

# 000 001 002 003 004 005 INTERNSPATIAL: A COMPREHENSIVE DATASET FOR 006 SPATIAL REASONING IN VISION-LANGUAGE MODELS 007 008 009

010 **Anonymous authors**  
011

012 Paper under double-blind review  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027

## ABSTRACT

028 Recent benchmarks and datasets have been proposed to improve spatial reasoning  
029 in vision-language models (VLMs), yet existing open resources remain limited in  
030 scale, visual diversity, and instruction expressiveness. In this work, we introduce  
031 InternSpatial, the largest open-source dataset for spatial reasoning in VLMs, along  
032 with InternSpatial-Bench, a corresponding evaluation benchmark designed to as-  
033 sess spatial understanding under diverse instruction formats. InternSpatial com-  
034 prises 12 million QA pairs spanning both single-view and multi-view settings,  
035 drawn from diverse visual environments and supporting 19 instruction formats  
036 that reflect varied query styles. For evaluation, we propose InternSpatial-Bench  
037 for single-view tasks and expand multi-view reasoning by introducing a novel ro-  
038 tation angle prediction task that has not been explored in prior work. Experimental  
039 results show that models trained on InternSpatial achieve 12.1% improvement on  
040 InternSpatial-Bench and 10.7% on VSI-Bench, while maintaining strong perfor-  
041 mance on general-purpose benchmarks. We hope these resources will support the  
042 development of spatially capable VLMs in practical applications such as robotics  
043 and embodied AI.  
044

## 1 INTRODUCTION

045 Vision-language models (VLMs) have achieved remarkable progress across a range of multimodal  
046 tasks such as visual question-answering (VQA), image captioning, and grounding, demonstrating  
047 their ability to align and reason over visual and textual inputs. Nonetheless, they still struggle with  
048 spatial reasoning, both in single-view settings (*e.g.*, identifying object position or size from a static  
049 image) and in multi-view scenarios (*e.g.*, estimating distances or tracking appearance order across  
050 dynamic video frames). Enhancing spatial reasoning capabilities in VLMs is crucial for real-world  
051 applications, including robotics, autonomous navigation, and augmented reality, where accurate  
052 spatial understanding is essential for interaction with complex environments.  
053

054 Recent efforts have introduced spatially-relevant VQA datasets and corresponding evaluation bench-  
055 marks to enhance and assess VLMs' spatial reasoning capabilities (Cai et al., 2025; Cheng et al.,  
056 2024; Chen et al., 2024a; Yang et al., 2024). While these works have advanced the field, they still  
057 exhibit several notable limitations. (1) *Limited scene diversity*: existing datasets are typically drawn  
058 from narrow sources, primarily indoor or outdoor scenes, and fail to capture a broader spectrum of  
059 scenarios. (2) *Restricted instruction formats*: SpatialVLM (Chen et al., 2024a) and SpatialQA (Cai  
060 et al., 2025) rely exclusively on natural language, and OSD (Cheng et al., 2024) uses region masks.  
061 These limited formats fail to reflect the diversity of instruction types required for practical spa-  
062 tial reasoning tasks. (3) *Narrow training scope*: existing spatial training data primarily focus on  
063 single-view settings and cover only basic spatial concepts from a single static image, such as object  
064 position or existence, without providing multi-view supervision that captures spatial relationships  
065 across different viewpoints or temporal sequences. These limitations underscore the need for a  
066 more comprehensive dataset paired with a corresponding evaluation benchmark to advance spatial  
067 reasoning in VLMs.  
068

069 To address these limitations, we propose the largest open-source spatial reasoning dataset, *Int-  
070 ernSpatial*, and a corresponding evaluation benchmark, *InternSpatial-Bench*, specifically designed  
071 to enhance spatial reasoning capabilities in VLMs. InternSpatial comprises 9.5M single-view and  
072 2.5M multi-view question-answer pairs, sourced from a broad spectrum of visual environments, in-

054 Table 1: Comparison of our InternSpatial with existing spatial reasoning datasets. W: in-the-wild, I:  
 055 indoor, D: drive, E: embodied, O: object-centric

056 <b>Dataset</b>	057 <b># of QA</b>	058 <b>Scenario</b>	059 <b>Open-source</b>	060 <b>View Type</b>	061 <b>Instruction format</b>
SpatialVLM (Chen et al., 2024a)	2B	W	✗	Single-view	Single-format
SpatialQA (Cai et al., 2025)	0.9M	W,E	✓	Single-view	Single-format
OSD (Cheng et al., 2024)	8.7M	W	✓	Single-view	Single-format
InternSpatial	12M	W,I,D,E,O	✓	Single-view, Multi-view	Multiple-format

062 cluding in-the-wild scenes (Lin et al., 2014; Wang et al., 2024c; Krishna et al., 2017), structured  
 063 indoor spaces (Wald et al., 2019; Dai et al., 2017; Mao et al., 2022), urban streetscapes (Cordts  
 064 et al., 2016), object-centric scenes (Deitke et al., 2022), and embodied navigation contexts (Ander-  
 065 son et al., 2018). To enrich instruction formats, we incorporate a diverse set of query represen-  
 066 tations, including masks, bounding boxes, and numerical indicators embedded in images, as well  
 067 as coordinate-based references and spatial cues expressed through textual instructions. In total, our  
 068 dataset supports 19 distinct instruction formats, enabling broader coverage of spatial reasoning query  
 069 types. We further introduce a novel multi-view task, rotation angle prediction, with 2.46M newly  
 070 collected training question-answer pairs, which has not been addressed in prior spatial reasoning  
 071 benchmarks. To facilitate evaluation, we construct InternSpatial-Bench with 6,008 question-answer  
 072 pairs, serving as a comprehensive diagnostic benchmark for single-view spatial reasoning tasks. For  
 073 multi-view evaluation, we extend the existing VSI benchmark by adding 1,000 additional question-  
 074 answer pairs for the rotation angle prediction task. As shown in Table 1, our InternSpatial signifi-  
 075 cantly expands scene coverage, instruction format diversity, and multi-view supervision compared  
 076 to existing benchmarks.

077 In summary, our contributions are threefold:

- 078 (1) We present InternSpatial, the largest open-source spatial reasoning dataset for VLMs, designed  
 079 for supervised fine-tuning. It contains single-view and multi-view samples across diverse scenes and  
 080 supports 19 instruction formats to support varied spatial query forms.
- 081 (2) To support evaluation, we introduce InternSpatial-Bench for single-view tasks and extend the  
 082 VSI benchmark for multi-view evaluation, incorporating a novel rotation angle prediction task not  
 083 addressed in existing datasets.
- 084 (3) Extensive experimental results demonstrate the effectiveness of InternSpatial, showing that it  
 085 substantially improves spatial reasoning in VLMs, achieving a 12.1% improvement on InternSpatial-  
 086 Bench and 10.7% on VSI-Bench while preserving general multimodal performance.

## 088 2 RELATED WORK

### 091 2.1 SPATIAL REASONING VIA VISION LANGUAGE MODELS

093 Recently, numerous large language models (LLMs) (Brown et al., 2020; Achiam et al., 2023; Tou-  
 094 vron et al., 2023) and vision-language models (VLMs) (Zhu et al., 2022; Li et al., 2023a; Zhu et al.,  
 095 2023a; Wang et al., 2023; Liu et al., 2023; Li et al., 2023b; Wang et al., 2024c; Chen et al., 2024c).  
 096 have been developed. However, growing evidence indicates that VLMs still struggle with spatial rea-  
 097 soning tasks (Cai et al., 2025; Chen et al., 2024a; Cheng et al., 2024; Yang et al., 2024). To alleviate  
 098 these issues on 2D images, recent works explicitly inject spatial awareness: Ferret-v2 (Zhang et al.,  
 099 2024a) improves fine-grained regional referring and grounding with any-resolution visual encod-  
 100 ing, LocVLM (Ranasinghe et al. (2024) enhances spatial reasoning via coordinate-based instruction  
 101 tuning, and Shikra (Chen et al. (2023) represents spatial coordinates purely in natural language to  
 102 support referential dialogue. Beyond 2D localization, several approaches further incorporate addi-  
 103 tional supervision signals. For example, 3D-LLM (Hong et al., 2023b) and 3D-CLR (Hong et al.,  
 104 2023a) introduce 3D representations and dense features; SpatialRGPT (Cheng et al., 2024) incorpo-  
 105 rates mask-based supervision; and SpatialBot (Cai et al., 2025) leverages depth information. Despite  
 106 these efforts, current methods have not succeeded in enabling VLMs to perform end-to-end spatial  
 107 reasoning effectively. Despite these efforts, current methods still fall short of comprehensive, end-  
 108 to-end spatial reasoning across diverse single-view and multi-view scenarios, which motivates the  
 109 design of our InternSpatial dataset.

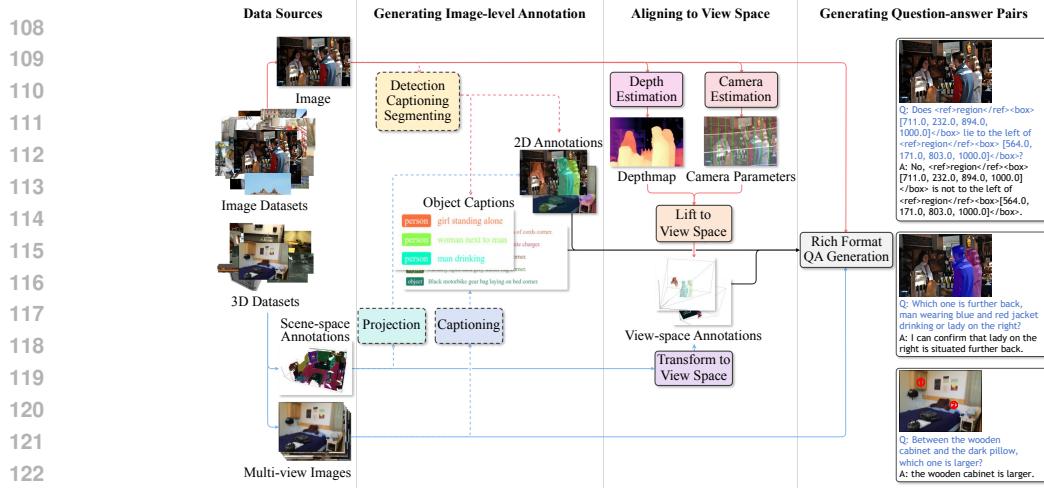


Figure 1: Generation pipeline for InternSpatial. The optional flows (represented by dashed lines and boxes) are only performed when the relevant annotations does not exist in the data source.

## 2.2 SPATIAL REASONING DATASETS

To evaluate and improve the spatial reasoning capabilities of VLMs, several datasets and benchmarks have been proposed to cover a range of tasks and scenarios. One such benchmark, Spatial-Eval (Wang et al., 2024a), targets 2D spatial reasoning across tasks such as relation understanding, navigation, and counting. Another line of work explores spatial reasoning from a top-down perspective, emphasizing the need to enhance VLM performance in top-view settings (Li et al., 2024c). To enable VLMs to understand 3D spatial relationships from images, several datasets have been introduced that focus on answering 3D spatial reasoning questions (Cheng et al., 2024; Cai et al., 2025; Li et al., 2024d). However, these datasets are primarily tailored to specific models and often rely on additional inputs, such as segmentation masks or depth maps. An automatic data generation framework has also been developed to construct a large-scale 3D spatial VQA dataset using Internet images (Chen et al., 2024a), demonstrating that with appropriate training data, VLMs can infer spatial relationships without relying on auxiliary inputs. Nevertheless, the dataset is not publicly available. Spatial reasoning over image sequences or videos presents additional challenges. To assess such capabilities, the VSI benchmark (Yang et al., 2024) was proposed, evaluating a range of open-source and proprietary VLMs. Results show that current models still struggle with multi-frame spatial reasoning tasks. Our work addresses these limitations by introducing a dataset that integrates both single-view and multi-view tasks, significantly enhancing the spatial reasoning ability of VLMs across diverse contexts and highlighting their potential for deeper spatial understanding.

## 3 DATASET

### 3.1 DATA ENGINE FOR INTERNSPATIAL

We construct InternSpatial, a large-scale dataset comprising nearly 12 million Question-Answer(QA) pairs, to enable VLMs to perform 3D spatial reasoning through supervised fine-tuning. InternSpatial aggregates data from a wide range of sources, including in-the-wild scenes (Lin et al., 2014; Wang et al., 2024c; Krishna et al., 2017), structured indoor spaces (Wald et al., 2019; Dai et al., 2017; Mao et al., 2022), urban streetscapes (Cordts et al., 2016), object-centric scenes (Deitke et al., 2022), and embodied navigation contexts (Anderson et al., 2018).

To handle the heterogeneity of source data and support large-scale QA generation, we develop a fully automated and modular data engine that consolidates intermediate annotation extraction and QA synthesis into a unified pipeline applicable across diverse data sources. As illustrated in Figure 1, the pipeline begins by generating necessary annotations at the image level, followed by transforming the annotations into a canonical view space. Finally, QA pairs are constructed using a template-based approach that supports a wide variety of task types and instruction formats.

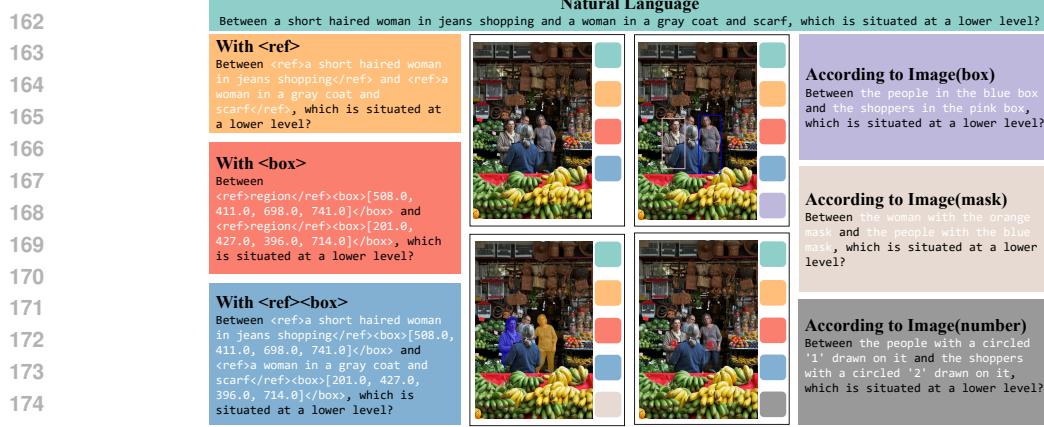


Figure 2: Examples of diverse instruction formats in text and image. The four images illustrate different visual formats: original (top-left), bounding boxes (top-right), segmentation masks (bottom-left), and numbered regions (bottom-right). Surrounding the images are seven corresponding text instruction formats. The color blocks beside each image indicate whether the corresponding image-text pair is included in InternSpatial and InternSpatial-Bench. Best viewed in color.

