A Compound 3D-Informed Design toward Spatially-Intelligent Large Multimodal Models

Wufei Ma[1], Luoxin Ye[1], Celso M de Melo[2], Alan Yuille[1], Jieneng Chen[1]
[1]Johns Hopkins University [2]DEVCOM Army Research Laboratory

Abstract

Humans naturally understand 3D spatial relationships, enabling complex reasoning like predicting collisions of vehicles from different directions. Current large multimodal models (LMMs), however, lack of this capability of 3D spatial reasoning. This limitation stems from the scarcity of 3D training data and the bias in current model designs toward 2D data. In this paper, we systematically study the impact of 3D-informed data, architecture, and training setups, introducing 3DI-LMM, an LMM with advanced 3D spatial reasoning abilities. To address data limitations, we develop two types of 3D-informed training datasets: (1) 3D-informed probing data focused on object's 3D location and orientation, and (2) 3D-informed conversation data for complex spatial relationships. Notably, we are the first to curate VQA data that incorporate 3D orientation relationships. Furthermore, we systematically integrate these two types of training data with the architectural and training designs of LMMs, providing a roadmap for optimal design aimed at achieving superior 3D reasoning capabilities. Our 3DI-LMM advances machines toward highly capable 3D-informed reasoning, surpassing GPT-4o performance by 8.7%. Our systematic empirical design and the resulting findings offer valuable insights for future research in this direction.[1]

1. Introduction

When humans observe a scene, they perceive more than isolated objects; they intuitively understand the spatial relationships among these objects [53], enabling complex reasoning and interaction with the environment. For example, recognizing that a car is approaching an obstacle and might collide involves interpreting the 3D object orientations and spatial relationships. This innate ability to comprehend and reason about 3D space is crucial for navigating and interacting with the physical world.

Despite significant advancements, current LMMs [1, 4, 38, 40, 55], exhibit limited capability of 3D spatial reasoning. These models excel at identifying objects and generating descriptive captions but struggle with tasks that require intricate 3D reasoning, such as predicting object interactions or understanding physical events.

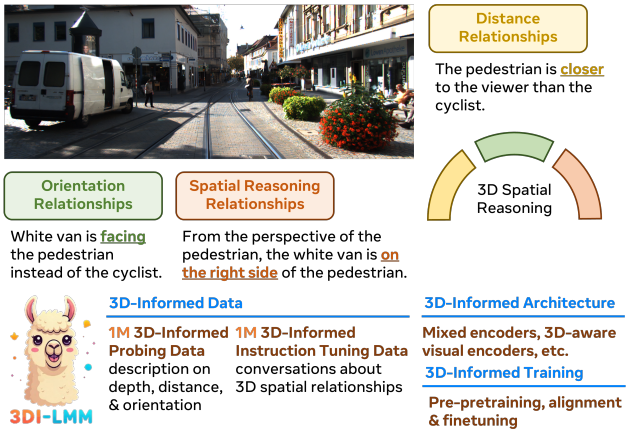


Figure 1. 3D spatial reasoning is crucial for LMMs to ground objects in 3D space and infer their 3D spatial relationships, such as distance, orientation, and spatial interactions. We systematically study the impact of 3D-informed data, architecture, and training setups, introducing 3DI-LMM, an LMM with advanced 3D spatial reasoning abilities.

The primary challenge lies in the scarcity of high-quality 3D spatial relationship data, which prevents models from learning effective 3D spatial reasoning. Collecting a small set of high-quality 3D-aware data to tackle the first challenge is feasible, albeit labor-intensive, using readily available tools. Recent efforts [2, 14] leverage vision foundation models to recognize [64], detect [41], segment [30] and estimate the depth [10, 60] in images. These approaches form semi-automated pipelines to construct pseudo-2.5D datasets for training LMMs with spatial reasoning capabilities, such as depth ordering [14, 56].

However, a significant gap remains: previous works [2, 14, 16] have primarily focused on 3D distance relationships, overlooking the crucial role of 3D object orientation. This

[1]Approved for public release: distribution is unlimited.

is primarily due to the absence of robust 6D pose estimators. Understanding object orientation is essential for complex spatial reasoning tasks, such as determining if two cars are facing the same direction to avoid dangerous collision. Addressing this gap, we aim to incorporate 3D orientation relationships—converted from ImageNet3D [17]—into our data engine, making us the first to enable complex spatial reasoning involving 3D orientation relationships.

Even with 3D-informed data in hand, fully unlocking its potential remains challenging due to the intricate, multi-stage development process required for an LMM. For example, SpatialVLM [14] utilizes 3D-informed data during the instruction tuning stage, but it remains uncertain whether such data can enhance other stages like vision encoder pre-training or multimodal alignment. This limitation suggests that a more holistic approach is necessary.

In this paper, we present a systematic investigation into building LMMs with enhanced 3D spatial reasoning capabilities through a compound design approach, as illustrated in Fig. 1. Our key contributions are threefold:

First, we identify three fundamental types of 3D spatial relationships that LMMs must master: distance relationships, orientation relationships, and complex spatial reasoning that combines both. To evaluate these capabilities, we introduce SpatialVQA, a comprehensive benchmark containing 1,323 questions that target 3D spatial understanding, built upon diverse urban and indoor scenes from Omni3D [11].

