GRAID: ENHANCING SPATIAL REASONING OF VLMS THROUGH HIGH-FIDELITY DATA GENERATION

Karim Elmaaroufi[1,2], **Liheng Lai**[1], **Justin Svegliato**[1], **Yutong Bai**[1],
Sanjit A. Seshia[1], **Matei Zaharia**[1]
[1]University of California, Berkeley [2]Embodied Science

ABSTRACT

Vision Language Models (VLMs) achieve strong performance on many vision-language tasks but often struggle with spatial reasoning—a prerequisite for applications such as medical imaging and robotics. We present GRAID, a data generation pipeline that generates high-fidelity spatial reasoning data from images through qualitative analysis of 2D geometry by using object detectors. By avoiding the use of single-image 3D reconstruction pipelines and generative hallucinations, GRAID produces datasets with higher accuracy, as confirmed by our human study. Crucially, we demonstrate that training on GRAID generated QA pairs leads to learning transferable concepts and improved reasoning across general visual reasoning problems. We fine-tune several VLM families on GRAID data and compare against models tuned on data from current methods. We find that GRAID-tuned models result in significant accuracy gains in both spatial reasoning and general visual reasoning benchmarks such as BLINK, A-OKVQA, and RealWorldQA. GRAID is publicly available at our website.

1 INTRODUCTION

Vision Language Models (VLMs) have already shown promise in a wide variety of applications, such as medical diagnosis (Jin et al., 2024), biology (Maruf et al., 2025), and engineering design (Picard et al., 2025). However, a key failure mode of VLMs is that they are poor spatial reasoners, that is, they struggle to understand how objects are located in space and the spatial relationships between them. For example, in medical image analysis, Jin et al. (2024) found that VLMs were unable to recognize that skin lesions shown at different angles were the same pathology. Similarly, in robotics, Wang et al. (2025) found that without integrating explicit spatial relationships, VLMs were unable to produce high-level, executable robotic task plans. As a result, without spatial reasoning, VLMs cannot be reliably deployed in embodied domains such as robotics or non-embodied domains such as medical image analysis.

While VLMs have been trained on internet-scale data, Deitke et al. (2024) found that commonly used datasets, such as COCO (Chen et al., 2015) and Localized Narratives (Pont-Tuset et al., 2020), on average only contain 11 and 37 words per description of an image, despite averaging 7.7 objects Lin et al. (2015) and 10.8 nouns per image (Pont-Tuset et al., 2020). In other words, the captions often contain the name of an object, but may not even contain a description, let alone a rich one describing the object's properties, placement, and relationships with one another.

In response, there have been several recent approaches focused on generating datasets to improve the spatial reasoning of VLMs. Chen et al. (2024a) proposed SpatialVLM to generate 2 billion visual question–answer (VQA) pairs in metric space, yet our human evaluation reveals that only 57.6% of questions are valid (Section 4), with errors stemming from compounded uncertainties in depth estimation, camera calibration, and scene geometry. Cheng et al. (2025) introduced SpatialRGPT, which similarly requires 3D representations but also architectural changes to the VLM. In addition, their region-based architecture requires region-based prompting, which eliminates localization as a learned skill. SpaRE (Ogezi & Shi, 2025) generates question–answer pairs using Large Language Models (LLMs) from hyper-detailed captions but is limited in scalability since it requires extensive human effort to create the captions and inherits hallucinations from the generative models.

We introduce GRAID (**G**enerating **R**easoning questions from **A**nalysis of **I**mages via **D**iscriminative Artificial Intelligence), built on the key insight that *qualitative* spatial relationships can be reliably

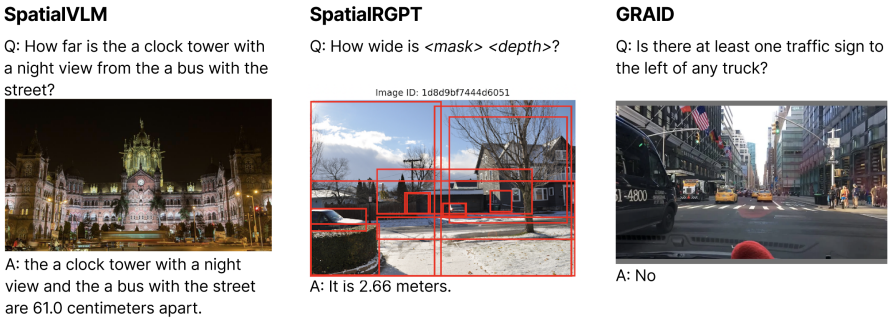


Figure 1: Example VQA pairs from the community implementation of SpatialVLM and SpatialRGPT, showing typical errors and issues in current synthetic data generation methods, and an example from GRAID.

determined through 2D geometric analysis of bounding boxes, avoiding the metric errors and generative hallucinations commonly found in existing methods. GRAID requires only images and object detection outputs—no architectural changes, no hyper-detailed captions, and no 3D reconstruction. Table 1 offers a comparison of the differences between GRAID and prior methods.

Unlike SpatialVLM's 57.6% human validation rate, GRAID has over 91.16% (Section 4). Consistent with recent benchmark findings (Ogezi & Shi, 2025), our human study implicates low-fidelity training data as the cause of a model underperforming its size class. We further validate these findings in Section 5, by fine-tuning several VLMs on the various datasets and show that VLMs perform better when tuned on GRAID data versus when they are tuned on data produced by current methods.

We demonstrate GRAID at scale by applying it to images from Berkeley Deep Drive 100k (BDD) (Yu et al., 2020), NuImages (Caesar et al., 2019), and Waymo Open Perception (Ettinger et al., 2021). We implement 22 VQA templates spanning spatial relations, counting, ranking/extrema, localization, and size/aspect, thus generating over 8.5M pairs. *GRAID is domain-agnostic; we instantiate on driving datasets because their annotations contain fewer errors than datasets such as COCO Schubert et al. (2024), not due to any AV-specific assumption in the method. Additionally, our 22 exemplar templates are merely to demonstrate GRAID's effectiveness as a framework; they are by no means the only VQA templates possible.*

In addition to our human review of generated question and answer accuracy, we conduct a series of quantitative experiments to demonstrate the effectiveness of GRAID's datasets. Our experiments demonstrate that training on GRAID data leads to improved VQA performance over training on datasets generated by current methods (RQ3). Additional experiments show cross-GRAID dataset generalization (RQ1) and learning simple spatial concepts that combine and lead to enhanced performance on more complex problems (RQ2). We fine-tune and benchmark several VLM models across a variety of tasks in existing VQA benchmarks (A-OKVQA (Schwenk et al., 2022), RealWorldQA (xAi, 2024), BLINK (Fu et al., 2024), NaturalBench (Li et al., 2024a), and VSR (Liu et al., 2023)) that challenge VLMs in both indoor and outdoor scenes far beyond the driving scenes from our exemplar source datasets. Overall, despite training on only 22 exemplar templates based on 2D data, GRAID-tuned models consistently outperform their counterparts tuned on data from existing methods, with significant gains in performance across multiple spatial reasoning and general visual reasoning benchmarks.

In summary, this paper makes the following contributions:

1. **GRAID:** a framework that uses 2D geometry to generate qualitative spatial VQA data, avoiding errors from single-view 3D reconstruction and hallucinations from generative models.
2. **High fidelity dataset:** over 8.5M VQA pairs generated. Fine-tuning experiments across a variety of model families and a human study confirm the datasets' 91.16% accuracy; significantly higher than datasets generated by current methods (see Sec. 4).
3. **Evaluation of generalization:** fine-tuning on GRAID data improves VLM performance on held-out question types and on non-template tasks as well as external benchmarks, outperforming models fine-tuned on existing datasets and, demonstrate knowledge transfer far beyond our question templates (see Sec. 5).

Table 1: Comparison of spatial reasoning data generation frameworks

Feature	GRAID	SpatialVLM	SpatialRGPT	SpaRE
Can operate on images only	✓	✓	✓	✗
No VLM architecture changes needed	✓	✓	✗	✓
No lengthy captions required	✓	✓	✓	✗
Avoids single-view 3D reconstruction	✓	✗	✗	✓
Avoids LLM-based QA gen.	✓	✓	✓	✗
Open-source implementation by authors	✓	✗	✓	✗

2 Related work

Whether analyzing MRI anatomical scans or planning robotic navigation, spatial reasoning is a prerequisite for embodied and non-embodied VLM deployment. Recent investigations across medical imaging (Jin et al., 2024), robotics (Wang et al., 2025), and autonomous vehicles (Jiang et al., 2025) reveal a consistent pattern: VLMs leave much to be desired in spatial understanding. To better understand these failures, recent works have investigated if VLMs can understand concepts such as physical domain understanding (Li et al., 2023), geometric understanding (Kosoy et al., 2025), and object states (Newman et al., 2024). These real-world concepts have also inspired many benchmarks like solving problems in the blink of an eye (Fu et al., 2024), naturally adversarial examples (Li et al., 2025b), physical world understanding for embodied agents (Chow et al., 2025), complex multi-step spatial concepts (Zhang et al., 2025b), spatial reasoning through occlusions Pothiraj et al. (2025), and even games (Tang et al., 2025a; Lyu et al., 2025). Yin et al. (2025) Find that VLMs exhibit near random performance in spatial reasoning even when presented with multiple view of the same scene. The common finding is that VLMs require significant improvement in spatial understanding and learning the laws that govern the physical world, i.e., physical reasoning Balazadeh et al. (2025); Puyin et al. (2025); Sreekumar & Boddeti (2025); Hu et al. (2025); Sun et al. (2025).

