Towards Foundation Models for 3D Vision: How Close Are We?

Yiming Zuo*, Karhan Kayan*, Maggie Wang, Kevin Jeon, Jia Deng, Thomas L. Griffiths
Princeton University
{zuoym,karhan,maggiewang,kevinjeon,jiadeng,tomg}@princeton.edu

Abstract

Building a foundation model for 3D vision is a complex challenge that remains unsolved. Towards that goal, it is important to understand the 3D reasoning capabilities of current models as well as identify the gaps between these models and humans. Therefore, we construct a new 3D visual understanding benchmark named UniQA-3D. UniQA-3D covers fundamental 3D vision tasks in the Visual Question Answering (VQA) format. We evaluate state-of-the-art Vision-Language Models (VLMs), specialized models, and human subjects on it. Our results show that VLMs generally perform poorly, while the specialized models are accurate but not robust, failing under geometric perturbations. In contrast, human vision continues to be the most reliable 3D visual system. We further demonstrate that neural networks align more closely with human 3D vision mechanisms compared to classical computer vision methods, and Transformer-based networks such as ViT [17] align more closely with human 3D vision mechanisms than CNNs. We hope our study will benefit the future development of foundation models for 3D vision. Code is available at https://github.com/princeton-vl/UniQA-3D.

1. Introduction

In recent years, impressive improvements in model accuracy and generalization ability on 2D vision tasks have been achieved with the introduction of *foundation models*. Vision-Language Models (VLMs) such as GPT4 [61] and LLaVA [42] can solve a wide range of visual understanding tasks, including Visual Question Answering (VQA) and image captioning on diverse datasets [48]. To thoroughly evaluate the performance of these foundation models, several benchmarks have been proposed [6, 20, 25, 40, 41, 50, 56]. Among them, a few benchmarks compare the VLM performance against humans [20, 26, 94]. Such comparison is crucial for understanding the robustness of the models and their alignment with human judgments in related tasks.

In comparison, foundation models and benchmarks are largely absent for 3D vision. On one hand, researchers often focus on training *specialized models* that can solve only a single task, such as depth estimation [36, 64] or optical flow [79]. On the other hand, each task has its own evaluation metrics and there is no benchmark for comparing a single model across different tasks, since the output space of each task is vastly different (*e.g.*, pixel-wise for depth estimation versus $SE(3)$ for camera pose estimation). The pixel-wise dense output required by many tasks also poses challenges in evaluating human performance, making it difficult to study and understand the differences and similarities between models and the human visual system.

To develop a foundation model for 3D vision, we must first understand the 3D vision capabilities of the existing models. Therefore, in this paper, we thoroughly evaluate the 3D understanding capabilities of a wide variety of models and focus on answering the following core questions:

- Do 2D VLMs show emergent 3D understanding capability allowing them to solve 3D tasks?
- What are the accuracy and robustness of specialized models trained on each task?
- Do humans remain the most accurate and robust 3D vision system? How do the error patterns of VLMs and specialized models compare to those of humans?

To answer the above questions, we construct a new 3D visual understanding benchmark UniQA-3D (**UNI**fied Visual-**Q**uestion-**A**nswering for **3D** Vision), which covers fundamental 3D vision tasks including depth estimation, spatial reasoning, camera pose estimation, and keypoint matching. While our benchmark is based on existing datasets with accurate ground truth, its key feature is a *unified output space* across tasks. We formulate all the questions in VQA format so that they can be easily answered by both VLMs and humans, enabling fair comparisons. For example, our depth estimation benchmark asks for binary depth relationships between two pixels rather than a dense depth map. Furthermore, we construct challenging cases with *geometric perturbation*, such as flipping or rotating the images, to make them diverge from the gravity-aligned views on which the models were trained. Those geometric perturbations are common in applications such as robotics,

*These authors contributed equally (random order).

making it important to test the robustness of models in such scenarios. Compared to existing benchmarks, UniQA-3D is the first to focus entirely on 3D, as shown in Tab. 1.

We conduct extensive experiments on UniQA-3D by evaluating a variety of models (Sec. 4). We test state-of-the-art VLMs including GPT4-Turbo, GPT4-Omni, and Gemini-1.5. We focus on closed-source VLMs because they have the best performance on the 2D leaderboards [6, 50]. We also test the performance of state-of-the-art specialized models on each task. Finally, we record human performance using Amazon Mechanical Turk (MTurk).

Here we present our key findings and our answers to the above questions:

Can 2D VLMs solve 3D tasks well? No. While VLMs achieve impressive performance on existing VQA benchmarks focusing on 2D, we find that they have *poor* 3D understanding abilities. None of the existing VLMs achieves human-level performance or performs on par with the specialized model. Surprisingly, on some tasks (*e.g.,* depth estimation), the VLMs perform only marginally better than random guessing, and on geometrically perturbed images, they perform even worse than random guessing.

Are specialized models accurate and robust? They are accurate but not robust. The specialized models have high accuracy in general, and interestingly we find that they are even better than humans on some tasks (*e.g.*, depth estimation and spatial reasoning) in the zero-shot setting. However, they are not as robust as humans against geometric perturbations. For example, on depth estimation, the accuracy of MiDaS [64] drops significantly (from 90.4% to 73.9%) on the upside-down images, while human accuracy remains the same (about 83%). Our finding suggests that the specialized models are still vulnerable, despite being trained on a large collection of diverse data [47, 64].