Second, we propose a novel compound 3D-informed design that introduces improvements across multiple dimensions, leading to our proposed **3DI-LMM** model. On the data front, we develop two types of 3D-informed training datasets: (1) 3D-informed probing data focused on object-level 3D properties, and (2) 3D-informed conversation data for for complex spatial relationships. Architecturally, we explore mixing different visual encoders and investigate their 3D awareness capabilities. For training, we introduce 3D-informed data at multiple stages: pre-pretraining, multimodal alignment, and instruction tuning.

Third, we present the first comprehensive search over the LMM design space for spatial reasoning tasks and propose a roadmap towards developing state-of-the-art models in this domain. The experiment demonstrates that our proposed **3DI-LMM** achieves state-of-the-art performance on spatial reasoning tasks, surpassing both GPT-4o [1] (by 8.7%) and spatial-centric models like SpatialVLM [14] (by 10.5%). Our systematic analysis reveals that while architectural improvements contribute a 1.3% gain, the strategic incorporation of 3D-informed data across different training stages yields a substantial 13.7% performance improvement.

Our findings highlight the importance of a holistic approach to improving LMMs' 3D spatial reasoning capabilities, demonstrating that both architectural choices and carefully curated open-source 3D-informed training data play crucial roles in achieving superior performance. This work establishes a new benchmark for spatial reasoning in LMMs and provides valuable insights for future research in this direction.

2. Related Works

Large multimodal models (LMMs). Recently, there has been a significant surge in research on multimodal LLMs, exemplified by products like GPT-4 [1] and Gemini [55], as well as open-source models such as LLaVA [33, 38, 40] and Qwen-VL [8]. These open-source models combine pre-trained vision encoders [15, 26, 45, 46] with LLMs [7, 27, 57], enabling the language models to handle a wide range of tasks that involve images as input. Typically, these methods has the alignment stage that aligns image features with the input embedding space of the language model using connectors [13, 34, 38], trained with paired images and captions [51]. The instruction fine-tuning [40] stage uses data that contains both images and instructions to enhances LMMs. However, current LMMs are unsuccessful in spatial visual reasoning, due to limited spatial-aware data and stagnant 3D-informed architecture design. We aim to bridge this research gap.

3D spatial reasoning with LMMs. Understanding about spatial relationships between objects in images is crucial for visual reasoning. While this comes naturally to humans, LMMs have been found to struggle with recognizing spatial relations [47], due to the lack of 3D-aware spatial-related data during training. Fortunately, with foundation models like SAM [30] and depth estimators [10, 60], semi-automatic annotation for 3D-aware VQA data has been adopted in recent studies [14, 16]. SpatialVLM [14] uses the pseudo-annotated 3D-aware VQA data for training only while SpatialRGPT [14, 16] further augments LLaVA with extra input of regional mask and depth map during inference, which is not our focus. Although human-annotated 3D-aware VQA data is rare, we aim to leverage human-annotated 3D object-level datasets (like ImageNet3D [43]) by simply converting image-pose pairs into image-text pair data. However, even with the collected data, a major challenge remains in effectively utilizing 3D-aware data to train optimal models. Current approaches like SpatialVLM [14] use the standard LLaVA and two-stage training. To date, no work has addressed when (which training stage) and where (which network component) are optimal to inject the 3D-awareness to LMMs. We bridge this gap by conducting a compound design that simultaneously considers 3D-aware data, architecture, and the training, leading to a best-performing design. Rather than competing, our study has the potential to complement prior works [14, 16] by offering new compound design recipe.

3. Methods

In this section we start by reviewing LMMs in Sec. 3.1, specifically the architecture and the choice of training data and strategies. We present the task of reasoning 3D spatial relationships and explain the challenges LMMs face when answering these questions in Sec. 3.2. Then we study how to address these challenging problems with LMMs and introduce our 3DI-LMM in Sec. 3.3.

3.1. Preliminary of LMMs

Data. The success of deep learning in vision is largely due to well-curated data [20, 54], and LMMs similarly depend on this foundation, albeit with diverse data formats below.

- Image-only data, such as ImageNet [17], is commonly used for MAE [23]/DINOv2 [45] pretraining with a self-supervised objective. This step enables the model to learn rich visual features solely from visual signals.
- Noisy image-text pairs: Large-scale image-text pairs [20, 36, 50] are used for pretraining with contrastive loss (e.g., CLIP [46, 62]) to align visual and textual representations.
- Multimodal alignment: At this stage, image-caption pairs (e.g., CC3M [51]) are utilized to align visual features with the language model's word embedding space, typically through image-captioning tasks. This pretraining ensures that the model can interpret and relate visual content to textual queries in a shared semantic space.
- Multimodal instruction-following data. Visual instruction tuning data is essential but challenging to collect, as it rarely exists in its natural form online. Prior works [38, 40, 44, 56] repurpose VQA benchmarks [22, 31] into instruction-tuning datasets.

Architecture. A standard LMM [38, 40] consists of a visual encoder to process the image, a multimodal connector to transform the visual feature to visual token, and a LLM for reasoning. However, the extent to which these architectural components contribute to the spatial understanding of LMMs remains unknown. We review these architectural components as below.