3D Reconstruction Hong et al. (2023) were among the first to teach spatial reasoning to VLMs by performing 3D scene reconstruction from multiple views then using a 3D feature extractor to connect to an LLM. While such methods worked, they required lots of data *per* scene. They did not specify how many images per scene were necessary but popular methods at the time such as Nerfstudio (Tancik et al., 2023) would have required tens to a few hundred images from *known* camera poses per scene. Later works avoided the requirement of many images by instead constructing implicit scene graphs: predicting depth from RGB images and using instance segmentation models to refine masks of detected objects, to lift 2D images to 3D point clouds and finally performing semantic grouping. However, these approaches come at the cost of compounding errors. Gu et al. (2024) introduces ConceptGraphs but are admittedly prone to LLM and VLM hallucinations in addition to missing small and thin objects which, "impacts downstream planning". Rather than trying to create spatial reasoning data, Zhang et al. (2024) propose Agent3D-Zero, a framework that allows VLMs to query for multiple viewing positions and angles before attempting to answer a question. The caveat is that the full scene must already have a 3D representation. Chen et al. (2024a) introduce SpatialVLM and propose a wide acceptance metric of [50%, 200%] to account for inaccuracies of their quantitative (metric-based) questions. Ma et al. (2025) also proposes a full 3D reconstruction pipelines but faces the same issue of compounding errors. Cheng et al. (2025) avoids many of these issues by generating their dataset from labeled 3D data, however, they propose a region-based VLM, which requires architectural changes and eliminates localization as a core competency of the VLM, i.e., the user must select the object of interest rather than describe it and let the VLM find it. Similarly, Ma et al. (2025); Yang et al. (2025) also propose further architectural changes to enhance the 3D awareness of VLMs.

Leveraging existing data is another popular approach in which recent works proposed enhancing spatial reasoning by explicitly training VLMs on bounding boxes (Wang et al., 2023; Yang et al., 2023b; Peng et al., 2023; Rasheed et al., 2024; Zhang et al., 2025a). Additionally, some methods have trained on point data (You et al., 2023; Deitke et al., 2024) thus becoming less dependent on bounding boxes, which may encompass with unwanted objects in object-dense scenes. However, many of these approaches leverage COCO related datasets and as Deitke et al. (2024) discovered, the sparsity of words in such source datasets is too little to contain spatial reasoning data. This led to their key insight that significantly longer human annotations are required to explicitly express spatial relationships.

3 GRAID

GRAID (**G**enerating **R**easoning questions from **A**nalysis of **I**mages via **D**iscriminative Artificial Intelligence) is an extensible framework that generates large-scale Visual-Question-Answering (VQA) datasets. The datasets are of higher quality than existing tools that produce similar datasets because GRAID produces valid questions and correct answers far more frequently than existing methodologies, as validated by human evaluations. GRAID does this by way of two components: Scene Understanding and SPARQ (Sieve Predicates And Realize Questions). We discuss each in turn.

3.1 SCENE UNDERSTANDING

GRAID's key insight into reducing hallucinations in both questions and answers, is to avoid performing single-image-view 3D reconstruction —the key feature in many existing works. Instead, GRAID does nearly all of its analysis in the 2D space. In particular, GRAID merely assumes the usage of object detection models, which provide class names and bounding boxes of objects in an image. Modern object detection models have achieved sufficiently high accuracy on prior global challenges such as ImageNet, and are robust enough for practical deployment, with both governments and private entities deploying popular single-stage detectors like YOLO for diverse real-world applications. Furthermore, there exist several widely accepted interpretability methods such as Saliency Maps (Li & Wong, 2024; Simonyan et al., 2014), Grad-CAM (Selvaraju et al., 2019), Grad-CAM++ (Chattopadhay et al., 2018), Score-CAM (Wang et al., 2020), SuperPixels (Hartley et al., 2021) and many more. This level of widespread deployment and tools for analysis, has yet to be achieved in the other components required to single-image-view 3D reconstruction which include but not are not limited to models such as depth perception, pose estimation, and plane estimation.

The problem of object detection can be formally described as follows: given an input image $I \in \mathbb{R}^{H \times W \times C}$ where H, W, and C denote the height, width, and number of channels, object detection models predict a set of up to N bounding boxes $\mathcal{B} = \{b_i\}_{i=1}^{N}$ and their corresponding class labels $\mathcal{Y} = \{y_i\}_{i=1}^{N}$. GRAID supports several representations of bounding boxes but for convenience, we will refer to one where each bounding box, $b_i = (x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}})$, with $(x_{\text{min}}, y_{\text{min}})$ and $(x_{\text{max}}, y_{\text{max}})$ correspond to the top-left and bottom-right corners of the bounding box. Within each box, the model must also assign a class label $y_i \in \{1, \ldots, C\}$ where C is the total number of class labels or object categories. This is typically formulated as a probability distribution over the label space, $p(y_i|I) = \text{softmax}(z_i)$ where $z_i \in \mathbb{R}^C$ are the raw logits from the discriminative model for class scores. Observe that C is a parameter of the underlying object detection datasets and models and can easily be changed by swapping models. For example, models trained on the COCO dataset (Lin et al., 2015) have $C = 80$, whereas models trained to compete ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015) have $C = 1000$.

Rather than designing a general-purpose object detector or assuming a single foundational model, we build GRAID to support three of the most widely used computer vision packages: Detectron2, MMDetection, and Ultralytics. We define a standard interface thus allowing users to either bring in labeled data or use their own prior trained object detection models. Note that segmentation models can also be used, as they often share the same backbone as an object detection model.

3.2 SIEVE PREDICATES AND REALIZE QUESTIONS (SPARQ)

Given an image and a list of detection objects, we can now construct questions and answers based on the relationships of those bounding boxes. However, for an image with many detected objects, checking spatial relationships between objects quickly becomes expensive as this is a quadratic operation that can require comparing every object to every other object. Thus to scalably generate millions of questions in under a few hours, we design SPARQ (Sieve Predicates And Realize Questions).

Predicates are designed to be lightweight sanity checks before performing the full realization of a question which are more computationally expensive. For example, in the base question, `RightOf`, implemented as, *"Is there at least one {object_1} to the right of any {object_2}?"*, we can immediately check to see if there are at least two different object classes before checking spatial relationships. We can also check to see if there exists at least one pair of objects whose classes are different and whose bounding boxes do not intersect (i.e., their boxes' $IoU = 0$). While these two checks are simple, their savings are significant. When generating the graid-bdd100k dataset, we find that these two predicates

Algorithm 1: RIGHT-OF Question Realizer

Input: Image I of width W and height H; detections $\mathcal{D}$, each comprising a class label and bounding box
Output: A list of (*question*, *answer*) pairs (possibly empty)

1. Precondition validation and grouping

Construct a map $\mathcal{C}$: label $\mapsto \{b = (x_{\min}, y_{\min}, x_{\max}, y_{\max})\}$ grouping detections by class.
If $|\text{keys}(\mathcal{C})| < 2$, **return** $\emptyset$. ; `// Fewer than two distinct classes`
If no pair of detections in $\mathcal{D}$ has $\text{IoU} = 0$, **return** $\emptyset$.; `// No non-overlapping pairs exist`

2. Evaluate ordered class pairs

Initialize QA $\leftarrow [\,]$.
foreach *ordered pair of distinct classes* $(c_1, c_2) \in \text{keys}(\mathcal{C})^2,\ c_1 \neq c_2$ **do**
  found $\leftarrow$ `False`;
  foreach $b_1 \in \mathcal{C}[c_1]$ **do**
    foreach $b_2 \in \mathcal{C}[c_2]$ **do**
      if $x_{\min}^{(1)} > x_{\max}^{(2)}$ **and** $\text{IoU}(b_1, b_2) = 0$ **then**
        Append $\big(Q(c_1, c_2),\ $ `"Yes"`$\big)$ to QA;
        found $\leftarrow$ `True`;
        break ; `// Exit both inner loops`
    if found **then break**;
  if $\neg$ found **then**
    Append $\big(Q(c_1, c_2),\ $ `"No"`$\big)$ to QA;

3. Return QA

complete on average in 5.17ms, while realizing the question takes 46.95ms—nine times slower. In other questions such as `LargestAppearance`, which uses just the former predicate, the savings are more pronounced: over 1407 times. Furthermore, we find that predicates not only saving time, but they often result in sufficient conditions for the questions to be realized. In `LargestAppearance`, the predicate completes on average in 0.02ms, and 78.8% of the time results in a question being realized. In the appendix, we provide a table of GRAID-BDD dataset that reports average predicate timing, realization time, and the share of cases where predicate success implied realization success. For other datasets, we refer the reader to each dataset's respective README after the review period.

Realize Questions. Once all predicates for a base question have succeeded, we *apply* the question—that is, we attempt to realize a question-answer pair for the image and its detected objects. One algorithm to solve the previously mentioned, `RightOf` question is to first find the left-most instance of every class of object in the image. Next, for each object found, we iterate over the remaining classes of objects in the image and check for the following: 1) the bounding boxes of each potential pair should be non-overlapping, and 2) they should lie on similar planes. Observe that the second condition is necessary in the process of realizing a question as we could encounter a case where we find out that the question could be ambiguous (e.g. is an object truly the right of another if they are also on different heights?). In such instances, the questions `apply` method returns an empty list. Otherwise, when we locate a potential pair, we save it as a candidate pair until we have completed all objects in the image. The full algorithm of the `RightOf` question, is provided in Algorithm 1.

As evidence of GRAID's effectiveness, we implement 22 base questions and apply them to 3 source datasets to generate over 8.5M VQA pairs. We discuss the resulting data in the next section and refer the reader to Appendix B for further details of these base questions, including the class name, a brief description of its predicates, and a one-line explanation of the corresponding realization algorithm.