Are humans the most accurate and robust 3D visual system? How do the error patterns of models compare to those of humans? Yes, humans remain the most accurate and robust 3D visual system. Compared to humans, the error patterns of different models vary significantly according to the model types and architectures. Here are our findings: 1) specialized models have better alignment with humans compared to VLMs according to Cohen's κ [58]; 2) the error patterns of neural networks are more similar to humans than hand-crafted algorithms such as SIFT [53] in the context of keypoint detection and correspondence matching; and 3) in terms of model architecture, the error patterns of the Transformer-based ones (*e.g.*, ViT [17]) are more similar to humans compared to the CNN-based ones. While Tuli *et al.* [82] found that Transformers are more similar to humans on 2D classification tasks than CNNs, the generalization of this finding to 3D is valuable and non-trivial. This may be because the mechanisms for processing 2D and 3D in the human brain are different: 2D vision is handled by the ventral stream in the brain while 3D vision is performed by the dorsal stream [16, 88].

Table 1. UniQA-3D has comparable size (in terms of the number of images) to modern benchmarks for testing VLMs, such as MMBench [50]. Moreover, UniQA-3D provides human accuracy and contains challenging samples with geometric perturbation.

Benchmarks	Modality	#Images	Human Acc.	Geom. Perturb
VQA [1]	2D	250,000	Yes	No
MSCOCO [46]	2D	328,000	No	No
MMVet [93]	2D	200	No	No
MMBench [50]	2D	2,590	No	No
SeedBench [41]	2D	19,242	No	No
VisIT [6]	2D	592	No	No
BLINK [20]	2D+3D	7,358	Yes	No
UniQA-3D (Ours)	3D	2,450	Yes	Yes

Our main contributions are as follows:

- We propose a new benchmark UniQA-3D with a unified output space for fair comparison of the 3D understanding capability of different models and humans.
- We evaluate the performance of state-of-the-art VLMs, specialized models, and humans on our benchmark, and we compare the error patterns of different models against humans under multiple criteria.

We hope our results will benefit the future development of foundation models for 3D vision by providing insights for improving their robustness and ability to generalize.

2. Related Work

2.1. VLMs and Benchmarks

Vision Language Models (VLMs) are typically trained on large-scale datasets of text-image pairs and achieve impressive performance and generalization on image understanding [6, 56]. There are both open-sourced VLMs such as LLava [42], CogVLM [87], and MiniCPM [30], and closed-sourced VLMs that one can only access through paid API, such as GPT4 [61], Claude [3], and Gemini [78]. In our evaluation, we focus mainly on closed-sourced models because they have better performance in general [6].

Several works combine VLMs with 3D understanding. For instance, 3D-LLM [29] and 3D-VisTA [96] train an LLM that can take 3D point clouds as input and complete high-level tasks (*e.g.*, navigation) by leveraging paired 3D-language datasets. In contrast, our paper focuses on benchmarking existing models instead of training. Additionally, we use 2D images as input modality and focus more on low-level 3D vision tasks. Banani *et al.* [19] trains linear probes on the features of large models to solve 3D tasks such as depth estimation and matching. Compared to them, evaluating on our UniQA-3D requires no training and can be applied to closed-sourced VLMs and humans.

Various benchmarks have been proposed to test the scene understanding and visual question answering capability of 2D VLMs [6, 20, 20, 25, 40, 41, 50, 56]. Datasets from a decade ago such as MSCOCO [46] and VQA [1] are large in scale but lack robust and fine-grained evaluation. To resolve this, MMBench [50] proposes an objective evaluation scheme with fine-grained classes of questions. VisIT [6] covers 70 families of instructions and proposes an automatic LLM-based evaluation aligned with human preferences. WildVision [56] generates open-ended questions automatically by leveraging classification datasets. All of these datasets focus on testing the 2D capability of VLMs, whereas UniQA-3D focuses entirely on 3D.

BLINK [20] evaluates on a diverse set of tasks. Some of the tasks are in 3D (*e.g.,* relative depth), while others are in 2D (*e.g.,* IQ test). In contrast to BLINK, our paper focuses entirely on 3D, covering more 3D tasks such as keypoint detection and including more challenging samples with geometric perturbations. Moreover, BLINK evaluates human performance on only 2 subjects (the coauthors), whereas we evaluate human performance with 162 subjects in depth estimation, 109 subjects in camera pose estimation, 449 subjects in VQA, and 143 subjects in keypoint matching. The large sample size of our UniQA-3D helps us discover statistically significant differences between models and humans.

2.2. Cognitive Science with Neural Networks

Cognitive scientists have extensively studied the relationship between computer vision models and the human visual system in the context of 2D vision, particularly object recognition [13, 38, 63, 72, 89]. However, a similar comparison is missing for 3D vision. Most research on the human 3D visual system has focused on Marr's implementation level [57], such as the discovery of grid cells, place cells, and head-direction cells explaining which parts of the brain activate for visual localization [5], or the stereogram and fMRI based works explaining which parts of the brain activate for depth perception [88]. However, these works are unlikely to provide direct insight into constructing 3D foundation models.

Our paper aims to provide such insight by quantitatively comparing the performance and error patterns of 3D vision models to human subjects. In the context of object recognition, Geirhos et al. [24] introduced Cohen's κ analysis in order to compare neural network error patterns to humans, and Tuli et al. [82] showed that human error patterns are more similar to ViTs than CNNs. However, the brain understands 2D and 3D through two different streams [16, 88], suggesting that these findings might not be simply extrapolated to 3D vision. Our findings show that human error patterns are, indeed, more similar to Transformers than CNNs in the case of depth perception. This is surprising considering the biological plausibility of convolution [21, 66, 89], particularly for depth perception [60]. Perhaps more surprisingly, we find that neither Transformers, nor CNNs perform camera pose estimation similar to humans, with the highest Cohen's κ being 0.16.