**Generating Image-level Annotation.** To generate 3D spatial reasoning QAs grounded in objects, we first obtain the necessary image-level annotations, including 2D bounding boxes, region descriptions, segmentation masks, etc. For image datasets that already provide such annotations, we directly utilize the existing labels. When annotations are missing, we employ pretrained models to generate them automatically. Specifically, we use open-source VLMs to extract object-level 2D boxes and associated textual descriptions, and apply the SAM2 model (Ravi et al., 2024) to generate segmentation masks within these boxes. These masks are subsequently lifted into 3D space to facilitate the construction of 3D bounding boxes. The prompts we used in this step can be found in Appendix C. In the case of 3D datasets, which typically include global 3D annotations and per-view camera parameters, we project the 3D information onto the image plane to obtain the corresponding 2D annotations. Although this projection is not strictly required for generating QAs, as the underlying 3D annotations are already available, it is necessary for supporting visual reference forms in prompts, such as bounding boxes and segmentation masks. **In order to reduce potential ambiguity in questions, We further apply several filtering strategies.** For in-the-wild images, we filtered out objects without clear boundaries by object category, such as the sky and grass. For indoor scenes with 3D annotations, we detected the occlusions and objects beyond the image boundary through the projection of 3D models and bounding boxes on the image plane, and excluded QA pairs that were ambiguous due to these situations.

**Aligning to View Space.** To determine spatial relationships between objects, it is essential to obtain their positions and dimensions within a well-defined 3D coordinate system. We adopt a canonical view space as the reference frame, defined as a 3D Cartesian coordinate system centered at the camera’s optical center. In this space, the y-axis aligns with the viewing direction, and the z-axis is perpendicular to the scene’s horizontal plane, pointing upward. For 3D datasets, which provide global annotations and per-view camera parameters, transforming annotations into the canonical view space is straightforward. In contrast, image-only datasets contain only 2D visual information, requiring estimation of both camera parameters and depth maps. To address this, we follow the pipeline of SpatialRGPT (Cheng et al., 2024), leveraging WildCamera (Zhu et al., 2023b) for intrinsic parameter estimation, PerspectiveFields (Jin et al., 2023) for extrinsic parameter inference, and Metric3Dv2 (Hu et al., 2024) to predict dense depth maps. By combining the outputs of these models, we lift 2D annotations into the canonical 3D space, enabling accurate reasoning over object-level spatial relationships.

**Template-based QA Generation.** While prompting a large language model (LLM) to generate QA pairs directly for each image can produce diverse instructions, this approach is prohibitively expensive at scale in terms of computation and time. Instead, we adopt a template-based generation

216 strategy that avoids invoking the LLM during QA construction. This approach not only improves efficiency  
 217 but also facilitates flexible expansion to multiple prompt styles, such as object references via  
 218 bounding boxes or segmentation masks. To ensure sufficient instruction diversity, we first prompt  
 219 an LLM to generate several question-answer templates for each task type and answer format. These  
 220 templates contain placeholders for object references and other variable content. During genera-  
 221 tion, we randomly select a subset of tasks and object instances (or pairs) for each image, derive the  
 222 corresponding answers using the previously constructed annotations, and instantiate the templates  
 223 accordingly. We then filter out low-quality QA pairs, such as those involving ambiguous spatial  
 224 relationships caused by occlusion, and balance the number of positive and negative examples to produce  
 225 a well-structured dataset. We generate templates for 4 single-view tasks, covering the position/size  
 226 relationship of two objects, as well as relationship-constrained count and existence tasks. The list of  
 227 templates are shown in Appendix B.

228 **Extending Instruction Formats.** To enhance dataset diversity and better reflect real-world  
 229 usage scenarios, we extend each QA pair into multiple instruction formats. Specifically, we  
 230 generate up to five textual formats and up to four image formats per QA pair. The im-  
 231 age formats include: (1) the original image, (2) the image annotated with bounding boxes,  
 232 (3) the image with segmentation masks, and (4) the image annotated with numbers over  
 233 key objects. The textual formats include: (1) natural language descriptions, (2) text with  
 234 `<ref>{caption}</ref>` (3) text with `<ref>region</ref><box>{bbox}</box>` (4)  
 235 text with `<ref>{caption}</ref><box>{bbox}</box>` and (5) text automatically gen-  
 236 erated based on image content. Representative examples of these visual and textual formats are shown  
 237 in Figure 2. As a result, each QA pair can produce up to 19 training samples, from which only  
 238 suitable ones are retained. Additionally, certain prompt types, such as images with numbers on key  
 239 objects, may not directly indicate the correct object. Therefore, in these cases, we utilize the pos-  
 240 ition information from the segmentation mask to correctly identify and reference the target object.  
 241 **During training, we uniformly sample across all instruction formats.**

242 **Generating Multi-view QA Pairs.** To develop a comprehensive multi-view dataset for spatial  
 243 understanding, we systematically collected and integrated multi-view data derived from the training  
 244 splits of the ScanNet (Dai et al., 2017), MultiScan (Mao et al., 2022), R2R (Anderson et al., 2018),  
 245 and Objaverse (Deitke et al., 2022), subsequently formulating temporally-agnostic training samples  
 246 that encapsulate inter-object relational attributes such as relative properties, scale variations, and  
 247 spatial distances, and cross-view relationships of objects such as rotation. Scene-level geometric  
 248 priors were established by estimating room dimensions via the Alpha Shape algorithm (Akkiraju  
 249 et al., 1995) applied to the point clouds, with the room centroid defined as the geometric center of the  
 250 minimal axis-aligned bounding box enclosing the scene. We meticulously cataloged instance counts  
 251 for each object semantic category. For unambiguous objects within the point clouds exhibiting a  
 252 principal dimension exceeding 15cm, annotations were standardized to the *OrientedBoundingBox*  
 253 format using Open3D (Zhou et al., 2018). For remaining objects or those with initial ambiguities, we  
 254 leveraged existing annotations to reduce the risk of shortcut learning by language models. Plausible  
 255 alternative options were constructed by extracting distractors from other items within the dataset,  
 256 thereby forming a corresponding multiple-choice question training set.

257  
 258 **Human validation.** Due to the huge number of QAs in our dataset, it's almost impossible to check  
 259 all generated QAs by human. Instead, we conducted a manual verification on a randomly sampled  
 260 subset with 500 items, including both final QAs and intermediate steps. With this validation, we  
 261 ensure the accuracy of QAs in our dataset was over 95%.

### 263 3.2 INTERNSPATIAL-BENCH

264 To evaluate the performance of VLMs on 3D spatial reasoning tasks, particularly under diverse  
 265 instruction formats, we propose InternSpatial-Bench, a novel multi-task benchmark that features a  
 266 broad range of input types. Existing benchmarks such as SpatialRGPT-Bench (Cheng et al., 2024)  
 267 and SpatialBench (Cai et al., 2025) present several limitations. First, the question formats are overly  
 268 simplistic and do not reflect real-world application scenarios. Second, these benchmarks are tailored  
 269 to specialized models and require auxiliary inputs such as region masks or depth maps. As a result,

270 many tasks are incompatible with general-purpose VLMs that operate solely on images and text.  
 271 Furthermore, SpatialBench suffers from a limited number of QA pairs, reducing its effectiveness as  
 272 a comprehensive evaluation suite.

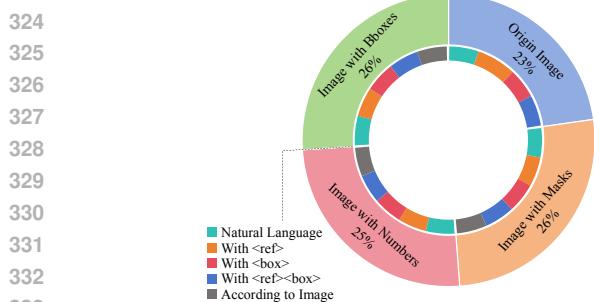
273 InternSpatial-Bench expands and refines both SpatialRGPT-Bench and SpatialBench to overcome  
 274 these limitations. Specifically, we enrich instruction formats and introduce 3,000 carefully curated  
 275 QA pairs, resulting in a total of 5,300 high-quality examples that span diverse task types and input  
 276 modalities. Certain tasks from the original benchmarks, such as reachability prediction and quan-  
 277 titative estimation of spatial extent, are excluded because they are unsuitable for general-purpose  
 278 VLMs when only a single-view image is provided. In the absence of additional information, such as  
 279 depth or camera parameters, these tasks become severely under-constrained and often ambiguous,  
 280 even for human annotators.

281 **Refining and Expanding SpatialRGPT-Bench and SpatialBench.** Since SpatialRGPT-  
 282 Bench (Cheng et al., 2024) already provides a sufficient number of QA pairs, our focus is on expanding  
 283 the diversity of question formats rather than increasing the dataset size. Specifically, we augment  
 284 the instruction styles of the original questions that do not involve numerical reasoning, following the  
 285 format extension strategy described in subsection 3.1. However, to avoid ambiguity caused by du-  
 286 plicate object labels, we exclude formats that rely on natural language references or textual content  
 287 containing `<ref>caption</ref>`. For each selected question, we randomly sample three dif-  
 288 ferent formats and leverage both object mask and bounding box annotations to construct the final  
 289 benchmark entries. SpatialBench (Cai et al., 2025) contains QA pairs exclusively in natural language  
 290 form. To diversify its instruction formats, we first manually extract reference phrases correspond-  
 291 ing to the mentioned objects and convert the questions into templates with placeholders. Next, we  
 292 prompt the VLM to ground the objects based on these phrases and apply SAM2 to segment the  
 293 corresponding regions. Using the resulting question templates, along with object bounding boxes  
 294 and masks, we apply the format extension method described in subsection 3.1 to generate diverse  
 295 instruction variants for each QA. Finally, all generated QA pairs are manually verified to ensure  
 296 quality, with erroneous answers corrected and ambiguous or ill-formed questions removed

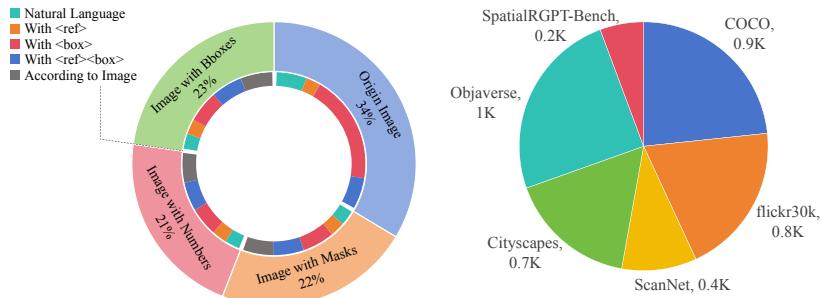
297 **Extending the Benchmark with Curated QA Pairs.** Unlike the large-scale training dataset, the  
 298 benchmark is relatively small but demands higher annotation quality. To this end, we implement  
 299 a dedicated pipeline for generating high-quality QAs used in the benchmark. This pipeline oper-  
 300 ates without relying on any pre-annotated information, making it applicable to any image-only data  
 301 source. To encourage diversity and expressiveness in question formulation, we prompt the VLM  
 302 to generate questions directly. Finally, we introduce a manual verification step to review all au-  
 303 tomatically constructed questions and answers, ensuring the overall quality and correctness of the  
 304 benchmark data. Details of the construction process are provided in Appendix D.

### 306 3.3 DATASET STATISTICS

308 **Statistics of InternSpatial.** Our proposed dataset, InternSpatial, encompasses a diverse set of  
 309 tasks and instruction formats to comprehensively enhance spatial reasoning capabilities. It consists  
 310 of a total of 12,035,415 question-answer pairs, covering both single-view and multi-view spatial  
 311 reasoning tasks. Specifically, the single-view tasks include *Position Comparison*, *Size Comparison*,  
 312 *Existence Estimation*, and *Object Counting*, while the multi-view tasks include *Rotation Estima-  
 313 tion*, *Object Counting*, *Room Size Estimation*, *Object Size Estimation*, *Route Planning*, and *Ap-  
 314 pearance Order*. Detailed task descriptions and corresponding statistics are provided in Appendix A,  
 315 and visual examples are shown in Appendix F. In addition, InternSpatial incorporates images from  
 316 various sources to enhance the robustness of the model. As illustrated in Figure 3, the dataset in-  
 317 cludes COCO (Lin et al., 2014), AS-1B (Wang et al., 2024c), and Visual Genome (VG) (Krishna  
 318 et al., 2017) for in-the-wild imagery; 3RScan (Wald et al., 2019), ScanNet (Dai et al., 2017), and  
 319 MultiScan (Mao et al., 2022) for indoor scenes; Cityscapes (Cordts et al., 2016) for street scenes;  
 320 Objaverse (Deitke et al., 2022) for single-object scenarios; and R2R (Anderson et al., 2018) for em-  
 321 bodied navigation tasks. Moreover, InternSpatial emphasizes diversity in instruction formats. As  
 322 shown in Figure 3, the number of samples across different formats is carefully balanced to avoid  
 323 bias and ensure uniform coverage during training. In summary, InternSpatial provides a large-scale,  
 324 diverse resource spanning task types, visual domains, and instruction formats, making it well-suited  
 325 for training VLMs to handle real-world spatial reasoning tasks effectively.



337 **Statistics of InternSpatial-Bench** Following Spatial-Bench and Spatial-RGPT, our proposed  
338 benchmark, InternSpatial-Bench, includes five tasks—*Position Estimation*, *Size Estimation*, *Ro-  
339 tation Estimation*, *Existence Estimation*, and *Object Counting*—designed to systematically evaluate  
340 the spatial reasoning capabilities of VLMs. In total, InternSpatial-Bench consists of 6,008 QA pairs.  
341 Detailed task statistics are provided in Appendix A, and visual examples are shown in Appendix G.  
342 To ensure robustness and diversity, InternSpatial-Bench incorporates images from a broad range of  
343 domains. As shown in Fig 4, in addition to the sources used in Spatial-Bench and Spatial-RGPT, we  
344 include samples from the test sets of COCO, Flickr30K, Objaverse, ScanNet, and Cityscapes. This  
345 diverse image collection spans a wide range of real-world contexts, from indoor and outdoor envi-  
346 ronments to single-object scenarios and in-the-wild imagery. We apply the same instruction format  
347 expansion strategy as used in InternSpatial, with one exception: for the Rotation Estimation task,  
348 since each image contains only a single object, we only use the original image format and natural  
349 language instructions. Consequently, these formats have a higher proportion in this task compared to  
350 others. By combining diversity in task types, visual domains, and instruction formats, InternSpatial-  
351 Bench offers a comprehensive and realistic benchmark for evaluating the spatial reasoning abilities  
352 of VLMs across a wide range of practical scenarios.



