paradigm from accidental discovery to deliberate cultivation. While our work focuses on capability, the textual data used for pre-training contains societal biases.

A significant risk is that these models could learn and reinforce harmful visual stereotypes, which could then manifest in downstream multimodal systems.

Nevertheless, our findings primarily help researchers and developers understand the nature and origins of these visual priors. We demonstrate that these priors are not a single, uniform block but are composed of separable perception and reasoning components, each cultivated by different types of text.

This deeper understanding provides a clear, actionable path for more efficiently cultivating these abilities. Instead of relying on serendipity scaling, teams can now strategically curate their text-only pre-training data to deliberately build a stronger foundation for vision tasks before multimodal training even begins. This targeted approach not only improves the final model's multimodal performance but also reduces the computational resources required, offering a more sustainable methodology for creating the next generation of vision-language models.

H Conceptual Language Data Classification

This section provides the detailed conceptual classification setting and results for the key pre-training data sources discussed in the main paper. We use a 32B dense LLM (Yang et al., 2025a) to perform a multi-label classification on 1024-token segments from each data source. Below is the full prompt provided to the LLM for the conceptual data classification task. The prompt instructs the model to perform a multi-label classification on text segments, assigning one or more predefined categories that describe the content.

Prompt for LLM-based Conceptual Data Classification

Analyze the provided text paragraph. Classify its content by identifying the primary concepts and domains using only the categories listed below. Select categories that represent the text's significant content.

- `visual concept:` Language for naming visual entities (e.g., objects, people, places, actions, scenes).
- `visual attribute:` Language describing visual properties (e.g., color, size, shape, texture, style).
- `visual relationship:` Language describing spatial or part-whole relations between entities.
- `code reasoning:` Content centered on algorithmic problem-solving, logical coding implementation, and software engineering challenges.
- `math reasoning:` Content focused on logical math reasoning, proof construction, and the application of mathematical principles to solve problems.
- `science reasoning:` Focuses on scientific reasoning, including hypothesis testing, data analysis, and modeling of complex systems.

If none of the above categories apply, output None.

The percentages in Table 4 represent the proportion of text segments within each data source that were assigned a given conceptual label.

I Robust Parsing for VQA Evaluations

A significant challenge in the automated evaluation of VQA is that models often generate conversational or free-form text instead of a single-letter answer. A naive parsing strategy that only checks for an exact match to the ground-truth letter (e.g., "B") would unfairly penalize models that provide a correct but differently formatted response.

To illustrate, consider a simple VQA task:

- Question: "What is the primary object in the image?"

	Reasoning Categories (%)				Visual Categories (%)			
Data sources	code reasoning	math reasoning	science reasoning	reasoning combination	visual concept	visual attribute	visual relationship	visual combination
web-crawl	3.5	3.6	8.6	10.0	27.7	14.1	13.5	26.9
encyclopedia	0.5	2.3	3.0	3.4	12.2	2.9	6.0	12.4
academia	21.0	68.2	74.4	83.3	5.2	0.7	5.1	5.3
literature	0.3	1.3	8.9	9.1	33.4	8.4	27.9	33.5
math	31.1	81.7	83.2	92.9	7.8	2.8	6.6	8.0
code	96.7	13.3	6.8	97.3	3.3	2.0	2.4	3.7

Table 4: **Conceptual categories of key pre-training data corpus (%).** The table shows the percentage of text segments from each data corpus classified into one of the conceptual categories.

- Options: `(A) A bicycle`, `(B) A car`, `(C) A tree`
- Ground Truth: `B`

A model could correctly answer in multiple ways, such as *"The answer is (B)"*, *"A car"*, or *"The image shows a car"*. To capture all these valid responses, our evaluation protocol employs a robust, hierarchical parsing strategy. The logic is executed as a sequence of prioritized steps, stopping as soon as a valid answer is found:

1. **Explicit letter extraction**: The parser first searches for high-confidence patterns that directly indicate the chosen option letter. It uses regular expressions to find formats like:
   - *"The correct answer is (B)"*
   - *"Answer: B"*
   - Outputs starting or ending with *'(B)'*
   - An output that is simply *'B'* or *'B.'*
2. **Exact option text matching**: If the first step fails, the parser extracts the text associated with each option from the prompt (e.g., "A bicycle", "A car", "A tree"). It then checks if the model's generated text is an *exact, case-insensitive match* for any of these option strings.
   - Example caught: A model output of *"A car"* would be correctly mapped to option `B`.
3. **Substring matching**: As a final fallback, the parser checks if the text of any option appears as a substring within the model's generated output. This handles more verbose, conversational answers.
   - Example caught: A model output of *"The image features a car driving down the street"* would be correctly mapped to option `B` because "a car" is present.
   - To prevent ambiguity (e.g., if one option was "car" and another was "race car"), this step returns the longest matching option text found in the response.

This multi-tiered parsing strategy ensures a comprehensive and fair evaluation across all models, regardless of their verbosity or adherence to specific formatting instructions. It allows us to more accurately measure the model's underlying visual capabilities rather than its ability to follow formatting rules.

J Multi-Level Existence Benchmark Construction

This section describes how we constructed our benchmark using publicly available SA-1B and ADE20K datasets. We selected images with ground-truth segmentation masks and calculated the proportion of the image area each object occupied. Based on this, we created three splits: 0–30 for small objects, 30–60 for medium, and 60–100 for large, dominant objects. For each image, we created a multiple-choice query to test object existence, sampling distractors from the dataset vocabulary and filtering them to exclude objects present in the ground truth. As this process is open-vocabulary, we use an LLM to filter out distractors that correspond to objects already present in the image but under different names. This ensures the distractors remain plausible but incorrect, providing a granular evaluation of a model's ability to identify objects across a wide range of sizes.