- Vision encoder: Most LMMs rely on language-supervised models like CLIP [46], leveraging web-scale noisy image-text data. Recent LMMs [42, 56] integrate vision models trained only on visual signals, including self-supervised [23, 45, 66] and generative models [49].
- Connector: Features from a visual encoder are mapped into the LLM token space by connectors [3, 13, 34, 40] such as a MLP [38, 40].
- LLMs: LLMs [1, 7, 57] serve as the foundational backbone of LMMs, driving their reasoning capabilities. Open-source LLMs like Llama series [18, 57] are frequently adopted to enhance the reasoning capability.

Training. The aforementioned data supports pre-pretraining for strong visual representation learning, alignment pretraining to synchronize visual and language representations, and instruction tuning to enable multimodal instruction-following and reasoning capabilities

- Pre-pretraining. This stage focuses on developing foundational visual representations, often with reconstruction-based objectives (e.g., MAE [23], DINOv2 [45]). This step prepares the model to handle more complex multimodal tasks by first refining its visual understanding.
- Multimodal alignment pretraining. At this stage, the model is trained to describe images in details to align visual and language representations in the same space.
- Visual instruction tuning. The final stage involves refining the model's ability to process complex multimodal instructions using visual instruction tuning data. This often incorporates transformed VQA datasets, enhancing the model's capacity in handling tasks that demand coordinated reasoning across language and vision.

3.2. Reasoning 3D Spatial Relationships

3D spatial reasoning analyzes the spatial relationships between objects in the 3D world space, such as objects being far or close to each other, or one object lies to the left or right of another object [58]. It extends beyond standard 3D tasks, such as 3D object detection and 3D pose estimation that studies the 3D nature of individual objects, and focuses on understanding their mutual relationships in 3D. In this work we identify three main types of 3D spatial relationships, *i.e.*, distance, orientation, and spatial reasoning, and study how we can better address these problems with LMMs.

3D spatial relationships. For *distance spatial relationships*, we study if models can estimate and compare 3D distances between objects. These spatial relationships require more than depth awareness, which gives the distance between the viewer and the object and can be roughly estimated from the object scale, but also the 3D distances between two objects that are often harder to estimate. On the other hand, *orientation spatial relationships* require models to understand the 3D orientations of an object, *e.g.*, predicting if two objects are facing the same direction. Lastly, *spatial reasoning spatial relationships* look into 3D spatial relationships that need a combination of 3D awareness of location and orientation and then performing spatial reasoning on the 3D information.

3.2.1. Challenges of 3D spatial reasoning

Reasoning about 3D spatial relationships remains a challenging task for state-of-the-art open-sourced and proprietary LMMs. We attribute this to two main limitations in modern LMMs, the *lack of 3D awareness* and *inability to perform 3D spatial reasoning*.

3D awareness refers to the ability of a visual encoder to represent 3D-aware features such as representing the 3D ob-

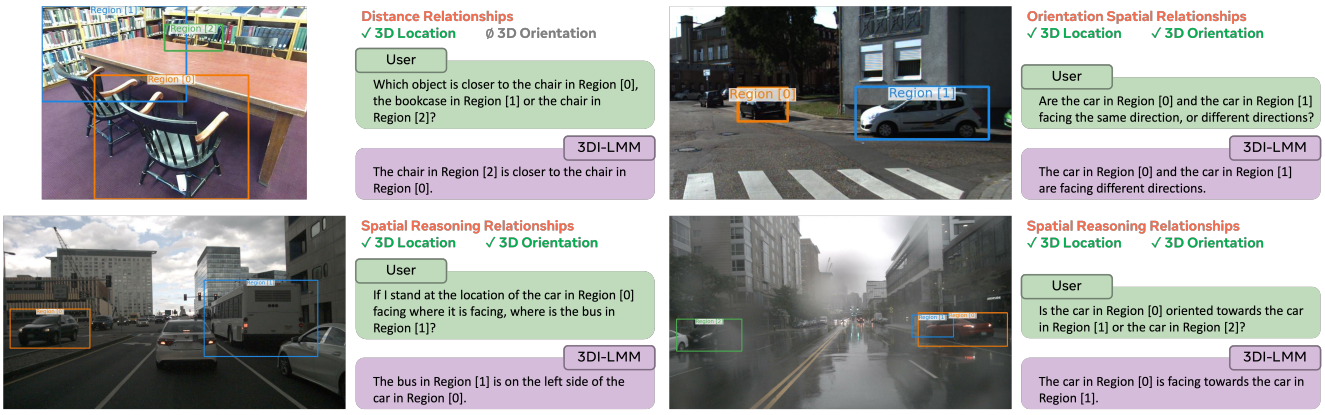


Figure 2. Examples from our SpatialVQA benchmark featuring a broad range of questions that require 3D spatial reasoning.

ject shapes and the 3D orientations. Previous studies resorted to proxy tasks, *e.g.*, part correspondence [19] and 3D pose estimation [43], to quantitatively evaluate the 3D awareness of visual foundation models with linear probing. Results showed that most visual encoder, specifically ones trained with language supervision, demonstrated limited 3D awareness of 3D object poses [43], which inevitably limits the 3D spatial reasoning capabilities of LMMs built upon these encoders.