4 GRAID DATASETS

The autonomous vehicle (AV) domain provides an ideal testbed for evaluating GRAID due to its exceptional wealth of high-quality, comprehensively labeled datasets that naturally capture diverse real-world scenarios. We select three prominent AV datasets —Berkeley Deep Drive (BDD) 100k, NuImages, and Waymo Open Perception—that collectively offer extensive ground truth annotations across varied driving conditions, geographical locations, and environmental factors. Additionally, the ground truth annotations in the AV space have been shown to have less human labeling errors

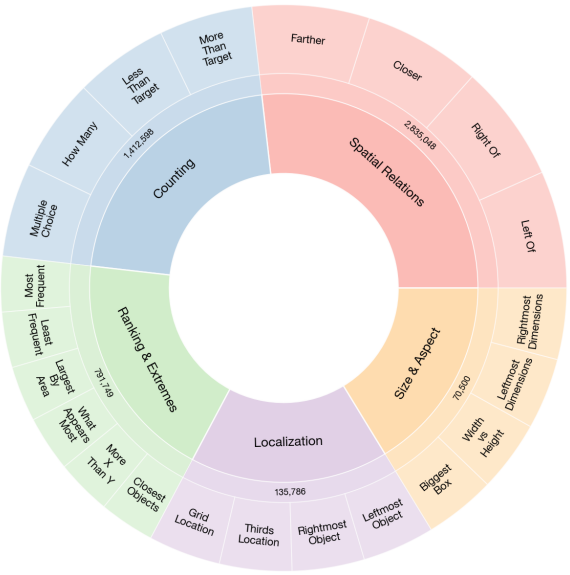


Figure 2: Hierarchical breakdown of 5.3M visual questions generated by GRAID using Berkeley Deep Drive as the source images. There are five cognitive categories: Spatial Relations (53.5%), Counting (26.7%), Ranking & Extremes (14.9%), Localization (2.6%), and Size & Aspect (1.3%). Question details including generation time, and `predicate` and `apply` methods, can be found in the Appendix.

Schubert et al. (2024) than more general datasets such as COCO. In the subsequent sections, we select to directly leverage these high-quality labels in GRAID's generation rather than train our own object detectors so that we can evaluate GRAID's effectiveness in isolation.

In total, we release six variants of GRAID generated datasets from the source datasets (see Table 2). Using BDD, we generate two variants: one without depth related questions yielding 18 classes of questions, and one with depth questions yielding 22 classes of questions. These depth questions are selected as a demonstration of GRAID's extensibility as a framework. In prior works such as SpatialVLM and SpatialRGPT, depth models are used to ask quantitative metric-based questions. Due to the inaccuracy of these models, the former proposed accepting answers that were within 50% and 200% of the estimated depth. Our human evaluators found that in 250 questions generated by the open source implementation of SpatialVLM, over half had incorrect ground-truth answers. This is one of the main motivations for why GRAID asks qualitative rather than quantitative questions, i.e., rather than asking how far an object is in terms of metric distance, it's easier to answer which object is closer, hence the **D**iscriminative in GRAID. To further account for inaccuracies in depth models, our depth questions, like most of our questions, are configurable with thresholds than can be set based on a models' confidence, a users' intuition, or domain expertise. For example, in `Closer`, we define `margin_ratio` as the configurable parameter, where the question will only be realized if the ratio of the predicted distances between the objects is $\geq$ `margin_ratio`. This eliminates questions that appear in existing datasets which can be deemed ambiguous.

Similarly, we release two variants using NuImages as the source images and Waymo Open Perception. However, in Waymo, rather than using the original images, we utilize a small subset. In the Waymo Open Perception dataset, there are a few hundred unique scenes. These scenes are actually videos across six cameras on a single vehicle and so many images are repeated with just a handful of objects changing location. Thus, in our Waymo variants we select one image from the front camera with a score that balances: (i) the number of detected objects and (ii) the ratio of the largest object area to the image area. We find this metric offers a good balance of generating more questions per image without sacrificing accuracy. Table 2 summarizes the various GRAID generated datasets.

Human Evaluation of Dataset Quality

To better compare VQA datasets, we perform several kinds of human evaluations. First, we examine Huggingface to identify the most popular VQA datasets which involve spatial reasoning. At the time of submission, under the VQA dataset category, three (Li et al. (2025a); Chen et al. (2024b); Li et al. (2024b)) of the top 30 datasets ranked by downloads explicitly test for spatial reasoning. However, all three are strictly datasets and not frameworks that are capable of generating additional data. In addition, all three utilize LLMs or VLMs in their dataset curation, leading to the question: if a VLM could already see something, is it that hard to test? A few of the remaining test for algebraic

Table 2: GRAID Generated Datasets Overview

Source Dataset	Question Types	# QA Pairs	Train QA	Val QA	# Train/Val Images
BDD100k	With Depth	**5.30M**	4.63M	672k	69.9k / 9.9k
	Without Depth	**3.82M**	3.34M	485k	
NuImages	With Depth	**3.29M**	2.65M	641k	60.7k / 14.9k
	Without Depth	**2.41M**	1.94M	478k	
Waymo	With Depth	**16.4k**	13.1k	3.33k	798/202
	Without Depth	**13.8k**	10.9k	2.79k	

reasoning from images via tests like geometric challenges (e.g., read the sides of a triangle and use Pythagorean's theorem to solve for the missing side), however, the vast majority test for document and chart understanding, or image captioning.

In the realm of VQA generation frameworks that explicitly test for spatial reasoning from just images, we find two candidates: SpatialRGPT and SpatialVLM. There are also works such as SpaRE (Ogezi & Shi, 2025) which generate VQA questions given image and caption pairs. However observe that in Deitke et al. (2024), the authors identify that human annotations are required for better image-caption pairs, as the average word count in captions for common pairs such as COCO is merely 11 words. With such little details, methods like SpaRE leave room for LLMs to hallucinate details of an object and scene.

Our human evaluators thus evaluated the OpenSpatialDataset, the only dataset produced by SpatialRGPT, and OpenSpaces one of the more popularly used datasets generated by the community implementation of SpatialVLM. VQA examples of each dataset are shown in Figure 1. Due to the masked region queries, our evaluators were unable to ascertain the quality of the examples. In some instances, it was possible to determine if the question and answer were correct as there were only one or two regions, however, in many others, there tens of regions which often lacked semantic meaning and so identifying the subject was not possible unless a region-based prompting technique such as Set-of-Mark (Yang et al., 2023a) was used. Our evaluators were able to evaluate 50 images with 5 questions per image in OpenSpaces. An example is shown in Figure 1. The evaluators noted that most questions were not grammatically correct. Despite their best attempts to understand the question, they found $\frac{104}{250} = 41.6\%$ were not valid questions, and $\frac{144}{250} = 57.6\%$ of answers in the dataset were incorrect. Of the questions that were valid, 25.2% of them had hallucinated answers. Our human evaluations corroborate recent findings from Ogezi & Shi (2025), who show that SpaceLLaVA, on average, performs the worst compared to other similarly-sized models on spatial reasoning benchmarks. Our results suggest that the poor quality of the data generated by the community implementation of SpatialVLM, which was used to train SpaceLLaVA, is a primary contributor to this performance gap. In Section 5, we further investigate this by benchmarking VLMs after fine-tuning on OpenSpaces and GRAID and share the full results in Table 6. While we find that GRAID-tuned models significantly outperform SpatialVLM tuned models, we also find that in some instances, SpatialVLM tuned models perform worse than the baseline model indicating a degradation of visual reasoning in the VLM due to the high hallucination rate in the SpatialVLM data.

Finally, we ask four humans to evaluate 317 VQA pairs from the GRAID-BDD dataset without depth questions. Each person is asked for their name, which is used to compute a seed for randomly sampling the VQA pairs. As with the two previous datasets, we asked our evaluators to determine if (i) a question was valid, and (ii) if the answer to the question is correct. Given that we are interested in the correctness of the question, we offer each person the option to view the image with and without bounding boxes. Without the boxes, they attempt to judge the difficulty of the questions on a Likert scale of 1 to 5. With the boxes, they can determine if the answer in the dataset is indeed correct, and if there are any labeling errors which led to a false answer. In total, our evaluators found 7 questions to be unclear, 2 questions to be invalid, and 5 labeling errors in the BDD dataset labels, i.e., over 95.58% of GRAID generated questions were valid. In terms of answers, 12 were found to be unclear and 8 were found to be invalid, hence over 93.69% of answers were valid. When we examine the unique instances (i.e., do not double count the VQA pairs with both question and answer concerns), we find that there are 28 unique instances and so in total less than 9% of the VQA pairs they evaluated were found to be either invalid or confusing. Using their feedback, we were able to address some of the ambiguities. The current public datasets have these corrections and thus, have even higher

validity. Lastly, our evaluators gave an average difficulty rating of 2.968, with a standard deviation of 1.146. 109 questions were marked as a 2 or less, while 95 were marked as a 4 or higher. These results confirm that GRAID generated datasets are of the highest accuracy VQA datasets made by automated generation pipelines, and that the questions generated are of a wide variety of difficulty levels, i.e., the data avoids being too easy or hard.

5 Vision Language Model Experiments

We conduct a series of fine-tuning experiments to determine how well a VLM can learn spatial reasoning concepts from our data. For all experiments we fine tune a VLM using LoRA Hu et al. (2021) with a learning rate of 2^{-4}, AdamW8bit optimizer, and a linear learning rate scheduler. We ask the following research questions:

RQ1: Does fine-tuning on spatial reasoning tasks enable cross-dataset generalization, demonstrating acquisition of transferable spatial concepts rather than dataset-specific overfitting?

RQ2: Can training on fundamental spatial reasoning primitives improve performance on more complex spatial reasoning tasks not seen during training?

