Rep2D-BEAT: Evaluating and Enhancing Spatial Understanding in Vision-Language Models

Jan Hutter [1] **Kshitiz Sharma** [1] **Maxim Voronin** [1] **Bhavesh Sood** [1] **Viktória Pravdová** [1] **Fabian Westerbeek** [1]

Abstract

Despite impressive advances in vision-language models (VLMs), their capacity for 3D spatial understanding from 2D inputs remains limited and underexplored. We introduce **Rep2D-BEAT**, a benchmark and dataset designed to evaluate and enhance the spatial reasoning abilities of VLMs in multi-view settings. Built from the Replica dataset, Rep2D-BEAT comprises three tasks—bounding box prediction, occlusion estimation, and camera pose inference—each aimed at probing different facets of 3D geometric understanding without relying on explicit 3D inputs. We assess state-of-the-art VLMs on Rep2D-BEAT as well as on two established 3D vision-language benchmarks: ScanQA and SQA3D. Our findings reveal that while current models demonstrate some latent 3D awareness, their performance significantly lags behind task-specific or 3D-augmented counterparts. Moreover, we show that increasing the number of input views offers diminishing returns, indicating current VLMs' limited capacity for integrating multi-view geometric context. Rep2D-BEAT provides a scalable framework for evaluating and fine-tuning VLMs toward deeper spatial understanding, bridging the gap between semantic recognition and true 3D reasoning.

1. Introduction

Vision-language models (VLMs) such as Gemini and GPT-4V have demonstrated strong 2D visual understanding by leveraging large-scale image–text pre-training (Alayrac et al., 2022; Li et al., 2023). However, their capacity to infer 3D geometry from 2D images—beyond simple semantic recognition—remains underexplored. In particular, view correspondence tasks, where a model must match identical objects (e.g., two indistinguishable chairs) across different camera angles, directly probe whether VLMs form an implicit spatial representation or rely solely on semantic information. Broadly, 3D correspondence refers to matching corresponding pixels or regions in different images, 3D shapes, or some other modality. Solving the problem of 3D correspondence would also solve multiple other problems in the computer vision domain. These include, but are not limited to, structure from motion, optical flow, part matching, and registration. More importantly, it would represent an important step towards getting VLMs to obtain a human-like understanding of the 3D world, moving from simple semantics to true geometric understanding.

Prior work has shown that VLMs can misidentify identical objects when semantic features are ambiguous, suggesting limited geometric reasoning (Chen et al., 2025; Eppel et al., 2025). This, however, does not imply that image models trained solely on 2D data have no 3D understanding. (Chen et al., 2023b), for instance, show that Latent Diffusion Models (LDMs) encode depth representations. Still, their abilities are limited. This is showcased by the inconsistencies in the generation of 3D objects. Additionally, (El Banani et al., 2024) find that models struggle with reconciling information from multiple views. They can usually match objects where the change in viewpoint is small. However, when these changes are large, the performance is poor, again suggesting a lack of 3D geometrical understanding. While these findings highlight the limitations of current models, one of the major bottlenecks is evaluation. There are no datasets and standardized benchmarks to test a model's 3D geometric understanding. Datasets which do exist with 3D information are either too simplistic for such a task, or require substantial preparation to be suitable for use. For instance, (Jampani et al., 2023) is a dataset with multiple objects and images, but the samples are too simplistic for correspondence tasks. Similarly, there are the (Nathan Silberman & Fergus, 2012) and (Straub et al., 2019a) datasets which contain multiple scenes with multiple objects but require elaborate preprocessing to generate samples to test the geometric understanding of models.

In this work, we address these gaps by first quantifying the performances of a suite of general-purpose VLMs on

[1]University of Amsterdam, Netherlands..

UvA Foundation Models Course 2025, Project Paper Submission.

a benchmark of multi-view scenes. Specifically, we use the ScanQA and SQA3D (Azuma et al., 2022), which are question answering datasets, to provide the models with 2D views of scenes and the questions as input.

In addition, we contribute a new dataset[1], created from Replica Dataset (Straub et al., 2019a), particularly suited for evaluating and fine-tuning VLMs for 3D geometrical understanding, moving away from pure semantics. Our dataset has three parts, each suited for a slightly different objective: (1) Bounding box prediction; (2) Estimating the amount of object occlusion; (3) Predicting the change in camera pose between different views.

To this end, we try to answer the question:
How effectively can general-purpose vision-language models infer 3D spatial structure and establish object correspondence across multiple 2D views without explicit 3D input?

2. Background

2.1. Vision Language Models

Vision Language Models (VLMs) are a type of multimodal generative models which are apt at handling image and text modalities. They typically have good zero-shot capabilities and can perform a variety of tasks, such as image recognition, visual question answering, image captioning, etc. There are certain multi-modal models, such as GPT-4o (OpenAI, 2024) which are capable of generating images as well, although these go beyond VLMs and can also process other modalities such as speech.

To process multimodal input, they typically employ a language model (LM) and a vision transformer (ViT) and merge the modalities using an MLP. They are capable of processing images of various resolutions. However, different models achieve this with different methods. Qwen2.5-VL (Qwen et al., 2025), for instance, breaks images into patches of a fixed size before using the ViT to process them, which makes it capable of handling images of different resolutions. Gemma 3 (Team et al., 2025), in contrast, uses SigLIP (Zhai et al., 2023) as the ViT and does not natively support different resolutions for the images. Instead, it relies on a separate algorithmic intervention to attain flexibility in image resolution. These architectural choices usually reflect a trade-off between performance and computational intensity. The sheer size of these models implies that they are expensive to train and use, and lower-performing models with lower computational requirements might be better suited to some situations than the best-performing models.

Beyond these architectural choices, there is a lot of diversity among the models in regard to the data they were trained on, how they were optimized, and whether or not they use specific optimizations during inference. This translates into different models performing better in some tasks than others. Overall, VLMs are a significant advancement in multi-modal learning, demonstrating impressive capabilities across a range of tasks.

2.2. 3D Awareness in VLMs

Defining 3D Awareness in VLMs. A vision-language model (VLM) can be considered 3D-aware if it possesses an internal representation of the three-dimensional structure of the world, enabling it to reason about depth, spatial relationships, and object permanence across varying viewpoints. This capability extends beyond surface-level semantic recognition, requiring a geometric understanding of how objects exist and relate in space.

At its core, 3D awareness implies the model can infer how scenes and objects transform under different camera perspectives, handle occlusions, estimate depth and pose, and maintain consistent object identity across views. Such abilities are fundamental to human-like visual perception and underpin a range of high-level cognitive and interactive tasks, from navigation and manipulation to embodied question answering.