Inability to perform 3D spatial reasoning. To infer the 3D spatial relationships, models must also be able to perform 3D spatial reasoning from 3D-aware visual features. Despite the abundant Internet-scale data for multi-modal training, we argue that high-quality 3D spatial reasoning data are scarce. Existing pretraining and visual instruction tuning data for LMMs [40, 56] focused on detailed descriptions and conversations about scenes, appearances, and actions, while being vague about the 3D spatial relationships that build up the high-level physical events. Despite their success on a wide range of image understanding tasks, these LMMs fails to learn the nature of 3D scenes or to perform 3D spatial reasoning.

3.2.2. SpatialVQA for Evaluation

Despite the rising interests in spatial reasoning, there lacks a comprehensive open source benchmark for reasoning 3D spatial relationships. Early datasets on spatial reasoning were built on 3D scans rather than images [6, 61]. Recent visual-language benchmarks on spatial reasoning either focused on 2D spatial relationships [14, 16], *e.g.*, left or right in the image plane, or only on distance spatial relationships [56].

To enable a thorough evaluation of the 3D spatial reasoning capabilities of LMMs, we introduce SpatialVQA, a visual question answering benchmark that covers a wide variety of 3D spatial relationships. Our SpatialVQA distinguishes itself from all previous spatial reasoning benchmarks in the sense that all questions require different levels of 3D awareness and cannot be answered from 2D spatial reasoning only. Specifically, SpatialVQA consists of 1,323 questions, ranging from questions can be answered from distance-awareness only, *e.g.*, which object is closer to a third object, to questions that require estimating objects' 3D orientations, *e.g.*, if two objects are facing the same direction, and eventually questions that require complex 3D spatial reasoning over a combination of object 3D locations and orientations, *e.g.*, an object is in the front/left/back/right direction of another object.

We build our SpatialVQA on images from Omni3D [11], which provide 3D bounding box annotations on diverse objects from both urban [12, 21] and indoor scenes [9, 48, 52]. We follow [16, 56] and develop rule-based methods to generate visual question-answer pairs from the 3D groundtruths. Please refer to Fig. 2 for some examples of SpatialVQA and to appendix Sec.B for details about our benchmark.

3.3. Compound 3D-Informed Design

Problem formulation. While previous approaches focused on generating 3D-informed instruction tuning data to enable LMMs with the abilities to estimate distances, depths, or spatial relationships [14, 16], it is largely understudied the best recipe for developing LMMs with strong 3D spatial reasoning capabilities. Motivated by the two aforementioned limitations of LMMs in Sec. 3.2.1, we consider two main aspects in our compound 3D-informed design – the architecture design that leads to visual encoders with strong 3D awareness and the training setup that advances the 3D spatial reasoning capabilities of LMMs.

3.3.1. Design space

We introduce the design space considered in our work, *i.e.*, choices of training data, model architecture, and training

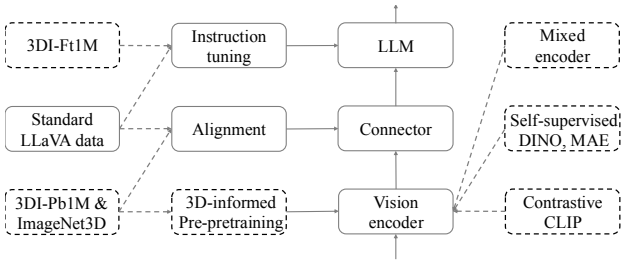


Figure 3. **Design space** for LMMs capable of spatial reasoning. The dashed boxes and lines highlight our new design space compared to LLaVA-v1.5. This compound design simultaneously considers 3D-informed data, architecture, and training methods to search for the best-performing models for spatial reasoning.

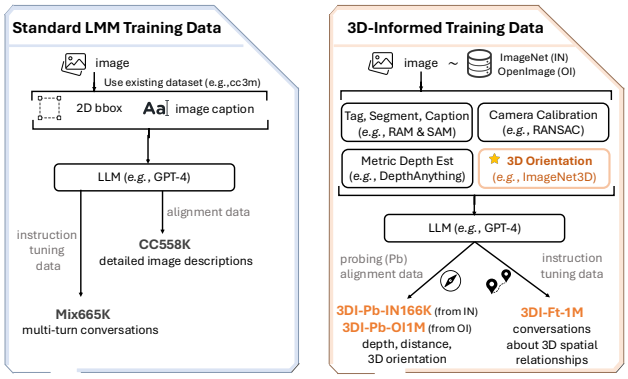


Figure 4. Comparison of data cards showcasing the curation process and data types: standard LLaVA data (left) vs. our 3D-informed data (right). For our 3D-informed data curation, we label depth and distance of objects in OpenImage [32] using semi-automatic tools, resulting in 3DI-Pb-OI1M. Beyond this, we introduce 3D orientation probing data derived from human-annotated ImageNet3D [43], resulting in 3DI-Pb-IN166K for alignment. We further augment the conversation about spatial relationship, leading to 3DI-Ft-1M for instruction tuning.

setup that advance the 3D spatial reasoning capabilities of LMMs. Refer to Fig. 3 for an overview of our design space.

