SpatialLLM: A Compound 3D-Informed Design towards Spatially-Intelligent Large Multimodal Models

Supplementary Material

A. 3D-Informed Data generation

Previous studies [14, 16] exploited visual foundation models for depth estimation, *e.g.*, DepthAnythingv2 [62], camera calibration *e.g.*, WildCamera [70] and PerspectiveFields [29], and object tagging and grounding, *e.g.*, RAM [66] and SAM [30]. By back projecting the points into 3D camera space, they estimated an axis-aligned 3D bounding box with category name or caption for each object. This enabled SpatialVLM [14] and SpatialRGPT [16] to explore 3D distance-based relationships or 2D spatial reasoning questions, *e.g.*, which is further left in the 2D image plane. Meanwhile, more complex 3D spatial reasoning questions are left unexplored, due to the lack of 3D orientation knowledge.

We address this key limitation by exploiting existing datasets with 3D pose annotations [45] or synthetic data with groundtruth 3D bounding boxes [5]. Besides 3D poses computed from the 3D annotations in [5], we adopt a 3D pose estimator pretrained on [45] and predict 3D orientations for rigid objects grounded in unannotated natural images. We aggregate all collected 3D information and produce oriented 3D bounding boxes, rather than axis-aligned bounding boxes considered in previous works [14, 16]. Lastly we generate diverse and rich 3D-informed data about various 3D spatial relationships.

3D-informed 3D probing data. Motivated by previous linear probing experiments on 3D awareness of visual foundation models [19, 45], we propose to exploit 3D-informed probing data and extract visual features with rich 3D awareness at the feature alignment stage. Specifically, 3D-informed probing data consists of fundamental 3D questions, *e.g.*, depth of an object, distances between two objects, and azimuth/elevation rotations. We pretrain the visual connector only while freezing other modules, such that the visual tokens would contain rich 3D information that benefit subsequent 3D spatial reasoning. Depending on the dataset used, we present: (i) *3DI-Pb-IN166K* with existing 3D annotations in ImageNet3D [45]; and (ii) *3DI-Pb-OI1M* with images from the OpenImages dataset [33] and 3D orientations estimated with a 3D pose estimator.

3D-informed instruction tuning data. We further present 3D-informed instruction tuning data, with 1 million question-answer data about 3D spatial relationships on images from [33], *i.e.*, *3DI-Ft1M*. By finetuning on our 3D-informed instruction tuning data, LMMs learn to aggregate the 3D-aware information from the visual tokens and to perform 3D spatial reasoning. This highlights the necessity of adopting 3D-informed data at both the pretraining and finetuning stage, which in our experiments, we find that LMMs pretrained with our 3D-informed probing data and finetuned with our 3D-informed instruction tuning data achieves the best 3D spatial reasoning capabilities.

Qualitative examples. We present some qualitative examples of our 3DI-Ft1M data in Fig. 8.

B. SpatialVQA

Previous spatial reasoning benchmarks were either built on 3D scans rather than images [6, 64], or focused on 2D spatial relationships [14, 16], *e.g.*, left or right in the image plane, or only on distance spatial relationships [58].

To quantitatively study the 3D spatial reasoning capabilities of LMMs, specifically on questions about object 3D orientations or questions that require reasoning over 3D locations and 3D orientations, we build a new evaluation dataset, *i.e.*, SpatialVQA. Please refer to Fig. 2 for some qualitative examples of SpatialVQA.

Dataset generation. We follow previous works [16, 58] and develop SpatialVQA based on the 3D annotations in Omni3D [11], with images from both urban [12, 21] and indoor scenes [9, 50, 54]. Specifically, we compute 3D locations, 3D orientations, 3D distances, and spatial relationships from the object-level 3D bounding box annotations. We generate visual question-answer pairs based on the 3D annotations using pre-defined rules for each question types.

Question types. Our SpatialVQA consists of three question types: (i) distance questions that can be answered from a 3D-awareness of distance only, (ii) orientation questions that require estimating objects' 3D orientations, and (iii) spatial relationship questions that require 3D spatial reasoning over various 3D information.

1. **Closer to camera [distance]:** determining which of the two objects is closer to the camera (viewer).
2. **Closer to object [distance]:** determining which of the two objects is closer to a third object.
3. **Facing camera [orientation]:** determining which side of the object is facing towards the camera (viewer).

Q: Consider the real-world 3D orientations of the objects. Are the sports car in Region [0] and the car in Region [1] facing the same (or similar) directions, or very different directions?

A: They are facing the same direction.

Q: Consider the real-world 3D locations and orientations of the objects. Which object is closer to the viewer, the chair in Region [0] or the chair in Region [1]?

A: The chair in Region [1] is closer to the viewer.

Q: Consider the real-world 3D orientations of the objects. Are the suv in Region [0] and the bus in Region [1] facing the same (or similar) directions, or very different directions?

A: They are facing different directions.

Q: Consider the real-world 3D locations and orientations of the objects. From the perspective of the player in Region [0], in which direction is the player in Region [1]?

A: The player in Region [1] is in the front direction of the player in Region [0].

Figure 8. Qualitative examples of our 3DI-Ft1M data.

4. **Facing object [orientation]:** determining which of the two object is the third object facing towards.
5. **Same direction [spatial relationship]:** determining if two objects are facing the same or different directions.
6. **Higher [spatial relationship]:** determining which of the two objects has a higher 3D location.
7. **On which side [spatial relationship]:** determining an object is on which side of another object, *e.g.*, front, left, *etc*.

Dataset statistics. Our SpatialVQA include a total of 1,323 questions, with around 240 questions in each question type. Moreover, the answers to the questions are roughly balanced, *e.g.*, for the 240 questions asking if two objects are facing the same directions, 120 questions have "yes" as the answer and 120 questions have "no" as the answer. Certain answers have a lower frequency, due to their natural scarcity in real images, *e.g.*, the bottom of an object is facing towards the camera.

C. Public Release

Data. To benefit future research on LMMs with strong 3D spatial reasoning capabilities, we will release all our training and testing data upon paper acceptance, including the 3D-informed probing data, *i.e.*, *3DI-Pb-IN166K* and *3DI-Pb-OI1M*, the 3D-informed instruction tuning data, *i.e.*, *3DI-Ft1M*, and our testing benchmark, *i.e.*, SpatialVQA.

Code. We provide the implementation of (i) our 3D-informed data generation, and (ii) our SpatialLLM in this anonymous link. All code will be made available upon acceptance of the paper.