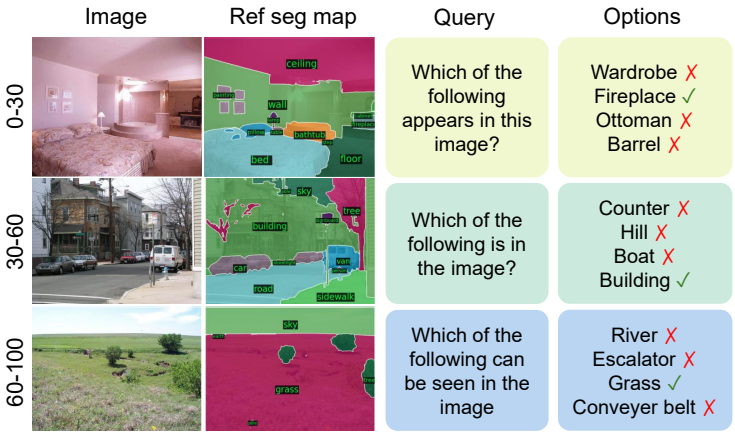

Figure 10: **MLE Benchmark Examples.** The figure provides examples from the MLE-Bench, illustrating how the dataset is partitioned based on the ground-truth object size from reference segmentation maps. For instance, in the 0-30 split, the target object (a fireplace) constitutes a small fraction of the image. In contrast, the 60-90 split features a correct object (grass) that covers a substantial portion of the image.

Model	0–30	30–60	60–100	Overall
`gpt-5`	73.63	90.69	86.31	82.97
`gpt-4o-mini`	68.72	85.10	87.94	79.32
`gemini-2.5-flash`	58.88	80.66	80.74	72.11
`claude-opus-4-20250514`	61.20	76.36	59.86	66.58

Table 5: **Model performance on the MLE-Bench.** Results are reported in three splits based on object size in percentage (0-30, 30-60, 60-100), along with a weighted overall accuracy, evaluating the ability of different models to identify objects of varying sizes.

K Multi-Level Existence Benchmark Results

This section presents the performance of top models on our Multi-Level Existence (MLE) benchmark, which evaluates their ability to identify objects of varying sizes. The results, detailed in Table 5, reveal distinct performance profiles and highlight that robust perception across all object scales remains a challenge.

gpt-5 achiees the highest overall performance with a weighted accuracy of 82.97%, demonstrating strong capabilities across all categories. It particularly excels at identifying medium-sized objects (90.69%), which form the largest portion of the visual scene, while still maintaining competent, albeit lower, performance on small objects (73.6%). In contrast, other models exhibit more pronounced trade-offs. gemini-2.5-flash struggles significantly with small objects, scoring only 58.88%, which is over 20 percentage points lower than its performance on medium and large objects. This indicates a potential weakness in fine-grained perception. Similarly, claude-opus-4-20250514 shows a notable drop in performance on large, dominant objects (59.86%) compared to smaller ones.

These findings underscore the utility of the MLE benchmark in diagnosing model weaknesses. While some models demonstrate strong general performance, universal recognition across different scales is not guaranteed. The benchmark effectively serves as a tool for driving progress toward more comprehensive and reliable visual understanding.

L Hallucinations in Blind VQA

To test for hallucinations, we present two randomly selected questions (with slight rephrasing to increase diversity) from our MLE-Bench to various models without the corresponding images. The examples below are from models also chosen randomly, primarily via LMArena (Chiang et al., 2024).

We observe that stronger models do not necessarily guarantee fewer hallucinations in this blind VQA setting.

Question 1

Question: Which of the following is visible in the image: Choose from: (A) purple and white striped wall, (B) a person's arm, (C) a purple triangle, (D) building. Please answer with a single phrase.

Answers:
gemini 2.5 pro:
A purple and white striped wall

gpt-5 thinking:
a person's arm

qwen-max-2025-08-15:
Red and white striped wall

o4-mini-2025-04-16:
a person's arm

mistral-medium-2508:
(D) building

gemma 3 27B:
a person's arm.

o4-mini-2025-04-16:
a person's arm

kimi-k2-0711-preview:
(B) a person's arm

qwq-32b:
(C) a purple triangle

gpt-oss-120b:
I'm sorry, but I can't see images. Could you describe what's in the picture?

kimi-k2-0711-preview:
(B) a person's arm

mistral-medium-2508:
(D) building

llama-4-scout-17b-16e-instruct:
B

o3-2025-04-16:
building

claude-sonnet-4-20250514:
I don't see any image attached to your message. Could you please share the image you'd like me to analyze?

phantom-0807-1:
(D) building

claude-opus-4-1-20250805:
(B) a person's arm

Question 2

Question: What can you see in the image: Choose one from: (A) a blue jersey, (B) the street, (C) araff logo, (D) a blue and white computer screen. Please answer directly.

Answers:
mistral-medium-2508:
D) a blue and white computer screen.

qwen3-235b-a22b-instruct-2507:
(C) araff logo

qwen3-235b-a22b:
(A) a blue jersey

amazon.nova-pro-v1:0:
(A) a blue jersey

phantom-0807-3:
Without access to the image, it is impossible to determine the correct answer definitively. However, based on the options provided, D) a blue and white computer screen is a common element in images and could be a plausible choice.

mistral-medium-2508:
(D) building

gpt-oss-20b:
I'm sorry, but I can't see the image.

qwen3-30b-a3b-instruct-2507:
(D) a blue and white computer screen.

gpt-4.1-2025-04-14:
(D) a blue and white computer screen.

phantom-0807-1:
(C) araff logo

claude-sonnet-4-20250514-thinking-32k:
I don't see any image attached to your message. Could you please share the image you'd like me to analyze? Once you upload it, I'll be happy to choose from the options you've provided.

gpt-5-chat:
(D) a blue and white computer screen.

mistral-medium-2508:
(D) building

claude-opus-4-1-20250805:
Looking at the image, I can see (B) the street. The image shows an urban street scene with buildings, trees, and what appears to be a road or pathway.

folsom-0811-1:
I cannot view or analyze images directly. However, if you describe the image or provide more context, I can help determine the correct answer. Since I must choose one option blindly, I'll note that (D) a blue and white computer screen is a common object that might align with typical visual descriptions, but this is a guess. For accuracy, please describe the image!