## 4 EXPERIMENTS

366 We begin in Section 4.1 by introducing the baseline model and outlining the evaluation benchmarks  
367 used in our experiments. Section 4.2 then presents results on InternSpatial-Bench to assess the  
368 spatial reasoning capabilities of vision-language models. Section 4.3 reports performance on VSI-  
369 Bench (Yang et al., 2024), which further evaluates the models’ multi-view spatial reasoning abilities.  
370 In Section 4.4, we conduct an ablation study to analyze the impact of different instruction formats  
371 on model performance. Finally, Section 4.5 evaluates whether training with InternSpatial affects  
372 general reasoning ability by benchmarking against a suite of standard vision-language tasks.

### 4.1 EXPERIMENT SETUP

374 **Baseline.** We construct our baselines based on InternVL2.5-8B (Chen et al., 2024c), a represen-  
375 tative traditional VLM. Following the training settings of InternVL2.5, we fine-tune our models from  
376 InternVL2.5-8B using a downsampled version of the general datasets employed in InternVL2.5,  
377

378 along with InternSpatial. For generality, we also utilize InternVL2.5-1B and Qwen2.5-VL-8B as  
 379 additional baselines to demonstrate the transferability of our training strategy. We refer to the models  
 380 fine-tuned on InternSpatial as InternVL-Spatial-8B (for InternVL2.5-8B), InternVL-Spatial-8B-  
 381 1B (for InternVL2.5-1B), and Qwen-Spatial-8B (Bai et al., 2025) (for Qwen2.5-VL-8B). Detailed  
 382 training configurations are provided in Appendix E.

384 **Evaluation.** We evaluate the models trained on InternSpatial using three types of benchmarks:  
 385 our proposed InternSpatial-Bench, the multi-view spatial reasoning benchmark VSI-Bench (Yang  
 386 et al., 2024), and several general-purpose benchmarks, including MathVision (Wang et al., 2024b),  
 387 OCRBench (Liu et al., 2024), TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022), and MM-  
 388 Star (Chen et al., 2024b). For InternSpatial-Bench, we follow the evaluation protocols of Spatial-  
 389 Bench (Cai et al., 2025) and Spatial-RGPT (Cheng et al., 2024), reporting relative error for counting  
 390 tasks, accuracy for multiple-choice questions, and GPT-4o-assigned (OpenAI, 2025) scores for quiz-  
 391 style questions. For VSI-Bench, we adopt the official evaluation protocol, with the only modification  
 392 being the use of 32 sampled frames per video during testing. For general benchmarks, we follow the  
 393 evaluation procedures provided by OpenCompass (Contributors, 2023).

## 395 4.2 EVALUATION ON INTERNSPATIAL-BENCH

398 Table 2: Results on InternSpatial-Bench. **Bold** indicates the best performance among all models,  
 399 while underline denotes the second-best performance.

Model	Position Comparison	Size Comparison	Rotation Estimation	Object Counting	Existence Estimation	Average
<b>Human Level</b>	<b>99.7</b>	<b>97.7</b>	<b>100.0</b>	<b>98.9</b>	<b>100.0</b>	<b>99.3</b>
GPT-4o-2024-11-20 (OpenAI, 2025)	71.2	71.5	26.7	63.5	74.9	61.6
Claude-3.7-Sonnet-20250219 (Anthropic, 2024)	73.2	72.3	25.9	59.2	70.5	60.2
Gemini-2.5-Flash(Comanici et al., 2025)	64.5	67.3	30.2	67.0	67.3	59.3
Llama-4-Scout(Meta Platforms, 2025)	42.2	45.0	20.8	44.0	25.7	35.5
Qwen2.5-VL-72B (Bai et al., 2025)	54.6	55.3	30.6	60.5	63.3	52.9
Pixtral-12B (Agrawal et al., 2024)	65.6	62.9	5.8	52.5	78.3	53.0
LLaVA-OneVision-72B(Li et al., 2024b)	77.8	77.0	25.8	64.5	77.6	64.5
<b>Qwen2.5-VL-8B (Bai et al., 2025)</b>	<b>57.1</b>	<b>60.8</b>	<b>26.9</b>	<b>58.0</b>	<b>66.7</b>	<b>53.9</b>
<b>Qwen-Spatial-8B</b>	<b>79.9(+22.8)</b>	<b>78.7(+17.9)</b>	<b>34.4(+7.5)</b>	<b>68.3(+10.3)</b>	<b>80.0(+13.3)</b>	<b>68.3(+14.4)</b>
<b>InternVL2.5-1B (Chen et al., 2024c)</b>	<b>42.9</b>	<b>43.3</b>	<b>23.8</b>	<b>21.3</b>	<b>59.9</b>	<b>38.2</b>
<b>InternVL-Spatial-1B</b>	<b>65.4(+22.5)</b>	<b>58.5(+15.2)</b>	<b>26.3(+2.5)</b>	<b>59.4(+28.1)</b>	<b>74.4(+14.5)</b>	<b>56.8(+18.6)</b>
InternVL2.5-8B (Chen et al., 2024c)	62.8	57.7	28.5	67.8	77.9	58.9
<b>InternVL-Spatial-8B</b>	<b>87.8(+25.0)</b>	<b>78.6(+20.9)</b>	<b>33.6(+5.1)</b>	<b>71.3(+3.5)</b>	<b>83.9(+6.0)</b>	<b>71.0(+12.1)</b>

414 To evaluate model performance in spatial reasoning, we conducted experiments on InternSpatial-  
 415 Bench. The accuracy computation follows the methodology of Spatial-Bench (Cai et al., 2025) and  
 416 Spatial-RPGT (Cheng et al., 2024), with a modification for the Object Counting task: since some  
 417 VLMs struggle to follow instructions precisely, we extract the last number mentioned in the response  
 418 as the predicted count and compute the relative error accordingly.

419 As shown in Table 2, our model, InternVL-Spatial-8B, outperforms the baseline InternVL2.5-  
 420 8B (Chen et al., 2024c) by 12% in average accuracy. Notably, it achieves a 25% improvement in the  
 421 Position Comparison task and a 20.9% gain in the Size Comparison task. Furthermore, InternVL-  
 422 Spatial-8B surpasses advanced proprietary models such as GPT-4o (OpenAI, 2025) and Claude 3.5  
 423 Sonnet (Anthropic, 2024) across all tasks, demonstrating the effectiveness of InternSpatial in en-  
 424 hancing the spatial reasoning capabilities of VLMs.

425 To demonstrate the generality and broad impact of our InternSpatial, we applied the same training  
 426 paradigm to two additional models with varying sizes: InternVL2.5-1B and Qwen2.5-VL-8B. The  
 427 results consistently confirm the effectiveness of our training data. Specifically, InternVL-Spatial-1B  
 428 improved its InternVL2.5-1B baseline by 18.6% (from 38.2% to 56.8% average accuracy), exhib-  
 429 iting a significant lift for a smaller model. Similarly, the Qwen-Spatial-8B elevated the Qwen2.5-VL-  
 430 8B baseline by 14.4% (from 53.9% to 68.3% average accuracy). These substantial and consistent  
 431 gains across different model families and sizes confirm that our proposed InternSpatial is highly  
 432 effective and transferable for universally enhancing the spatial-aware capability of VLMs.

432 Table 3: Results on VSI-Bench. **Bold** indicates the best performance among all models, while  
 433 underline denotes the second-best performance.

434 Model	435 Obj.Count	436 Abs.Dist.	437 Obj.size	438 Room Size	439 Rel.Dist.	440 Route Plan	441 Appr.Order	442 Average
GPT-4o (OpenAI, 2025)	46.2	5.3	43.8	38.2	37.0	31.5	28.5	32.9
Gemini-1.5 Flash (Reid et al., 2024)	49.8	30.8	<u>53.5</u>	<b>54.4</b>	37.7	31.5	37.8	42.3
Gemini-1.5 Pro (Reid et al., 2024)	56.2	30.9	<b>64.1</b>	43.6	<b>51.3</b>	<u>36.0</u>	34.6	<u>45.3</u>
VILA-1.5-40B (Lin et al., 2024)	22.4	24.8	48.7	22.7	40.5	31.5	32.9	32.0
LLaVA-NeXT-Video-72B (Zhang et al., 2024b)	48.9	22.8	57.4	35.3	42.4	35.0	48.6	41.5
LLaVA-OneVision-72B (Li et al., 2024a)	43.5	23.9	57.6	37.5	42.5	32.5	44.6	40.2
Qwen2.5-VL-8B (Bai et al., 2025)	<u>41.5</u>	<u>21.2</u>	50.7	<u>36.6</u>	<u>37.9</u>	30.4	34.0	36.0
Qwen-Spatial-8B	60.8(+19.3)	35.0(+13.8)	53.4(+2.7)	45.0(+8.4)	40.0(+2.1)	<b>36.6(+6.2)</b>	34.5(+0.5)	43.6(+7.6)
InternVL2.5-1B (Chen et al., 2024c)	51.8	3.9	24.8	13.7	25.6	32.5	7.6	22.8
InternVL-Spatial-1B	<u>66.4(+14.6)</u>	<u>25.4(+21.5)</u>	<u>42.0(+17.2)</u>	<u>48.5(+24.8)</u>	<u>34.1(+8.5)</u>	<u>34.0(+1.5)</u>	11.0(+3.4)	<u>37.3(+14.5)</u>
InternVL2.5-8B (Chen et al., 2024c)	51.7	32.9	45.1	42.3	40.8	27.8	<u>50.5</u>	41.6
InternVL-Spatial-8B	<b>68.7(+17.0)</b>	<b>40.9(+8.0)</b>	<u>63.1(+18.0)</u>	<u>54.3(+12.0)</u>	<u>47.7(+6.9)</u>	29.9(+2.1)	<b>60.5(+10.0)</b>	<b>52.3(+10.7)</b>

### 445 4.3 EVALUATION ON VSI-BENCH

446 To evaluate the additional multi-view spatial reasoning capabilities of InternVL-Spatial-8B trained  
 447 on InternSpatial, we conducted experiments on VSI-Bench (Yang et al., 2024). As shown in Table 3,  
 448 InternVL-Spatial-8B achieves notable improvements over the baseline InternVL2.5-8B (Chen et al.,  
 449 2024c) across all tasks in the benchmark. In particular, it surpasses the baseline by more than 10%  
 450 in Object Counting, Object Size Estimation, and Appearance Order tasks.

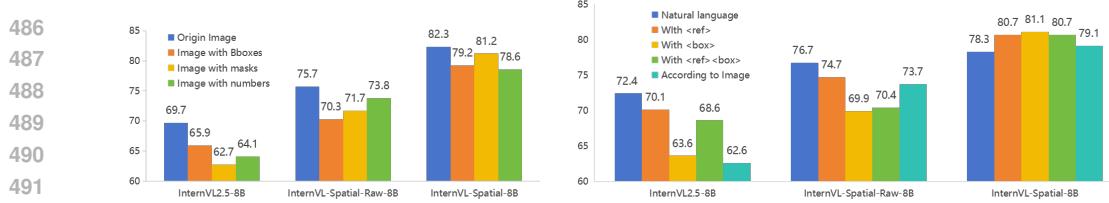
451 When compared against both open-source and proprietary models, InternVL-Spatial-8B delivers  
 452 top-tier performance: it ranks first in Object Counting, Absolute Distance Estimation, Object Size  
 453 Estimation and Appearance Order, and second in the remaining tasks. Overall, it achieves the high-  
 454 est average score among all evaluated models, including GPT-4o (OpenAI, 2025) and Gemini-1.5  
 455 Pro (Reid et al., 2024). These results demonstrate that InternSpatial substantially enhances the spa-  
 456 tial reasoning capabilities of vision-language models in multi-image scenarios.

457 **On VSI-BENCH, our InternSpatial yields consistent gains across architectures and scales.** The  
 458 InternVL-Spatial-1B gained a remarkable 14.5% in average accuracy (22.8% → 37.3%), with  
 459 dramatic improvements in Absolute Distance (+21.5%) and Room Size (+24.8%). Similarly, the  
 460 Qwen-Spatial-8B model elevated its baseline by a strong 7.6% (36.0% → 43.6%), including a  
 461 19.3% gain in Object Counting. These results on a challenging spatial benchmark further show  
 462 that InternSpatial can improve multi-view spatial reasoning in VLMs in a model-agnostic manner,  
 463 without any architecture-specific modifications.

### 464 4.4 EFFECT OF THE VARIOUS QUESTION FORMATS

465 We conduct an ablation study on InternSpatial-Bench to evaluate the impact of different instruction  
 466 formats in both the training step and evaluation step. Since the Rotation Estimation task does not  
 467 include instruction format expansion, we exclude it from this analysis. Additionally, we train a  
 468 variant of InternVL2.5-8B using InternSpatial-Bench without instruction format expansion, referred  
 469 to as **InternVL-Spatial-Raw-8B**.

470 As shown in Figure 5, the baseline model, InternVL2.5-8B (Chen et al., 2024c), performs best on  
 471 original images and natural language instructions, which are prevalent in general-purpose training  
 472 datasets. However, it performs significantly worse on formats involving elements such as  ${}_i box_i$ ,  
 473 which are rare in typical datasets. In contrast, InternVL-Spatial-8B, trained on InternSpatial with  
 474 diverse instruction format expansions, substantially narrows this performance gap across different  
 475 instruction styles. Furthermore, comparing InternVL2.5-8B with InternVL-Spatial-Raw-8B reveals  
 476 that even without instruction format expansion, InternVL-Spatial-Raw-8B consistently outperforms  
 477 the baseline across all instruction styles. This indicates that the model gains a degree of general-  
 478 ization and cross-format transfer ability, even without being explicitly trained on diverse instruction  
 479 forms. Finally, InternVL-Spatial-8B achieves the best performance across all instruction formats,  
 480 including natural language and original image styles. This demonstrates that instruction format ex-  
 481 pansion not only improves the model’s robustness to diverse input styles but also enhances its overall  
 482 spatial reasoning capability.

Figure 5: The results of the different image (**Left**) and text (**Right**) formats in the ablation study.

#### 4.5 GENERAL VQA

For fairness, we re-evaluated InternVL2.5-8B (Chen et al., 2024c) under our experimental setup instead of directly using the results reported in its technical report. As shown in Table 4, InternVL-Spatial-8B achieves comparable performance to the baseline InternVL2.5-8B on general reasoning benchmarks. Specifically, InternVL-Spatial-8B shows a performance gain of +1.8% on Math-Vista (Wang et al., 2024b), -0.1% on OCRBench (Liu et al., 2024), +0.9% on TextVQA (Singh et al., 2019), -1.6% on ChartQA (Masry et al., 2022), and +0.2% on MMStar (Chen et al., 2024b). These results indicate that training with InternSpatial does not compromise the model’s general reasoning capabilities, including mathematical reasoning, optical character recognition, visual question answering, and chart understanding.