Data. Standard LMMs follow LLaVA [40] and adopte feature alignment data for connector pretraining and visual instruction data for supervised instruction tuning. These datasets focus on detailed descriptions of images or machine-generated multi-turn conversations. Besides standard visual-language datasets, we further consider 3D-informed data, including 3D-aware probing data and 3D-informed instruction tuning data. Inspired by [43], the 3D-aware probing data involves *fundamental* object-level 3D properties, such as the azimuth and elevation 3D viewpoint of an object, depth of the object from the viewer, and 3D distances between two objects. The goal of the 3D-aware probing data it to effectively inject fundamental 3D awareness into the models and benefit subsequent reasoning. Meanwhile, the 3D-informed instruction tuning data focuses on *high-level* 3D spatial relationships between the objects, which are built on a combination of the fundamental 3D properties and 3D spatial reasoning.

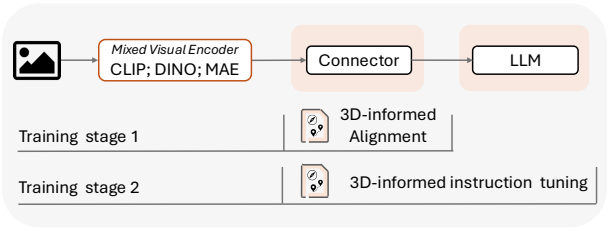
To generate the 3D-aware probing data and the 3D-informed instruction tuning data, we follow the previous data generation pipelines [14, 16] and exploit visual foundation models to extract and tag the objects in 3D space. Furthermore, we adopt a pretrained 6D pose estimator in [43] and estimate 3D orientations of the objects, which provides the crucial 3D information for generating 3D-informed data regarding various orientation-related 3D spatial relationships. We illustrate our 3D-informed data generation pipeline in Fig. 4 and refer the readers to appendeix Sec.A for details of our 3D-informed data generation.

Architecture. Inspired by Probe3D [19] which probes the 3D-awareness of visual foundation models, we study the design of visual encoder to enable better 3D spatial reasoning. We consider two types of visual encoders: (i) Frozen & pretrained visual encoder CLIP [46] following [40] but with the option to mix a wider variety of models, such as MAE [23] and DINOv2 [45], and (ii) 3D-aware visual encoders finetuned with our 3D-informed data, *e.g.*, contrastive pretraining or finetuning with LoRA using language modeling loss.

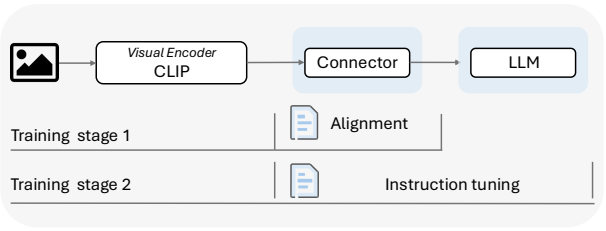
Training setup. To address the limitations of 3D-aware LMMs discussed in Sec. 3.2.1, we propose new training setups that aim to improve 3D awareness and advance the 3D spatial reasoning capabilities. We consider the following: (i) in Stage 0: use 3D-informed probing data (3DI-Pb-OI1M/1N166K) to tune the Lora layers of CLIP vision encoder (Pre-pretraining); (ii) in Stage 1: 3D-informed multimodal alignment using a combination of standard CC558K [40] and our 3D-informed probing data (3DI-Pb-OI1M/1N166K); and (iii) in Stage 2: 3D-informed instruction tuning with the standard Mix665K [40] and our 3D-informed visual instruction tuning data (3DI-Ft-1M).

3.3.2. Designing 3D-Informed LMM: A Roadmap

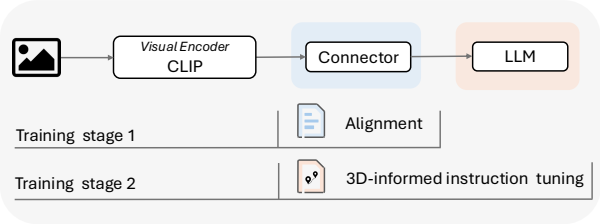
Design instantiation and comparison. The mentioned design space can result in several design instantiations, as depicted in Fig. 5. In a multimodal LLM, a pre-trained CLIP visual encoder extracts grid features from the input image. These features are converted into visual tokens via a connector and then processed by the LLM for visual reasoning. Fig. 5 (b) shows that LLaVA is a standard design space without any new spatial-centric design, meaning that the 3D-awareness is completely missing in LLaVA (Fig. 5 (b)). As in Fig. 5 (c), SpatialVLM improves LLaVA by collecting



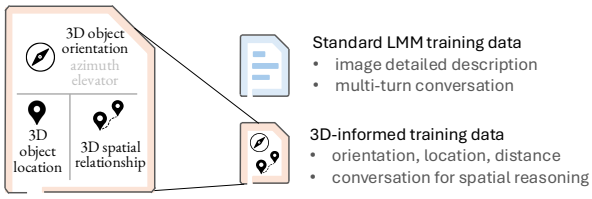
(a) our proposed design



(b) design of LLaVA-v1.5 (CVPR'24)



(c) design of SpatialVLM (CVPR'24)



(d) data cards

Figure 5. **Design instantiation and comparison**. (a) Architecture and Training of our proposed design. We investigate the 3D-awareness of mixed visual encoders, and incorporate 3D-informed data at each training stage across all architecture components. (b) LLaVA comparison, where only traditional data are used. (c) SpatialVLM comparison, where 3D-informed data are used in the final stage. (d) Traditional data vs. 3D-informed data. Note SpatialVLM lacks of 3D orientation data. Architectures trained with traditional data and 3D-informed data are marked with blue and orange backgrounds, respectively.