M Additional Qualitative Examples

To illustrate the differences between our models, we provide additional chat examples. These include one from `reasoning combination` (Figure 11) and one from `visual combination` (Figure 12). Each figure shows five different proportions of pre-training data.

Question: Two points are circled on the image, labeled by A and B beside each circle. Which point is closer to the camera?

Select from the following choices.

(A) A is closer
(B) B is closer

Answer the question with a detailed thinking process and explanations.

Answers

0% reasoning	25% reasoning	50% reasoning	75% reasoning	100% reasoning
The correct answer is (B) B.	The image shows a bird that is closer to the camera. The bird is positioned to the left of center of the frame.	The point of interest for the two points on the image is (B). B is closer to the camera. The point of interest is located at (B).	The correct answer is (B) B is closer. The image shows the bird perched on the edge of a rock, which is closer to the camera compared to the rock. ✓	Point B is located on the body of the bird, which is the main subject in the foreground of the image. Point A is on the background. ✓

Figure 11: **Qualitative impact of reasoning-centric data on visual spatial reasoning.** The figure shows answers from five models—pre-trained with 0% to 100% reasoning combination data—to a visual question requiring depth perception. The answers demonstrate a clear improvement in reasoning quality: the model with 0% reasoning data gives a blunt answer, while the model with 100% reasoning data provides a detailed explanation correctly applying concepts of foreground and background.

Question: Two points are annotated on the image, labeled by A and B. Consider the surface color of the points (the albedo of the surface, without the effect of shading). Which point has darker surface color, or the colors is about the same? Select from the following choices.

(A) A is darker
(B) B is darker
(C) About the same

Answer the question with a detailed thinking process and explanations.

Answers

0% visual	25% visual	50% visual	75% visual	100% visual
The question asks about the color of the points on the image. The correct answer is (C) About the same. ✗	The question asks to identify the color of the points. Point B is on a chair which is black. The answer is (B) B is darker. ✓	The image has darker surface color, which is the result of the surface color being lighter. The lighter shades are on the left side of the image. The answer is (A). ✗	The correct answer to the question is (C) About the same. The image shows a kitchen with a white countertop, a white refrigerator. ✗	The correct answer is (C) About the same. The image shows a kitchen with a white color scheme, which is a common choice for kitchens. ✗

Figure 12: **Qualitative impact of visual-world data on complex visual perception.** The figure displays answers from five models—pre-trained with 0% to 100% visual combination data—to a question requiring an understanding of color constancy. The results show that while the model with 25% visual data provides the correct answer with reasoning relevant to the core visual principle, models trained on more visual data offer incorrect answers and flawed explanations. This suggests that simply increasing descriptive visual text does not necessarily cultivate a deeper perceptual understanding.

N Samples of Language Pre-training Data

We present sample texts from our pre-training data, alongside examples from their corresponding conceptual categories. These sample texts have been slightly reformatted for clarity.

academic

The three Cartesian components of the dipole moment vector were determined using finite differences. Calculations were carried out at the CCSD(T) [coupled cluster with all single and double excitations and a perturbational estimate of connected triple excitations] level of theory with the augmented correlation consistent quadruple zeta basis set, aug-cc-pVQZ(+d for Cl), in the frozen core approximation. MOLPRO2012 was used for all calculations.

arts

What To Do If Youŕe Not Like Everybody Else (Radio 4) is misanthropic comedian Andrew Lawrenceś guide to fitting in with the rest of the world. This weekś monologue, the first of the showś second series, looks at special occasions. He begins in typically cynical style: "Birthdays, weddings, anniversaries, Christmas, new year: every month that goes by seems to slap us in the face with another contrived cause for celebration... How are the socially incompetent amongst us supposed to acquit ourselves in such circumstances without coming across as disagreeable?"
Lawrence speaks with the voice of a gnarled Lee Evans and looks at life through the eyes of a warped, sociopathic Jack Dee. In his writing he channels the vocabulary of Stewart Lee, describing Christmas as "a ludicrous social penury" and New Yearś Eve as "the zenith of nonsensical social situations".
It is, in essence, a standup set, and there are some cracking gags here, from an inventive riff on the phrase "the done thing", to an extended piece about the itinerant lives of the cheap bottles of wine guests are obliged to bring along to parties. Parties which, he repeatedly asserts, no one wants him to come to in the first place. The irony is that, on this evidence, Lawrence is very good company indeed

biology

Green ash (Fraxinus pennsylvanica) is a member of the olive family, and the species of ash tree with the largest geographic range in the United States. Green ash grows across almost the entire Eastern part of the nation and as far west as the Rocky Mountains. Green ash may prove difficult to tell apart from the other types of ash trees that grow in the U.S. The key to identification of green ash is to pay the utmost attention to even the smallest details of the tree.
Look for a tree that grows to a mature height of between 60 and 70 feet, with a spread of about 45 feet across in its upper canopy. Green ash is a medium-sized species and has an oval crown of branches. The trunk can be as wide as a foot and a half and the leaves fall from the tree every autumn.
Observe the foliage of a green ash, looking for a leaf that botanists term as being pinnately compound. Pinnate means that the separate leaflets that comprise the compound leaf grow in two rows on a central axis, called a rachis. Green ash has from seven to nine individual leaflets that grow on a rachis and compose one single leaf. The leaflets are as long as 5 inches, shaped like the sharp head of a spear, and a shiny shade of green. The entire compound leaf may be as long as a foot, says the University of Connecticut Plant Database website.
Watch for the leaves of green ash to change to yellow in the autumn. This trait adds to the appeal that makes the tree a very attractive ornamental species and one that often adorns lawns, streets and campuses.
Examine the flowers on a green ash. The species has male and female flowers that grow on separate trees. The flowers have no petals, are a light green or purple color and emerge after the leaves do in the spring. The female flowers develop in long clusters, according to the Virginia Department of Forestry website, while the male flowers grow in tighter groupings.