Evaluating 3D awareness. Tasks that challenge or reveal 3D awareness include view correspondence, where models must recognize an object seen from one angle when it appears again under a different viewpoint. These are non-trivial without an internalized geometric model, particularly when appearance-based features are insufficient or misleading due to lighting, occlusion, or orientation. Conversely, models that lack 3D awareness often falter in such scenarios, succeeding only in tasks with strong semantic cues or minimal spatial ambiguity, such as single-view captioning or attribute classification.

In contrast, tasks like object recognition in canonical views or identifying common object categories (e.g., “a chair,” “a tree”) are relatively easy for semantically rich, yet geometrically naive models. These models rely primarily on 2D pattern matching and do not require reasoning about depth or spatial consistency.

Improving 3D awareness in VLMs. True 3D understanding demands more than multi-view input aggregation. It necessitates the integration of geometric principles such as perspective projection, object permanence, and spatial contiguity. This is especially apparent in tasks like structure-from-motion, depth estimation, pose prediction, and scene reconstruction, where spatial reasoning is essential to effectively perform these tasks.

For VLMs to become 3D-aware, they must bridge the gap

[1]We will open-source our dataset. In the meantime, our code is available at

between vision as pattern recognition and vision as spatial inference. This entails developing inductive biases or learning strategies that allow models to form viewpoint-invariant representations. Some approaches inject explicit 3D information, such as point clouds or camera extrinsics, into model inputs; others, like ROSS3D (Wang et al., 2025), introduce 3D supervision through reconstructive objectives without embedding geometric priors at the input level. LLaVA-3D (Zhu et al., 2025), on the other hand, adds 3D position embeddings to the 2D patch tokens from multiple views and constructs 3D patches. These are then forwarded to the LLaVA LLM after projection to align the input with 3D-visual-language data.

2.3. ScanQA and SQA3D

ScanQA (Azuma et al., 2022) is a large-scale benchmark for 3D question answering (3D-QA), designed to evaluate spatial understanding in indoor environments. It comprises over 41,000 human-edited question–answer pairs grounded in 800 RGB-D scans from the ScanNet dataset. Each question is associated with free-form answers and annotated 3D object locations, requiring models to jointly reason over language and 3D geometry. The dataset supports both object localization and answer prediction tasks, making it a comprehensive testbed for evaluating 3D-aware vision-language models.

SQA3D (Ma et al., 2023) (Situated Question Answering in 3D Scenes) extends the 3D-QA paradigm by introducing agent-centric reasoning. Built on 650 ScanNet scenes, it includes 6,800 unique situations, 20,400 textual descriptions, and 33,400 diverse reasoning questions. Each situation specifies an agent's position and orientation within the scene, challenging models to interpret spatial context and answer questions involving spatial relations, navigation, and common-sense reasoning. SQA3D emphasizes embodied scene understanding and highlights the performance gap between current models and human-level reasoning.

Together, ScanQA and SQA3D provide complementary benchmarks for assessing and advancing 3D-aware vision-language models, focusing on object grounding and situated reasoning, respectively.

3. Related Work

3.1. General Capabilities and Limitations of Vision-Language Models

VLMs like Flamingo and BLIP-2 achieve state-of-the-art performance on image captioning and VQA by fusing vision encoders with large language models (Alayrac et al., 2022; Li et al., 2023). While they excel at object recognition and attribute description, they often falter on tasks requiring spatial relationships (e.g., left/right, counting), indicating a bias toward semantic correlations over explicit geometry (Chen et al., 2025; Eppel et al., 2025; Liu et al., 2024a).

3.2. Extending VLMs with 3D Spatial Reasoning

Efforts to infuse 3D reasoning include augmenting VLMs with point-cloud inputs or 3D detectors (Hong et al., 2023; Wang et al., 2023), and training on embodied 3D datasets (Chen et al., 2023a). These specialized models improve on occlusion and depth queries but require heavy retraining and 3D annotations. Multi-view 2D approaches, such as LLaVA-NeXT-Interleave, interleave image tokens from different perspectives to simulate 3D context, yet still struggle with novel viewpoints without explicit geometric supervision (Li et al., 2024a; Zuo et al., 2024). Chen et al. (2024) proposes a method for generating a synthetic spatial reasoning QA dataset, which allows fine-tuning VLMs for spatial question answering on 2D images, such as determining the distance between two objects in an image.

3.3. View Correspondence and Geometric Alignment in Multimodal Settings

Classical computer vision has long tackled feature matching for stereo and structure-from-motion, but VLMs' native ability to perform cross-view grounding is nascent. The Coarse Correspondences approach marks matched regions across frames to boost GPT-4V's spatial reasoning, yielding gains on benchmarks like ScanQA (Azuma et al., 2022). Shape–texture matching studies (Eppel et al., 2025) further reveal that VLMs' performance degrades when appearance factors vary, suggesting limited viewpoint invariance. Our work diverges by evaluating off-the-shelf VLMs on unmarked multi-view scenes, testing whether they inherently establish geometric correspondences or require external prompts. This focused benchmark fills a critical gap, assessing true 3D consistency in general-purpose vision-language models.

3.4. Inducing 3D into foundation models without inductive bias

The models showing a good performance on 3D benchmarks like ScanQA and SQA3D (Azuma et al., 2022), that is Llava-3D (Zhu et al., 2025) and LEO (Wang et al., 2023) have been beaten by an approach involving no inductive bias. ROSS3D's core method (Wang et al., 2025) stands out by avoiding input-level inductive biases that dominate most 3D LMMs. Instead of injecting 3D information through engineered representations (like 3D point encodings), it introduces output-level 3D supervision using cross-view reconstruction and global-view reconstruction.