2.3. Specialized 3D Vision Models

Monocular Depth Estimation. MiDaS [64] is the first to explore large-scale training on a mixture of datasets, followed by more recent works [36, 90]. Since absolute depth is challenging to predict, DIW [11] explores using humans to annotate the relative depth between two pixels. Our relative depth prediction task is inspired by DIW, but our relative depth ground-truth labels are generated from Lidar sensors rather than from human annotators, making them more reliable. We use MiDaS [64] in this study because it provides both CNN-based and ViT-based variants.

Visual Question Answering (VQA). Tranformer-based VQA methods can be classified into single-stream [12, 43, 44, 95] and two-streams [54, 55, 75, 77], depending on how the images and questions are processed. MDETR [35] explicitly detects the objects and fuses the vision and language stream with a Transformer. We use MDETR as the specialized model for its state-of-the-art accuracy.

Many datasets have been proposed as VQA benchmarks: some focus on semantics [1, 34, 65, 73, 85] while others focus on spatial relationship reasoning [31, 33, 92]. We build our benchmark on top of the CLEVR [33] dataset.

Camera Pose Estimation aims to estimate the 6DoF pose (translation and rotation) of a camera either with respect to a global coordinate system or relative to another camera. Matching-based methods estimate the relative pose using matched keypoints [49, 59, 62, 69, 69, 71, 80, 81], whereas pose regression methods [2, 9, 10, 37, 39, 74, 84] estimate the 6DoF pose directly based on the input images. Our experiments show that none of the models classify pose similar to humans with the highest Cohen's κ score being 0.16.

Keypoint Detection involves detecting salient keypoints from an image, typically followed by local feature extraction around those keypoints. Classical computer vision methods such as [27, 53, 67, 68] extract keypoints and features using local information such as image gradients. In contrast, recent deep neural network-based methods such as [4, 15, 18, 83] train CNNs to detect keypoints and extract local features. Our results suggest that humans follow a keypoint detection strategy more similar to neural networks than classical methods.

Keypoint Matching matches a pixel to the corresponding pixel in another view. Keypoint matching methods can be grouped into two classes: detector-free methods [8, 76, 79, 86], which perform dense matching and avoid the keypoint detection phase; and detector-based methods, which rely on a keypoint detector and local features extracted from those keypoints. Detector-based methods can be further divided

Table 2. Statistics of UniQA-3D. Our benchmark has 4 sub-tasks and is built on top of existing datasets with high-quality annotations. †: We train custom models with the ResNet [28], ViT [7], and Swin [51] backbone. ‡: For keypoint detection, we test SIFT [53], FAST [67], and SuperPoint(SP) [15]; for keypoint matching, we test ORB [68] and LightGlue [47].

Task	Data Source	#Images	Specialized Models
Relative Depth	KITTI [22]	750	MiDaS [64]
Spatial Reasoning	CLEVR [33]	500	MDETR [35]
Camera Pose	DTU [32]	750	Custom†
Keypoint-Matching	Megadepth [45]	450	SIFT,FAST,SP ORB,LighGlue‡

into two subclasses: classical methods [27, 52, 53, 67, 68], which use k-Nearest Neighbors in the feature space; and deep learning-based matchers [7, 47, 70], which train neural networks to match the extracted keypoints. Our experiments show that human keypoint matching is more similar to neural network-based methods than classical methods.

3. The UniQA-3D Benchmark

We develop a new benchmark UniQA-3D for 3D vision tasks based on existing public datasets (Tab. 2). The key feature of our benchmark is a *unified output space* for all sub-tasks, *i.e.,* we form all questions to be multiple-choice so that all models and humans can be easily and fairly compared. In addition to the specialized models, we collect the response from VLMs (GPT4-Turbo, GPT4-Omni, and Gemini-1.5) and humans.

We formally define the four tasks below. We use the term "subject" to refer to either a human or a model.

3.1. Relative Depth Estimation

Task Definition The subject is provided a single image with two markers annotating two pixels in the image. The subject is asked to determine which pixel is closer to the camera. See Fig. 1a for an example.

Dataset We use KITTI [23] for the relative depth estimation analysis. KITTI is a real-world autonomous driving dataset, and we choose it for two reasons: 1) KITTI has rich and accurate annotations, including accurate depth collected by Lidar, as well as semantic segmentation labels. 2) KITTI images contain both natural and man-made objects, providing a high diversity of visual content. Following DIW [11], we sample markers with a 50% probability of being placed either randomly or symmetrically along a horizontal line. We collect 500 regular images and another 250 geometrically perturbed images by flipping them upside-down.

3.2. Spatial Reasoning

Task Definition The format of the questions follows the VQA task, but they specifically require the subject to reason about the spatial relationships among objects in a scene. See Fig. 3a for an example.

Dataset We use a subset of CLEVR [33], which is a synthetic dataset with complex spatial relationships. We use the ground-truth questions from CLEVR but only keep the questions that contain certain spatial keywords (*e.g,* left, right, front, behind, top, bottom). Note our selected subset is significantly more challenging than the full CLEVR dataset: MDETR [35] achieves 99.7% accuracy on the original CLEVR test set but only 74.4% accuracy on our dataset. Our dataset contains 500 image-question pairs in total.

3.3. Relative Camera Pose Estimation

Task Definition We give the subject two views of the same scene with an object of focus. The subject is then asked to choose the most prominent motion of the camera, *i.e.,* move left, right, down, or up.

Dataset We use DTU [32], which has accurate ground-truth for the poses of the camera mounted on a robotic arm. Out of the original 49 views, we randomly choose view pairs but ensure a prominent movement axis by setting a threshold on the ratio of the major movement axis to the minor movement axis. This ensures that the movement between the two views is clear and pronounced along a single axis, making the task easier to evaluate by both models and human subjects. We sample 500 image pairs and an additional 250 upside-down image pairs as geometric perturbations.