Study the samaras into which the flowers on the female trees turn. The samaras are a seedpod that contains the green ash treeś seeds. The "National Audubon Field Guide to Trees" states that these samaras resemble keys, hanging in cluster from the female trees as they change from a green color to a tan hue. Once mature, the 1.5- to 2.5-inch-long samaras fall off the tree and create quite a mess. This makes the male tree the preference of many landscapers, as the males will not litter your property.

code

```
<?php\nRhaco::import("lang.Validate");\nRhaco::import("lang.
    StringUtil");\nRhaco::import("exception.ExceptionTrigger
    ");\nRhaco::import("exception.model.MaxLengthException")
    ;\nRhaco::import("exception.model.MinLengthException");\
    nRhaco::import("exception.model.RequireException");\
    nRhaco::import("exception.model.DataTypeException");\
    nRhaco::import("resources.Message");\n\nclass
    UrlsTableVerify{\n\tvar $valid = true;\n\n\tfunction
    UrlsTableVerify(){\n\t}\n\tfunction verify(&$tableObject)
    {\n\t\tforeach(get_class_methods($this) as $methodName){\
    n\t\t\tif(preg_match("/^verify(.+)$/i",$methodName))
    $this->$methodName($tableObject);\n\t\t}\n\t\treturn
    $this->valid;\t\n\t}\n\tfunction verifyId(&$tableObject)
    {\n\t\t$value = $tableObject->getId();\n\t\tif(!empty(
    $value) && !Validate::isIntegerLength($value,22)){\n\t\t\
    tExceptionTrigger::raise(new DataTypeException(array(
    $this->namedId())),$this->_validName("id"));\n\t\t}\n\t}\
    n\tfunction verifyUrl(&$tableObject){\n\t\t$value =
    $tableObject->getUrl();\n\t\tif($value === "" || $value
    === null){\n\t\t\tExceptionTrigger::raise(new
    RequireException(array($this->namedUrl())),$this->
    _validName("url"));\n\t\t}\n\t}\n\tfunction verifyPubdate
    (&$tableObject){\n\t\t$value = $tableObject->getPubdate()
    ;\n\t}\n\n\tfunction namedId(){\n\t\treturn Message::_("
    id");\n\t}\n\tfunction namedUrl(){\n\t\treturn Message::_
    ("url");\n\t}\n\tfunction namedPubdate(){\n\t\treturn
    Message::_("pubdate");\n\t}\n\tfunction _validName($name)
    {\n\t\t$this->valid = false;\n\t\treturn "Urls_".$name;\n
    \t}\n}\n\n?>
```

computer science

Microseismic Extends Competitive Edge in Passive Seismic Monitoring Using Panasas Activestor Parallel Storage.
Panasas Parallel Storage Improves I/O Throughput by 2X over Traditional SAN Products, Enables MicroSeismic to Deliver Accurate Seismic Data Faster to Their Customers FREMONT, Calif. —September 17, 2007— Panasas, Inc., the leader in parallel clustered storage solutions for the High Performance Computing (HPC) market, has added MicroSeismic, Inc. to its ever-growing roster of customers providing seismic imaging services for the oil & gas industry. MicroSeismic has boosted its passive seismic application performance by 2X since deploying the Panasas® ActiveStor™ Parallel Storage Cluster, enabling the geophysical data services company to significantly reduce the time it takes to process data and return high-quality results to their oil & gas customers who benefit from this time advantage in their drilling efforts.
Houston-based MicroSeismic is the industry leader in passive seismic data acquisition and analysis. Their passive seismic technology uses data from arrays of surface-located receivers to monitor reservoir response to simulation injection and other production-related activities, allowing companies to increase the efficiency and volume of produced oil & gas. As with

most seismic applications, specialists at MicroSeismic collect and manipulate vast amounts of data on a daily basis. The company uses 70 terabytes of Panasas parallel storage, which will allow MicroSeismic to take on more client projects in less time and improve overall IT productivity, truly giving them a competitive edge and higher profitability.
"Initially, we relied on the local storage provided with Linux servers, but when we introduced a new processing application, the servers became I/O bound because the prior storage architecture wasn't optimized to handle large files. By deploying Panasas ActiveStor parallel storage running the DirectFLOW® protocol, we increased the performance of the Linux servers ten-fold and had all of the throughput we needed to satisfy not only the existing applications, but also future applications that are in development," said Michael Thornton, vice president of Data Analysis at MicroSeismic. "We evaluated competitive storage, but Panasas ActiveStor storage was the only solution that would meet our objectives for performance, throughput, and ease of management."

751

economics

Item HT 42264 Business Card - Greg McKibbin, Chairman & Managing Director, Kodak Australasia Pty Ltd, 2005-2007 Greg McKibbin's Kodak Australasia Pty Ltd business card circa 2005 - 2007. The card identifies Greg as Chairman & Managing Director of Australia & New Zealand, based at Collingwood.
Greg McKibbin joined Kodak in 1964 and went on to build a 45 year career in sales, marketing and business management with the company, working in Australia, Asia, America and Europe. He was responsible for the dismantling of the Coburg factory complex during his time as Chairman and Managing Director from 2005 - 2007.
This document was donated as part of the Kodak Oral History Project. It complements the Kodak Heritage Collection of products, promotional materials, photographs and working life artefacts collected from Kodak Australasia in 2005, when the Melbourne manufacturing plant at Coburg closed down.
Cardboard business card printed with Kodak logo, photograph and black text. Acquisition Information Donation from Mr Greg McKibbin, May 2015 Person Depicted Mr Greg McKibbin - Kodak Australasia, Pty Ltd , Collingwood , Greater Melbourne , Victoria , Australia , 2005-2007 Organisation Named Kodak (Australasia) Pty Ltd , Collingwood , Greater Melbourne , Victoria , Australia , 2005-2007 Keywords Photography , Manufacturing , Staff , Business Cards , Marketing"