Table 4: General benchmark results for InternVL2.5-8B vs. InternVL-Spatial-8B.

Model	MathVision (Wang et al., 2024b)	OCRBench (Liu et al., 2024)	TextVQA (Singh et al., 2019)	ChartQA (Masry et al., 2022)	MMStar (Chen et al., 2024b)
InternVL2.5-8B	19.0	82.3	79.0	83.0	62.9
InternVL-Spatial-8B	20.8(+1.8)	82.2(-0.1)	79.9(+0.9)	81.4(-1.6)	63.1(+0.2)

## 5 CONCLUSIONS

We introduce InternSpatial, the largest open-source spatial reasoning dataset, and the benchmark InternSpatial-Bench, which together advance spatial understanding in VLMs through diverse scene coverage, rich instruction formats, and multi-view supervision. InternSpatial provides 12M high-quality QA pairs covering both single-view and multi-view settings, with broad scene diversity and 19 instruction formats that reflect the varied ways users express spatial queries. InternSpatial-Bench complements this with a diagnostic single-view benchmark and an extended multi-view evaluation via rotation angle prediction, a task not addressed in prior work. Extensive experiments show that training on InternSpatial yields substantial improvements on spatial reasoning benchmarks while maintaining strong performance on general multimodal tasks. Despite its scale and diversity, our template-based generation pipeline may underrepresent the full richness of natural language in real-world scenarios. Future work will explore more expressive QA generation and open-ended spatial reasoning in interactive environments. We anticipate that our dataset will support downstream applications such as robotics, embodied AI, and AR/VR, where spatial understanding is essential.

## REPRODUCIBILITY STATEMENT

All results reported in this paper are fully reproducible using the provided resources. The training configurations are detailed in Section 4 and Appendix E, while the dataset pipeline is described in Section 3, Appendix A, and Appendix B.

## ETHICS STATEMENT

Our work does not involve sensitive personal data. All dataset components were collected from open-source and publicly available sources, with careful filtering to exclude content that may be discriminatory or infringe copyright. We do not foresee any negative societal impacts arising from our methods or datasets.

540 REFERENCES  
541

542 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-  
543 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical  
544 report. *arXiv preprint arXiv:2303.08774*, 2023.

545 Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jes-  
546 sica Chudnovsky, Diogo Costa, Baudouin De Moncault, Saurabh Garg, Theophile Gervet, So-  
547 ham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix,  
548 Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Mar-  
549 shall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pel-  
550 lat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile  
551 Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock,  
552 Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral  
553 12b, 2024. URL <https://arxiv.org/abs/2410.07073>.

554 N. Akkiraju, H. Edelsbrunner, M. Facello, P. Fu, and C. Varela. Alpha shapes: Definition and  
555 software. In *GCG: International Computational Geometry Software Workshop*, 1995.

556 Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid,  
557 Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting  
558 visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Con-  
559 ference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

560 Anthropic. The claude 3 model family: Opus, sonnet, haiku. <https://www.anthropic.com>, 2024. URL [https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model\\_Card\\_Claude\\_3.pdf](https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf).

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

566 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,  
567 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are  
568 few-shot learners. *Advances in neural information processing systems*, 2020.

569 Wenzhao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and  
570 Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE  
571 International Conference on Robotics and Automation (ICRA)*, 2025.

572 Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia.  
573 Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings  
574 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–  
575 14465, June 2024a.

576 Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing  
577 multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

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

581 Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shen-  
582 glong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source  
583 multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*,  
584 2024c.

585 An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang,  
586 and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In A. Globor-  
587 son, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in  
588 Neural Information Processing Systems*, volume 37, pp. 135062–135093. Curran Associates, Inc.,  
589 2024.

594 2024. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/f38cb4cf9a5eaa92b3cfa481832719c6-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/f38cb4cf9a5eaa92b3cfa481832719c6-Paper-Conference.pdf).

595

596

597 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit

598 Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the

599 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-

600 bilities. *arXiv preprint arXiv:2507.06261*, 2025.

601 OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models.

602 <https://github.com/open-compass/opencompass>, 2023.

603

604 Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo

605 Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic

606 urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern*

607 *Recognition (CVPR)*, 2016.

608 Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias

609 Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the*

610 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

611 Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig

612 Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of anno-

613 tated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022.

614

615 Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. 3d

616 concept learning and reasoning from multi-view images. In *IEEE/CVF Conference on Computer*

617 *Vision and Pattern Recognition (CVPR)*, 2023a.

618 Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang

619 Gan. 3d-llm: Injecting the 3d world into large language models. *NeurIPS*, 2023b.

620

621 Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,

622 and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Con-*

623 *ference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeFYf9>.

624

625 Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu,

626 Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation

627 model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pat-*

628 *tern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024. doi: 10.1109/TPAMI.2024.

629 3444912.

630

631 Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen,

632 Matthew Sticha, and David F. Fouhey. Perspective fields for single image camera calibration. In

633 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,

634 pp. 17307–17316, June 2023.

635

636 Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie

637 Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting lan-

638 guage and vision using crowdsourced dense image annotations. *International journal of computer*

639 *vision*, 123:32–73, 2017.

640

641 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan

642 Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer,

643 2024a. URL <https://arxiv.org/abs/2408.03326>.

644

645 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan

646 Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint*

647 *arXiv:2408.03326*, 2024b.

648

649 Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs:

650 Vision-language models as top-view spatial reasoners, 2024c. URL <https://arxiv.org/abs/2406.02537>.

648 Jianing Li, Xi Nan, Ming Lu, Li Du, and Shanghang Zhang. Proximity qa: Unleashing the power  
 649 of multi-modal large language models for spatial proximity analysis, 2024d. URL <https://arxiv.org/abs/2401.17862>.

650

651 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-  
 652 image pre-training with frozen image encoders and large language models. *arXiv preprint*  
 653 *arXiv:2301.12597*, 2023a.

654

655 KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang,  
 656 and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*,  
 657 2023b.

658

659 Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-  
 660 training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer*  
 661 *Vision and Pattern Recognition*, pp. 26689–26699, 2024.

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

666

667 Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng  
 668 Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of  
 669 ocr in large multimodal models. *Science China Information Sciences*, 67(12), 2024. ISSN  
 670 1869-1919. doi: 10.1007/s11432-024-4235-6. URL <http://dx.doi.org/10.1007/s11432-024-4235-6>.

671

672 Zhaoyang Liu, Yinan He, Wenhui Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang,  
 673 Yang Yang, Qingyun Li, Jiashuo Yu, et al. Interngpt: Solving vision-centric tasks by interacting  
 674 with chatbots beyond language. *arXiv preprint arXiv:2305.05662*, 2023.

675

676 Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel Chang, and Manolis Savva. Multi-  
 677 scan: Scalable rgbd scanning for 3d environments with articulated objects. In S. Koyejo,  
 678 S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neu-*  
 679 *ral Information Processing Systems*, volume 35, pp. 9058–9071. Curran Associates, Inc.,  
 680 2022. URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/3b3a83a5d86e1d424daefed43d998079-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/3b3a83a5d86e1d424daefed43d998079-Paper-Conference.pdf).

681

682 Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A bench-  
 683 mark for question answering about charts with visual and logical reasoning. In *Proceedings of*  
 684 *the annual meeting of the Association for Computational Linguistics*, pp. 2263–2279, 2022.

685

686 Inc. Meta Platforms. Llama-4-scout-17b-16e-instruct, 2025. URL <https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct>. Accessed: 2024-05-15.

687

688 OpenAI. Gpt-4o system card. <https://openai.com/index/gpt-4o-system-card/>,  
 689 2025.

690

691 Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu  
 692 Lin. Learning to localize objects improves spatial reasoning in visual-lmms. In *Proceedings of the*  
 693 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12977–12987, 2024.

694

695 Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham  
 696 Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Va-  
 697 sudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Fe-  
 698 ichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*,  
 699 2024. URL <https://arxiv.org/abs/2408.00714>.

700

701 Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-  
 702 baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittweiser, et al. Gem-  
 703 ini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*  
 704 *arXiv:2403.05530*, 2024.

702 Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh,  
 703 and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF*  
 704 *Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

705 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée  
 706 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and  
 707 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

709 Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Niessner. Rio: 3d  
 710 object instance re-localization in changing indoor environments. In *Proceedings IEEE Interna-*  
 711 *tional Conference on Computer Vision (ICCV)*, 2019.

712 Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and  
 713 Neel Joshi. Is a picture worth a thousand words? delving into spatial re-  
 714 reasoning for vision language models. In A. Globerson, L. Mackey, D. Belgrave,  
 715 A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural In-*  
 716 *formation Processing Systems*, volume 37, pp. 75392–75421. Curran Associates, Inc.,  
 717 2024a. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/89cc5e613d34f90de90c21e996e60b30-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/89cc5e613d34f90de90c21e996e60b30-Paper-Conference.pdf).

719 Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multi-  
 720 modal mathematical reasoning with math-vision dataset, 2024b. URL <https://arxiv.org/abs/2402.14804>.

723 Weiyun Wang, Min Shi, Qingyun Li, Wenhui Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao  
 724 Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition  
 725 and understanding of the open world. In *ICLR*, 2024c.

726 Wenhui Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong  
 727 Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for  
 728 vision-centric tasks. *NeurIPS*, 2023.

730 Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking  
 731 in space: How multimodal large language models see, remember, and recall spaces, 2024. URL  
 732 <https://arxiv.org/abs/2412.14171>.

733 Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui  
 734 Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for  
 735 referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024a.

737 Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and  
 738 Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024b. URL  
 739 <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>.

740 Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing.  
 741 *arXiv:1801.09847*, 2018.

742 Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: En-  
 743 hancing vision-language understanding with advanced large language models. *arXiv preprint*  
 744 *arXiv:2304.10592*, 2023a.

746 Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera:  
 747 In-the-wild monocular camera calibration. In A. Oh, T. Naumann, A. Globerson,  
 748 K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Process-*  
 749 *ing Systems*, volume 36, pp. 45137–45149. Curran Associates, Inc., 2023b. URL  
 750 [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/8db9279f593652ee9bb2223b4a2c43fa-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/8db9279f593652ee9bb2223b4a2c43fa-Paper-Conference.pdf).

752 Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai.  
 753 Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot  
 754 tasks. In *CVPR*, 2022.

## APPENDIX

## A EXPLANATION AND STATISTICS OF TASKS

InternSpatial and InternSpatial-Bench covers a total of 10 spatial reasoning tasks. The explanations of each task are shown in Table 5. We also count the number of QAs for each task in InternSpatial and InternSpatial-Bench, which are shown in Table 6 and Table 7 respectively.

Table 5: Explanation of tasks

Task	Description
Position Comparison	Compare the position of two objects in an image, involving three pairs of positional relationship: left/right, above/below, near/far.
Size Comparison	Compare the size of two objects in an image, involving three pairs of size relationship: wider/thinner, taller/shorter, larger/smaller.
Existence Estimation	Determine whether there are objects in the image whose positional/size relationships with the specified object meet the constraint conditions.
Object Counting	Estimate how many objects that meet the constraint conditions there are in a single image or multiple images.
Rotation Estimation	Estimate the rotation angle of an object between two images.
Absolute Distance	Estimate the closest distance between two objects given a serial of images.
Room Size	Estimate the volume of the room(s) given a serial of images.
Object Size	Estimate the longest dimension of an object given a serial of images.
Route Plan	Given a serial of images, choose what action should be performed between a sequence of actions in order to route to from a start point to a target.
Appearance Order	Given a serial of images, determine the first-time appearance order of several objects.

Table 6: Statistics of tasks in InternSpatial

Task	Related Views	# of QAs
Position Comparison	Single	6,214,628
Size Comparison	Single	3,227,124
Existence Estimation	Single	50,845
Object Counting	Single/Multiple	53,866
Rotation Estimation	Multiple	2,464,500
Absolute Distance	Multiple	14,596
Room Size	Multiple	1,181
Object Size	Multiple	3,709
Route Plan	Multiple	4,966
Appearance Order	Multiple	8,562

## B TEMPLATES FOR GENERATING QAS IN INTERNSPATIAL

The QAs in InternSpatial are generated by template-based generation method. Here we provide the full list of templates. The "[...]" in templates are placeholders which will be replaced by object references in different formats, values, choices, and so on. Several candidates are provided to be randomly selected in generation process to enrich the structure of sentences.

810  
811  
812  
813  
814  
815  
816  
817  
818

Table 7: Statistics of tasks in InternSpatial-Bench

Task	Position Comparison	Size Comparison	Rotation Estimation	Object Counting	Existence Estimation
# of QAs	1845	1855	409	899	1000

Listing 1: Templates for task *Position Comparison*

```

819 above_predict_templates = {
820     "question_templates": [
821         "[A] is placed higher than [B], isn't it?",
822         "Can we say that [A] is positioned above [B]?",
823         "Is it correct to assume that [A] is located at a higher level than [B]?",
824         "Is [A] placed higher than [B]?"
825     ],
826     "positive_answer_templates": [
827         "Absolutely, [A] is clearly positioned above [B].",
828         "Without a doubt, [A] is situated at a higher elevation than [B].",
829         "Indeed, [A] is placed higher than [B].",
830         "Certainly, [A] is located above [B]."
831     ],
832     "negative_answer_templates": [
833         "Not at all, [A] is actually below [B].",
834         "Definitely not, [A] is positioned lower than [B].",
835         "Sorry, but [A] is not higher than [B].",
836         "Unfortunately, [A] is not placed above [B]."
837     ]
838 }
839 below_predict_templates = {
840     "question_templates": [
841         "[A] is placed lower than [B], right?",
842         "Can we say that [A] is positioned below [B]?",
843         "Is it correct to assume that [A] is situated lower than [B]?",
844         "Is [A] placed lower than [B]?"
845     ],
846     "positive_answer_templates": [
847         "Absolutely, [A] is clearly positioned below [B].",
848         "Without a doubt, [A] is located lower than [B].",
849         "Indeed, [A] is situated beneath [B].",
850         "Certainly, [A] is found at a lower level than [B]."
851     ],
852     "negative_answer_templates": [
853         "Not at all, [A] is actually higher than [B].",
854         "Definitely not, [A] is positioned above [B].",
855         "In fact, [A] is situated higher than [B].",
856         "Quite the opposite, [A] is at a higher level than [B]."
857     ]
858 }
859 left_predict_templates = {
860     "question_templates": [
861         "[A] is more to the left of [B], isn't it?",
862         "Can we say that [A] is positioned more to the left than [B]?",
863         "Is it correct to assume that [A] is situated to the left of [B]?",
864         "Is [A] more to the left of [B]?"
865     ],
866     "positive_answer_templates": [
867         "Absolutely, [A] is clearly positioned to the left of [B].",
868         "Without a doubt, [A] is more to the left compared to [B].",
869         "Indeed, [A] is located to the left of [B].",
870         "Certainly, [A] is on the left side when compared to [B]."
871     ],
872     "negative_answer_templates": [
873