RQ3: Does training on GRAID generated datasets improve performance on established benchmarks in general visual reasoning? How do GRAID-tuned models compare against other models tuned on other semi-synthetically generated training data?

RQ1 We perform supervised fine-tuning (SFT) on a limited sample of GRAID-BDD. We randomly select 10% from the training split without stratified sampling by question type. Using LoRA with rank of 16 and 200 training steps, we evaluate a Meta Llama-3.2-Vision-Instruct-11B model on two distinct test scenarios: (1) 1,000 held-out unstratified examples (Figure 2 provides the full distribution of questions of GRAID-BDD) from the same dataset (GRAID-BDD), and (2) 1,000 unstratified examples from an entirely different dataset (GRAID-NuImages). On the first, model performance improved dramatically from 31% to 80.7% (+49.7%), already demonstrating improved spatial reasoning capabilities. In the second, the model achieved substantial gains from 38% to 67.1% (+29.1%) on the completely unseen GRAID-NuImages dataset—which contains entirely different cities, scenes, objects, and visual contexts. These cross-dataset results strongly indicate that the model acquired transferable spatial reasoning representations rather than merely memorizing dataset and image-specific patterns.

RQ2 To evaluate whether a model is truly learning spatial concepts, we select six questions to serve as our training set for supervised fine-tuning of a Meta Llama-3.2-Vision-Instruct-11B VLM: `LeftOf`, `RightOf`, `HowMany`, `AreMore`, `LargestAppearance`, and `IsObjectCentered` (full definitions are provided in Appendix B. We use a LoRA with rank 32, batch size 2 with 4 gradient accumulation steps, 5 warmup steps, AdamW8bit optimizer with a linear schedule, weight decay of 0.01, and train for 200 steps. Observe that these six questions yield over 18,000 VQA pairs using just GRAID-BDD (much less than half of the total training examples available), but our SFT process completes only a fraction of an epoch (by design). We evaluate the model on all question types in GRAID-BDD, and GRAID-NuImages, with the latter dataset never seen in training. The results are shown in Figure 3. In nearly all questions and across both datasets, we observe wide performance increases despite only seeing six kinds of questions from only one of the datasets. These results are in agreement with findings by Tang et al. (2025b) who find that learning basic spatial concepts in simple simulated settings, leads to spatial reasoning in real world images. In both datasets, we observe a regression in `LessThanThresholdHowMany` and in GRAID-BDD, a slight regression in the same question's counterpart, `MoreThanThresholdHowMany`. Being that these two questions are some of the most common, we suspect that this is a symptom of overfitting.

RQ3 To evaluate whether GRAID can produce datasets that transfer to real-world spatial reasoning challenges of both indoor and outdoor scenes that extend far beyond driving scenes, we SFT four instruction tuned VLMs, Meta Llama 3.2 11B (Grattafiori et al., 2024), Gemma 3 4B (Team et al., 2025), Qwen2.5 VL 3B (Bai et al., 2025b), and Qwen3 VL 8B (Bai et al., 2025a), on GRAID-BDD (full training details are provided in Appendix D). For comparison purposes, we also perform the same SFT experiment using OpenSpaces, a dataset generated by the community implementation of SpatialVLM Chen et al. (2024a). We evaluate all SFT variants of all models on five established VQA benchmarks which contain a variety of indoor and outdoor scenes with varying spatial reasoning complexity: BLINK (Fu et al., 2024), NaturalBench (Li et al., 2024a), A-OKVQA (Schwenk et al., 2022), RealWorldQA (xAi, 2024), and VSR (Liu et al., 2023). Rather than using GRAID's built in

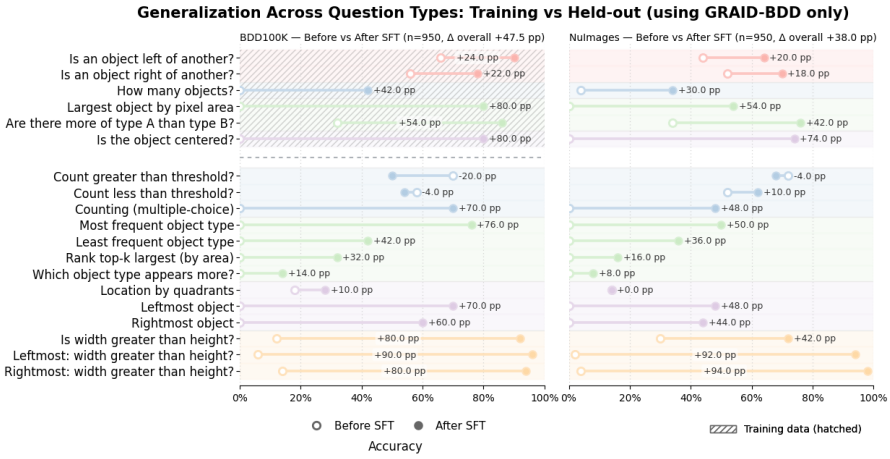


Figure 3: We fine-tune Llama 3.2 11B on only 6 questions from GRAID-BDD (hatched upper-left corner). Evaluations demonstrate a greater understanding across more difficult spatial reasoning questions in the GRAID-BDD validation set, generalization to a fifth topic not seen in training (Size & Aspect), and in all 19 question types never seen from GRAID-NuImages.

VLM evaluator which supports multiple prompting (zero, few-shot, etc.) and decoding techniques (constrained, greedy, etc.), we instead follow others (Ogezi & Shi, 2025) and use VLMEvalKit (Duan et al., 2024) with exact match grading, to ensure consistency with reported baselines.

The results in Appendix Table 4 (Llama), 5 (Gemma), and 6 (all four VLM models) provide further evidence that data from GRAID is of high quality as it enables substantial performance gains on VQA benchmarks across various tasks. For example, with the Llama model, we observe a significant 32.5% improvement on A-OKVQA and 15.94% overall improvement on BLINK, with particularly impressive gains on core spatial reasoning tasks: +41.13% on Relative Depth, +31.98% on Visual Correspondence, and +30.77% on Spatial Relations. We also see significant gains with the Gemma and older Qwen models, and less pronounced gains in Qwen 3VL. Despite our training data containing images of mostly cars and only 10 of 143 BLINK Spatial Relations questions contain the word ”car”, the results demonstrate that GRAID generated training data captures spatial reasoning concepts that are transferable across domains, objects, and scenes rather than driving-specific nuances. The strong improvements across all benchmarks further support that the spatial reasoning primitives learned are not specific to driving scenes and indeed apply to all kinds of scenes both indoor and outdoor. Moreover, across all four backbones, models fine-tuned on GRAID data consistently outperform those fine-tuned on the SpatialVLM dataset (OpenSpaces) and, unlike OpenSpaces tuned models, far less frequently incur large regressions on non-spatial tasks. Finally, the absence of overfitting to strictly driving concepts is further validated by stable performance on NaturalBench—a benchmark designed with completely adversarial examples. For further insights into the fine-tuning process, see the additional ablations in Appendix E.

6 Conclusion

In this work, we present GRAID, a framework for generating high-fidelity spatial reasoning VQA data from real images using 2D detector outputs and qualitative geometry. By explicitly avoiding single-view 3D reconstruction and caption-driven synthesis, GRAID reduces cascading modeling errors and generative hallucinations while remaining easy to adopt with any object detector. Instantiated with 22 templates on three large image corpora, GRAID yields one of the largest high-quality spatial VQA datasets to date—more than 8.5M VQA pairs with over 91.16% human-verified validity—significantly higher than prior works. Supervised fine-tuning on GRAID validates that models learn spatial concepts that *transfer* beyond our templates and datasets, with consistent gains on public evaluations. As the community improves other kinds of models such as segmentation, gaze target and pose estimation, GRAID is already prepared to support those kinds of models with future templates following the SPARQ predicate and question template library. By open sourcing GRAID, we hope to further accelerate improvements in spatial reasoning so that higher-level concepts such as spatio-physical reasoning (Han et al., 2025) could be better researched.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G. Krishnan. Physics context builders: A modular framework for physical reasoning in vision-language models, 2025. URL https://arxiv.org/abs/2412.08619.

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. *arXiv preprint arXiv:1903.11027*, 2019.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2018. doi: 10.1109/wacv.2018.00097. URL http://dx.doi.org/10.1109/WACV.2018.00097.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465. IEEE, June 2024a. doi: 10.1109/cvpr52733.2024.01370. URL http://dx.doi.org/10.1109/CVPR52733.2024.01370.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024b.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas, 2025. URL https://arxiv.org/abs/2503.01773.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. URL https://arxiv.org/abs/1504.00325.

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 models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding, 2025. URL https://arxiv.org/abs/2501.16411.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin

Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles Qi, Yin Zhou, Zoey Yang, Aurelien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving : The waymo open motion dataset, 2021. URL https://arxiv.org/abs/2104.10133.

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, 2024. URL https://arxiv.org/abs/2404.12390.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier

Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel

de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028, 2024. doi: 10.1109/ICRA57147.2024.10610243.

Tiancheng Han, Yunfei Gao, Yong Li, Wuzhou Yu, Qiaosheng Zhang, and Wenqi Shao. From diagnosis to improvement: Probing spatio-physical reasoning in vision language models, 2025. URL https://arxiv.org/abs/2508.10770.

Thomas Hartley, Kirill Sidorov, Christopher Willis, and David Marshall. Swag: Superpixels weighted by average gradients for explanations of cnns. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 423–432, 2021. doi: 10.1109/WACV48630.2021.00047.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models, 2023. URL https://arxiv.org/abs/2307.12981.

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, 2021. URL https://arxiv.org/abs/2106.09685.

Lanxiang Hu, Abhilash Shankarampeta, Yixin Huang, Zilin Dai, Haoyang Yu, Yujie Zhao, Haoqiang Kang, Daniel Zhao, Tajana Rosing, and Hao Zhang. Benchmarking scientific understanding and reasoning for video generation using videoscience-bench, 2025. URL https://arxiv.org/abs/2512.02942.

Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, Hao Ye, Zihao Sheng, Xin Zhao, Tuopu Wen, Zheng Fu, Sikai Chen, Kun Jiang, Diange Yang, Seongjin Choi, and Lijun Sun. A survey on vision-language-action models for autonomous driving, 2025. URL https://arxiv.org/abs/2506.24044.

Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M. Cheung, Robert Chen, Ronald M. Summers, Justin F. Rousseau, Peiyun Ni, Marc J. Landsman, Sally L. Baxter, Subhi J. Al'Aref, Yijia Li, Alexander Chen, Josef A. Brejt, Michael F. Chiang, Yifan Peng, and Zhiyong Lu. Hidden flaws behind expert-level accuracy of multimodal gpt-4 vision in medicine. *npj Digital Medicine*, 7(1), July 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01185-7. URL http://dx.doi.org/10.1038/s41746-024-01185-7.

Eliza Kosoy, Annya Dahmani, Andrew Kyle Lampinen, Iulia Maria Comsa, Soojin Jeong, Ishita Dasgupta, and Kelsey R Allen. Decoupling the components of geometric understanding. In *Workshop on Reasoning and Planning for Large Language Models*, 2025. URL https://openreview.net/forum?id=VSLCgwK5Az.

Ang Li, Charles Wang, Kaiyu Yue, Zikui Cai, Ollie Liu, Deqing Fu, Peng Guo, Wang Bill Zhu, Vatsal Sharan, Robin Jia, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision language reasoning, 2025a. URL https://arxiv.org/abs/2507.16746.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=Dx88A9Zgnv.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples, 2025b. URL https://arxiv.org/abs/2410.14669.

Kaia K.Y. Li and Nicole H. L. Wong. Topographical representation of saliency in the human visual and temporo-occipital cortex. *The Journal of Neuroscience*, 44(19):e0037242024, May 2024. ISSN 1529-2401. doi: 10.1523/jneurosci.0037-24.2024. URL http://dx.doi.org/10.1523/JNEUROSCI.0037-24.2024.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024b. URL https://arxiv.org/abs/2311.17005.

Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Xu Sun, Lingpeng Kong, and Qi Liu. Can language models understand physical concepts? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11843–11861, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.726. URL https://aclanthology.org/2023.emnlp-main.726/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning, 2023. URL https://arxiv.org/abs/2205.00363.

Zesen Lyu, Dandan Zhang, Wei Ye, Fangdi Li, Zhihang Jiang, and Yao Yang. Jigsaw-puzzles: From seeing to understanding to reasoning in vision-language models, 2025. URL https://arxiv.org/abs/2505.20728.

Wufei Ma, Luoxin Ye, Celso de Melo, Alan L Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025.

M. Maruf, Arka Daw, Kazi Sajeed Mehrab, Harish Babu Manogaran, Abhilash Neog, Medha Sawhney, Mridul Khurana, James P. Balhoff, Yasin Bakış, Bahadir Altintas, Matthew J Thompson, Elizabeth G Campolongo, Josef C. Uyeda, Hilmar Lapp, Henry L. Bart, Paula M. Mabee, Yu Su, Wei-Lun Chao, Charles Stewart, Tanya Berger-Wolf, Wasila Dahdul, and Anuj Karpatne. Vlm4bio: a benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

Kaleb Newman, Shijie Wang, Yuan Zang, David Heffren, and Chen Sun. Do pre-trained vision-language models encode object states?, 2024. URL https://arxiv.org/abs/2409.10488.

Michael Ogezi and Freda Shi. Spare: Enhancing spatial reasoning in vision-language models with synthetic data, 2025. URL https://arxiv.org/abs/2504.20648.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. URL https://arxiv.org/abs/2306.14824.

Cyril Picard, Kristen M. Edwards, Anna C. Doris, Brandon Man, Giorgio Giannone, Md Ferdous Alam, and Faez Ahmed. From concept to manufacturing: evaluating vision-language models for engineering design. *Artificial Intelligence Review*, 58(9), July 2025. ISSN 1573-7462. doi: 10.1007/s10462-025-11290-y. URL http://dx.doi.org/10.1007/s10462-025-11290-y.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. *Connecting Vision and Language with Localized Narratives*, pp. 647–664. Springer International Publishing, 2020. ISBN 9783030585587. doi: 10.1007/978-3-030-58558-7_38. URL http://dx.doi.org/10.1007/978-3-030-58558-7_38.

Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. In *ICCV*, 2025.

Li Puyin, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Li Fei-fei, and Ehsan Adeli. Quantiphy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models, 2025. URL https://arxiv.org/abs/2512.19526.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji Mullappilly, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model, 2024. URL https://arxiv.org/abs/2311.03356.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

Marius Schubert, Tobias Riedlinger, Karsten Kahl, Daniel Kröll, Sebastian Schoenen, Siniša Šegvić, and Matthias Rottmann. Identifying label errors in object detection datasets by loss inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4582–4591, January 2024.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pp. 146–162, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20073-1. doi: 10.1007/978-3-031-20074-8_9. URL https://doi.org/10.1007/978-3-031-20074-8_9.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Gautam Sreekumar and Vishnu Naresh Boddeti. Inphyre discovers: Large multimodal models struggle in inductive physical reasoning, 2025. URL https://arxiv.org/abs/2509.12263.

Haoran Sun, Qingying Gao, Haiyun Lyu, Dezhi Luo, Yijiang Li, and Hokin Deng. Probing mechanical reasoning in large vision language models, 2025. URL https://arxiv.org/abs/2410.00318.

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, 2023.

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. Lego-puzzles: How good are mllms at multi-step spatial reasoning?, 2025a. URL https://arxiv.org/abs/2503.19990.

Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to spatial reasoning, 2025b. URL https://arxiv.org/abs/2410.16162.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa

Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2020. doi: 10.1109/cvprw50498.2020.00020. URL http://dx.doi.org/10.1109/CVPRW50498.2020.00020.

Peng Wang, Minh Huy Pham, Zhihao Guo, and Wei Zhou. A spatial relationship aware dataset for robotics, 2025. URL https://arxiv.org/abs/2506.12525.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and Jifeng Dai. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks, 2023. URL https://arxiv.org/abs/2305.11175.

xAi. Grok-1.5 Vision Preview — xAI — x.ai. https://x.ai/news/grok-1.5v, 2024. [Accessed 24-09-2025].

Fan Yang, Sicheng Zhao, Yanhao Zhang, Hui Chen, Haonan Lu, Jungong Han, and Guiguang Ding. Llmi3d: Mllm-based 3d perception from a single 2d image, 2025. URL https://arxiv.org/abs/2408.07422.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, 2023a. URL https://arxiv.org/abs/2310.11441.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action, 2023b. URL https://arxiv.org/abs/2303.11381.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. Spatial mental modeling from limited views, 2025. URL https://arxiv.org/abs/2506.21458.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity, 2023. URL https://arxiv.org/abs/2310.07704.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning, 2020. URL https://arxiv.org/abs/1805.04687.

Sha Zhang, Di Huang, Jiajun Deng, Shixiang Tang, Wanli Ouyang, Tong He, and Yanyong Zhang. Agent3d-zero: An agent for zero-shot 3d understanding, 2024. URL https://arxiv.org/abs/2403.11835.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest, 2025a. URL https://arxiv.org/abs/2307.03601.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation, 2025b. URL https://arxiv.org/abs/2412.12693.

A APPENDIX

B QUESTIONS IMPLEMENTATION

`IsObjectCentered`
Base Question: "Divide the image into thirds. In which third does the {object_1} primarily appear? Respond with the letter only: A) left third, B) middle third, C) right third."
Predicate: Requires at least one object class to appear exactly once.
Apply: For each single-instance class, assigns A/B/C based on the bbox relative to thirds with a buffer; skips ambiguous or spanning cases.

`WidthVsHeight`
Base Question: "Is the width of the {object_1} appear to be larger than the height?"
Predicate: Requires at least one object class to appear exactly once.
Apply: For single-instance (optionally restricted) classes, compares width vs height; skips near-square within a threshold; returns Yes/No (supports an alternate reversed phrasing).

`LeftMost`
Base Question: "What is the leftmost object in the image?"
Predicate: At least one object class appears exactly once.
Apply: Finds the leftmost detection fully on the left half and separated from the second-leftmost by a margin; otherwise skips.

`RightMost`
Base Question: "What is the rightmost object in the image?"
Predicate: At least one object class appears exactly once.
Apply: Finds the rightmost detection fully on the right half and separated from the second-rightmost by a margin; otherwise skips.

`LargestAppearance`
Base Question: "If you were to draw a tight box around each object in the image, which type of object would have the biggest box?"
Predicate: Requires at least two different object classes.
Apply: Compares detection areas; returns the largest class only if its area exceeds the second by a margin.

`RankLargestK(k)`
Base Question: "Rank the {k} kinds of objects that appear the largest (by pixel area) in the image from largest to smallest. Provide your answer as a comma-separated list of object names only."
Predicate: Requires at least {k} different object classes.
Apply: Ranks classes by max single-instance area; asks only if each consecutive pair has a sufficient multiplicative gap.

`MostAppearance`
Base Question: "What kind of object appears the most frequently in the image?"
Predicate: Requires at least two different object classes.
Apply: Counts detections per class; returns the top class only if it exceeds the second by a margin.

`LeastAppearance`
Base Question: "What kind of object appears the least frequently in the image?"
Predicate: Requires at least two different object classes.
Apply: Counts detections per class; returns the least frequent class only if it is sufficiently below the second-least.