752

encyclopedia

Gleniffer Lake (Alberta)
Gleniffer Lake also known as Gleniffer Reservoir or originally Lake Gleniffer is an artificial lake in central Alberta, Canada created in 1983 by the construction of the Dickson Dam which impounded the Red Deer River, a major tributary of the South Saskatchewan River which flows into the Saskatchewan River Basin.
It lies at an elevation of 945 metres (3,100 ft), and is approximately 7 kilometres (4.3 mi) long and 2 kilometres (1.2 mi) wide. The lake is south of Highway 54 and east of the Cowboy Trail, 36 kilometres (22 mi) west of Innisfail, Alberta and 36 kilometres (22 mi) east of Caroline.
The lake has a surface of 17.6 square kilometres (6.8 sq mi), and a watershed of 5,610 square kilometres (2,170 sq mi). It has an average depth of 11.6 metres (38 ft), and reaches a maximum of 33 metres (108 ft).
Gleniffer Lake has day-use areas, cottages, a campground and resort developments including Carefree Resort and Gleniffer Lake Resort.
The lake reservoir is a source of drinking water for the surrounding area.
Dickson Dam regulates the flow of the Red Deer River to control for floods and low winter flows, to improve quality of the river, to create a recreational resource and to provide a reliable, year-round water supply sufficient for future industrial, regional and municipal growth.

753

food

Flavour molecules come in all sorts of weird and wonderful organic chemistry types, but one thing they tend to have in common is that they are somewhat volatile; happy to flit from the thing you're eating or drinking directly into your retronasal cavities where the majority of tasting happens. We know that the tongue only really discerns sweet, sour, bitter, salty and umami flavours, so when a sommelier is telling you that you'll definitely taste slightly under-ripe Californian peaches in your orange wine, what they really mean is that you'll smell them as you drink, and your pathetic brain will trick you into thinking it's a taste. Enough insults, onto the instructions.This one really is the most basic; take your material for infusing and whack it in your booze, then leave, strain out and you're done. It works really well with dried herbs, fruit, vegetables, teas and in the recipe below, for salted peanut infused rum.
When we steep something in alcohol, we're pulling out all those great flavours, but we're also letting the material start to break down, so it's very important to taste your infusion and stop it at the point it's ready. . . don't just leave some cucumber in a bottle of gin and come back in 6 months, it will be grim. In fact, most infusions are at their best 4-7 days in: primarily because of the breakdown of plant cells due to alcohol/osmosis stuff , which leaches chlorophyll into the booze, leaving it tasting bitter and 'stewed'. good rule of thumb is that you should have at least 10x the weight of booze as you do material for infusion, so for example, to make a very passable copy of Hendricks: peel of one cucumber (roughtly 50g) Add all that to a kilner jar, leave for 4 days, strain and have in a G&T. You just saved £20.
To make the EC Salted Peanut Cuba Libre, you'll need: Let the peanuts sit in the rum for 24 hours, then strain through a muslin cloth to remove any particles. Do not, under any circumstances, eat the peanuts that have been soaked, they have been memorably described as having what you imagine the texture of baby teeth to be. Leave to stand for 24 hours then carefully pour into another vessel, leaving any sediment. Taking advantage of Ethanol's solvent properties, one brilliant thing we can do is strip the flavours from fats and oils. This was something PDT in New York was famous for, making Bacon infused bourbons for old fashioneds served with maple syrup (they, quite simply, fucking rule). Hawksmoor did something similar with their Full Fat Old Fashioned, infusing butter into the bourbon. At Every Cloud, we infused some lovely olive oil in gin for our clean dirty martini. To do this, we're going to stick to the 10:1 ratio again. You're going to need some space in the freezer and something for straining the liquid when you're done. 500ml Gin 50ml Olive Oil (something very flavourful, you don't necessarily want extra-virgin here. Olive pomace can work brilliantly) Mix the gin and olive oil either in a bag you can seal (zip-lock bags are great for this) or a container you have a lid

law

Attorneys for a white Chicago police officer charged with murder in the 2014 fatal shooting of black teenager Laquan McDonald announced Friday they will stick with jurors for the trial rather than have a judge decide the case. Jury selection wrapped up on Thursday. The court vetted and selected 12 jurors and five alternates during the past week. The jury is made up of seven whites, three Hispanics, one African-American and one Asian-American.
Judge Vincent Gaughan had set a Friday deadline for Jason Van Dyke to say whether he wanted to switch to a bench trial in which the judge would have decided the officerś fate. Under Illinois law, Van Dyke could unilaterally switch to a bench trial, CBS Chicago reported. Opening statements are now planned for Monday, although the judge still must decide on a defense request to move the trial outside of Cook County, where Chicago is located. The legal team representing Van Dyke has been pushing to move the case out of the county, according to CBS Chicago. Jury selection wrapped up much more quickly than expected.Most of the prospective jurors said they had seen police video of the shooting. Video shows Van Dyke shooting McDonald 16 times as the teen seems to be walking away from police with a knife in his hand. It will be one of the centerpieces at the trial.Some jurors who were excused said they could not be impartial after what they had seen the video.The release of the video, in November 2015, sparked large protests, the ouster of the police superintendent and demands for police reform.Even those who were picked for the panel expressed concern, with the last male juror saying that he thought the officer had "gone too far" when he shot the 17-year-old.