3D-informed data that is applied to instruction tuning for LLM, boosting the spatial reasoning capability. Compared to SpatialVLM, our design (in Fig. 5 (a)) inject the 3D-awareness by introduce 3D-informed alignment with ImageNet3D data. Additionally, our 3D-informed intruction tuning data contain extra 3D orientation data that is missing in spatialVLM.

Design roadmap. We present the roadmap going from a standard LLaVA to our final model in Fig. 6. We start with LLaVA-v1.5 which only achieves a 47.7% accuracy on our SpatialVQA. This will be our baseline. We then study a series of design decisions.

- + *mixed vision encoder*. To build stronger visual encoder, we mix the original CLIP vision encoder [46] with the DINOv2 [45] vision encoder to extract the hybrid feature maps, lifting a 0.3% performance gain. This show that the self-supervised feature can enhance spatial reasoning.
- *Vicuna1.5→Llama3*. We switch our large language model from Vicuna-v1.5-7B [65] to Llama3-8B [18], as the latter achieves 20% higher accuracy on MMLU. This change results in 1% gain on spatial reasoning. We analyze that improved language reasoning may enhance spatial reasoning as well.
- + *3DI-Ft1M data for stage 2 instruction tuning*. We collect the new spatial-aware training set and use it in the instruction tuning stage. Impressively, this 3D-informed instruction tuning results in a 10.7% performance improvement, highlighting the critical importance of 3D-informed data.
- + *3DI-Pb-1N166K data for stage 1 multimodal alignment*. We propose leveraging object-level 3D-informed annotation (e.g., 3D location and 3D orientation of objects) to align visual features with 3D awareness. 3DI-Pb-1N166K are converted from human-annotated Ima-

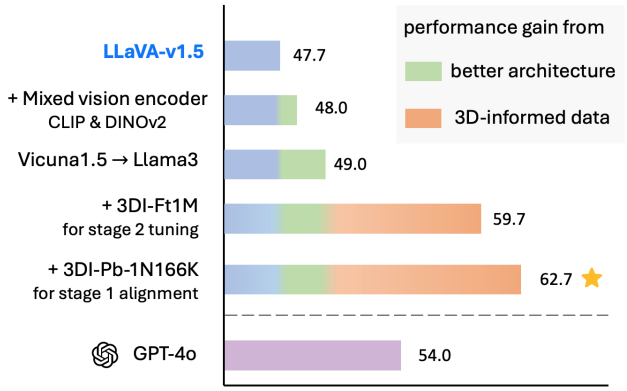


Figure 6. We modernize a standard LLaVA-v1.5 towards the design of a 3D-informed LMM. The bars are the answer accuracies on the SpatialVQA benchmark, indicating the spatial reasoning capability of each model; results for the GPT-4o is shown with the purple bar.

geNet3D [43]. This experiment results in an additional 3% performance improvement, underscoring the importance of 3D-informed feature alignment.

In a nutshell, the performance gain can be summarized as 1) better architecture and 2) 3D-informed data. This roadmap leads to our 3DI-LMM.

4. Experiments

In this section we start by outlining the experimental setup in Sec. 4.1, followed by presenting the results in Sec. 4.2.

4.1. Experimental setup

We prepare our training data as described in Fig. 4, with 1M paired image and 3D-informed text for pre-pretraining and alignment, and another 1M multi-turn 3D-informed conversation for instruction tuning. Data distributions are supplemented in appendix. Our training setup is built upon LLaVA-v1.5 [38] and all hyperparameters remain unchanged unless explicitly stated otherwise. We briefly overview the training and data setup for the ablated design choices in Sec. 3.3.1. All experiments are conducted in 8 Nvidia A100 GPUs. We evaluate the spatial reasoning capability of LMMs on SpatialVQA as introduced in Sec. 3.2.2.

4.2. Results

Comparison with the state-of-the-art. Table 1 presents the performance of our model compared to state-of-the-art proprietary systems such as GPT-4o and Claude 3.5 Sonnet, along with top open-source models. Remarkably, our model achieves a performance of 62.7%, outperforming the top proprietary model by 8.7% and the best open-source model by 10.5%. Besides, our model excels in all aspects of spatial reasoning involving 3D distance, 3D orientation and 3D spatial relationship. Moreover, our model not only enhances spatial reasoning performance, but still preserves general visual reasoning (detailed in appendix). Interestingly, although SpatialVLM [14] (implemented in SpaceLLaVA [2]) outperforms other open-source models in overall performance, it falls short in 3D orientation reasoning compared to LLaVA, due to the lack of tuning with 3D orientation-specific data. We exclude SpatialRGPT from this comparison as it relies on additional inputs such as object bounding box coordinates, segmentation masks, and depth maps. In contrast, our 3DI-LMM, along with LLaVA and Cambrian, focuses on learning spatial awareness directly from raw images. We will consider models with additional inputs in future work. In sum, our model achieves state-of-the-art performance on spatial reasoning, surpassing API models [1, 4] and also spatial-centric VLM [14].