```

```

864     "Not at all, [A] is not more to the left of [B].",
865     "Actually, [A] is not positioned to the left of [B].",
866     "Contrary to that, [A] is not situated to the left of [B].",
867     "In fact, [A] is not on the left side when compared to [B]."
868   ]
869 }
870 right_predict_templates = {
871   "question_templates": [
872     "[A] is more to the right of [B], isn't it?",
873     "Can we say that [A] is positioned further to the right than [B]?",
874     "Is it correct to assume that [A] is located to the right side of [B]?",
875     "Is [A] more to the right of [B]?"
876   ],
877   "positive_answer_templates": [
878     "Absolutely, [A] is clearly positioned to the right of [B].",
879     "Indeed, [A] is noticeably more to the right compared to [B].",
880     "Without a doubt, [A] is situated further to the right than [B].",
881     "Certainly, [A] is distinctly to the right of [B]."
882   ],
883   "negative_answer_templates": [
884     "Not at all, [A] is actually to the left of [B].",
885     "Definitely not, [A] is not positioned to the right of [B].",
886     "In fact, [A] is on the left side of [B].",
887     "Contrary to that, [A] is not further to the right than [B]."
888   ]
889 }
890 near_predict_templates = {
891   "question_templates": [
892     "Is [A] positioned in front of [B]?",
893     "Does [A] precede [B] in this arrangement?",
894     "Is [A] in front of [B]?",
895     "Is [A] closer to the observer than [B]?"
896   ],
897   "positive_answer_templates": [
898     "Without a doubt, [A] stands nearer to the viewer than [B].",
899     "Definitely, [A] is more proximate to the observer than [B].",
900     "Indeed, [A] is in front of [B].",
901     "Absolutely, [A] is before [B]."
902   ],
903   "negative_answer_templates": [
904     "Not at all, [A] is not closer to the observer than [B].",
905     "No, [A] is not in front of [B].",
906     "Unfortunately, [A] is not ahead of [B].",
907     "Definitely not, [A] is not closer to the observer than [B]."
908   ]
909 }
910 far_predict_templates = {
911   "question_templates": [
912     "Is [A] situated behind [B]?",
913     "Does [A] lie behind [B]?",
914     "Is [A] to the rear of [B]?",
915     "Is [A] farther from the observer than [B]?"
916   ],
917   "positive_answer_templates": [
918     "Indeed, [A] is behind [B].",
919     "Yes, [A] is behind [B].",
920     "Without a doubt, [A] maintains a greater distance from the observer
921     than [B].",
922     "Certainly, [A] is positioned further away from the observer than [B].
923   ],
924   "negative_answer_templates": [
925     "No, [A] is not behind [B].",
926     "Incorrect, [A] is not behind [B].",

```

```

918     "That's wrong. [A] is not positioned behind [B].",
919     "Unfortunately, [A] is not to the rear of [B]."
920   ]
921 }
922 above_choice_templates = {
923   "question_templates": [
924     "Which one is positioned at a higher elevation, [A] or [B]?",
925     "In terms of altitude, which comes first, [A] or [B]?",
926     "Who stands taller, [A] or [B]?",
927     "Which is placed higher, [A] or [B]?"
928   ],
929   "answer_templates": [
930     "[O] is the one that is placed higher.",
931     "The higher position belongs to [O].",
932     "[O] occupies the superior location.",
933     "It is [O] that is situated at a greater height."
934   ]
935 }
936 below_choice_templates = {
937   "question_templates": [
938     "Which is positioned closer to the ground, [A] or [B]?",
939     "Which one is situated at a lower elevation, [A] or [B]?",
940     "Which of these is nearer to the base level, [A] or [B]?",
941     "Which is placed lower, [A] or [B]?"
942   ],
943   "answer_templates": [
944     "[O] is placed lower.",
945     "The lower position belongs to [O].",
946     "[O] occupies the lower spot.",
947     "Lower down, you'll find [O]."
948   ]
949 }
950 left_choice_templates = {
951   "question_templates": [
952     "Which is positioned further to the left, [A] or [B]?",
953     "In terms of leftward placement, which comes first, [A] or [B]?",
954     "When considering the left side, which one is closer, [A] or [B]?",
955     "Which is more to the left, [A] or [B]?"
956   ],
957   "answer_templates": [
958     "[O] is located more to the left.",
959     "The position of [O] is further to the left.",
960     "In comparison, [O] stands out as being more on the left.",
961     "It is evident that [O] is situated more towards the left."
962   ]
963 }
964 right_choice_templates = {
965   "question_templates": [
966     "Which is positioned further to the right, [A] or [B]?",
967     "In terms of horizontal alignment, which one is more to the right, [A]
968     ] or [B]?",
969     "When comparing their positions, which one is situated more to the
970     right, [A] or [B]?",
971     "Which is more to the right, [A] or [B]?"
972   ],
973   "answer_templates": [
974     "[O] is clearly more to the right.",
975     "The position of [O] is further to the right.",
976     "Comparing the two, [O] is definitively more to the right.",
977     "It is evident that [O] is positioned more to the right."
978   ]
979 }
980 near_choice_templates = {
981   "question_templates": [
982     "Which one is positioned further forward, [A] or [B]?",
983   ]

```

```

972     "Between [A] and [B], which object is closer to the observer?",  

973     "Can you identify which of the two, [A] or [B], is in the foremost  

974     position?",  

975     "Of the two, [A] and [B], which is closer to the front?"  

976   ],  

977   "answer_templates": [  

978     "[O] is in front.",  

979     "The frontmost object is [O].",  

980     "[O] is situated at the foremost position.",  

981     "Among the options, [O] is the one that is most ahead."  

982   ]  

983 far_choice_templates = {  

984   "question_templates": [  

985     "Which one is further back, [A] or [B]?",  

986     "Can you tell me which is positioned more towards the back, [A] or [B]  

987     ]?",  

988     "Between [A] and [B], which is more distant in the rear aspect?",  

989     "Comparing [A] and [B], which is more behind?"  

990   ],  

991   "answer_templates": [  

992     "[O] is definitely more behind.",  

993     "I can confirm that [O] is situated further back.",  

994     "[O] is clearly more behind than the other.",  

995     "There is no question that [O] is more behind."  

996   ]  

997 above_below_choice_templates = {  

998   "question_templates": [  

999     "Is [A] positioned higher or lower than [B]?",  

1000     "Does [A] lie above or beneath [B]?",  

1001     "Is [A] situated over or under [B]?",  

1002     "Is [A] above or below [B]?"  

1003   ],  

1004   "above_answer_templates": [  

1005     "[A] is above [B].",  

1006     "[A] is positioned higher than [B].",  

1007     "[A] lies over [B].",  

1008     "[A] is situated above [B]."  

1009   ],  

1010   "below_answer_templates": [  

1011     "[A] is below [B].",  

1012     "[A] is positioned lower than [B].",  

1013     "[A] lies under [B].",  

1014     "[A] is situated below [B]."  

1015   ]  

1016 }  

1017 left_right_choice_templates = {  

1018   "question_templates": [  

1019     "Is [A] relatively farther to the left or right than [B]?",  

1020     "Does [A] lie on the left or right side of [B]?",  

1021     "Is [A] to the left or right of [B]?"  

1022   ],  

1023   "left_answer_templates": [  

1024     "[A] is to the left of [B].",  

1025     "[A] occupies the left side relative to [B].",  

1026     "[A] lies on the left side of [B]."  

1027   ],  

1028   "right_answer_templates": [  

1029     "[A] is to the right of [B].",  

1030     "[A] occupies the right side relative to [B].",  

1031     "[A] lies on the right side of [B]."  

1032   ]  

1033 }  

1034 near_far_choice_templates = {  

1035   "question_templates": [  

1036     "Is [A] closer to the observer than [B]?",  

1037     "Does [A] occupy a closer position to the observer than [B]?",  

1038     "Is [A] situated closer to the observer than [B]?"  

1039   ],  

1040   "answer_templates": [  

1041     "[O] is closer to the observer.",  

1042     "The closer object is [O].",  

1043     "[O] is situated closer to the observer.",  

1044     "Among the options, [O] is the one that is closer to the observer."  

1045   ]  

1046 }  

1047

```

```

1026 "question_templates": [
1027     "Is [A] relatively nearer or farther from the observer than [B]?", 
1028     "Can you determine if [A] is closer or farther from the observer 
1029     compared to [B]?", 
1030     "Is [A] in front of or behind [B]?" 
1031 ],
1032 "near_answer_templates": [
1033     "[A] is closer to the observer than [B].", 
1034     "[A] is more proximate to the observer than [B].", 
1035     "[A] comes before [B].", 
1036     "[A] is in front of [B]."
1037 ],
1038 "far_answer_templates": [
1039     "[A] is farther from the observer than [B].", 
1040     "[A] is less proximate to the observer than [B].", 
1041     "[A] is behind [B]."
1042 ]
1043 }
```

Listing 2: Templates for task *Size Comparison*

```

1044 wide_predict_templates = {
1045     "question_templates": [
1046         "Is [A] broader than [B]?", 
1047         "Does [A] have a larger width compared to [B]?", 
1048         "Can we say that [A] spans more horizontally than [B]?", 
1049         "Is [A] wider than [B]?" 
1050     ],
1051     "positive_answer_templates": [
1052         "Yes, [A] is noticeably broader than [B].", 
1053         "Indeed, [A] has a significantly larger width than [B].", 
1054         "Absolutely, [A] spans more horizontally than [B].", 
1055         "Certainly, [A] is wider than [B]."
1056     ],
1057     "negative_answer_templates": [
1058         "No, [A] is not broader than [B].", 
1059         "In fact, [A] does not have a larger width compared to [B].", 
1060         "Sorry, but [A] does not span more horizontally than [B].", 
1061         "Unfortunately, [A] is not wider than [B]."
1062     ]
1063 }
1064 narrow_predict_templates = {
1065     "question_templates": [
1066         "Is [A] thinner than [B]?", 
1067         "Does [A] have a smaller width compared to [B]?", 
1068         "Is the width of [A] less than that of [B]?", 
1069         "Is [A] narrower than [B]?" 
1070     ],
1071     "positive_answer_templates": [
1072         "Yes, [A] is noticeably narrower than [B].", 
1073         "Indeed, [A] has a significantly smaller width than [B].", 
1074         "Absolutely, the width of [A] is less than that of [B].", 
1075         "Certainly, [A] is thinner than [B]."
1076     ],
1077     "negative_answer_templates": [
1078         "No, [A] is not narrower than [B]; in fact, it's wider.", 
1079         "Definitely not; [A] has a larger width than [B].", 
1080         "Not at all; the width of [A] exceeds that of [B].", 
1081         "No way; [A] is thicker than [B]."
1082     ]
1083 }
1084 tall_predict_templates = {
1085     "question_templates": [
1086         "[A] is taller than [B], isn't it?", 
1087         "Can we say that [A] surpasses [B] in height?", 
1088     ]
1089 }
```

```

1080     "Is it correct to assume that [A] is taller than [B]?",  

1081     "Is [A] taller than [B]?"  

1082   ],  

1083   "positive_answer_templates": [  

1084     "Absolutely, [A] towers over [B].",  

1085     "Without a doubt, [A] is significantly taller than [B].",  

1086     "Indeed, [A] outshines [B] in terms of height.",  

1087     "Unquestionably, [A] is taller than [B]."  

1088   ],  

1089   "negative_answer_templates": [  

1090     "Not at all, [B] is actually taller than [A].",  

1091     "Sorry, but [A] does not exceed [B] in height.",  

1092     "In fact, [B] surpasses [A] in height.",  

1093     "Regrettably, [A] falls short when compared to [B]'s height."  

1094   ]  

1095 vshort_predict_templates = {  

1096   "question_templates": [  

1097     "Is [A] shorter than [B] in vertical direction?",  

1098     "Does [A] have less height than [B]?",  

1099     "Is the vertical length of [A] smaller than that of [B]?",  

1100     "Is [A] shorter than [B] in vertical direction?"  

1101   ],  

1102   "positive_answer_templates": [  

1103     "Yes, [A] is indeed shorter than [B] in the vertical direction.",  

1104     "Absolutely, [A] has less height compared to [B].",  

1105     "Certainly, the vertical length of [A] is smaller than that of [B].",  

1106     "Without a doubt, [A] is shorter than [B] vertically."  

1107   ],  

1108   "negative_answer_templates": [  

1109     "No, [A] is not shorter than [B] in the vertical direction.",  

1110     "Definitely not, [A] does not have less height than [B].",  

1111     "Not at all, the vertical length of [A] is not smaller than that of [B].",  

1112     "Certainly not, [A] is not shorter than [B] vertically."  

1113   ]  

1114 large_predict_templates = {  

1115   "question_templates": [  

1116     "[A] is larger than [B], isn't it?",  

1117     "Can we say that [A] has a bigger size compared to [B]?",  

1118     "Is it correct to assume that [A] surpasses [B] in size?",  

1119     "Is [A] larger than [B]?"  

1120   ],  

1121   "positive_answer_templates": [  

1122     "Absolutely, [A] is noticeably larger than [B].",  

1123     "Without a doubt, [A] outsizes [B] significantly.",  

1124     "Indeed, [A] is clearly more expansive than [B].",  

1125     "Definitely, [A] dwarfs [B] in terms of size."  

1126   ],  

1127   "negative_answer_templates": [  

1128     "Not at all, [B] is actually larger than [A].",  

1129     "Quite the opposite, [B] surpasses [A] in size.",  

1130     "In fact, [B] is the larger one when compared to [A].",  

1131     "Sorry, but [B] is bigger than [A]."  

1132   ]  

1133 }  

1134 small_predict_templates = {  

1135   "question_templates": [  

1136     "[A] is smaller than [B], isn't it?",  

1137     "Can we say that [A] is smaller than [B]?",  

1138     "Is it true that [A] is smaller than [B]?",  

1139     "Is [A] smaller than [B]?"  

1140   ],  

1141   "positive_answer_templates": [  

1142     "Yes, [A] is indeed smaller than [B] in size.",  

1143     "Absolutely, [A] has less size compared to [B].",  

1144     "Certainly, the size of [A] is smaller than that of [B].",  

1145     "Without a doubt, [A] is smaller than [B] vertically."  

1146   ]  

1147 }  

1148