literature

erta y se sentó al volante. Las tres figuras avanzaban hacia ellos. El hombre y la mujer retrocedieron hasta la casa mientras el motor del coche se ponía en marcha. Durante un instante, las ruedas giraron sobre la nieve hasta que dejaron de resbalar y el coche se alejó. Las figuras echaron a correr y pasaron frente al hombre y la mujer sin prestarles atención, pendientes solo del coche que bajaba deslizándose por la calle nevada. El hombre de pelo blanco aferraba el volante con ambas manos. Por suerte era tarde, Nochebuena y nevaba, por lo que no había tráfico que ralentizara su marcha. Sin embargo, aunque el hombre conducía a gran velocidad, las figuras negras cada vez estaban más cerca. Corrían tan sigilosamente que resultaba sobrecogedor; a cada zancada cubrían doce metros y las puntas de sus abrigos negros ondeaban tras sí. Al doblar una esquina, el coche topó con una furgoneta estacionada y dos figuras se elevaron de un salto por los aires, asiéndose a las fachadas de las casas que bordeaban la calle. El hombre miró por el retrovisor y vio que sus perseguidores avanzaban pegados a las fachadas como gárgolas que se hubieran desprendido de los tejados. Aunque su mirada no denotaba sorpresa, pisó a fondo el acelerador. El coche cruzó a toda velocidad una plaza y pasó como una exhalación junto a un grupo de feligreses que salían de la iglesia a medianoche. Se adentró en el casco antiguo de la ciudad, y a pesar del estruendo de las ruedas rebotando en las calles adoquinadas, los niños seguían durmiendo en el asiento trasero. Una de las figuras se impulsó contra la fachada rojiza de una de las casas y aterrizó con gran estruendo sobre el coche y, acto seguido, su mano pálida rompía el techo de un puñetazo y empezaba a arañar la chapa.

mathematics

^{0} is an optimized solution for this problem. Now ... Intuition for orientation of a simplex (in 3 dimensions)
In trying to begin to learn basic homological algebra, i am confronted with orientation of simplices. The definition seems unmotivated and unintuitive: for n-simplices with $n \in \{-1, 0, 1, 2\}$, it ...
Covering n-simplex with k-subsets to produce a lower m-simplex, $m < n$?
Let vertices of an n-simplex be labeled $\{x_1, x_2, \ldots, x_n\}$ and let the k-subsets or k-intersections ($k \leq n$) be identified as $x_{i_1} \cap x_{i_2} \cap \ldots \cap x_{i_k} = x_{i_1} x_{i_2} \ldots x_{i_k} \ldots$
Do we distinguish two singular simplices if they have different vertex orders?
We define a singular n-simplex in X to be a continuous map $\sigma : \Delta^n \to X$ where Δ^n is the standard n-simplex. Now, as an example, Let X be a singleton $\{p\}$.

medicine

Mohamed I. Fayad, D.D.S., M.S., Ph.D. Dr. Fayad received his DDS in 1985 from the Collage of Dentistry, Cairo University. Dr. Fayad received his Master's in Oral Sciences in 1994 from the University of Buffalo. He received his PhD in 1996 as a joint supervision between University of Buffalo and Cairo University. He received his Endodontic training at the college of Dentistry at UIC. Currently he is the director of endodontic research, and a clinical associate professor in the Endodontic department at College of Dentistry at UIC, dividing his time between teaching, research, and private practice. Dr. Fayad is the co-author of the Periradicular Surgery chapter in Pathways of the pulp 10th edition (2011) and 11th edition 2015. Dr. Fayad is the co-editor of the text book "3-D imaging in Endodontics".
He is a Diplomate of the American Board of Endodontics and gave numerous presentations nationally and internationally. 25 E Washington St. Ste 1833 3-D Imaging in Endodontics: A new era in diagnosis and treatment Diagnostic information directly influences clinical decisions. Accurate data lead to better treatment-planning decisions and potentially more predictable outcomes. CBVT is an emerging technology that can offer the clinician clinically relevant information that cannot be gathered from conventional radiography. The ability to assess an area of Interest in 3 dimensions eliminates the superimposition that is inherent in conventional radiographic imaging. Cone-beam technology currently has numerous

applications in the dental field. CBVT is having great impact and is changing dramatically case diagnosis, treatment planning and treatment outcomes in the daily practice. Half day, full day, or two day lecture, workshop, or live demo.

philosophy

Light of the eyes : homilies on the Torah, The Hasidism Talmud RELIGION / Judaism / Sacred Writings Hasidism is an influential spiritual revival movement within Judaism that began in the eighteenth century and continues to thrive today. One of the great classics of early Hasidism, The Light of the Eyes is a collection of homilies on the Torah, reading the entire Five Books of Moses as a guide to spiritual awareness and cultivation of the inner life. This is the first English translation of any major work from Hasidism's earliest and most creative period. Arthur Green's introduction and annotations survey the history of Hasidism and outline the essential religious and moral teachings of this mystical movement. The Light of the Eyes, by Rabbi Menahem Nahum of Chernobyl, offers insights that remain as fresh and relevant for the contemporary reader as they were when first published in 1798.

politics

There are huge forces arrayed against following the British vote to leave. There are the E.U., determined to make sure that leaving is neither easy nor cheap, lest anyone else get the idea; assorted left-wingers dreaming of socialist paradise; and the slavishly like-minded press. They all hope for more of the same. On the other side, seemingly outnumbered, are those who were tired of not being represented, of being told what cucumbers to buy and of imminent doom if they were to dare defy the powers that be. For the sake of the United Kingdom and, ultimately, Europe itself, here is hoping that on March 29, 2019, the E.U. will be one member less.

q-a forum

Question: Assigning one value to another value in rails. I have two models din & company, on one of the form I am trying to create a functionality in which one can assign Directors to the respected company & company can be assigned to the directors. Here one company can have many directors & director can be on board of many companies. I have implemented the relation between the above two models using HABTM,I have created one model coDir in which the relation between two will be saved as a combination of din_id & company_id. I am trying to use two drop down menus where multiple selection is to be used.Now the problem is that how can I assign the directors to the companies & vice versa. If any one has any idea will save my weekend.
Answer: I would suggest you use has_many :through in this case. And use checkboxes for your desired functionality(although you could probably use a multiple select if you wish. Here are a couple of example tutorials to get you started.

web-crawl

Vieques offers a lot of options for adventurers, from horseback riding on deserted beaches, to the one-of-a-kind bio bay experience, to scuba diving at an abandoned military pier (Mosquito Pier or Rompeolas) a mile of the northern coast. The pier is about a mile long. It was constructed in 1941 and was supposed to connect Vieques with the mainland of Puerto Rico. But those plans were abandoned in 1943. At the end of the pier is a massive dock supported by pillars going deep to the ocean floor. The dock is on the calm, protected side of the pier and over the years the dock has become a shelter for all kinds of marine life. The beautiful thing is, it's basically an off-shore dive. The scuba operators take the divers and all the gear there by pick-up truck. If you are into diving you have to check it out. We are happy to help you arrange the dive trip. Below is a video that Mike Corey did earlier this year about

diving at the pier. It gives you a great idea of what to expect, but it doesn't come close to the real experience. We are sure you are going to love it.