`LeftOf`
Base Question: "Is there at least one {object_1} to the left of any {object_2}?"
Predicate: Requires at least two classes and non-overlapping detections.
Apply: Answers Yes if any non-overlapping pair has {object_1}'s right edge strictly left of {object_2}'s left edge; otherwise No.

`RightOf`
Base Question: "Is there at least one {object_1} to the right of any {object_2}?"

Predicate: Requires at least two classes and non-overlapping detections.
Apply: Answers Yes if any non-overlapping pair has {object_1} strictly to the right of {object_2}; otherwise No.

`HowMany`
Base Question: "How many {object_1}(s) are there in this image?"
Predicate: At least one object class present.
Apply: Counts instances per class and returns (class, count) pairs.

`AreMore`
Base Question: "Are there more {object_1}(s) than {object_2}(s)?"
Predicate: Requires at least two object classes.
Apply: Pairwise compares counts; asks only when the larger exceeds the smaller by a margin; returns Yes/No accordingly.

`WhichMore`
Base Question: "What appears the most in this image: {object_1}s, {object_2}s, or {object_3}s?"
Predicate: Requires at least two object classes.
Apply: Evaluates all 3-class combinations and returns the winner only when it exceeds the runner-up by a margin.

`Quadrants(N, M)`
Base Question: "Divide the image into a grid of {N} rows x {M} columns. Number the cells from left to right, then top to bottom, starting with 1. In what cell does the {object_1} appear?"
Predicate: Requires a single-instance object detection.
Apply: Returns the 1-indexed cell if the bbox fits wholly inside one cell with margins (supports up to 12 cells); otherwise skips.

`LeftMostWidthVsHeight`
Base Question: "Does the leftmost object in the image appear to be wider than it is tall?"
Predicate: At least one object class appears exactly once.
Apply: Uses the leftmost single-instance fully on the left half; requires separation from the second-leftmost and no overlap; compares aspect ratio with a threshold; returns Yes/No (also supports reversed phrasing).

`RightMostWidthVsHeight`
Base Question: "Does the rightmost object in the image appear to be wider than it is tall?"
Predicate: At least one object class appears exactly once.
Apply: Uses the rightmost single-instance fully on the right half; requires separation from the second-rightmost and no overlap; compares aspect ratio with a threshold; returns Yes/No (also supports reversed phrasing).

`MoreThanThresholdHowMany`
Base Question: "Are there {target} or more {object_1}(s) in this image? Respond Yes/No."
Predicate: At least one object class present.
Apply: For each class with count $N > 0$, asks two targets (below and above N) to yield one Yes and one No, using a multiplicative threshold.

`LessThanThresholdHowMany`
Base Question: "Are there less than {target} {object_1}(s) in this image? Respond Yes/No."
Predicate: At least one object class present.
Apply: For each class with count $N > 0$, asks two targets (above and below N) to yield one Yes and one No; special-cases target 1 as a presence question.

`MultiChoiceHowMany`
Base Question: "How many {object_1}(s) are in the image? Choose one: A) {range_a}, B) {range_b}, C) {range_c}, D) Unsure / Not Visible. Respond with the letter only."
Predicate: At least one object class present.
Apply: For classes with $N \geq 4$, builds contiguous low/mid/high buckets (variance-adjusted), shuffles them across A/B/C, and returns the correct letter; D is provided as a fallback option.

`ObjectsInRow`
Base Question: "Are there any objects arranged in a row?"

Predicate: At least 3 detections.
Apply: Slides windows of 3+ centers, fits a line, and returns Yes if normalized vertical residual variance is below a threshold; otherwise No.

`ObjectsInLine`
Base Question: "Which objects appear to be arranged in a row? A) {option_a}, B) {option_b}, C) {option_c}, D) No clear row arrangement. Respond with the letter only."
Predicate: At least 3 detections.
Apply: Finds the best low-variance row of 3+ detections via linear regression; builds two distractors and returns the letter of the correct option.

`MostClusteredObjects`
Base Question: "Which group of objects appears most tightly clustered? A) {option_a}, B) {option_b}, C) {option_c}, D) No clear clusters. Respond with the letter only."
Predicate: Requires at least 9 detections.
Apply: Runs DBSCAN on centers with eps proportional to image diagonal; selects the most compact cluster, constructs distractors, and returns the correct letter.

`Closer`
Base Question: "Is there at least one {object_1} that appears closer to the camera than any {object_2}?"
Predicate: Requires at least two classes and non-overlapping detections.
Apply: Uses SAM masks and a monocular depth map to compare non-overlapping pairs; answers Yes if any {object_1} is estimated in front of a {object_2} by a margin; otherwise No.

`Farther`
Base Question: "Is there at least one {object_1} that appears farther from the camera than any {object_2}?"
Predicate: Requires at least two classes and non-overlapping detections.
Apply: As above but checks if {object_2} is in front; answers Yes when a {object_1} is farther than a {object_2} by a margin; otherwise No.

`DepthRanking(k)`
Base Question: "Rank the {k} kinds of objects that appear the closest to the camera in the image from closest to farthest. Provide your answer as a comma-separated list of object names only."
Predicate: Requires at least {k} different object classes.
Apply: Uses SAM masks and a depth map to estimate per-class closest depth; returns the top-k order only if each consecutive pair differs by a sufficient margin.

C GRAID-BDD WITHOUT DEPTH REALIZATION STATISTICS

Table 3: Performance and Hit Rate metrics for different question types of GRAID-BDD without depth realization statistics

Question Type	is_applicable Avg (ms)	apply Avg (ms)	Predicate → QA Hit Rate	Empty cases
Divide the image into thirds. In which third does the {object_1} primarily appear? Respond with the letter only: A) left third, B) middle third, C) right third.	0.03	1.82	71.7%	11535
Is the width of the {object_1} appear to be larger than the height?	0.02	2.66	16.7%	34017
Divide the image into a grid of {N} rows x {M} columns. Number the cells from left to right, then top to bottom, starting with 1. In what cell does the {object_1} appear?	0.02	12.28	42.5%	93933
If you were to draw a tight box around each object in the image, which type of object would have the biggest box?	0.02	69.74	78.8%	15593
Rank the {k} kinds of objects that appear the largest (by pixel area) in the image from largest to smallest. Provide your answer as a comma-separated list of object names only.	0.03	72.25	87.0%	16663
What kind of object appears the most frequently in the image?	0.02	0.01	87.5%	9182
What kind of object appears the least frequently in the image?	0.01	0.01	72.6%	20133
Is there at least one {object_1} to the left of any {object_2}?	16.86	228.16	100.0%	0
Is there at least one {object_1} to the right of any {object_2}?	16.09	206.98	100.0%	0
What is the leftmost object in the image?	0.03	10.13	18.0%	33486
What is the rightmost object in the image?	0.02	10.05	20.3%	32526
How many {object_1}(s) are there in this image?	0.02	0.02	100.0%	0
Are there more {object_1}(s) than {object_2}(s) in this image?	0.01	0.02	97.7%	1708
What appears the most in this image: {object_1}s, {object_2}s, or {object_3}s?	0.01	0.02	69.5%	22432
Does the leftmost object in the image appear to be wider than it is tall?	0.01	7.41	9.0%	37131
Does the rightmost object in the image appear to be wider than it is tall?	0.02	6.61	6.6%	38108
Are there more than {target} {object_1}(s) in this image? Respond Yes/No.	0.02	0.02	100.0%	0
Are there less than {target} {object_1}(s) in this image? Respond Yes/No.	0.01	0.02	100.0%	0
How many {object_1}(s) are in the image? Choose one: A) {range_a}, B) {range_b}, C) {range_c}, D) Unsure / Not Visible. Respond with the letter only.	0.01	0.15	94.3%	4504

Notes:

- `is_applicable` checks if a question type can be applied to an image
- `apply` realizes the actual question-answer pairs
- Predicate → QA Hit Rate = Percentage of applicable cases that generated at least one QA pair
- Empty cases = Number of times predicates passed but `apply` realized no QA pairs

D Research Question 3 Training Details

Here we discuss the details of our SFT using the GRAID-BDD dataset. We identify the rarest kind of question in GRAID-BDD, then randomly select that many questions from each question type yielding 51,546 training examples. We fit a LoRA of rank 32, with a batch size of 2 and 4 gradient accumulation steps. We train with a learning rate of 2^{-4} for 200 steps, with 5 warm up steps, with a linear schedule, weight decay of 0.01, and use the AdamW8bit optimizer.

Table 4: Performance comparison between the baseline model (Meta Llama 3.2 11B Vision Instruct), the same model fine-tuned on the OpenSpaces dataset produced by the community implementation of SpatialVLM, and the same model fine-tuned using only the GRAID-BDD dataset. All benchmarks are evaluated with VLMEvalKit using its exact match protocol.