```

```

1134     "Absolutely, [A] is noticeably smaller than [B].",
1135     "Yes, [A] is indeed smaller than [B].",
1136     "Without a doubt, [A] is smaller than [B].",
1137     "Definitely, [A] is smaller than [B]."
1138 ],
1139 "negative_answer_templates": [
1140     "Not at all, [A] is actually larger than [B].",
1141     "No, [A] is not smaller than [B].",
1142     "Quite the opposite, [A] is bigger than [B].",
1143     "False, [A] is not smaller than [B]."
1144 ],
1145 wide_choice_templates = {
1146     "question_templates": [
1147         "Which has a greater width, [A] or [B]?",
1148         "In terms of width, which one is larger, [A] or [B]?",
1149         "When comparing widths, which one comes out on top, [A] or [B]?",
1150         "Which is wider, [A] or [B]?"
1151     ],
1152     "answer_templates": [
1153         "[O] is wider.",
1154         "The width of [O] is greater.",
1155         "Comparing the two, [O] has the larger width.",
1156         "In terms of width, [O] surpasses the other."
1157     ]
1158 },
1159 narrow_choice_templates = {
1160     "question_templates": [
1161         "Which has a smaller width, [A] or [B]?",
1162         "In terms of width, which one is less, [A] or [B]?",
1163         "When comparing widths, which one comes out smaller, [A] or [B]?",
1164         "Which is narrower, [A] or [B]?"
1165     ],
1166     "answer_templates": [
1167         "[O] is the narrower one.",
1168         "The narrower object is [O].",
1169         "[O] has the lesser width.",
1170         "Comparing the two, [O] is clearly narrower."
1171     ]
1172 },
1173 tall_choice_templates = {
1174     "question_templates": [
1175         "Which has a greater height, [A] or [B]?",
1176         "In terms of height, which one is superior, [A] or [B]?",
1177         "When comparing heights, which comes out on top, [A] or [B]?",
1178         "Which is taller, [A] or [B]?"
1179     ],
1180     "answer_templates": [
1181         "[O] is the taller one.",
1182         "The height of [O] surpasses the other.",
1183         "[O] stands out as the taller between the two.",
1184         "Comparatively speaking, [O] is taller."
1185     ]
1186 },
1187 vshort_choice_templates = {
1188     "question_templates": [
1189         "Which has a shorter vertical length, [A] or [B]?",
1190         "In terms of vertical measurement, which one is shorter, [A] or [B]?"
1191     ],
1192     "answer_templates": [
1193         "[O] is shorter in the vertical direction."
1194     ]
1195 }

```

```

1188     "The vertical length of [O] is less than the other.",
1189     "Comparing vertically, [O] comes out shorter.",
1190     "In terms of height, [O] is the shorter one."
1191   ]
1192 }
1193 large_choice_templates = {
1194   "question_templates": [
1195     "Which has a greater size, [A] or [B]?",
1196     "In terms of size, which one is bigger, [A] or [B]?",
1197     "When comparing sizes, which one comes out on top, [A] or [B]?",
1198     "Which is larger, [A] or [B]?"
1199   ],
1200   "answer_templates": [
1201     "[O] is the larger one.",
1202     "The bigger size belongs to [O].",
1203     "[O] surpasses the other in size.",
1204     "Comparatively speaking, [O] is the larger."
1205   ]
1206 }
1207 small_choice_templates = {
1208   "question_templates": [
1209     "Which has a smaller size, [A] or [B]?",
1210     "In terms of size, which one is smaller, [A] or [B]?",
1211     "When comparing sizes, which one comes out smaller, [A] or [B]?",
1212     "Which is smaller, [A] or [B]?"
1213   ],
1214   "answer_templates": [
1215     "[O] is the smaller one.",
1216     "The smaller of the two is [O].",
1217     "[O] has the smaller size.",
1218     "Comparing the two, [O] is the smaller."
1219   ]
1220 }
1221 wide_narrow_choice_templates = {
1222   "question_templates": [
1223     "Is [A] relatively wider or narrower than [B]?",
1224     "How does the width of [A] compare to [B]?",
1225     "Can you tell me if [A] has a greater or lesser width than [B]?",
1226     "Is [A] wider or narrower than [B]?"
1227   ],
1228   "wide_answer_templates": [
1229     "[A] is wider than [B].",
1230     "The width of [A] exceeds that of [B].",
1231     "[A] has a larger width compared to [B].",
1232     "In terms of width, [A] surpasses [B]."
1233   ],
1234   "narrow_answer_templates": [
1235     "[A] is narrower than [B].",
1236     "The width of [B] is greater than that of [A].",
1237     "[A] has a smaller width compared to [B].",
1238     "In terms of width, [B] surpasses [A]."
1239   ]
1240 }
1241 tall_short_choice_templates = {
1242   "question_templates": [
1243     "Is [A] relatively taller or shorter than [B]?",
1244     "How does the height of [A] compare to [B]?",
1245     "Can you determine if [A] is taller or shorter than [B]?",
1246     "Is [A] taller or shorter than [B]?"
1247   ],
1248   "tall_answer_templates": [
1249     "[A] is taller than [B].",
1250     "The height of [A] exceeds that of [B].",
1251     "[A] surpasses [B] in height.",
1252     "Compared to [B], [A] is definitely taller."
1253 ]

```

```

1242 ],
1243 "short_answer_templates": [
1244   "[A] is shorter than [B].",
1245   "In terms of height, [A] falls below [B].",
1246   "[B] is taller than [A].",
1247   "[A]'s height is less than that of [B]."
1248 ]
1249 }
1250 large_small_choice_templates = {
1251   "question_templates": [
1252     "Is [A] relatively larger or smaller than [B]?",
1253     "How does the size of [A] compare to [B]?",
1254     "Can you determine if [A] is bigger or smaller than [B]?",
1255     "Is [A] larger or smaller than [B]?"
1256   ],
1257   "large_answer_templates": [
1258     "[A] is larger than [B].",
1259     "The size of [A] exceeds that of [B].",
1260     "[A] surpasses [B] in size.",
1261     "Compared to [B], [A] is bigger."
1262   ],
1263   "small_answer_templates": [
1264     "[A] is smaller than [B].",
1265     "The size of [A] is less than that of [B].",
1266     "[B] is larger than [A].",
1267     "In comparison to [B], [A] is smaller."
1268   ]
1269 }
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295

```

Listing 3: Templates for task *Existence Estimation*

```

existence_left_templates = {
  "question_templates": [
    "Does [B] exist to the left of [A]?",
    "Is there [B] to the left of [A]?",
    "Is there [B] more to the left than [A]?"
  ],
  "positive_answer_templates": [
    "Yes."
  ],
  "negative_answer_templates": [
    "No."
  ]
}
existence_right_templates = {
  "question_templates": [
    "Is there [B] positioned more to the right than [A]?",
    "Does [B] exist to the right of [A]?",
    "Is there [B] locating to the rightsde of [A]?"
  ],
  "positive_answer_templates": [
    "Yes."
  ],
  "negative_answer_templates": [
    "No."
  ]
}
existence_above_templates = {
  "question_templates": [
    "Does [B] exist at higher elevation than [A]?",
    "Can you find [B] above [A]?",
    "Is there [B] that is located above [A]?"
  ],
  "positive_answer_templates": [
    "Yes."
  ]
}

```

```

1296     ],
1297     "negative_answer_templates": [
1298       "No."
1299     ]
1300   }
1301   existence_below_templates = {
1302     "question_templates": [
1303       "Is there [B] that is situated below [A]?",
1304       "Does [B] exist below [A]?",
1305       "Is there [B] positioned lower than [A]?"
1306     ],
1307     "positive_answer_templates": [
1308       "Yes."
1309     ],
1310     "negative_answer_templates": [
1311       "No."
1312     ]
1313   }
1314   existence_near_templates = {
1315     "question_templates": [
1316       "Does [B] exist near [A]?",
1317       "Is there [B] that is in front of [A]?",
1318       "Can you find [B] that is closer than observer than [A]?"
1319     ],
1320     "positive_answer_templates": [
1321       "Yes."
1322     ],
1323     "negative_answer_templates": [
1324       "No."
1325     ]
1326   }
1327   existence_far_templates = {
1328     "question_templates": [
1329       "Does [B] exist far from [A]?",
1330       "Is there [B] that is behind [A]?",
1331       "Does [B] exist behind [A]?"
1332     ],
1333     "positive_answer_templates": [
1334       "Yes."
1335     ],
1336     "negative_answer_templates": [
1337       "No."
1338     ]
1339   }
1340   existence_wide_templates = {
1341     "question_templates": [
1342       "Can you find [B] that is wider than [A]?",
1343       "Is there [B] that is wider than [A]?",
1344       "Is there [B] that has a larger extent in horizontal than [A]?"
1345     ],
1346     "positive_answer_templates": [
1347       "Yes."
1348     ],
1349     "negative_answer_templates": [
1350       "No."
1351     ]
1352   }
1353   existence_narrow_templates = {
1354     "question_templates": [
1355       "Is there [B] that is narrower than [A]?",
1356       "Can you find [B] that is narrower than [A]?",
1357       "Does [B] with smaller width than [A] exist?"
1358     ],
1359   }

```

```

1350
1351     "positive_answer_templates": [
1352         "Yes."
1353     ],
1354     "negative_answer_templates": [
1355         "No."
1356     ]
1357 }
1358 existence_tall_templates = {
1359     "question_templates": [
1360         "Is there [B] that is taller than [A]?",
1361         "Can you find [B] that has a larger height than [A]?"
1362         "Is there [B] that is larger in vertical than [A]?"
1363     ],
1364     "positive_answer_templates": [
1365         "Yes."
1366     ],
1367     "negative_answer_templates": [
1368         "No."
1369     ]
1370 }
1371 existence_vshort_templates = {
1372     "question_templates": [
1373         "Is there [B] that is shorter than [A] in vertical?",
1374         "Does [B] shorter than [A] exists?",
1375         "Is there [B] that has a smaller height than [A]?"
1376     ],
1377     "positive_answer_templates": [
1378         "Yes."
1379     ],
1380     "negative_answer_templates": [
1381         "No."
1382     ]
1383 }
1384 existence_large_templates = {
1385     "question_templates": [
1386         "Can you find [B] that is larger than [A]?",
1387         "Is there [B] that is larger than [A]?",
1388         "Does [B] exist that has a larger volume than [A]?"
1389     ],
1390     "positive_answer_templates": [
1391         "Yes."
1392     ],
1393     "negative_answer_templates": [
1394         "No."
1395     ]
1396 }
1397 existence_small_templates = {
1398     "question_templates": [
1399         "Is there [B] that is smaller in size than [A]?",
1400         "Does [B] with smaller size than [A] exist?",
1401         "Does [B] exist that is smaller than [A]?"
1402     ],
1403     "positive_answer_templates": [
1404         "Yes."
1405     ],
1406     "negative_answer_templates": [
1407         "No."
1408     ]
1409 }
1410

```

Listing 4: Templates for task *Object Counting*

```

1402 count_above_templates = {
1403     "question_templates": [
1404         "How many [B] are located higher than [A]?",
```

```

1404         "How many [B] are positioned higher than [A]?",  

1405         "How many [B] are above [A]?"  

1406     ],  

1407     "answer_templates": [  

1408         "[V]."  

1409     ]  

1410 }
1411 count_below_templates = {  

1412     "question_templates": [  

1413         "How many [B] are lower than [A]?",  

1414         "How many [B] are situated below [A]?",  

1415         "How many [B] are positioned lower than [A]?"  

1416     ],  

1417     "answer_templates": [  

1418         "[V]."  

1419     ]  

1420 }
1421 count_left_templates = {  

1422     "question_templates": [  

1423         "How many [B] are positioned to the left of [A]?",  

1424         "How many [B] are more to the left than [A]?",  

1425         "How many [B] are on the leftside of [A]?"  

1426     ],  

1427     "answer_templates": [  

1428         "[V]."  

1429     ]  

1430 }
1431 count_right_templates = {  

1432     "question_templates": [  

1433         "How many [B] are found to the right of [A]?",  

1434         "How many [B] lie to the rightsdie of [A]?",  

1435         "How many [B] are more to the right than [A]?"  

1436     ],  

1437     "answer_templates": [  

1438         "[V]."  

1439     ]  

1440 }
1441 count_near_templates = {  

1442     "question_templates": [  

1443         "How many [B] are closer to the observer than [A]?",  

1444         "How many [B] are in front of [A]?",  

1445         "How many [B] are located nearer to the observer than [A]?"  

1446     ],  

1447     "answer_templates": [  

1448         "[V]."  

1449     ]  

1450 }
1451 count_far_templates = {  

1452     "question_templates": [  

1453         "How many [B] are positioned farther from the observer than [A]?"  

1454     ],  

1455         "How many [B] are located behind [A]?"  

1456         "How many [B] are farther from the observer than [A]?"  

1457     ],  

1458     "answer_templates": [  

1459         "[V]."  

1460     ]  

1461 }
1462 }
1463 count_wide_templates = {  

1464     "question_templates": [  

1465         "How many [B] have a larger width compared to [A]?",  

1466         "How many [B] are wider than [A]?",  

1467         "How many [B] have a larger extent in horizontal than [A]?"  

1468     ],  

1469     "answer_templates": [  

1470

```

```

1458     "[V]."
1459 ]
1460 }
1461 count_narrow_templates = {
1462     "question_templates": [
1463         "How many [B] are narrower than [A]?",
1464         "How many [B] have a less width than that of [A]?",
1465         "How many [B] are thinner than [A]?"
1466     ],
1467     "answer_templates": [
1468         "[V]."
1469     ]
1470 }
1471 count_tall_templates = {
1472     "question_templates": [
1473         "How many [B] are taller than [A]?",
1474         "How many [B] surpass [A] in height?",
1475         "How many [B] have a larger extent in vertical than [A]?"
1476     ],
1477     "answer_templates": [
1478         "[V]."
1479     ]
1480 }
1481 count_vshort_templates = {
1482     "question_templates": [
1483         "How many [B] have less height than [A]?",
1484         "How many [B] are shorter than [A]?",
1485         "How many [B] have a smaller vertical length than that of [A]?"
1486     ],
1487     "answer_templates": [
1488         "[V]."
1489     ]
1490 }
1491 count_large_templates = {
1492     "question_templates": [
1493         "How many [B] are larger than [A]?",
1494         "How many [B] have a bigger size compared to [A]?",
1495         "How many [B] surpass [A] in size?"
1496     ],
1497     "answer_templates": [
1498         "[V]."
1499     ]
1500 }
1501 }
1502
1503
1504 Listing 5: Templates for multi-view tasks
1505 object_rotation_predict_templates = {
1506     "question_templates": [
1507         "Here are two images of the same object:\nImage 1:\n<image>\nImage 2:\n<image>\nPlease estimate how [A] in image 2 is rotated relative to image 1?",
1508         "Here are two images of the same object:\nImage 1:\n<image>\nImage 2:\n<image>\nIn what direction and by what angle has [A] in image 2 been rotated from its position in image 1?"
1509     ],
1510
1511