visual concept

Take a riverboat up the Wailua—which translates as the "river of the great sacred spirit"—fed by the Mount Waialeale shield volcano, one of the wettest spots on the planet. Seven temples once stood along Hawaiiś longest and only navigable freshwater passage. Today, the remains of four are still visible, alongside petroglyphs and rocks where the islandś alii (royalty) would give birth. Stretch your legs at the stunning Fern Grotto: Verdant plants blanket the roof of the volcanic-rock cave there. Smith's offers 80-minute tours there on open-air boats, which include the songs and stories of ancient HawaiÍ, plus a bonus hula lesson (smithskauai.com). Take a riverboat up the Wailua—which translates as the "river of the great sacred spirit"—fed by Mount Waialeale, one of the wettest spots on the planet. Seven temples once stood along Hawaiiś longest and only navigable freshwater passage. Today, the remains of four are still visible, alongside petroglyphs and rocks where the islandś alii (royalty) would give birth. Stretch your legs at the Fern Grotto: Verdant plants blanket the roof of the volcanic-rock cave there.

visual attribute

These Spring Play Dough Mats are perfect for the upcoming weather! Is anyone else ready for winter to just be done?I'm so ready for spring, yet it is snowing outside right now, not cool. It seems as this winter weather has literally gone on for forever. . . and I'm ready for Spring to be here now. Luckily, in just a few short weeks, Spring will be a reality! In order to prepare for that glorious nature change, we're celebrating with these Printable Spring Play Dough Mats. You'll love the ease of being able to print out these play dough mats and watch your little ones create. It's so much fun to watch their excitement grow with their learning! Print out the play dough mats and either laminate them or place them inside page protectors and they can be used with play dough right away. My daughter loves covering up the pictures with different colored play dough. My son rolls out the play dough to place them on top of the letters. It's a great activity for preschoolers to learn their letters and it gives them a multi-sensory approach to letters. Especially if they're not writing that well yet. Don't forget to get your Spring Play Dough Mat Printables here! We also love games in our home, so while the weather is still just a little chilly outside, here are some fun family games as well! Visit these amazing bloggers and see what their favorite games that they enjoy in their home.

visual relationship

I do like to amuse myself with the titles of my blog posts. How can a box be both tall and small??? When it's one of mine it can be! OK, so it's not super tall, it's 5.5" (14cm) tall, but its definitely small at 1.25" (3cm) wide. So yes, I reckon it's tall and small at the same time. . . . I really enjoyed doing the tone on tone stamping a few weeks ago and am quite passionate about the Beautiful Bouquet stamp set right now, so I decided to see what they'd look like when put together as a set and a technique. I haven't used the matching framelits all that much, so I added just a tiny delicate detail on the front with a little bit of sponging around the edges. Also quite cute! And how gorgeous is the Island Indigo tone on tone? I genuinely don't think I've used this colour in the 12 months since we moved to Cambridgeshire. The reason I know this is because when I set up my new office, I set out my cardstock in a particular way with a space for full sheets of cardstock and a space for off cuts, of which I ready prepped half a dozen card blanks in each colour. I haven't used or added to the Island Indigo space. Oops, bad demo here! So it was worth making this project even if just to remind myself what a fabulous colour it is! I adore the flower detail on the ribbon!! Very cute, love the flower. Very pretty. I love that blue, such a pretty color. Thanks for sharing this project with us.

References

AI@Meta. Llama 3 model card. 2024. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Alon Albalak, Liangming Pan, Colin Raffel, and William Yang Wang. Efficient online data mixing for language model pre-training. *arXiv preprint arXiv:2312.02406*, 2023.

Zeyuan Allen-Zhu. Icml 2024 tutorial: Physics of language models. *Project page: https://physics. allen-zhu. com*, 2024.

Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. To code, or not to code? exploring impact of code in pre-training. *arXiv preprint arXiv:2408.10914*, 2024.

Kumar Ashutosh, Yossi Gandelsman, Xinlei Chen, Ishan Misra, and Rohit Girdhar. Llms can see and hear without any training. *arXiv preprint arXiv:2501.18096*, 2025.

Sören Auer, Dante A. C. Barone, Cassiano Bartz, Eduardo G. Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, Ivan Shilin, Markus Stocker, and Eleni Tsalapati. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240, May 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-33607-z.

Lichen Bai, Zixuan Xiong, Hai Lin, Guangwei Xu, Xiangjin Xie, Ruijie Guo, Zhanhui Kang, Hai-Tao Zheng, and Hong-Gee Kim. Frozen language models are gradient coherence rectifiers in vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1817–1825, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Tianyi Bai, Ling Yang, Zhen Hao Wong, Jiahui Peng, Xinlin Zhuang, Chi Zhang, Lijun Wu, Jiantao Qiu, Wentao Zhang, Binhang Yuan, et al. Multi-agent collaborative data selection for efficient llm pretraining. *arXiv preprint arXiv:2410.08102*, 2024.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, 2024.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *CVPR*, 2023.

Mayee F Chen, Michael Y Hu, Nicholas Lourie, Kyunghyun Cho, and Christopher Ré. Aioli: A unified optimization framework for language model data mixing. *arXiv preprint arXiv:2411.05735*, 2024.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. Bring reason to vision: Understanding perception and reasoning through model merging. *arXiv preprint arXiv:2505.05464*, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*, 2024.

Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. Evev2: Improved baselines for encoder-free vision-language models. *arXiv preprint arXiv:2502.06788*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Junhao Du, Chuqin Zhou, Ning Cao, Gang Chen, Yunuo Chen, Zhengxue Cheng, Li Song, Guo Lu, and Wenjun Zhang. Large language model for lossless image compression with visual prompts. *arXiv preprint arXiv:2502.16163*, 2025.