3.4. Keypoint Matching

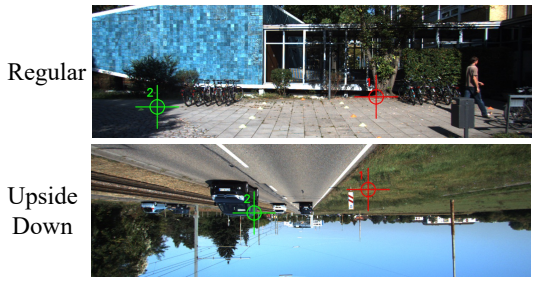
Task Definition We give the subject two views of a scene. We then ask the subject to choose five points in one image and match the corresponding points in the other image.

Dataset We use the Megadepth-1500 dataset [45, 76]. The scenes are historic locations with wide-baseline images. It provides accurate depth and camera pose ground-truth. For the keypoints chosen by the subject, we use the depth and camera pose to find the ground-truth correspondence in the other image, and we run additional forward-backward consistency checks to ensure co-visibility. We randomly sample 450 image pairs and ask the human subject to annotate 5 keypoint pairs on each, resulting in 2,250 keypoint pairs.

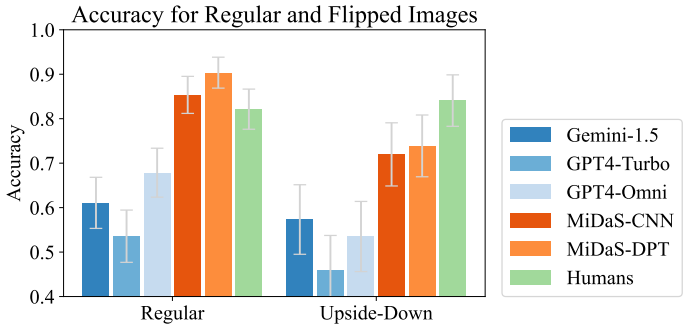
4. Experiment Results and Analysis

4.1. Human Annotation Collection

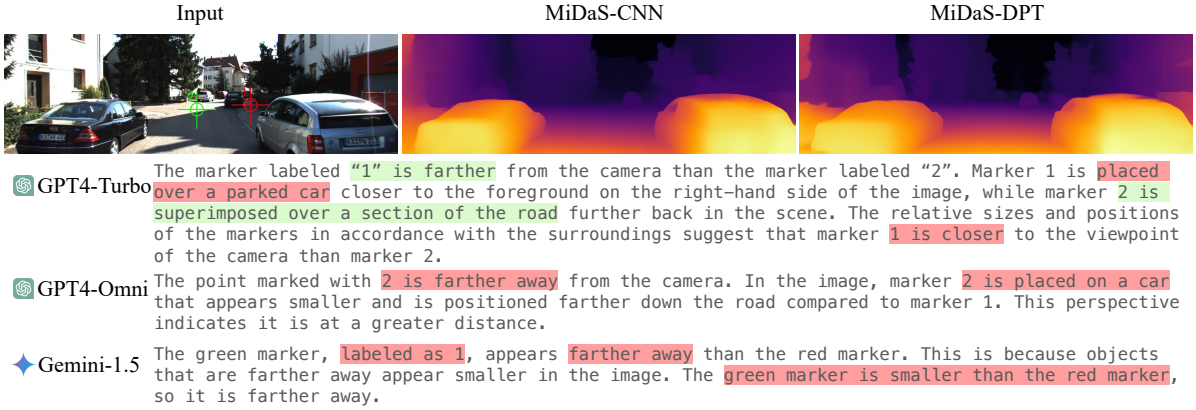
We collected human annotations using Amazon Mechanical Turk (MTurk) with IRB approval. Since there are bots/spammers on MTurk, it is important to filter them out to ensure high answer quality. We maintain strict quality control by: 1) requiring the HIT approval rate $\geq$ 95% and that the number of HITs $\geq$ 1000; 2) when possible, structuring the multiple-choice questions to require clicking in designated locations, such as asking users to click inside a



(a) Each image has two markers and the subject decides which one is closer to the camera.



(b) Accuracy on regular and upside-down images. MiDaS has comparable performance to humans while VLMs perform poorly.



(c) Output visualization. For VLM responses, we mark the correct statements green and the incorrect ones red. GPT4-Turbo and GPT4-Omni wrongly localize the markers, while Gemini-1.5 generates self-contradicting answers.

Figure 1. (a) We sample images from the KITTI dataset and flip to create upside-down images. (b) Comparison of accuracy of different methods. MiDaS-DPT works the best in general, and both MiDaS models are slightly better than humans. All the VLMs perform poorly, with GPT4-Omni performing the best on regular inputs. (c) VLMs have multiple failure modes. See text for details.

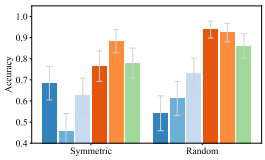
checkbox painted on the image. This allows us to filter out responses that do not follow the requested format; and 3) leveraging consensus scoring by assigning each HIT to 3 different users and only considering the result valid when all 3 users provide the same answer.

4.2. Relative Depth Estimation

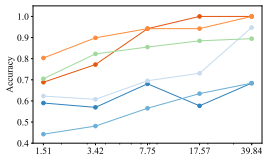
Models Compared We use the same prompt for the VLMs as for the humans: "There are two markers on the image. Which is farther away from the camera?" We also compare two variants of state-of-the-art specialized depth estimation model MiDaS [64], *i.e.*, MiDaS-CNN and MiDaS-DPT. We extract the relative depth relationship from the dense depth predictions. Note that MiDaS is not trained on KITTI.