Model	Avg.	3D Dist.	3D Orient.	3D Spatial Rel.
Proprietary				
GPT-4o [1]	54.0	59.4	52.4	50.8
GPT-4o Mini [1]	45.4	46.9	39.3	48.5
Claude 3.5 Sonnet [4]	49.3	50.8	44.5	51.5
Open-sourced				
LLaVA-v1.5-7B [38]	47.7	50.6	46.4	46.2
LLaVA-NeXT-8B [39]	50.6	63.5	49.2	41.5
Cambrian-8B [56]	51.7	59.8	46.4	48.9
Spatial-centric				
SpatialVLM-13B [14]	52.2	64.3	48.5	45.8
Ours	**62.7**	**86.3**	**52.9**	**50.8**

Table 1. **Comparison with the state-of-the-arts** including proprietary and open source models.

Thorough empirical exploration of the design space. We comprehensively explore the design space of our approach as shown in Table. 2. We analyze the results and elaborate the findings detailed below.

Upgrading architectures enable better spatial reasoning. In terms of architecture, integrating a mixed vision encoder can improve overall performance especially for the 3D orientation. Then, upgrading the LLM to the more advanced LLama3-8B enhances spatial reasoning performance by 1%, leveraging its improved language reasoning capabilities.

3D-informed alignment and tuning matter a lot. In terms of 3D-informed data and training, we find that 3D-informed instruction tuning with our proposed 3DI-Ft1M dataset yields a substantial performance boost of +10.7%. To further refine the framework, we explore three data types for stage 1 multimodal alignment: curated 3DI-Ft1M, 3DI-Pb-OI1M, and 3DI-Pb-IN166K. The model trained with 3DI-Pb-IN166K performs best. We attribute this to the use of human-annotated ImageNet3D data, as opposed to the semi-automated methods employed for 3DI-Pb-OI1M. The human-annotated ImageNet3D provides more accurate 3D orientation of objects. This highlights the importance of high-quality data in the alignment stage.

3D-informed pretraining of vision encoder? The pretrained CLIP vision encoder, originally designed for image-text contrastive learning, lacks 3D-awareness [19], We hypotheszie that the vision encoder has a chance to benefit from 3D-informed data. To verify this, we conduct 3D-informed pretraining for the vision encoder using curated 3DI-Pb-OI1M probing data. We explore both CLIP [25] and next token prediction (NTP) [40] objectives to fine-tune Lora layers [24] of the ViT. We observe that pre-pretraining in stage 0 reduces performance, suggesting that tuning visual foundation models may compromise feature generalizability. This aligns with prior works advocating for frozen or locked visual encoders [35, 37].

Roadmap progression towards the best-performing

Model		Stage 0	Stage 1	Stage 2	SpatialVQA			
Vision Encoder	LLM	Encoder pretraining	Alignment	Instruction Tuning	Avg.	Dist.	Orient.	Spatial Rel.
***Baseline** (LLaVA-v1.5)*								
CLIP [46]	Vicuna-7B [65]	–	CC558K [38]	Mix665K [38]	47.7	50.6	46.4	46.2
Mixed Encoders								
CLIP [46]+MAE [23]	Vicuna-7B [65]	–	CC558K	Mix665K	47.9	52.2	45.7	46.1
CLIP [46]+DINOv2 [45]	Vicuna-7B [65]	–	CC558K	Mix665K	48.0	51.9	46.9	45.6
Advanced LLM								
CLIP [46]+DINOv2 [45]	Llama3-8B [18]	–	CC558K [38]	Mix665K [38]	49.0	56.3	51.7	41.5
3D-informed Tuning for Stage 2								
CLIP [46]+DINOv2 [45]	Llama3-8B [18]	–	CC558K [38]	Mix665K [38] + 3DI-Ft1M	59.7	85.2	52.1	44.8
3D-informed Alignment for Stage 1								
CLIP [46]+DINOv2 [45]	Llama3-8B [18]	–	CC558K [38] + 3DI-Ft1M	Mix665K [38] + 3DI-Ft1M	57.5	80.8	46.9	46.4
CLIP [46]+DINOv2 [45]	Llama3-8B [18]	–	CC558K [38] + 3DI-Pb-OI1M	Mix665K [38] + 3DI-Ft1M	60.0	82.1	46.9	51.6
CLIP [46]+DINOv2 [45]	Llama3-8B [18]	–	CC558K [38] + 3DI-Pb-IN166K	Mix665K [38] + 3DI-Ft1M	**62.7**	**86.3**	**52.9**	**51.0**
3D-informed pretraining for vision encoder in stage 0								
CLIP [46]+DINOv2 [45]	Llama3-8B [18]	3DI-Pb-OI1M (CLIP [25])	CC558K [38] + 3DI-Pb-IN166K	Mix665K [38] + 3DI-Ft1M	60.3	85.0	50.5	47.5
CLIP [46]+DINOv2 [45]	Llama3-8B [18]	3DI-Pb-OI1M (NTP [33])	CC558K [38] + 3DI-Pb-IN166K	Mix665K [38] + 3DI-Ft1M	62.6	85.2	52.9	51.5

Table 2. **Thorough exploration of the design space and roadmap progression.** We systematically examine the 3D-informed design space from the aspects of data, architecture and training. The non-gray rows, listed from top to bottom, illustrate the step-by-step progression of our design roadmap. Gray rows correspond to ablations across the full design space. Our final model is highlighted with a light orange background. Our final model's superior performance over other variants confirms the necessity of a 3D-informed compound design.

model. After a thorough empirical exploration of the design space, we can devise our roadmap progressed towards the best-performing model. The non-gray rows in Table. 2 align with the roadmap illustrated in Fig. 6, highlighting our systematic approach to upgrade the architecture and incorporate 3D-informed data into the training process. Overall, the improved architecture achieves a 1.3% performance gain, while 3D-informed alignment and instruction tuning further enhance performance by 13.7%.