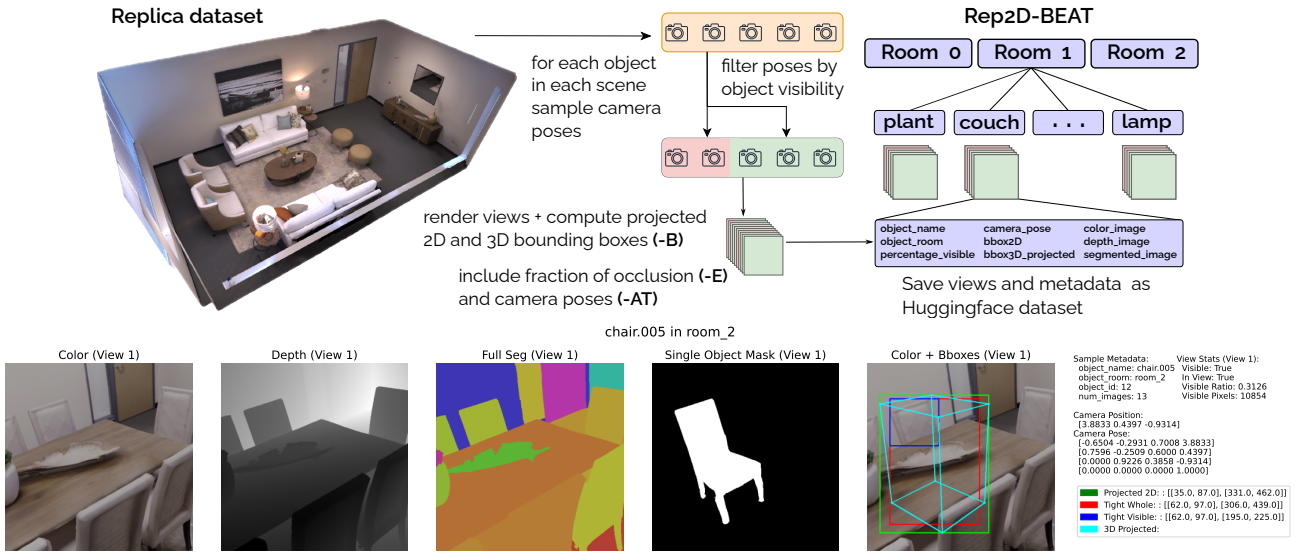


Figure 1. Rep2D-BEAT

4. Method

4.1. Rep2D-BEAT Dataset

We introduce *Rep2D-BEAT*, a dataset for fine-tuning a VLM to improve 3D reasoning capabilities, solely utilizing 2D images. This dataset is based on the 3D scenes from Replica (Straub et al., 2019b).

To create the dataset, we followed the methodology outlined below and illustrated in 1:

1. We sampled random camera positions within the room, ensuring the minimal distance to other positions

2. Oriented the camera at the object, adding a small rotation perturbation to ensure different positions of the object within the frame.

3. Rendered images and depth maps, exporting the projected 3D bounding boxes, which were then used to determine the loose 2D bounding boxes within the frame. We saved the camera extrinsics as well.

4. Exported the segmentation masks of the whole room, as well as the segmentation masks of just the single object (while having all the objects hidden). We used this to determine: a) tight bounding boxes of the whole object in the frame (including occluded parts), b) bounding boxes of just the visible part of the object and c) visibility proportion of the part of the object within the frame.

For each object we saved 10 views where the object is visible in the scene and 2 views where it isn't. If it took too many tries to sample enough suitable positions, we skipped the object. For resolution, we chose 512x512 pixels.

We utilized our dataset for three different tasks: bounding box prediction (Rep2D-B), using the loose projected bounding boxes, target object extent of occlusion estimation (Rep2D-E), and across-view transformation prediction (Rep2D-AT).

Problem formulation. Given a set of observations from an environment, denoted as $[I_1, \ldots, I_k, \ldots, I_n]$, and a target object specified in the first amount of observations $[I_1, \ldots I_k]$, the VLM must identify the same target object in each of the remaining observations I_{k+1} through I_n. Our goal is to enable reasoning about the 3D structure of the scene and to equip the VLM with the ability to establish correspondence between the observations. Below, a more in-depth description of the three tasks is given.

1. **Rep2D-B.** (**B**ounding box). Given a set of target observations $[I_1, \ldots, I_k]$, annotated with bounding boxes, the model has to output a single bounding box for the next observation I_{k+1}. The model output will be evaluated using Intersection over Union (IoU) score, which is used as objective during fine-tuning. An example input for this task is located in Appendix G.

2. **Rep2D-E.** (**E**xposure). Given a set of annotated observations, the model must perform a regression task to estimate the visible amount of the target object in a single unannotated target observation. The model output will be evaluated using mean squared error (MSE).

3. **Rep2D-AT.** (**CA**mera ex**T**rinsics). Given the set of annotated observations with their camera extrinsics (a 3x1 vector), the model must output the current camera

extrinsics in a single unannotated target observation. This model output will be evaluated using mean L2 error.

All tasks support varying the number of example observations, facilitating straightforward multi-shot prompting. Table 1 shows additional information about our dataset.

Dataset	Scenes	Objects	#Img/Obj	#Images
Rep2D-BEAT	16	604	10+2	7248

Table 1. Rep2D-BEAT statistics.

4.2. Evaluation

Datasets. Similar to earlier research on 3D understanding, we utilize the ScanQA and SQA3D datasets (Mo & Liu, 2024; Liu et al., 2024a; Zhu et al., 2024) to evaluate our models. Both datasets are derived from ScanNet, which provides 968x1296 video frames of various scenes (Azuma et al., 2022; Ma et al., 2023). The frames used for model evaluation are randomly sampled from these videos, and their resolution is kept the same. Some information about these datasets can be found in Table 2

Dataset	Scenes	Questions	Answers per question	Split
ScanQA	71	4682	2.01	Val
SQA3D	65	3261	1.00	Test

Table 2. Information about the two spatial question-answering sets.

Baselines. The performance of various VLMs is evaluated on both ScanQA and SQA3D and compared to the results of task-specific models, and 3D models, which utilize 3D scene information like point clouds from the scenes; the results for these task-specific and 3D models are taken from their respective research papers. We compare the performance of Gemma-3-27B-IT (Team et al., 2025), LLaVA-Next-7B (Liu et al., 2024b), Qwen2.5-VL-7B-Instruct (Qwen et al., 2025), and SpaceThinker-Qwen-2.5VL-3B, a Qwen model finetuned on data obtained through an open-source implementation [2] of SpatialVLM (Chen et al., 2024), against ScanQA (Azuma et al., 2022) and LLaVA-3D (Zhu et al., 2024). We also evaluate the general multi-modal understanding of these modals, and the results are presented in Appendix A.

Metrics. Following prior work, we evaluate performance on the SQA3D dataset using EM@1, and on the ScanQA dataset using CIDEr (Vedantam et al., 2015), BLEU-4 (Papineni et al., 2002), METEOR (Lavie & Denkowski, 2009), and ROUGE (Lin, 2004). CIDEr measures the consensus between the target and the generated answer using a weighted n-gram overlap. BLEU-4 evaluates the precision of 1 to 4-gram matches between the generation and the target. Similarly, METEOR aligns words based on exact matches, stemmed forms, paraphrases, and synonyms. ROUGE, which also had different n-gram variants like BLEU, focuses on recall by assessing the overlap between generated and reference sentences.