Model Accuracy The primary results are shown in Fig. 1b. The two MiDaS models achieve the best performance, with the Transformer variant (MiDaS-DPT) being slightly better than the CNN variant. Surprisingly, both MiDaS variants outperform humans. This shows that state-of-the-art neural network models have competitive 3D understanding capabilities when trained on a large dataset, and can beat humans in unseen environments for 3D understanding tasks.

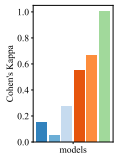
All VLMs have significantly lower accuracy compared



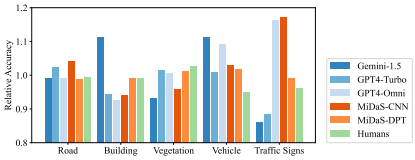
(a) The model accuracy on symmetrically and randomly sampled pixel pairs.



(b) The model accuracy against the absolute difference in depth (meters) between the paired pixels.

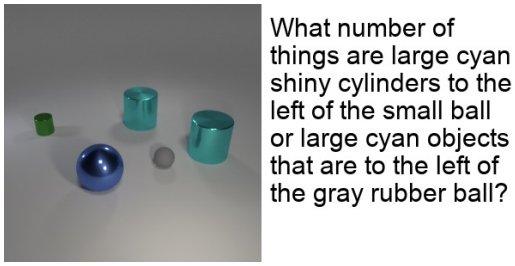


(c) Cohen's κ against humans.

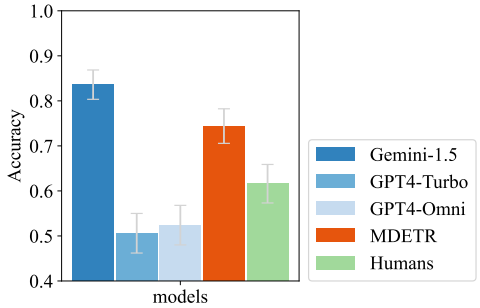


(d) Accuracy against different semantic labels. The y-axis is normalized by the overall accuracy of each model.

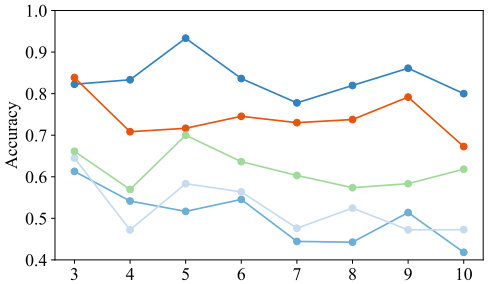
Figure 2. We compare the similarity between humans and different models using different metrics, including (a) pair sampling strategy, (b) relative depth difference, (c) Cohen's κ, and (d) semantic labels. Best viewed zoomed-in and in colors.



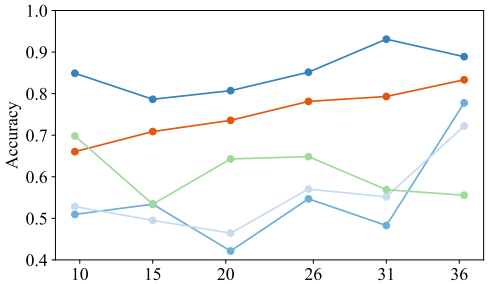
(a) Example image and question pair.



(b) Accuracy on different models and humans.



(c) Accuracy against the number of objects in the scene.



(d) Accuracy against the length of the question.

Figure 3. Results on the spatial reasoning task. (a) Our benchmark requires a strong spatial reasoning ability and is very challenging. (b) Even the specialized VQA model MDETR can only achieve 74.4% accuracy. (c) model accuracy drops as the scene complexity grows (more objects). (d) longer questions don't necessarily lead to worse performance. See text for detailed analysis.

to MiDaS and humans, where the best-performing GPT4-Omni only achieves 67.9% accuracy. We visualize the output of different methods in Fig. 1c. We find several failure modes in the VLM outputs: 1) *localization*: markers placed on the road are classified as placed on a car; 2) *scene understanding*: markers are understood as real objects in the scene and the depth is reasoned using relative size; and 3) *reasoning*: the response is self-contradicting.

Robustness Results are shown in Fig. 1b. The accuracy of all machine learning models drops significantly on the flipped images compared to regular ones (GPT4-Omni: 67.9% $\rightarrow$ 53.5%; MiDaS-DPT: 90.4% $\rightarrow$ 73.9%). In contrast, the human performance remains on par (82.1% $\rightarrow$ 84.1%) and is better than all models, showing the superior robustness of the human depth perception system.

Alignment with Humans We measure the similarity of the answers of different models to the answers of humans, both qualitatively and quantitatively, as show in Fig. 2.

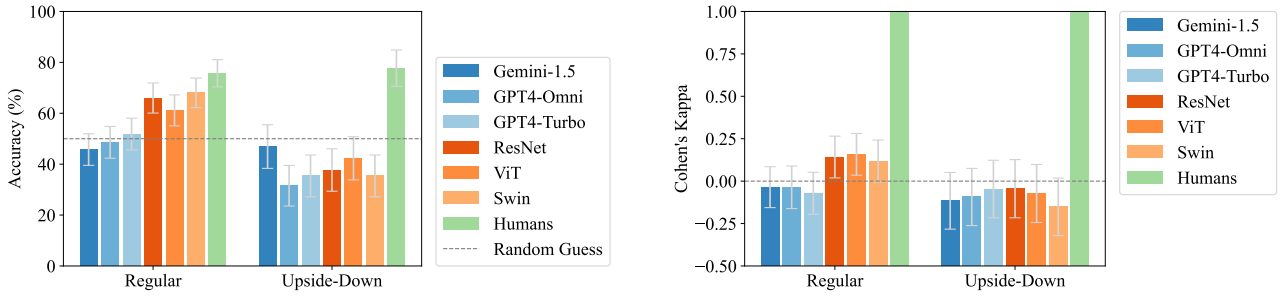
Quantitative comparisons are shown in Fig. 2c. We use Cohen's κ [58] to compare the consistency between each model and humans. Note that Cohen's κ rules out the effect of the model's accuracy, *i.e.,* MiDaS won't have a higher alignment score just because it has higher overall accuracy. MiDaS-DPT achieves the best consistency, with Cohen's κ of 0.66, followed by MiDaS-CNN of 0.56. All VLMs have very low consistency with humans.