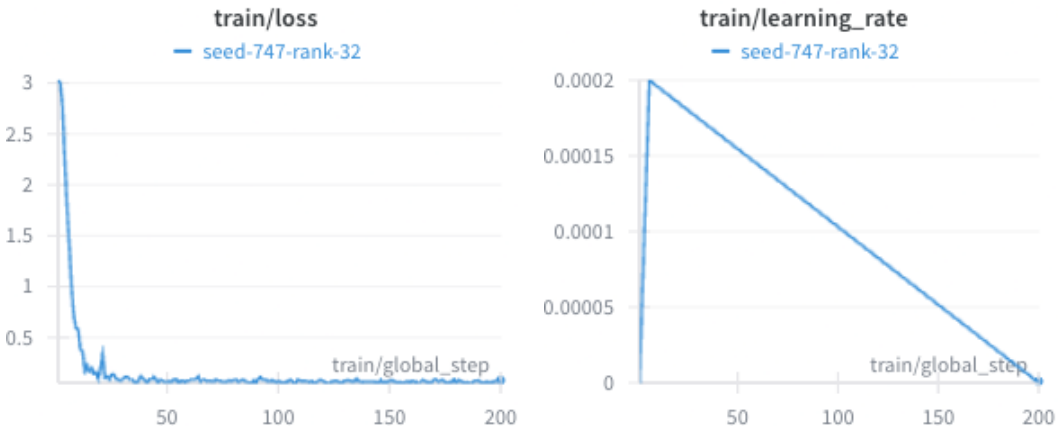
Dataset	Llama 3.2	Llama+OpenSpaces	Llama+GRAID
A-OKVQA	64.02%	55.37% (-8.65)	83.67% (+19.65)
RealWorldQA	36.73%	21.31% (-15.42)	59.48% (+22.75)
NaturalBench			
Q_Acc	48.97%	15.21% (-33.76)	50.29% (+1.32)
I_Acc	52.82%	15.79% (-37.03)	53.36% (+0.54)
Acc	73.40%	49.25% (-24.15)	74.28% (+0.88)
G_Acc	23.42%	3.63% (-19.79)	25.42% (+2.00)
BLINK			
Overall	25.72%	25.46% (-0.26)	42.13% (+16.41)
Art Style	47.86%	20.51% (-27.35)	47.01% (-0.85%)
Counting	25.00%	13.33% (-11.67)	52.50% (+25.50)
Forensic Detection	25.76%	26.52% (+0.76)	26.51% (+0.75%)
Functional Correspondence	3.08%	16.92% (+13.84)	24.61% (+21.53)
IQ Test	6.67%	25.33% (+18.66)	18.00% (+11.33)
Jigsaw	52.00%	27.33% (-24.67)	52.67% (+0.67)
Multi-view Reasoning	35.34%	18.05% (-17.29)	44.36% (+9.02)
Object Localization	61.48%	25.41% (-36.07)	63.11% (+1.63)
Relative Depth	10.48%	50.00% (+39.52)	52.42% (+41.94)
Relative Reflectance	0.75%	24.63% (+23.88)	31.34% (+30.59)
Semantic Correspondence	12.23%	23.02% (+10.79)	35.97% (+23.74)
Spatial Relation	36.36%	18.88% (-17.48)	72.02% (+35.66)
Visual Correspondence	5.23%	25.00% (+19.77)	29.06% (+23.83)
Visual Similarity	46.67%	41.48% (-5.19)	47.41% (+0.74)
VSR-zeroshot			
Precision	57.35%	54.44% (-2.91)	52.50% (-4.85)
Recall	95.55%	21.46% (-74.09)	98.57% (+3.02)
Accuracy	61.13%	41.98% (-19.15)	53.36% (-7.77)
F1	71.68%	30.79% (-40.89)	69.00% (-2.68)

Table 5: Performance comparison between the baseline model (Gemma 3 4B IT), the same model fine-tuned on the OpenSpaces dataset produced by the community implementation of SpatialVLM, and same model fine-tuned using only the GRAID-BDD dataset. All benchmarks are evaluated with VLMEvalKit using its exact match protocol.

Dataset	**Gemma 3**	**Gemma+OpenSpaces**	**Gemma+GRAID**
A-OKVQA	1.57%	53.01% (+51.44)	76.07% (+74.50)
RealWorldQA	13.33%	34.90% (+21.57)	49.02% (+35.69)
NaturalBench			
Q_Acc	42.76%	19.74% (-23.02)	33.76% (-8.00)
I_Acc	47.03%	19.97% (-27.06)	36.32% (-10.71)
Acc	70.05%	48.74% (-21.31)	63.13% (-6.92)
G_Acc	17.95%	3.84% (-14.11)	10.68% (-7.27)
BLINK			
Overall	4.21%	29.72% (+25.51)	38.72% (+34.51)
Art Style	35.90%	48.72% (+12.82)	50.43% (+14.53)
Counting	10.00%	14.17% (+4.17)	29.17% (+19.17)
Forensic Detection	13.64%	13.64% (0.0)	29.55% (+15.91)
Functional Correspondence	0.00%	15.38% (+15.38)	17.69% (+17.69)
IQ Test	0.67%	14.00% (+13.33)	21.33% (+20.66)
Jigsaw	3.33%	42.67% (+39.34)	54.67% (+51.34)
Multi-view Reasoning	0.75%	36.09% (+35.34)	39.10% (+38.35)
Object Localization	0.00%	22.13% (+22.13)	48.36% (+48.36)
Relative Depth	0.00%	52.42% (+52.42)	51.61% (+51.61)
Relative Reflectance	0.00%	34.33% (+34.33)	28.36% (+28.36)
Semantic Correspondence	0.72%	20.14% (+19.42)	31.65% (+30.93)
Spatial Relation	0.00%	48.25% (+48.25)	58.74% (+58.74)
Visual Correspondence	0.00%	20.93% (+20.93)	33.72% (+33.72)
Visual Similarity	0.00%	36.30% (+36.30)	49.63% (+49.63)
VSR-zeroshot			
Precision	54.74%	55.56% (+0.82)	54.15% (-0.49)
Recall	93.64%	27.82% (-65.82)	78.86% (-14.78)
Accuracy	56.87%	48.85% (-8.02)	54.75% (-2.12)
F1	69.00%	37.00% (-32.00)	64.00% (-5.00)

Table 6: Performance comparison between the baseline models, the same model fine-tuned on OpenSpaces, and the same model fine-tuned on the GRAID-BDD dataset. All benchmarks are evaluated using VLMEvalKit and its exact match protocol. Results are shown for four model families: Llama-3.2-11B-Vision-Instruct, Gemma-3-4B-IT, Qwen2.5-VL-3B-Instruct, and Qwen3-VL-8B-Instruct. Each cell contains four values corresponding to these model families in order.

Dataset	Base	OpenSpaces-SFT	GRAID-SFT
A-OKVQA	64.02% / 1.57% / 85.32% / 86.72%	55.37% / 53.01% / 57.03% / 77.38%	83.67% / 76.07% / 81.92% / 87.34%
RealWorldQA	36.73% / 13.33% / 65.50% / 72.03%	21.31% / 34.90% / 39.74% / 53.59%	59.48% / 49.02% / 61.44% / 71.76%
NaturalBench			
Q_Acc	48.97% / 42.76% / 51.39% / 61.89%	15.21% / 19.74% / 21.34% / 39.42%	50.29% / 33.76% / 47.45% / 58.97%
I_Acc	52.82% / 47.03% / 55.23% / 63.87%	15.79% / 19.97% / 21.74% / 41.16%	53.36% / 36.32% / 50.37% / 61.08%
Acc	73.40% / 70.05% / 74.46% / 80.09%	49.25% / 48.74% / 52.39% / 65.76%	74.28% / 63.13% / 71.21% / 78.50%
G_Acc	23.42% / 17.95% / 25.63% / 37.37%	3.63% / 3.84% / 5.42% / 15.74%	25.42% / 10.68% / 23.05% / 35.05%
BLINK			
Overall	25.72% / 4.21% / 49.18% / 56.71%	25.46% / 29.72% / 37.30% / 42.98%	42.14% / 38.72% / 44.45% / 62.28%
Art Style	47.86% / 35.90% / 56.41% / 43.59%	20.51% / 48.72% / 46.15% / 50.43%	47.01% / 50.43% / 56.41% / 72.65%
Counting	25.00% / 10.00% / 68.33% / 65.00%	13.33% / 14.17% / 48.33% / 45.83%	52.50% / 29.17% / 61.67% / 64.17%
Forensic Detection	25.76% / 13.64% / 32.57% / 89.39%	26.52% / 13.64% / 21.21% / 28.03%	26.52% / 29.55% / 20.45% / 75.76%
Functional Correspondence	3.08% / 0.00% / 23.84% / 3.08%	16.92% / 15.38% / 18.46% / 28.46%	24.62% / 17.69% / 29.23% / 36.15%
IQ Test	6.67% / 0.67% / 26.00% / 0.00%	25.33% / 14.00% / 27.33% / 28.00%	18.00% / 21.33% / 18.67% / 26.67%
Jigsaw	52.00% / 3.33% / 50.00% / 69.33%	27.33% / 42.67% / 54.00% / 39.33%	52.67% / 54.67% / 48.67% / 62.67%
Multi-view Reasoning	35.34% / 0.75% / 48.12% / 54.14%	18.05% / 36.09% / 46.62% / 45.86%	44.36% / 39.10% / 46.62% / 50.38%
Object Localization	61.48% / 0.00% / 54.91% / 68.03%	25.41% / 22.13% / 35.25% / 59.02%	63.11% / 48.36% / 50.00% / 67.21%
Relative Depth	10.48% / 0.00% / 70.96% / 87.90%	50.00% / 52.42% / 57.26% / 51.61%	52.42% / 51.61% / 60.48% / 86.29%
Relative Reflectance	0.75% / 0.00% / 39.55% / 32.84%	24.63% / 34.33% / 32.84% / 38.06%	31.34% / 28.36% / 40.30% / 33.58%
Semantic Correspondence	12.23% / 0.72% / 31.65% / 17.99%	23.02% / 20.14% / 24.46% / 29.50%	35.97% / 31.65% / 29.50% / 47.48%
Spatial Relation	36.36% / 0.00% / 83.21% / 86.01%	18.88% / 48.25% / 48.25% / 54.55%	72.03% / 58.74% / 75.52% / 82.52%
Visual Correspondence	5.23% / 0.00% / 40.11% / 86.63%	25.00% / 20.93% / 20.93% / 49.42%	29.07% / 33.72% / 36.05% / 84.30%
Visual Similarity	46.67% / 0.00% / 70.37% / 87.41%	41.48% / 36.30% / 47.41% / 56.30%	47.41% / 49.63% / 56.30% / 82.22%
VSR-zeroshot			
Precision	57.35% / 54.74% / 78.08% / 88.58%	54.44% / 55.56% / 55.66% / 65.44%	52.50% / 54.15% / 68.34% / 80.35%
Recall	95.55% / 93.64% / 80.44% / 85.06%	21.46% / 27.82% / 18.76% / 31.00%	98.57% / 78.86% / 84.42% / 88.39%
Accuracy	61.13% / 56.87% / 78.31% / 86.67%	41.98% / 48.85% / 50.33% / 56.06%	53.36% / 54.75% / 71.85% / 82.90%
F1	71.68% / 69.09% / 79.24% / 86.78%	30.79% / 37.08% / 28.06% / 42.07%	69.00% / 64.21% / 75.53% / 84.18%