Beyond these standard metrics, we incorporate an LLM-as-a-judge (Li et al., 2024b) evaluation to measure the semantic equivalence between model predictions and reference answers. This approach enables a more intuitive assessment of answer quality that goes beyond surface-level similarity. The prompt used for this evaluation is provided in Appendix F.

Setup. Due to computational constraints, each of our evaluated VLMs is run using a single random seed. A batch size of 2 was used for all experiments. These were run on NVIDIA A100 and NVIDIA H100 GPUs.

5. Experiments

5.1. ScanQA & SQA3D

We evaluate four VLMs—Gemma3, LLaVA, Qwen-2.5, and SpaceThinker-Qwen-2.5VL-3B (ST)—on two spatial question-answering datasets. A specified number of frames are sampled randomly from the ScanNet video to provide models with multi-view information and assess how effectively they process and combine visual information for spatial reasoning. Gemma3, LLaVA, and Qwen process images at the original resolution of 968×1296, whereas SpaceThinker (ST) operates on downscaled inputs with a resolution of 512×685.

5.2. Rep2D-BEAT

We assess the effectiveness of our dataset by evaluating Qwen-2.5-VL-7B, Gemma3-it-27B, and SpaceThinker-Qwen-2.5VL-3B (ST) across the three defined tasks. Initial experiments showed that the LLaVA model failed to produce meaningful outputs for any of the three tasks, and thus it was excluded from our Rep2D experiments. Specific prompts that were utilized in our research, can be found in Appendix E.

6. Results

6.1. Main Results

Evaluating the 2D VLMs on the random views of ScanQA and SQA3D scenes, a general decrease in performance can be observed in Table 3 compared to Task-specific and 3D

[2] https://github.com/remyxai/VQASynth

VLMs. Particularly, the CIDEr and BLEU-4 scores decrease significantly for the 2D VLMs. Qwen-2.5-VL achieves strong CIDEr performance, largely due to its consistent generation of concise, one-word or short answer phrases, resulting in a larger match with the target answer. An interesting phenomenon is that, Gemma-3, although the biggest VLM model, shows the worst performance across all metrics. An investigation into this is left for future work.

		ScanQA					SQA3D	
Model	Frame	C	B-4	M	R	LLM	EM@1	LLM
Task-specific								
ScanQA	-	64.9	10.1	13.1	33.3	-	47.2	-
3D VLMs								
LLaVA-3D	-	103.1	16.4	20.8	49.6	-	60.1	-
2D VLMs								
Gemma-3	15	3.9	4.5	12.1	23.6	33.9	37.6	37.2
LLaVA	30	3.0	1.6	**19.0**	12.7	**43.7**	**42.5**	**44.6**
Qwen-2.5-VL	15	**46.2**	**4.9**	14.0	**27.8**	39.6	32.8	28.8
SpaceThinker	0	13.6	2.5	8.9	15.1	-	34.9	-

Table 3. Comparison of non-finetuned 2D VLLMs with task-specific models and 3D LLMs on the ScanQA and SQA3D datasets. Metrics include CIDEr (C), BLEU-4 (B-4), METEOR (M), ROUGE (R), Exact Match at 1 (EM@1), and LLM-as-a-judge scores (LLM).

6.2. Varying Frame Counts

We ablate the number of input frames to study how increasing the quantity of views affects 2D VLMs performance. Table 4 illustrates how this performance changes with varying frames on the ScanQA and SQA3D datasets. The optimal number of frames seems to be 5 for Gemma-3 and Qwen-2.5 and provides the best balance between performance, context length, and computational cost at inference time. Contrary to these results, SpaceThinker performs best when not given any additional context, and the optimal frame count for LLaVA varies based on the metric being measured. Due to hardware limitations, we were unable to perform inference on Gemma-3 and Qwen-2.5-VL using 30 input images, and could not run inference on SpaceThinker-Qwen-2.5VL-3B with more than 5 images.

6.3. Prompt ablation

Interleaving the images in the input on both ScanQA and SQA3D with simple pieces of text reduces question-answering performance, as shown in Table 5. This interleaving is achieved by appending 'Image: i' to the input, where i is filled in with the actual index of the image in the prompt.

		ScanQA					SQA3D	
Model	Frame	C	B-4	M	R	LLM	EM@1	LLM
Gemma-3	0	3.9	4.3	11.7	22.8	33.4	22.4	18.02
	1	2.3	4.2	11.4	20.7	30.6	27.0	24.9
	5	**4.1**	**5.5**	**15.4**	**27.8**	**43.1**	**38.1**	**42.3**
	15	3.9	4.5	12.1	23.6	33.9	37.6	37.2
LLaVA	0	1.7	2.6	7.2	14.5	20.2	30.8	33.5
	1	**4.6**	**4.4**	13.1	**23.2**	30.2	34.2	39.4
	5	2.9	1.5	18.6	12.5	41.6	40.1	**44.6**
	15	2.5	1.4	**19.0**	12.2	**44.5**	40.7	44.5
	30	3.0	1.6	**19.0**	12.7	43.7	**42.5**	**44.6**
Qwen-2.5	0	6.0	3.5	9.8	20.2	32.1	20.7	15.9
	1	5.0	4.4	12.4	24.5	36.1	28.1	25.1
	5	**54.8**	**5.9**	**16.1**	**31.7**	**45.4**	**39.6**	**39.6**
	15	46.2	4.9	14.0	27.8	39.6	32.8	28.8
ST	0	**13.6**	**2.5**	**8.9**	**15.1**	-	**34.9**	-
	1	4.8	1.4	6.9	8.9	-	28.6	-
	5	2.5	0.8	2.4	4.5	-	17.5	-

Table 4. Performance comparison of non-finetuned 2D VLLMs on the ScanQA and SQA3D datasets across varying numbers of frames per scene.

	ScanQA					SQA3D	
Model	C	B-4	M	R	LLM	EM@1	LLM
LLaVA	**3.0**	**1.6**	**19.0**	**12.7**	**43.7**	**42.5**	**44.6**
+ interleave	0.6	0.7	10.8	6.5	17.0	23.3	21.0
Qwen-2.5	**46.2**	**4.9**	**14.0**	**27.8**	**39.6**	**32.8**	**28.8**
+ interleave	45.4	4.8	13.6	27.2	31.0	31.9	27.0

Table 5. Comparison of LLaVA and Qwen when interleaving images with text on the ScanQA and SQA3D datasets.