Qualitatively, we compare the distribution of the accuracy over different factors. In Fig. 2a, we compare the accuracy on symmetric and randomly sampled depth point pairs. The performance gap between the two cases is larger for the MiDaS-CNN compared to humans and MiDaS-DPT, showing the stronger alignment between Transformers and humans. Fig. 2b shows the accuracy against the difference in depth (meters) for the two-point sampled. The pairs with larger differences in depth are easier to tell apart, so slopes are positive. Comparing the shape of the curve, MiDaS-DPT is more similar to humans than MiDaS-CNN. Finally, in Fig. 2d, we compare the accuracy of each model on different classes. We merge the original KITTI classes into 5 super-classes. Note the accuracies are normalized by the overall accuracy of each model. The pattern of MiDaS-DPT is the most similar to humans, whereas MiDaS-CNN performs significantly better on traffic signs and worse on buildings and vegetation. In conclusion, MiDaS-DPT has the error pattern most similar to humans under multiple criteria, suggesting that Transformers are more similar to humans than CNNs in depth perception.

4.3. Spatial Reasoning

Models Compared In addition to the three VLMs and humans, we evaluate against MDETR [35] as the specialized model on our benchmark. MDETR is a state-of-the-art VQA model trained on CLEVR [33].

Model Accuracy The overall accuracy of different mod-



(a) Relative camera pose classification accuracy.

(b) Cohen's κ similarity to human relative pose classification.

Figure 4. Comparison between specialist neural networks, VLMs, and humans on relative camera pose classification. The bars are 95% confidence intervals.

els and humans are shown in Fig. 3b. Human accuracy is 61.6%. The accuracy is low for two reasons: 1) Our benchmark is challenging, requiring reasoning through a long logical chain. See Fig. 3a for an example. 2) While we believe humans can do better if they pay full attention and are given enough time, the numbers we report are the "average" humans on MTurk, instead of the performance upper bound.

Among the VLMs, Gemini-1.5 achieves an accuracy of 83.6%, which is surprisingly even better than MDETR (74.4%) which is trained on the CLEVR training set. Although the 3D understanding capability of VLMs may be limited as shown in the other tasks, their reasoning capability seems to be very strong, achieving good performance on the reasoning-oriented VQA task. The two GPT4 models perform relatively poorly, with GPT4-Omni having a slightly better accuracy of 52.4%.

Alignment with Humans We compare the accuracy of different models against the scene complexity, measured by the number of objects in the scene. Results are shown in Fig. 3c. The accuracy of all models drops as the scene complexity increases, while human accuracy is not much affected. We also measure the complexity of the question. While it is difficult to define the exact complexity, we use the number of words in a sentence as a proxy. The results shown in Fig. 3d are quite counter-intuitive: the accuracy of all models grows as the question becomes more complex. In contrast, human accuracy drops slightly. None of the models show a strong correlation with humans, highlighting the potentially different ways that models and humans approach the spatial reasoning task.

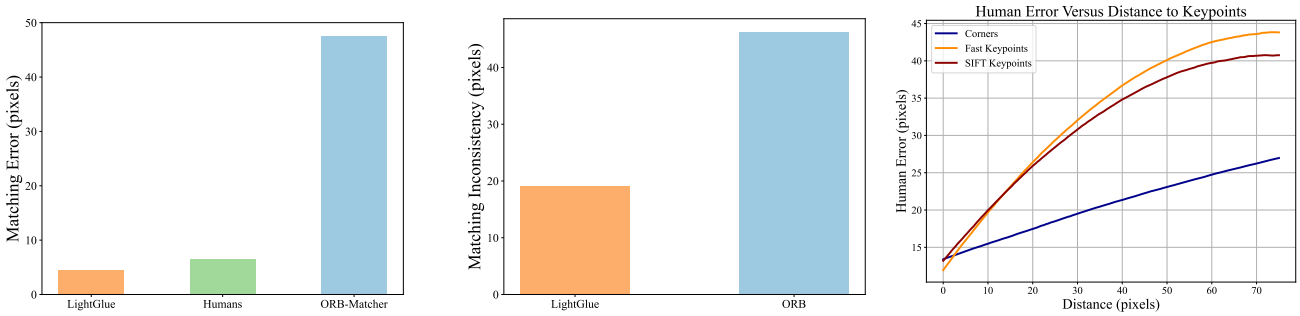
4.4. Relative Camera Pose Estimation

Models Compared We train our custom specialized models on the BlendedMVS [91] dataset. We use ResNet-50 [28], ViT [17], and Swin Transformer [51] with pretrained ImageNet [14] weights as our neural network backbone for comparison. We provide the most prominent movement axis (up/down vs. right/left) to the subject as input and formulate the question as a two-way classification task. We ask the same question to both human subjects and VLMs.

Model Accuracy The main results are shown in Fig. 4. Humans achieve the highest accuracy (75.7%) followed by the specialized model with a Swin Transformer backbone (68%). Interestingly, the VLM models perform approximately the same as random guess, the most accurate one being GPT4-Turbo with 51.8% accuracy.