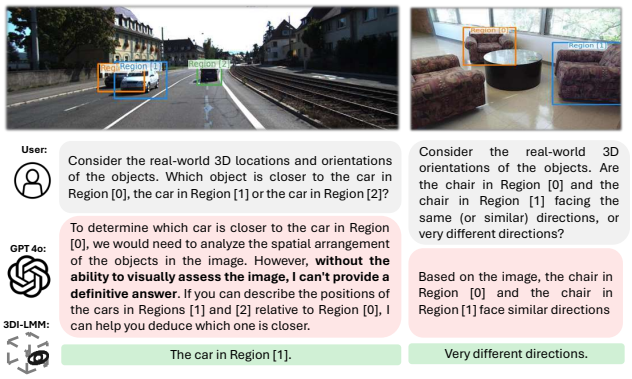


Figure 7. Our model is capable of answering question correctly that needs accurate reasoning on **spatial distance** and **3D orientation**, while GPT-4o either needs more spatial information or gives wrong answer.

Qualitative comparison. Fig. 7 illustrates a scenario involving spatial reasoning. GPT-4o struggles to reason about the spatial relationships between objects due to its limited ability to comprehend their 3D orientation. In contrast, our model successfully understands these spatial relationships and provides accurate answers, demonstrating superior 3D spatial reasoning capability.

5. Conclusions

In this work, we systematically enhance the 3D spatial reasoning capabilities of LMMs through a compound design approach. By introducing 3DI-LMM, we demonstrated that integrating 3D-informed data, architectural innovations, and tailored training setups significantly improves an LMM's ability to understand and reason about complex 3D spatial relationships. To tackle data scarcity, we develop 3D-informed training datasets focused on 3D locations, orientations and complex spatial relationship. Our systematic integration of these datasets with architectural and training designs provide a roadmap for optimizing LMMs for superior 3D reasoning capabilities. Experimental results on the SpatialVQA benchmark highlight 3DI-LMM's superiority over state-of-the-art models, setting a new standard for 3D spatial reasoning. We invite the community to build upon our design roadmap and findings to push the boundaries of what LMMs can achieve in understanding the complex 3D spatial relationships of the real world.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2, 3, 7

[2] Remyx AI. SpaceLLaVA. https://huggingface.co/remyxai/SpaceLLaVA. 1, 7

[3] 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. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[4] Anthropic. Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. 1, 7

[5] Apollo Team. Apollo syntheic dataset, 2019. 1

[6] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 4, 1

[7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2, 3

[8] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2

[9] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 4, 1

[10] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1, 2

[11] 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*, Vancouver, Canada, 2023. IEEE. 2, 4, 1

[12] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 4, 1

[13] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827, 2024. 2, 3

[14] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 1, 2, 4, 5, 7

[15] Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Vitamin: Designing scalable vision models in the vision-language era. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[16] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*, 2024. 1, 2, 4, 5

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 3

[18] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 6, 8

[19] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 4, 5, 7, 1

[20] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 3

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 4, 1

[22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3, 5, 8

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 7

[25] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Han-

naneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 7, 8

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2

[27] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 2

[28] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17307–17316, 2023. 1

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2

[31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 3

[32] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 5, 1

[33] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 8

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2, 3

[35] 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 7

[36] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024. 3

[37] Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15148–15158, 2023. 7

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 2, 3, 7, 8

[39] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 7

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2, 3, 4, 5, 7

[41] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[42] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 3

[43] Wufei Ma, Guanning Zeng, Guofeng Zhang, Qihao Liu, Letian Zhang, Adam Kortylewski, Yaoyao Liu, and Alan Yuille. Imagenet3d: Towards general-purpose object-level 3d understanding. In *NeurIPS*, 2024. 2, 4, 5, 7, 1

[44] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 3

[45] 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. *TMLR*, 2024. 2, 3, 5, 6, 8

[46] 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. 2, 3, 5, 6, 8

[47] Navid Rajabi and Jana Kosecka. Gsr-bench: A benchmark for grounded spatial reasoning evaluation via multimodal llms. *arXiv preprint arXiv:2406.13246*, 2024. 2

[48] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021. 4, 1

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image

synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[50] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[51] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2, 3

[52] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 4, 1

[53] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007. 1

[54] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 3

[55] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2

[56] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1, 3, 4, 7

[57] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 3

[58] Xingrui Wang, Wufei Ma, Zhuowan Li, Adam Kortylewski, and Alan L Yuille. 3d-aware visual question answering about parts, poses and occlusions. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[59] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 1

[60] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1, 2

[61] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. *arXiv preprint arXiv:2112.08359*, 2021. 4, 1

[62] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3

[63] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 1

[64] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 1

[65] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 6, 8

[66] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 3

[67] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. In *NeurIPS*, 2023. 1