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. *arXiv preprint arXiv:2504.01017*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: a comprehensive evaluation benchmark for multimodal large language models. corr abs/2306.13394 (2023), 2023.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou, Yi-Hao Peng, Sanjay Subramanian, Qinyue Tan, Maarten Sap, Alane Suhr, Daniel Fried, Graham Neubig, et al. Autopresent: Designing structured visuals from scratch. *arXiv preprint arXiv:2501.00912*, 2025.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023.

Google. Gemini, 2023. `https://blog.google/technology/ai/google-gemini-ai/`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Junlin Han, Huangying Zhan, Jie Hong, Pengfei Fang, Hongdong Li, Lars Petersson, and Ian Reid. What images are more memorable to machines? *arXiv preprint arXiv:2211.07625*, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

William Held, Bhargavi Paranjape, Punit Singh Koura, Mike Lewis, Frank Zhang, and Todor Mihaylov. Optimizing pretraining data mixtures with llm-estimated utility. *arXiv preprint arXiv:2501.11747*, 2025.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55:661–688, 2021.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *ICML*, 2024a.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024b.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X Morris. Harnessing the universal geometry of embeddings. *arXiv preprint arXiv:2505.12540*, 2025.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Gurucharan Marthi Krishna Kumar, Aman Chadha, Janine Mendola, and Amir Shmuel. Medvisionllama: Leveraging pre-trained large language model layers to enhance medical image segmentation. *arXiv preprint arXiv:2410.02458*, 2024.

Zhixin Lai, Jing Wu, Suiyao Chen, Yucheng Zhou, and Naira Hovakimyan. Residual-based language models are free boosters for biomedical imaging tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5086–5096, 2024.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

Qiao Liang, Yanjiang Liu, Weixiang Zhou, Ben He, Yaojie Lu, Hongyu Lin, Jia Zheng, Xianpei Han, Le Sun, and Yingfei Sun. Expanding the boundaries of vision prior knowledge in multi-modal large language models. *arXiv preprint arXiv:2503.18034*, 2025.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024b. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, et al. X-reasoner: Towards generalizable reasoning across modalities and domains. *arXiv preprint arXiv:2505.03981*, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024c.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Frozen pretrained transformers as universal computation engines. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 7628–7636, 2022a.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2023.

Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. At which training stage does code data help llms reasoning? *arXiv preprint arXiv:2309.16298*, 2023.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Sayak Mallick, Matthias Bethge, and Wieland Brendel. Llms on the line: Data determines loss-to-loss scaling laws. *arXiv preprint arXiv:2502.12120*, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

OpenAI. gpt4o, 2024. `https://openai.com/index/hello-gpt-4o/`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. In *TMLR*, 2023.

Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. Finding and editing multi-modal neurons in pre-trained transformers. *arXiv preprint arXiv:2311.07470*, 2023.

Ziqi Pang, Ziyang Xie, Yunze Man, and Yu-Xiong Wang. Frozen transformers in language models are effective visual encoder layers. *ICLR*, 2024.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pages 3403–3417, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, et al. Magistral. *arXiv preprint arXiv:2506.10910*, 2025.

Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. Multimodal neurons in pretrained text-only transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2862–2867, 2023.

Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, and Antonio Torralba. A vision check-up for language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14410–14419, 2024.

Zachary Shinnick, Liangze Jiang, Hemanth Saratchandran, Anton van den Hengel, and Damien Teney. Transformers pretrained on procedural data contain modular structures for algorithmic reasoning. *arXiv preprint arXiv:2505.22308*, 2025.

Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures, 2025a. `https://arxiv.org/abs/2507.09404`.

Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. *arXiv preprint arXiv:2504.07951*, 2025b.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021.

Chunyi Sun, Junlin Han, Weijian Deng, Xinlong Wang, Zishan Qin, and Stephen Gould. 3d-gpt: 3d modeling with large language models. In *International Conference on 3D Vision (3DV)*, 2025.

Chenxin Tao, Shiqian Su, Xizhou Zhu, Chenyu Zhang, Zhe Chen, Jiawen Liu, Wenhai Wang, Lewei Lu, Gao Huang, Yu Qiao, et al. Hovle: Unleashing the power of monolithic vision-language models with holistic vision-language embedding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14559–14569, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024a.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024b.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024c.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. 2023.

Gaurav Verma, Minje Choi, Kartik Sharma, Jamelle Watson-Daniels, Sejoon Oh, and Srijan Kumar. Cross-modal projection in multimodal llms doesn't really project visual attributes to textual space. *arXiv preprint arXiv:2402.16832*, 2024.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jisheng Yin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Zejia Weng, et al. Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning. *arXiv preprint arXiv:2507.05255*, 2025.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

xAI. grok, 2024. `https://x.ai/blog/grok-1.5v`.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36:34201–34227, 2023.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025b.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *arXiv preprint arXiv:2403.16952*, 2024.

Zheng-Xin Yong, M Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H Bach, and Alham Fikri Aji. Crosslingual reasoning through test-time scaling. *arXiv preprint arXiv:2505.05408*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Chi Zhang, Huaping Zhong, Kuan Zhang, Chengliang Chai, Rui Wang, Xinlin Zhuang, Tianyi Bai, Jiantao Qiu, Lei Cao, Ju Fan, et al. Harnessing diversity for important data selection in pretraining large language models. *arXiv preprint arXiv:2409.16986*, 2024.

Xinlu Zhang, Zhiyu Zoey Chen, Xi Ye, Xianjun Yang, Lichang Chen, William Yang Wang, and Linda Ruth Petzold. Unveiling the impact of coding data instruction fine-tuning on large language models reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25949–25957, 2025.

Boyang Zheng, Jinjin Gu, Shijun Li, and Chao Dong. Lm4lv: A frozen large language model for low-level vision tasks. *arXiv preprint arXiv:2405.15734*, 2024.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. In *NeurIPS*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.