6.4. Rep2D-BEAT

The results for Qwen-2.5-VL-7B, Gemma-3-it-27B, and SpaceThinker-Qwen-2.5VL-3B are visualized in Tables 6, 8, and 9.

Rep2D-B. Table 6 shows a general increase in IoU score on the bounding box task, when increasing the amount of input images. Specifically, going from a zero-shot (1 frame) to a 1-shot setting (2 frames) yields a significant increase in performance for both models. We also found that increasing the number of input images for this task may lead to more output artifacts, such as incorrectly formatted JSON. These faulty outputs might have influenced these observed results. These results on failure rate can be found in Appendix C.

By modifying the input configuration for the Rep2D-B task, Table 7 demonstrates the significance of including text-based example bounding boxes. Incorporating only example bounding boxes into the input yields a 375% increase in IoU score (from 0.04 to 0.19) compared to omitting them. Additional results for paceThinker-Qwen-2.5VL-3B, when utilizing Chain-of-Thought (CoT) reasoning, are provided in Appendix B. These results indicate lower performance

Images	Qwen-2.5	Gemma-3	ST
1	0.04	0.17	0.03
2	0.20	0.24	0.11
5	0.25	0.27	0.14
10	0.32	0.32	0.22

Table 6. Bounding box prediction (IoU) performance comparison of Qwen-2.5-VL-7B, Gemma-3-27B, and SpaceThinker-Qwen-2.5VL-3B (ST) on the Rep2D dataset. Rows indicate the number of images per scene and object which is used as input.

Setting	IoU
1 Image	0.04
+ 9 example bboxes	0.19
10 Images	0.06
+ 9 annotated bboxes	0.32

Table 7. Performance of Qwen-2.5-VL-7B on Rep2D-B with varying numbers of frames and inclusion of text-annotated bounding boxes. The first row in each setting represents performance using image-only inputs. The second row includes both images and corresponding text-annotated bounding boxes.

on Rep2D-B.

Rep2D-E. Table 8 shows an unexpected change in mean squared error when estimating the visible percentage of the target object. Adding more examples of this percentage, and the corresponding image where the target object is annotated with a bounding box, decreases performance compared to the 0-shot setting.

Images	Qwen-2.5	ST
1	0.153	0.453
2	0.188	0.629
5	0.184	0.605
10	0.166	0.611

Table 8. E (MSE) performance comparison of Qwen-2.5-VL-7B, and SpaceThinker-Qwen-2.5VL-3B (ST) on the Rep2D dataset. Rows indicate the number of images per scene and object used as input.

Rep2D-AT. The camera intrinsics vector prediction task yields expected results, as shown in Table 9. Increasing the number of input images annotated with camera intrinsics vectors leads to a lower L2 error when compared to the ground truth camera intrinsics vector. Furthermore, we observe a significant performance improvement when moving from a 0-shot to a 1-shot task on Qwen-2.5.

7. Discussion

3D awareness and number of frames. Our experimental results on the ScanQA and SQA3D benchmarks for different models differ from each other, but we see a consistent trend, which is that after a certain number of frames, the performance of the model drops. None of the models improve performance when provided with more than five frames with small exceptions on some specific metrics. Intuitively, this is unexpected because more frames would usually equip the models with additional visual context, which should improve performance in spatial reasoning tasks. However, the stagnation in the benchmark results suggests that the evaluated VLMs cannot effectively extract this additional information from larger sequences of images. We hypothesize that this limitation stems from the models not having been explicitly trained to handle multi-image prompts that extend beyond a few frames. As a result, they may lack the architectural experience or the training experience necessary to exploit added visual context, leading them to neglect potentially valuable information. Additionally, more views typically also imply more variation in the semantics of the scene, which the models would find difficult to process if they primarily rely on semantic cues to make judgements about the scene configuration. These can introduce noise in the model's understanding and have a detrimental effect on performance. This is also reinforced by the results we obtain on our own bounding box prediction dataset. Even when the models receive 10 images in total, the scores remain very low. In fact, except for SpaceThinker, the performance jump when going from 5 images to 10 is very low. For SpaceThinker, the performance improves substantially but it is still worse than the other models.

Images	Qwen-2.5	Gemma-3	ST
1	32.23	4.37	3.70
2	3.31	3.62	3.00
5	2.86	2.73	2.83
10	2.65	2.03	2.66

Table 9. Exposure prediction (L2) performance comparison of Qwen-2.5-VL-7B, Gemma-3-27B, and SpaceThinker-Qwen-2.5VL-3B (ST) on the Rep2D dataset. Rows indicate the number of images per scene and object used as input.

Unexpected Rep2D-E performance. Qwen-2.5 and SpaceThinker achieved lower MSE on Rep2D-E when no extra views were provided for additional context, compared to the scenarios when the models were provided extra 3D information. We see it decreases again as we increase the images, but the case with no extra images remains the best result. We hypothesize that this is because the model relies on misleading cues, which degrade performance. This strengthens our original hypothesis that these models lack a suitable mechanism to leverage meaningful 3D information from multiple views of the scene and instead rely on different cues, such as semantics, to make decisions.

Future research. Although our research introduced a

new dataset, time constraints limited the scope of exploration. For example, using this dataset for direct end-to-end optimization of a VLM—without incorporating additional layers—remains an unexplored direction. Furthermore, due to the aforementioned reason and limited computational resources, we only included a limited range of VLMs in our experiments. Lastly, including 3D VLMs as baselines on the Rep2D-BEAT tasks enables direct comparison with our evaluated 2D VLMs. These models are readily testable thanks to the availability of camera extrinsics and depth maps in our dataset, but were ultimately left out due to time and computational resource constraints.

8. Conclusion

This work addressed the research question: How effectively can general-purpose vision-language models infer 3D spatial structure and establish object correspondence across multiple 2D views without explicit 3D input? To this end, we introduced Rep2D-BEAT, a novel benchmark grounded in the Replica dataset, comprising three spatial tasks designed to probe different facets of implicit 3D reasoning using only 2D inputs. Our evaluation of state-of-the-art VLMs on Rep2D-BEAT, ScanQA, and SQA3D revealed that while some models exhibit latent spatial awareness, their ability to establish geometric consistency across views remains limited and inconsistent. Increasing the number of input frames yielded diminishing returns, and sometimes even a degradation in performance, suggesting that current VLMs are not fully leveraging multi-view information. Our results underscore a significant performance gap between 2D VLMs and task-specific or 3D-augmented counterparts, especially in complex spatial reasoning tasks. By providing a scalable and targeted benchmark, Rep2D-BEAT offers a practical pathway toward enhancing 3D understanding in general-purpose VLMs without the need for explicit 3D supervision, bridging the divide between semantic perception and geometric reasoning.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Azuma, D., Miyanishi, T., Kurita, S., and Kawanabe, M. Scanqa: 3d question answering for spatial scene understanding, 2022. URL https://arxiv.org/abs/2112.10482.