Robustness Although specialized models have similar accuracy to humans with regular images, they perform significantly worse than random guess when the images are flipped (ResNet: 65.9% $\rightarrow$ 37.7%; ViT: 61.1% $\rightarrow$ 42.3%; Swin: 68% $\rightarrow$ 35.4%). This observation holds for VLMs as well (GPT4-Omni: 48.6% $\rightarrow$ 31.5%; GPT4-Turbo: 51.8% $\rightarrow$ 35.4%), with the exception of Gemini-1.5 (45.7% $\rightarrow$ 46.9%) which has similar accuracy. None of the VLMs perform better than random guess when the images are flipped. In contrast, the human accuracy is almost invariant to geometric perturbation (75.7% $\rightarrow$ 77.7%). This suggests that the human visual system is way more robust compared to both VLMs and specialized models in the context of camera pose estimation.

Alignment with Humans We measure Cohen's κ to compare the models in terms of how similar they are to humans in the answers they output. Fig. 4b shows that ViT is slightly more similar to humans than CNNs in this task, although the 95% confidence intervals highly overlap. VLMs have negative Cohen's κ with humans, showing that their answers are highly dissimilar even controlling for accuracy. When the images are flipped, all models have negative Cohen's κ, which demonstrates that all models are dissimilar to humans in terms of output under geometric perturbation. In conclusion, none of the models significantly align with humans, suggesting that they are unsuitable as models of human perception in this task.



(a) Matching error made by humans, LightGlue, and ORB matcher measured EPE.

(b) Matching inconsistency score with humans. LightGlue is more similar to human matching.

(c) Human error and distance to detector keypoints.

Figure 5. Matching experiment results. Transformer-based LightGlue is more similar to human matching than the classical ORB matcher.

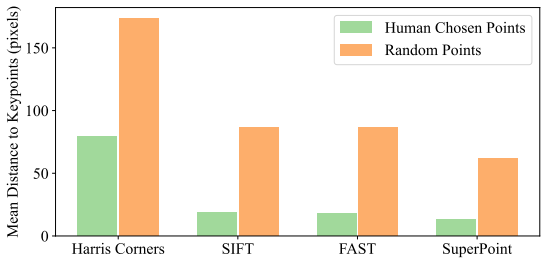


Figure 6. The average distance of human chosen vs random keypoint to the nearest detector keypoint. Humans are more likely to choose salient points than randomly guess.

4.5. Keypoint Matching

Models Compared The keypoint matching task is further divided into two sub-tasks, *i.e.,* keypoint detection and correspondence matching, each of which typically use specialized models [15, 70].

- **Keypoint Detection** We compare against a set of classical hand-cratfed detectors, including Harris Corner Detector [27] with the top k corners, Difference of Gaussians (DoG) used in SIFT descriptors [53], and the FAST detectors [67] used in ORB descriptors [68]. We also compare the neural network based detector SuperPoint [15]. All detectors are configured to choose the same number of best candidates for a fair comparison.
- **Matching** We use LightGlue [47] as our deep learning based method and ORB-matcher as our classical vision based method. LightGlue uses DISK [83] for feature description, and uses Transformer as its backbone.

Model Accuracy We evaluate human matches, LightGlue, and ORB matching on the average end-point error (EPE) from the ground truth correspondence. Fig. 5a shows that LightGlue makes the least matching errors, followed closely by humans, while the ORB matcher is much less accurate, suggesting that the state-of-the-art can achieve comparable performance to humans.

Alignment with Humans

- **Keypoint Detection** We find that humans tend to choose salient points. Fig. 6 shows that the subjects were much more likely to choose near corners, SIFT keypoints, FAST keypoints, or SuperPoint keypoints than random choice. Among the detectors, SuperPoint aligns the most with humans. Thus, the neural network model acts more similar to humans than hand-crafted detectors.
- **Matching** We evaluate the consistency of LightGlue and ORB matcher with human matches. We measure consistency by computing the average difference in EPE made by humans and models on each keypoint. As shown in Fig. 5b, LightGlue (diff=19.0px) is more consistent with human matches compared to ORB (diff=44.3px). This suggests that human 3D visual system is more similar to a Transformer-based neural network architecture that performs global reasoning compared to a classical algorithm that keeps track of local statistical information.
  Human subjects make more matching errors as they deviate farther from salient regions, regardless of the saliency metric used (corners, FAST, or SIFT; Fig. 5c). This finding is consistent with the common belief that pixels with rich textures such as corners are easier to match.
  Linear regression analysis shows that around 15% of the variance in human error is explained by the distance from SIFT and FAST keypoints with $p < 0.001$. We also find that approximately 10% of the variance is explained by the distance from a corner with $p < 0.05$.

5. Limitations

While we focus on four fundamental 3D vision tasks (relative depth, spatial reasoning, camera pose estimation, and keypoint matching), evaluating additional tasks like surface normals and open-source VLMs would provide a more comprehensive analysis.

Acknowledgements

This work was partially supported by the National Science Foundation (2331763). We thank Thomas A. Langlois for insightful discussions.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2, 3

[2] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition, May 2016. arXiv:1511.07247 [cs]. 3

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022. 2

[4] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters, October 2019. arXiv:1904.00889 [cs]. 3

[5] C. Barry and N. Burgess. Neural mechanisms of self-location. *Current biology: CB*, 24(8):R330–339, April 2014. 3

[6] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023. 1, 2, 3

[7] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to Match Features with Seeded Graph Matching Network, August 2021. arXiv:2108.08771 [cs]. 4

[8] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. ASpanFormer: Detector-Free Image Matching with Adaptive Span Transformer, August 2022. arXiv:2208.14201 [cs]. 3

[9] Shuai Chen, Xinghui Li, Zirui Wang, and Victor Adrian Prisacariu. DFNet: Enhance Absolute Pose Regression with Direct Feature Matching, July 2022. arXiv:2204.00559 [cs]. 3