```

```

1512 "clockwise_answer_templates": [
1513     "[A] rotates about [D] degrees clockwise.",
1514     "[A] turns clockwise by about [D] degrees.",
1515     "[A] undergoes approximately a [D] degree clockwise rotation."
1516 ],
1517 "counterclockwise_answer_templates": [
1518     "[A] rotates about [D] degrees counterclockwise.",
1519     "[A] turns counterclockwise by about [D] degrees.",
1520     "[A] undergoes approximately a [D] degree counterclockwise rotation."
1521 ],
1522 "rotate_180_answer_templates": [
1523     "[A] rotates about [D] degrees.",
1524     "[A] turns by about [D] degrees.",
1525     "[A] undergoes approximately a [D] degree rotation."
1526 ]
1527 route_plan_templates = {
1528     "question_templates": [
1529         "Image-1: <image>\nImage-2: <image>\nImage-3: <image>\nImage-4: <
1530         image>\nImage-5: <image>\nImage-6: <image>\nImage-7: <image>\nImage
1531         -8: <image>\nImage-9: <image>\nImage-10: <image>\nImage-11: <image>\n
1532         Image-12: <image>\nImage-13: <image>\nImage-14: <image>\nImage-15: <
1533         image>\nImage-16: <image>\nImage-17: <image>\nImage-18: <image>\n
1534         Image-19: <image>\nImage-20: <image>\nImage-21: <image>\nImage-22: <
1535         image>\nImage-23: <image>\nImage-24: <image>\nYou are a robot
1536         beginning at the column and facing the staircase. You want to
1537         navigate to the grand staircase. You will perform the following
1538         actions (Note: for each [please fill in], choose either 'turn back,'
1539         'turn left,' or 'turn right.'): 1. Go forward until the columns. 2. [please
1540         fill in]. 3. Go forward until the steps. 4. Stop on the
1541         landing.\nA. Turn Right\nB. Turn Left\nC. Turn Back\nAnswer with the
1542         option's letter from the given choices directly."
1543     ],
1544     "answer_templates": [
1545         "[O]"
1546     ]
1547 }
1548 abs_dist_templates = {
1549     "question_templates": [
1550         "Image-1: <image>\nImage-2: <image>\nImage-3: <image>\nImage-4: <
1551         image>\nImage-5: <image>\nImage-6: <image>\nImage-7: <image>\nImage
1552         -8: <image>\nImage-9: <image>\nImage-10: <image>\nImage-11: <image>\n
1553         Measuring from the closest point of each object, what is the
1554         distance between [A] and [B] (in meters)?\nPlease answer the question
1555         using a single word or phrase."
1556     ],
1557     "answer_templates": [
1558         "[V]"
1559     ]
1560 }
1561 obj_count_templates = {
1562     "question_templates": [
1563         "Image-1: <image>\nImage-2: <image>\nImage-3: <image>\nImage-4: <
1564         image>\nImage-5: <image>\nImage-6: <image>\nImage-7: <image>\nImage
1565         -8: <image>\nImage-9: <image>\nImage-10: <image>\nImage-11: <image>\n
1566         nThese are frames of a video.\nHow many [A] (s) are in this room?\n
1567         nPlease answer the question using a single word or phrase."
1568     ],
1569     "answer_templates": [
1570         "[V]"
1571     ]
1572 }
1573 room_size_templates = {
1574     "question_templates": [

```

```

1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
  "Image-1: <image>\nImage-2: <image>\nImage-3: <image>\nImage-4: <
  image>\nImage-5: <image>\nImage-6: <image>\nImage-7: <image>\nImage
  -8: <image>\nImage-9: <image>\nImage-10: <image>\nImage-11: <image>\n
  These are frames of a video.\nWhat is the size of this room (in
  square meters)?\nIf multiple rooms are shown, estimate the size of
  the combined space.\nPlease answer the question using a single word
  or phrase."
],
"answer_templates": [
  "[V]"
]
}
rel_dist_templates = {
  "question_templates": [
    "Image-1: <image>\nImage-2: <image>\nImage-3: <image>\nImage-4: <
    image>\nImage-5: <image>\nImage-6: <image>\nImage-7: <image>\nImage
    -8: <image>\nImage-9: <image>\nImage-10: <image>\nThese are frames of
    a video.\nMeasuring from the closest point of each object, which of
    these objects ([B0], [B1], [B2], [B3]) is the closest to [A]?nA. [B0]\n
    nB. [B1]\nC. [B2]\nD. [B3]\nAnswer with the option's letter from the
    given choices directly."
],
"answer_templates": [
  "[O]"
]
}
object_size_templates = {
  "question_templates": [
    "Image-1: <image>\nImage-2: <image>\nImage-3: <image>\nImage-4: <
      image>\nImage-5: <image>\nImage-6: <image>\nImage-7: <image>\nImage
      -8: <image>\nImage-9: <image>\nImage-10: <image>\nImage-11: <image>\n
      These are frames of a video.\nWhat is the length of the longest
      dimension (length, width, or height) of [A], measured in centimeters
      ?nPlease answer the question using a single word or phrase."
],
"answer_templates": [
  "[V]"
]
}
appear_order_templates = {
  "question_templates": [
    "Image-1: <image>\nImage-2: <image>\nImage-3: <image>\nImage-4: <
      image>\nImage-5: <image>\nImage-6: <image>\nImage-7: <image>\nImage
      -8: <image>\nImage-9: <image>\nImage-10: <image>\nImage-11: <image>\n
      These are frames of a video.\nWhat will be the first-time appearance
      order of the following categories in the video: [A0], [A1], [A2], [A3]?nA. [B0]\nB. [B1]\nC. [B2]\nD. [B3]\nAnswer with the option's
      letter from the given choices directly."
],
"answer_templates": [
  "[O]"
]
}
}

```

## C VLM-ASSISTED ANNOTATION FOR INTERNSPATIAL

As described in Dataset section, we involved open-source VLM to do the object detection, captioning, and grounding in the pipeline of InternSpatial generation. We use QWen2.5-VL 72B(Bai et al., 2025) as the assistant. For each process, we design corresponding prompt to make the VLM understand what should do and what should output. Here we provide the prompts for these processes.

Listing 6: Prompts for detecting objects in images

```
messages = [{"role": "system", "content": f"""}
```

```

1620 You are an object detector. Given an image, you should find all objects
1621 in image with grounding. The term "object" includes all living and
1622 non-living things. For each detected object, you should assign a
1623 label, which represents what the object is. You should also describe
1624 each detected object in detail with a phrase. The description can
1625 contain appearance, function, action, etc.
1626
1627 Output format: The response should be in json format, which contains a
1628 list of dicts. Each dict is for an detected object and has three keys
1629 : "label" for the label, "caption" for the description and "box" for
1630 the grounding. The description should be lowercases and no period at
1631 the end. The grounding should be a list of four ints [x1, y1, x2, y2]
1632 ], where (x1, y1) is the top-left coord and (x2, y2) is the bottom-
1633 right coord. Compact the responded json in one line.
1634 """
1635 messages.append({"role": "user", "content": '\n'.join(query)})
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673

```

Listing 7: Prompts for captioning objects given bounding boxes

```

messages = [{"role": "system", "content": """
You are an language assistant. You will be given an array dict. Each
dict contains a field "box" for grounding box and an optional field "
label" for label of a object in image. Your task is to generate brief
descriptions with less than ten words for these objects. Output one
description per line.

Here is an example:

Input:
[{"box": [10, 20, 300, 400], "label": "bus"}, {"box": [42, 512, 64,
890]}]

Output:
a blue bus seat with a suitcase partially resting on it
a red car on right side
"""}]
messages.append({"role": "user", "content": '\n'.join(query)})

```

Listing 8: Prompts for grounding objects given captions

```

messages = [{"role": "system", "content": """
You are a professional image annotator. I will give you an image and a
phrase about one or more objects in the image. Please detect all
objects matching the phrase. The response should be a JSON object,
containing a field "boxes". "boxes" is a list of grounding boxes [x1,
y1, x2, y2].

Example Output:
{
  "boxes": [[192, 29, 321, 49], [19, 65, 392, 569], [59, 102, 439,
139]]
}
"""}]
messages.append({"role": "user", "content": '\n'.join(query)})

```

## D MORE DETAILS ABOUT INTERNSPATIAL-BENCH GENERATION PIPELINE

The generation pipeline of InternSpatial-Bench does not rely on existing annotations. Starting from the images, we carried out four steps including image filtering, image captioning, question design, and object grounding to obtain the necessary 2D annotations for the questions of the benchmark and the generation of answers. In this steps, we design prompts respectively to enable the Visual Language Model (VLM) to automatically generate intermediate results. These prompts are presented in Listing 9, Listing 10, Listing 11, Listing 12, Listing 13, and Listing 8. Subsequently, we reused

1674 the processes in the second and third stages of the training dataset pipeline to generate answers  
 1675 and expand the instruction formats. After generated the QA pairs, we invited experienced human  
 1676 annotators to conduct manual verification of all the pairs to ensure the quality of the benchmark.  
 1677

1678 **Listing 9: Prompts for filtering images**

```
1679 messages = [{"role": "system", "content": f"""
1680     You are a helpful visual assistant. Please determine whether the
1681     following conditions are met:
1682     1. 5-or-more-objects: There are at least 5 objects in the image.
1683     2. cartoon: This is a cartoon image.
1684     3. group-of-images: This is a group of images.
1685     4. screenshot: This is a screenshot.
1686     5. realistic-image: This is a realistic image captured by camera.

1687     For each condition, answer true of false. Response in JSON dict with
1688     five fields: "5-or-more-objects", "cartoon", "group-of-images", "
1689     screenshot", "realistic-image
1690 """]}
1691 messages.append({"role": "user", "content": '\n'.join(query)})
```

1692 **Listing 10: Prompts for captioning images**

```
1693 messages = [{"role": "system", "content": f"""
1694     You are a helpful visual assistant. Please describe the image as detail
1695     as possible. Then detect all top-level objects and return their
1696     detailed descriptions (top-level means it's not a part of another
1697     object).
1698 """]}
1699 messages.append({"role": "user", "content": '\n'.join(query)})
```