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., and Xia, F. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024.

Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., and Chen, T. Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning, 2023a. URL https://arxiv.org/abs/2311.18651.

Chen, S., Zhu, T., Zhou, R., Zhang, J., Gao, S., Niebles, J. C., Geva, M., He, J., Wu, J., and Li, M. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas, 2025. URL https://arxiv.org/abs/2503.01773.

Chen, Y., Viegas, F., and Wattenberg, M. Beyond surface statistics: Scene representations in a latent diffusion model. *arXiv preprint arXiv:2306.05720*, 2023b.

El Banani, M., Raj, A., Maninis, K.-K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., and Jampani, V. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Eppel, S., Bismut, M., and Faktor, A. Shape and texture recognition in large vision-language models, 2025. URL https://arxiv.org/abs/2503.23062.

Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., and Gan, C. 3d-llm: Injecting the 3d world into large language models, 2023. URL https://arxiv.org/abs/2307.12981.

Jampani, V., Maninis, K.-K., Engelhardt, A., Karpur, A., Truong, K., Sargent, K., Popov, S., Araujo, A., Martin-Brualla, R., Patel, K., Vlasic, D., Ferrari, V., Makadia, A., Liu, C., Li, Y., and Zhou, H. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *NeurIPS*, 2023. URL https://navidataset.github.io/.

Lavie, A. and Denkowski, M. J. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23:105–115, 2009.

Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., and Li, C. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024a. URL https://arxiv.org/abs/2407.07895.

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., and Liu, Y. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024b. URL https://arxiv.org/abs/2412.05579.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Liu, B., Dong, Y., Wang, Y., Ma, Z., Tang, Y., Tang, L., Rao, Y., Ma, W.-C., and Krishna, R. Coarse correspondences boost spatial-temporal reasoning in multimodal language model, 2024a. URL https://arxiv.org/abs/2408.00754.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.-C., and Huang, S. Sqa3d: Situated question answering in 3d scenes, 2023. URL https://arxiv.org/abs/2210.07474.

Mo, W. and Liu, Y. Bridging the gap between 2d and 3d visual question answering: A fusion approach for 3d vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4261–4268, 2024.

Nathan Silberman, Derek Hoiem, P. K. and Fergus, R. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

OpenAI. Gpt-4o. https://openai.com/gpt-4o, 2024. Accessed: 2025-05-29.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H. M., Nardi, R. D., Goesele, M., Lovegrove, S., and Newcombe, R. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019a.

Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H. M., Nardi, R. D., Goesele, M., Lovegrove, S., and Newcombe, R. The replica dataset: A digital replica of indoor spaces, 2019b. URL https://arxiv.org/abs/1906.05797.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A. M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A. S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C. L., Choquette-Choo, C. A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D. S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H. T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P. K., Culliton, P., Schmid, P., Sessa, P. G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S. R., Eiger, S., Zhang, S., Liu, T.,

Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L. G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., and Hussenot, L. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Wang, H., Zhao, Y., Wang, T., Fan, H., Zhang, X., and Zhang, Z. Ross3d: Reconstructive visual instruction tuning with 3d-awareness, 2025. URL https://arxiv.org/abs/2504.01901.

Wang, Z., Huang, H., Zhao, Y., Zhang, Z., and Zhao, Z. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes, 2023. URL https://arxiv.org/abs/2308.08769.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

Zhu, C., Wang, T., Zhang, W., Pang, J., and Liu, X. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024.

Zhu, C., Wang, T., Zhang, W., Pang, J., and Liu, X. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness, 2025. URL https://arxiv.org/abs/2409.18125.

Zuo, Y., Kayan, K., Wang, M., Jeon, K., Deng, J., and Griffiths, T. L. Towards foundation models for 3d vision: How close are we?, 2024. URL https://arxiv.org/abs/2410.10799.

A. Performance on Multi-Modal Understanding

To investigate the general abilities of state-of-the-art VLMs , we evaluate performance of Gemma-3-27B-IT (Team et al., 2025), LLaVA-Next-7B (Liu et al., 2024b), and Qwen2.5-VL-7B-Instruct (Qwen et al., 2025) on Massive Multi-discipline Multimodal Understanding (MMMU) (Yue et al., 2024) benchmark, a comprehensive benchmark for evaluating general multi-modal reasoning across disciplines. All models are evaluated in a zero-shot setting. Table 10 shows the breakdown of the results. Qwen2.5-VL-7B-Instruct shows excellent performance in multimodal understanding, while being while being much smaller than Gemma-3-27B-IT. Due to compute constraints, the evaluation was limited to this benchmark.

Model	**MMMU (val) Score**
Qwen2.5-VL-7B-Instruct	50%
LLaVA-Next-7B	36%
Gemma-3-27B-IT	51.3%

Table 10. MMMU benchmark scores for evaluated VLMs.

B. SpaceThinker with CoT

SpaceThinker demonstrated lower performance on our Rep2D-B task when employing Chain-of-Thought (CoT) reasoning in its output. On ScanQA, CoT demonstrated no significant performance gain or loss. A comparison with its performance without CoT reasoning is presented in Table 11.

Dataset	Setting	METEOR
ScanQA	SpaceThinker	2.4
	+ CoT	2.5
Dataset	Setting	IoU
Rep2D-B	SpaceThinker	0.22
	+ CoT	0.14

Table 11. Impact of Chain-of-Thought (CoT) reasoning on SpaceThinker-Qwen-2.5VL-3B performance on ScanQA (5 images) and Rep2D-B (10 images).

C. Failure to respond

We observed that the models used in our Rep2D-BEAT tasks frequently failed to respond correctly. They often produced invalid JSON, hallucinated responses, or repeated the same tokens. This failure rate is visualized in Figure 2.