[10] Shuai Chen, Zirui Wang, and Victor Prisacariu. Direct-PoseNet: Absolute Pose Regression with Photometric Consistency, October 2021. arXiv:2104.04073 [cs]. 3

[11] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. 3, 4

[12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 3

[13] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):27755, June 2016. Publisher: Nature Publishing Group. 3

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919. 7

[15] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description, April 2018. arXiv:1712.07629 [cs]. 3, 4, 8

[16] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, February 2012. 2, 3

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs]. 1, 2, 7

[18] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, May 2019. arXiv:1905.03561 [cs]. 3

[19] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 2

[20] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 1, 2, 3

[21] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980. 3

[22] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The kitti dataset. *Int. J. Rob. Res.*, 32(11):1231–1237, sep 2013. 4

[23] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 4

[24] Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency, December 2020. arXiv:2006.16736 [cs, q-bio]. 3

[25] Simon Ging, María A Bravo, and Thomas Brox. Open-ended vqa benchmarking of vision-language models by exploiting classification datasets and their semantic hierarchy. *arXiv preprint arXiv:2402.07270*, 2024. 1, 3

[26] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[27] Christopher G. Harris and M. J. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988. 3, 4, 8

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. arXiv:1512.03385 [cs]. 4, 7

[29] 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. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023. 2

[30] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. 2

[31] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3

[32] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014. 4

[33] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 3, 4, 6

[34] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*, pages 1965–1973, 2017. 3

[35] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 3, 4, 6

[36] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation, December 2023. 1, 3

[37] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization, February 2016. arXiv:1505.07427 [cs]. 3

[38] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, 10(11):e1003915, November 2014. Publisher: Public Library of Science. 3

[39] Zakaria Laskar, Iaroslav Melekhov, Surya Kalia, and Juho Kannala. Camera Relocalization by Computing Pairwise Relative Poses Using Convolutional Neural Network, August 2017. arXiv:1707.09733 [cs]. 3

[40] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 1, 3

[41] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1, 2, 3

[42] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M^3 it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023. 1, 2

[43] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[44] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 3

[45] Zhengqi Li and Noah Snavely. MegaDepth: Learning Single-View Depth Prediction from Internet Photos, April 2018. 4

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3

[47] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed, June 2023. arXiv:2306.13643 [cs]. 2, 4, 8

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

[49] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2391–2400, 2017. 3

[50] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 1, 2, 3

[51] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, August 2021. arXiv:2103.14030 [cs]. 4, 7

[52] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004. 4

[53] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999. 2, 3, 4, 8

[54] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[55] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10437–10446, 2020. 3

[56] Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*, 2024. 1, 2, 3

[57] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press, July 2010. 3

[58] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012. 2, 6

[59] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardos. ORB-SLAM: a Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31(5):1147–1163, October 2015. arXiv:1502.00956 [cs]. 3

[60] Izumi Ohzawa, Gregory C. DeAngelis, and Ralph D. Freeman. Stereoscopic Depth Discrimination in the Visual Cortex: Neurons Ideally Suited as Disparity Detectors. *Science*, 249(4972):1037–1041, August 1990. Publisher: American Association for the Advancement of Science. 3

[61] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023. 1, 2

[62] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. MeshLoc: Mesh-Based Visual Localization, July 2022. arXiv:2207.10762 [cs]. 3

[63] Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *Journal of Neuroscience*, 38(33):7255–7269, August 2018. Publisher: Society for Neuroscience Section: Research Articles. 3

[64] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 1, 2, 3, 4, 5

[65] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015. 3

[66] Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsay, Kenneth D. Miller, Richard Naud, Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema, João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, November 2019. Publisher: Nature Publishing Group. 3

[67] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, 2006. 3, 4, 8

[68] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. Orb: An efficient alternative to sift or surf. *2011 International Conference on Computer Vision*, pages 2564–2571, 2011. 3, 4, 8

[69] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale, April 2019. arXiv:1812.03506 [cs]. 3

[70] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks, 2020. 4, 8

[71] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[72] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?, January 2020. Pages: 407007 Section: New Results. 3

[73] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884, 2019. 3

[74] Yoli Shavit, Ron Ferens, and Yosi Keller. Learning Multi-Scene Absolute Pose Regression with Transformers, July 2021. arXiv:2103.11468 [cs]. 3

[75] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 3

[76] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-Free Local Feature Matching with Transformers, April 2021. arXiv:2104.00680 [cs]. 3, 4

[77] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

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

[79] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow, March 2020. 1, 3

[80] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras, February 2022. arXiv:2108.10869 [cs]. 3

[81] Zachary Teed, Lahav Lipson, and Jia Deng. Deep Patch Visual Odometry, May 2023. arXiv:2208.04726 [cs]. 3

[82] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision? *ArXiv*, abs/2105.07197, 2021. 2, 3

[83] Michał J. Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning local features with policy gradient, October 2020. arXiv:2006.13566 [cs]. 3, 8

[84] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using LSTMs for structured feature correlation, August 2017. arXiv:1611.07890 [cs]. 3

[85] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017. 3

[86] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. MatchFormer: Interleaving Attention in Transformers for Feature Matching, September 2022. arXiv:2203.09645 [cs, eess]. 3

[87] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 2

[88] Andrew E. Welchman. The Human Brain in Depth: How We See in 3D. *Annual Review of Vision Science*, 2:345–376, October 2016. 2, 3

[89] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014. Publisher: Proceedings of the National Academy of Sciences. 3

[90] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 3

[91] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[92] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 3

[93] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2

[94] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5014–5022, 2016. 1

[95] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 3

[96] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023. 2