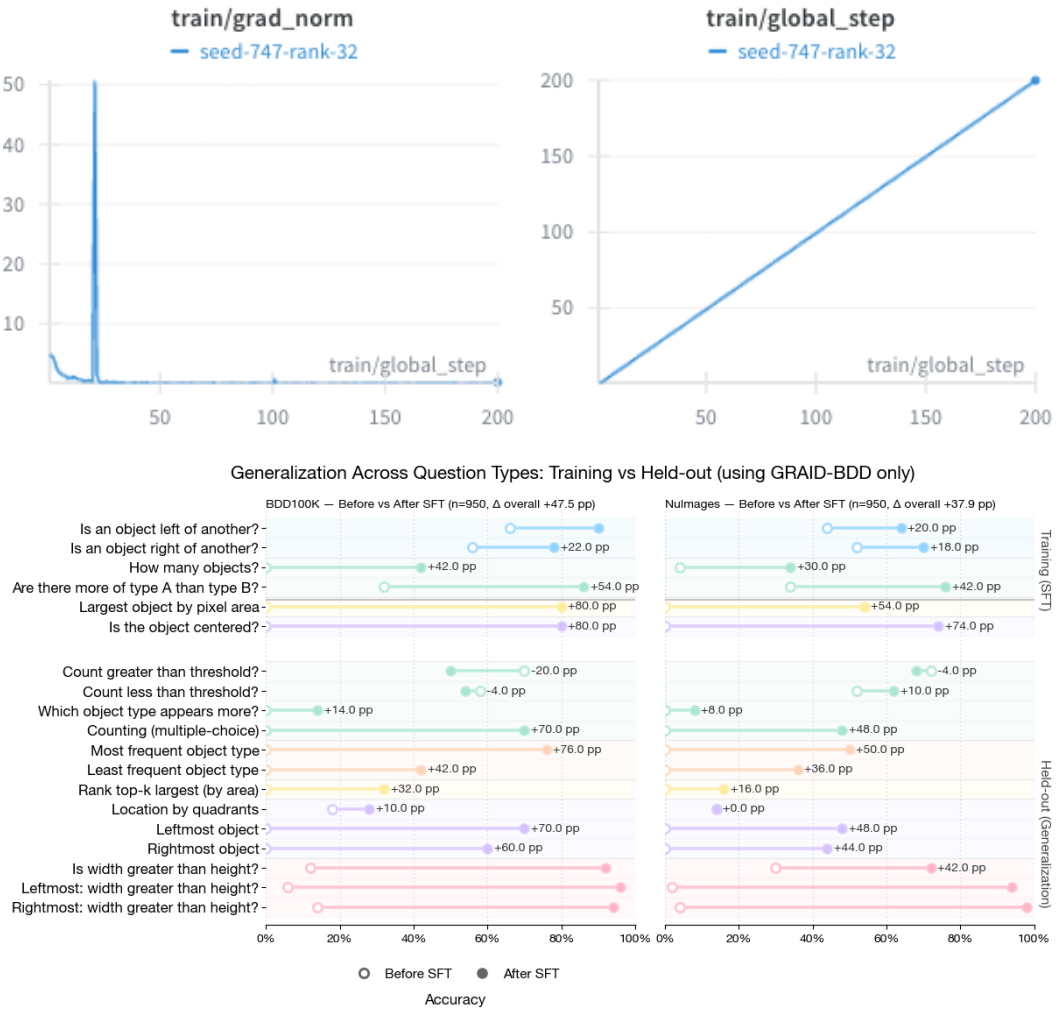
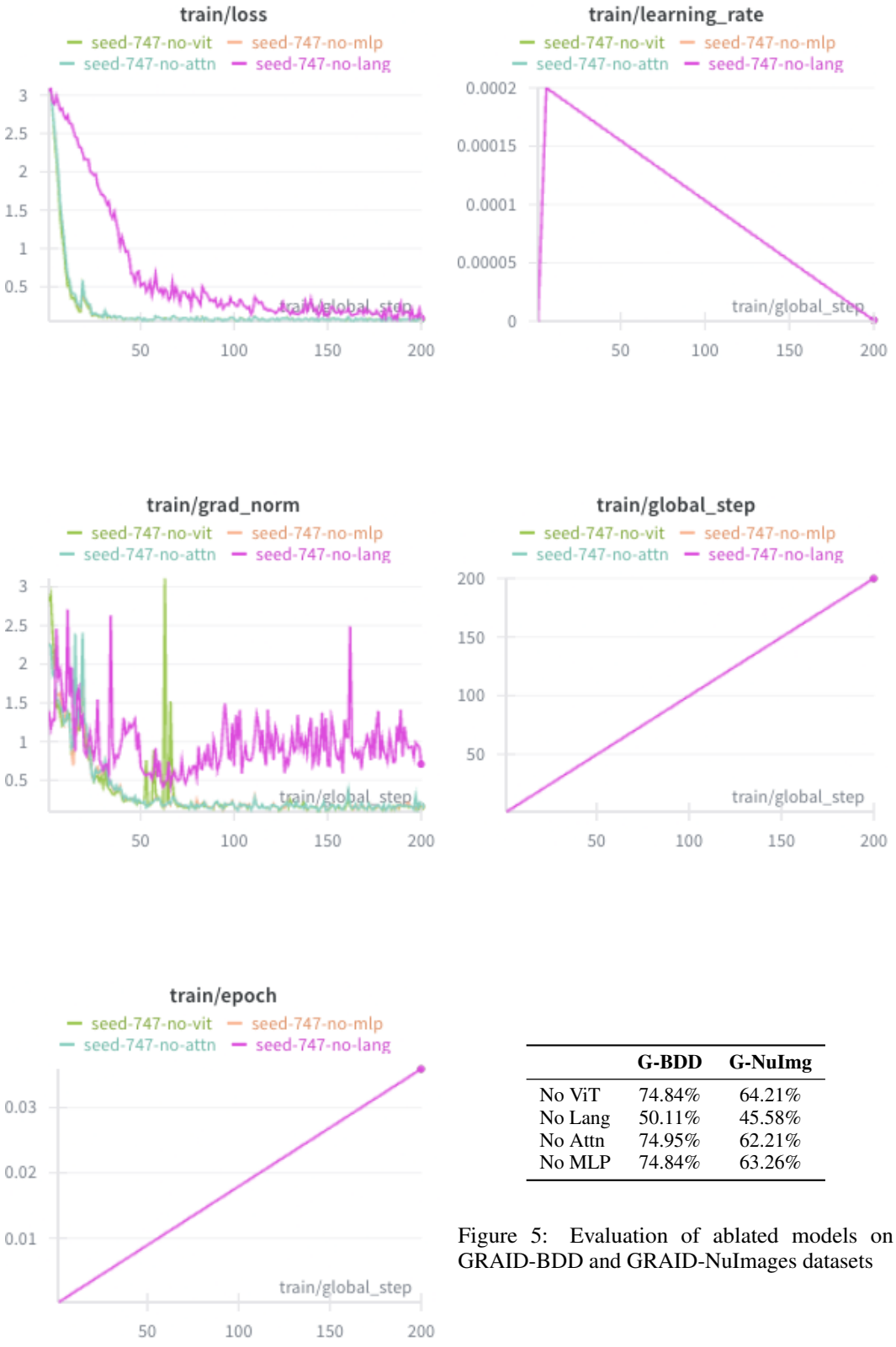


Figure 4: After supervised fine-tuning on the GRAID-BDD dataset, we can see improvements in the model's ability to answer questions on the held-out questions of GRAID-BDD, and a dataset with a different distribution of scenes, GRAID-NuImages.

E VLM Training Ablations

In all our supervised fine-tuning experiments, we use `unsloth` to training our LoRA adapters. In this section, we discuss ablations on the various components we can have LoRA adapters for: vision layers, language layers, attention modules, and mlp modules. In each of the experiments, we enable SFT of all components except one at a time. All experiments use a rank of 16, batch size of 2, 4 gradient accumulation steps, 5 warmup steps, 200 steps, a learning rate of 2^{-4}, a linear scheduler, AdamW8bit optimizer, and 0.01 weight decay. In the charts below, we see that the training loss curves for all experiments except one are identical: disallowing the fine-tuning of the language layers. In this setting, we are unable to train the model as well as in the others, and the gradient norm remains relatively high. These results hint that the vision layers of a VLM can use some improvement, however, the vast majority of spatial reasoning is still occurring in the language space of the model, and not in vision space. Observe that this finding concurs with prior work in that "while image tokens take approximately 90% of the sequence length, they receive only about 10% of the model's total attention", i.e., VLMs mostly reason about images in the text space and not image space Chen et al. (2025).



	G-BDD	G-NuImg
No ViT	74.84%	64.21%
No Lang	50.11%	45.58%
No Attn	74.95%	62.21%
No MLP	74.84%	63.26%

Figure 5: Evaluation of ablated models on GRAID-BDD and GRAID-NuImages datasets