D. ScanQA & SQA3D Prompts

Below, the prompt are listed which are used for the ScanQA and SQA3D datasets
ScanQA
If examples are provided:

```
'You are an AI with the ability to analyze a series of images, each representing a
    different perspective of a single scene. Your task is to construct a 3D understanding
    based on these images. A user will provide you with a query which you should answer.
    Output a very concise answer to the question; can be a single word or short phrase.
    Output digits as numbers, not words (e.g. 3, not three). '
```

If no example images are provided:

```
'You are an AI with the ability to answer questions about a single scene. Your task is to
    quess an answer which seems most likely. A user will provide you with a query which
    you should answer. Output a very concise answer to the question; can be a single word
    or short phrase. Output digits as numbers, not words (e.g. 3, not three).
```

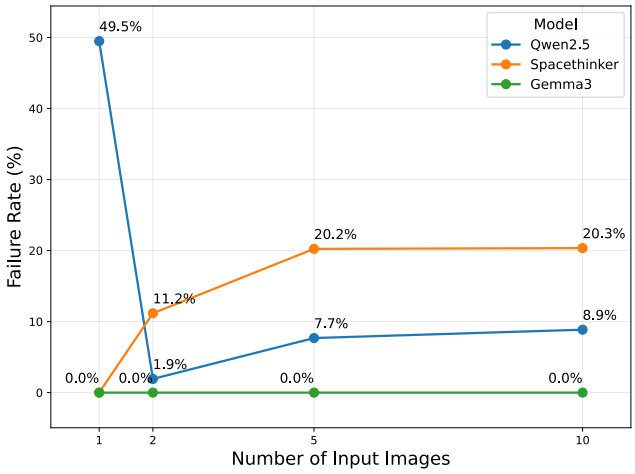


Figure 2. Model failure rate on the Rep2D-B task, measured by inconsistent or invalid responses across varying numbers of input images.

SQA3D

If examples are provided:

```
"You are an AI with the ability to analyze a series of images, each representing a
    different perspective of a single scene. Your task is to construct a 3D understanding
    based on these images. A user will provide you with a query and a position in the
    scene from which you should reason about the answer. Output a very concise answer to
    the question; can be a single word or short phrase. Output digits as words, not
    numbers (e.g. three, not 3). If the question is not answerable, output 'unknown'. "
```

If no example images are provided

```
"You are an AI with the ability to answer questions about a single scene. Your task is to
    quess an answer which seems most likely. A user will provide you with a query and a
    position in the scene from which you should reason about the answer. Output a very
    concise answer to the question; can be a single word or short phrase. Output digits as
     words, not numbers (e.g. three, not 3). If the question is not answerable, output '
    unknown'. "
```

E. Rep2D Prompts

Below, the prompt are listed which are used for all of our tasks.

Rep2D-B

```
"You are an AI with the ability to analyze a series of images, each representing a
    different perspective of a single scene. Your task is to construct a 3D understanding
    based on these images. A user will provide you with a set of images and 2D bounding
    boxes in each image of a single specfic object. The user will also provide a single
    image of the scene without a bounding box. Your task is to generate a 2D bounding box
    around the same object that can also be seen in this last image. Only output the
    bounding box without any other text. So do not give context about the input images or
    the bounding box. The bounding box is normalized to the image resolution, and the
    origin is the top left. For example the bottom right corner is (1000, 1000). The
    bounding box should be in a json format, exactly like: {'x_min': x_min, 'y_min': y_min,
     'x_max': x_max, 'y_max': y_max}. but where the actual values are filled in. "
```

Rep2D-E

```
'You are an AI with the ability to analyze a series of images, each representing a
    different perspective of a single scene. Your task is to construct a 3D understanding
    based on these images. A user will provide you with a set of images and percentages
    that the object in the bounding box is visible in each image. The user will also
    provide a single image of the scene without the bounding box for that same object.
    Your task is to generate this percentage of visibility of that specific object in the
    final image. Only output the percentage as a float without any other text. So do not
    give context about the input images or the percentage. '
```

Rep2D-AT

```
'You are an AI with the ability to analyze a series of images, each representing a
    different perspective of a single scene. Your task is to construct a 3D understanding
    based on these images. A user will provide you with a set of images and 3x1 (x,y,z)
    vectors that indicate the position of the camera in the room in each image. The user
    will also provide a single image of the scene without the corresponding camera
    position. Your task is to generate this camera position vector. Only output the camera
     position vector without any other text. So do not give context about the input images
     or the camera translation vector. '
```

F. LLM-as-a-judge prompt

```
You are an evaluator for LLM outputs. Compare the actual output with multiple possible
    expected outputs and determine semantic equivalence.

Input Question: <input question>
Actual Output: <answer prediction>
Expected Outputs (any of these is acceptable):
<expected outputs/labels>

Rate the semantic equivalence on a scale from 0 to 1, where:
- 0 means the output doesn't match any of the expected outputs in meaning
- 1 means the output perfectly matches at least one of the expected outputs in meaning

Your task is to determine whether the actual output conveys the same meaning as any of the
     expected outputs.
Score should be high if it matches well with ANY of the expected outputs.
```

G. Rep2D-B example

Below, in Figure 3, an example input is given to our Rep2D-B task, where the last image is unannotated and for which the model should generate a bounding box.

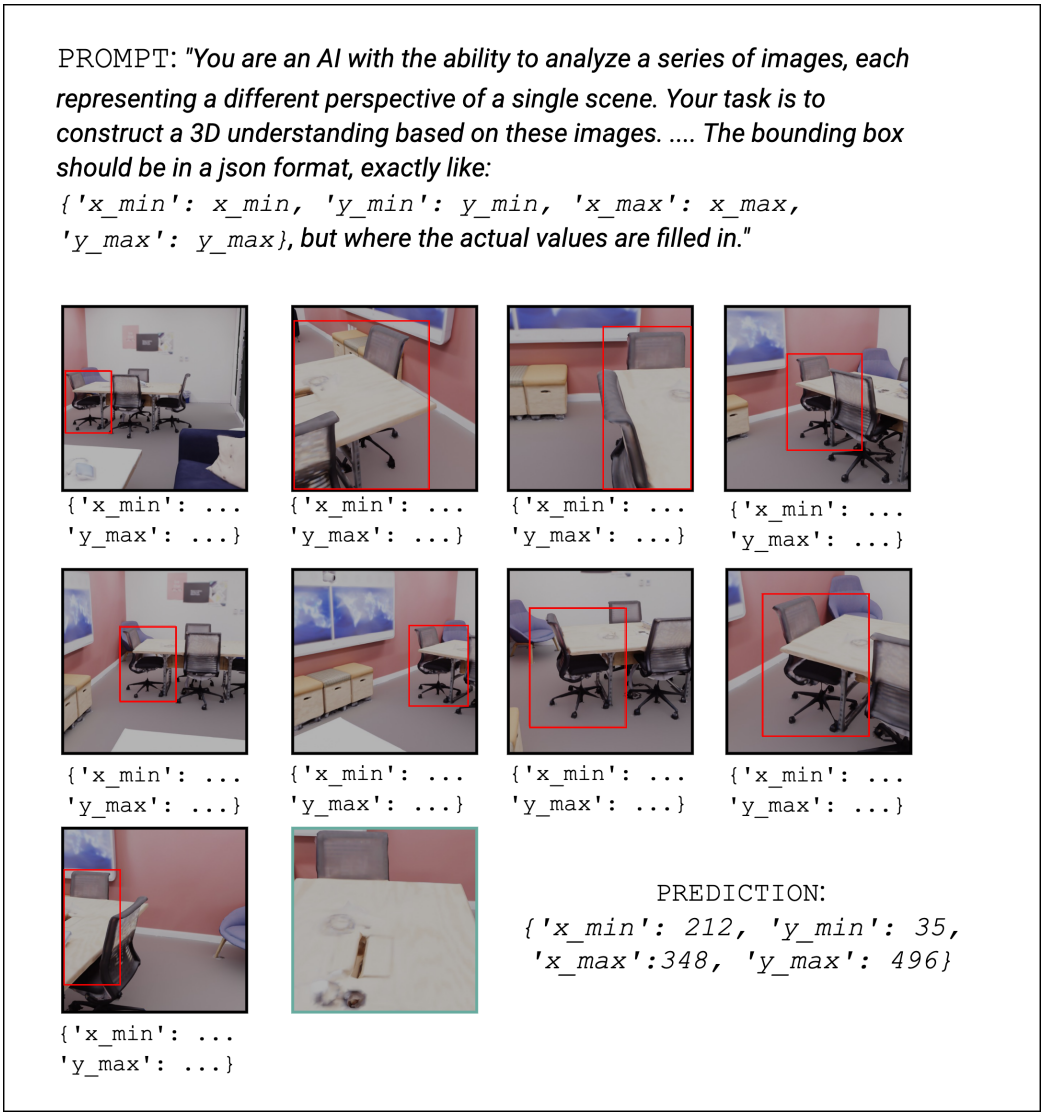
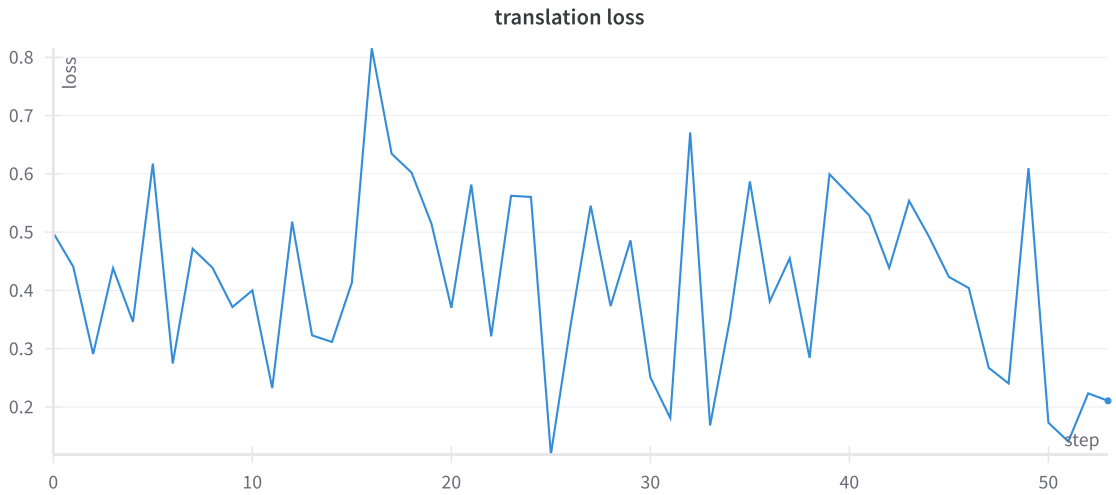


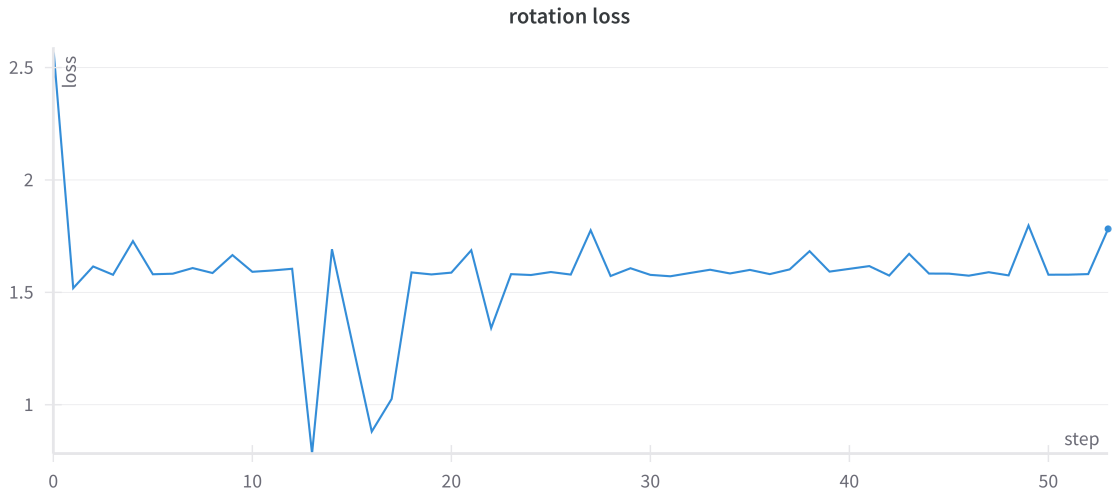
Figure 3. Example input for our Rep2D-B task.

H. Fine-tuning Setup and Observations

We fine-tuned LLaVA-NeXT (7B) to regress relative 3D poses between image pairs using a lightweight LoRA adaptation. Each sample included two images, a prompt and a position ⟨ image⟩⟨ image ⟩⟨ POS ⟩. A regression head predicted rotation and translation from the hidden state at the ⟨ POS ⟩ token. The CLIP vision encoder was frozen; only LoRA-adapted decoder layers and the pose head were updated. Unfortunately, total training loss plateaued at around 2. This may result from limited expressiveness of a single token representation, freezing of visual features, or mismatch between generative pretraining and regression objectives or general inability of LLMs to extract useful features suitable for regression.

Future work may explore pooling strategies, partial unfreezing, or intermediate alignment tasks.


translation loss
loss
0.8
0.7
0.6
0.5
0.4
0.3
0.2
step
0
10
20
30
40
50


rotation loss
loss
2.5
2
1.5
1
step
0
10
20
30
40
50


total loss
loss
3
2.5
2
1.5
step
0
10
20
30
40
50