1700 **Listing 11: Prompts for design questions of task *Position Comparison***

```
1701 messages = [{"role": "system", "content": f"""
1702     I will give you an image and a description about the image. You should
1703     design 2 questions regarding position judgments around top-level
1704     objects in the image (top-level means it's not a part of another
1705     object). The question should involve two object (anchor, target) and
1706     a type of relationship:
1707
1708     - *The anchor and target object* should be randomly chosen from the top
1709     - level objects. You possibly need to add the attributions about
1710     - appearance, behavior, posture, position in the image, etc. to the
1711     - description of the anchor and target objects so that they can be
1712     - distinguished from others.
1713     - *The relationship* should be randomly selected from: more to the left
1714     - , more to the right, closer (to the observer), farther (from the
1715     - observer), higher, lower.
1716     - Only design questions about top-level objects. Ignore those not in
1717     - top-level objects list. Ignore environment objects such as water, sky
1718     - , grass, cloud, etc.
1719
1720     After that, generate 2 more questions based on the designed questions
1721     by choose another relationship and keep other parts unchanged.
1722
1723     Please respond in JSON format. All content should be in English. Here
1724     is an example of output:
1725
1726     {
1727         "questions": [
1728             {
1729                 "question": "Is the wooden chair positioned higher than the blue
1730                 table?",
1731                 "anchor": "blue table",
1732                 "target": "wooden chair",
1733                 "relationship": "higher",
1734                 "task": "position"
1735             }
1736         ]
1737     }
1738
1739     [
1740         {
1741             "question": "Is the wooden chair positioned higher than the blue
1742             table?",
1743             "anchor": "blue table",
1744             "target": "wooden chair",
1745             "relationship": "higher",
1746             "task": "position"
1747         }
1748     ]
1749
1750     [
1751         {
1752             "question": "Is the wooden chair positioned higher than the blue
1753             table?",
1754             "anchor": "blue table",
1755             "target": "wooden chair",
1756             "relationship": "higher",
1757             "task": "position"
1758         }
1759     ]
1760
1761     [
1762         {
1763             "question": "Is the wooden chair positioned higher than the blue
1764             table?",
1765             "anchor": "blue table",
1766             "target": "wooden chair",
1767             "relationship": "higher",
1768             "task": "position"
1769         }
1770     ]
1771
1772     [
1773         {
1774             "question": "Is the wooden chair positioned higher than the blue
1775             table?",
1776             "anchor": "blue table",
1777             "target": "wooden chair",
1778             "relationship": "higher",
1779             "task": "position"
1780         }
1781     ]
1782
1783     [
1784         {
1785             "question": "Is the wooden chair positioned higher than the blue
1786             table?",
1787             "anchor": "blue table",
1788             "target": "wooden chair",
1789             "relationship": "higher",
1790             "task": "position"
1791         }
1792     ]
1793
1794     [
1795         {
1796             "question": "Is the wooden chair positioned higher than the blue
1797             table?",
1798             "anchor": "blue table",
1799             "target": "wooden chair",
1800             "relationship": "higher",
1801             "task": "position"
1802         }
1803     ]
1804
1805     [
1806         {
1807             "question": "Is the wooden chair positioned higher than the blue
1808             table?",
1809             "anchor": "blue table",
1810             "target": "wooden chair",
1811             "relationship": "higher",
1812             "task": "position"
1813         }
1814     ]
1815
1816     [
1817         {
1818             "question": "Is the wooden chair positioned higher than the blue
1819             table?",
1820             "anchor": "blue table",
1821             "target": "wooden chair",
1822             "relationship": "higher",
1823             "task": "position"
1824         }
1825     ]
1826
1827     [
1828         {
1829             "question": "Is the wooden chair positioned higher than the blue
1830             table?",
1831             "anchor": "blue table",
1832             "target": "wooden chair",
1833             "relationship": "higher",
1834             "task": "position"
1835         }
1836     ]
1837
1838     [
1839         {
1840             "question": "Is the wooden chair positioned higher than the blue
1841             table?",
1842             "anchor": "blue table",
1843             "target": "wooden chair",
1844             "relationship": "higher",
1845             "task": "position"
1846         }
1847     ]
1848
1849     [
1850         {
1851             "question": "Is the wooden chair positioned higher than the blue
1852             table?",
1853             "anchor": "blue table",
1854             "target": "wooden chair",
1855             "relationship": "higher",
1856             "task": "position"
1857         }
1858     ]
1859
1860     [
1861         {
1862             "question": "Is the wooden chair positioned higher than the blue
1863             table?",
1864             "anchor": "blue table",
1865             "target": "wooden chair",
1866             "relationship": "higher",
1867             "task": "position"
1868         }
1869     ]
1870
1871     [
1872         {
1873             "question": "Is the wooden chair positioned higher than the blue
1874             table?",
1875             "anchor": "blue table",
1876             "target": "wooden chair",
1877             "relationship": "higher",
1878             "task": "position"
1879         }
1880     ]
1881
1882     [
1883         {
1884             "question": "Is the wooden chair positioned higher than the blue
1885             table?",
1886             "anchor": "blue table",
1887             "target": "wooden chair",
1888             "relationship": "higher",
1889             "task": "position"
1890         }
1891     ]
1892
1893     [
1894         {
1895             "question": "Is the wooden chair positioned higher than the blue
1896             table?",
1897             "anchor": "blue table",
1898             "target": "wooden chair",
1899             "relationship": "higher",
1900             "task": "position"
1901         }
1902     ]
1903
1904     [
1905         {
1906             "question": "Is the wooden chair positioned higher than the blue
1907             table?",
1908             "anchor": "blue table",
1909             "target": "wooden chair",
1910             "relationship": "higher",
1911             "task": "position"
1912         }
1913     ]
1914
1915     [
1916         {
1917             "question": "Is the wooden chair positioned higher than the blue
1918             table?",
1919             "anchor": "blue table",
1920             "target": "wooden chair",
1921             "relationship": "higher",
1922             "task": "position"
1923         }
1924     ]
1925
1926     [
1927         {
1928             "question": "Is the wooden chair positioned higher than the blue
1929             table?",
1930             "anchor": "blue table",
1931             "target": "wooden chair",
1932             "relationship": "higher",
1933             "task": "position"
1934         }
1935     ]
1936
1937     [
1938         {
1939             "question": "Is the wooden chair positioned higher than the blue
1940             table?",
1941             "anchor": "blue table",
1942             "target": "wooden chair",
1943             "relationship": "higher",
1944             "task": "position"
1945         }
1946     ]
1947
1948     [
1949         {
1950             "question": "Is the wooden chair positioned higher than the blue
1951             table?",
1952             "anchor": "blue table",
1953             "target": "wooden chair",
1954             "relationship": "higher",
1955             "task": "position"
1956         }
1957     ]
1958
1959     [
1960         {
1961             "question": "Is the wooden chair positioned higher than the blue
1962             table?",
1963             "anchor": "blue table",
1964             "target": "wooden chair",
1965             "relationship": "higher",
1966             "task": "position"
1967         }
1968     ]
1969
1970     [
1971         {
1972             "question": "Is the wooden chair positioned higher than the blue
1973             table?",
1974             "anchor": "blue table",
1975             "target": "wooden chair",
1976             "relationship": "higher",
1977             "task": "position"
1978         }
1979     ]
1980
1981     [
1982         {
1983             "question": "Is the wooden chair positioned higher than the blue
1984             table?",
1985             "anchor": "blue table",
1986             "target": "wooden chair",
1987             "relationship": "higher",
1988             "task": "position"
1989         }
1990     ]
1991
1992     [
1993         {
1994             "question": "Is the wooden chair positioned higher than the blue
1995             table?",
1996             "anchor": "blue table",
1997             "target": "wooden chair",
1998             "relationship": "higher",
1999             "task": "position"
2000         }
2001     ]
2002
2003     [
2004         {
2005             "question": "Is the wooden chair positioned higher than the blue
2006             table?",
2007             "anchor": "blue table",
2008             "target": "wooden chair",
2009             "relationship": "higher",
2010             "task": "position"
2011         }
2012     ]
2013
2014     [
2015         {
2016             "question": "Is the wooden chair positioned higher than the blue
2017             table?",
2018             "anchor": "blue table",
2019             "target": "wooden chair",
2020             "relationship": "higher",
2021             "task": "position"
2022         }
2023     ]
2024
2025     [
2026         {
2027             "question": "Is the wooden chair positioned higher than the blue
2028             table?",
2029             "anchor": "blue table",
2030             "target": "wooden chair",
2031             "relationship": "higher",
2032             "task": "position"
2033         }
2034     ]
2035
2036     [
2037         {
2038             "question": "Is the wooden chair positioned higher than the blue
2039             table?",
2040             "anchor": "blue table",
2041             "target": "wooden chair",
2042             "relationship": "higher",
2043             "task": "position"
2044         }
2045     ]
2046
2047     [
2048         {
2049             "question": "Is the wooden chair positioned higher than the blue
2050             table?",
2051             "anchor": "blue table",
2052             "target": "wooden chair",
2053             "relationship": "higher",
2054             "task": "position"
2055         }
2056     ]
2057
2058     [
2059         {
2060             "question": "Is the wooden chair positioned higher than the blue
2061             table?",
2062             "anchor": "blue table",
2063             "target": "wooden chair",
2064             "relationship": "higher",
2065             "task": "position"
2066         }
2067     ]
2068
2069     [
2070         {
2071             "question": "Is the wooden chair positioned higher than the blue
2072             table?",
2073             "anchor": "blue table",
2074             "target": "wooden chair",
2075             "relationship": "higher",
2076             "task": "position"
2077         }
2078     ]
2079
2080     [
2081         {
2082             "question": "Is the wooden chair positioned higher than the blue
2083             table?",
2084             "anchor": "blue table",
2085             "target": "wooden chair",
2086             "relationship": "higher",
2087             "task": "position"
2088         }
2089     ]
2090
2091     [
2092         {
2093             "question": "Is the wooden chair positioned higher than the blue
2094             table?",
2095             "anchor": "blue table",
2096             "target": "wooden chair",
2097             "relationship": "higher",
2098             "task": "position"
2099         }
2100     ]
2101
2102     [
2103         {
2104             "question": "Is the wooden chair positioned higher than the blue
2105             table?",
2106             "anchor": "blue table",
2107             "target": "wooden chair",
2108             "relationship": "higher",
2109             "task": "position"
2110         }
2111     ]
2112
2113     [
2114         {
2115             "question": "Is the wooden chair positioned higher than the blue
2116             table?",
2117             "anchor": "blue table",
2118             "target": "wooden chair",
2119             "relationship": "higher",
2120             "task": "position"
2121         }
2122     ]
2123
2124     [
2125         {
2126             "question": "Is the wooden chair positioned higher than the blue
2127             table?",
2128             "anchor": "blue table",
2129             "target": "wooden chair",
2130             "relationship": "higher",
2131             "task": "position"
2132         }
2133     ]
2134
2135     [
2136         {
2137             "question": "Is the wooden chair positioned higher than the blue
2138             table?",
2139             "anchor": "blue table",
2140             "target": "wooden chair",
2141             "relationship": "higher",
2142             "task": "position"
2143         }
2144     ]
2145
2146     [
2147         {
2148             "question": "Is the wooden chair positioned higher than the blue
2149             table?",
2150             "anchor": "blue table",
2151             "target": "wooden chair",
2152             "relationship": "higher",
2153             "task": "position"
2154         }
2155     ]
2156
2157     [
2158         {
2159             "question": "Is the wooden chair positioned higher than the blue
2160             table?",
2161             "anchor": "blue table",
2162             "target": "wooden chair",
2163             "relationship": "higher",
2164             "task": "position"
2165         }
2166     ]
2167
2168     [
2169         {
2170             "question": "Is the wooden chair positioned higher than the blue
2171             table?",
2172             "anchor": "blue table",
2173             "target": "wooden chair",
2174             "relationship": "higher",
2175             "task": "position"
2176         }
2177     ]
2178
2179     [
2180         {
2181             "question": "Is the wooden chair positioned higher than the blue
2182             table?",
2183             "anchor": "blue table",
2184             "target": "wooden chair",
2185             "relationship": "higher",
2186             "task": "position"
2187         }
2188     ]
2189
2190     [
2191         {
2192             "question": "Is the wooden chair positioned higher than the blue
2193             table?",
2194             "anchor": "blue table",
2195             "target": "wooden chair",
2196             "relationship": "higher",
2197             "task": "position"
2198         }
2199     ]
2200
2201     [
2202         {
2203             "question": "Is the wooden chair positioned higher than the blue
2204             table?",
2205             "anchor": "blue table",
2206             "target": "wooden chair",
2207             "relationship": "higher",
2208             "task": "position"
2209         }
2210     ]
2211
2212     [
2213         {
2214             "question": "Is the wooden chair positioned higher than the blue
2215             table?",
2216             "anchor": "blue table",
2217             "target": "wooden chair",
2218             "relationship": "higher",
2219             "task": "position"
2220         }
2221     ]
2222
2223     [
2224         {
2225             "question": "Is the wooden chair positioned higher than the blue
2226             table?",
2227             "anchor": "blue table",
2228             "target": "wooden chair",
2229             "relationship": "higher",
2230             "task": "position"
2231         }
2232     ]
2233
2234     [
2235         {
2236             "question": "Is the wooden chair positioned higher than the blue
2237             table?",
2238             "anchor": "blue table",
2239             "target": "wooden chair",
2240             "relationship": "higher",
2241             "task": "position"
2242         }
2243     ]
2244
2245     [
2246         {
2247             "question": "Is the wooden chair positioned higher than the blue
2248             table?",
2249             "anchor": "blue table",
2250             "target": "wooden chair",
2251             "relationship": "higher",
2252             "task": "position"
2253         }
2254     ]
2255
2256     [
2257         {
2258             "question": "Is the wooden chair positioned higher than the blue
2259             table?",
2260             "anchor": "blue table",
2261             "target": "wooden chair",
2262             "relationship": "higher",
2263             "task": "position"
2264         }
2265     ]
2266
2267     [
2268         {
2269             "question": "Is the wooden chair positioned higher than the blue
2270             table?",
2271             "anchor": "blue table",
2272             "target": "wooden chair",
2273             "relationship": "higher",
2274             "task": "position"
2275         }
2276     ]
2277
2278     [
2279         {
2280             "question": "Is the wooden chair positioned higher than the blue
2281             table?",
2282             "anchor": "blue table",
2283             "target": "wooden chair",
2284             "relationship": "higher",
2285             "task": "position"
2286         }
2287     ]
2288
2289     [
2290         {
2291             "question": "Is the wooden chair positioned higher than the blue
2292             table?",
2293             "anchor": "blue table",
2294             "target": "wooden chair",
2295             "relationship": "higher",
2296             "task": "position"
2297         }
2298     ]
2299
2300     [
2301         {
2302             "question": "Is the wooden chair positioned higher than the blue
2303             table?",
2304             "anchor": "blue table",
2305             "target": "wooden chair",
2306             "relationship": "higher",
2307             "task": "position"
2308         }
2309     ]
2310
2311     [
2312         {
2313             "question": "Is the wooden chair positioned higher than the blue
2314             table?",
2315             "anchor": "blue table",
2316             "target": "wooden chair",
2317             "relationship": "higher",
2318             "task": "position"
2319         }
2320     ]
2321
2322     [
2323         {
2324             "question": "Is the wooden chair positioned higher than the blue
2325             table?",
2326             "anchor": "blue table",
2327             "target": "wooden chair",
2328             "relationship": "higher",
2329             "task": "position"
2330         }
2331     ]
2332
2333     [
2334         {
2335             "question": "Is the wooden chair positioned higher than the blue
2336             table?",
2337             "anchor": "blue table",
2338             "target": "wooden chair",
2339             "relationship": "higher",
2340             "task": "position"
2341         }
2342     ]
2343
2344     [
2345         {
2346             "question": "Is the wooden chair positioned higher than the blue
2347             table?",
2348             "anchor": "blue table",
2349             "target": "wooden chair",
2350             "relationship": "higher",
2351             "task": "position"
2352         }
2353     ]
2354
2355     [
2356         {
2357             "question": "Is the wooden chair positioned higher than the blue
2358             table?",
2359             "anchor": "blue table",
2360             "target": "wooden chair",
2361             "relationship": "higher",
2362             "task": "position"
2363         }
2364     ]
2365
2366     [
2367         {
2368             "question": "Is the wooden chair positioned higher than the blue
2369             table?",
2370             "anchor": "blue table",
2371             "target": "wooden chair",
2372             "relationship": "higher",
2373             "task": "position"
2374         }
2375     ]
2376
2377     [
2378         {
2379             "question": "Is the wooden chair positioned higher than the blue
2380             table?",
2381             "anchor": "blue table",
2382             "target": "wooden chair",
2383             "relationship": "higher",
2384             "task": "position"
2385         }
2386     ]
2387
2388     [
2389         {
2390             "question": "Is the wooden chair positioned higher than the blue
2391             table?",
2392             "anchor": "blue table",
2393             "target": "wooden chair",
2394             "relationship": "higher",
2395             "task": "position"
2396         }
2397     ]
2398
2399     [
2400         {
2401             "question": "Is the wooden chair positioned higher than the blue
2402             table?",
2403             "anchor": "blue table",
2404             "target": "wooden chair",
2405             "relationship": "higher",
2406             "task": "position"
2407         }
2408     ]
2409
2410     [
2411         {
2412             "question": "Is the wooden chair positioned higher than the blue
2413             table?",
2414             "anchor": "blue table",
2415             "target": "wooden chair",
2416             "relationship": "higher",
2417             "task": "position"
2418         }
2419     ]
2420
2421     [
2422         {
2423             "question": "Is the wooden chair positioned higher than the blue
2424             table?",
2425             "anchor": "blue table",
2426             "target": "wooden chair",
2427             "relationship": "higher",
2428             "task": "position"
2429         }
2430     ]
2431
2432     [
2433         {
2434             "question": "Is the wooden chair positioned higher than the blue
2435             table?",
2436             "anchor": "blue table",
2437             "target": "wooden chair",
2438             "relationship": "higher",
2439             "task": "position"
2440         }
2441     ]
2442
2443     [
2444         {
2445             "question": "Is the wooden chair positioned higher than the blue
2446             table?",
2447             "anchor": "blue table",
2448             "target": "wooden chair",
2449             "relationship": "higher",
2450             "task": "position"
2451         }
2452     ]
2453
2454     [
2455         {
2456             "question": "Is the wooden chair positioned higher than the blue
2457             table?",
2458             "anchor": "blue table",
2459             "target": "wooden chair",
2460             "relationship": "higher",
2461             "task": "position"
2462         }
2463     ]
2464
2465     [
2466         {
2467             "question": "Is the wooden chair positioned higher than the blue
2468             table?",
2469             "anchor": "blue table",
2470             "target": "wooden chair",
2471             "relationship": "higher",
2472             "task": "position"
2473         }
2474     ]
2475
2476     [
2477         {
2478             "question": "Is the wooden chair positioned higher than the blue
2479             table?",
2480             "anchor": "blue table",
2481             "target": "wooden chair",
2482             "relationship": "higher",
2483             "task": "position"
2484         }
2485     ]
2486
2487     [
2488         {
2489             "question": "Is the wooden chair positioned higher than the blue
2490             table?",
2491             "anchor": "blue table",
2492             "target": "wooden chair",
2493             "relationship": "higher",
2494             "task": "position"
2495         }
2496     ]
2497
2498     [
2499         {
2500             "question": "Is the wooden chair positioned higher than the blue
2501             table?",
2502             "anchor": "blue table",
2503             "target": "wooden chair",
2504             "relationship": "higher",
2505             "task": "position"
2506         }
2507     ]
2508
2509     [
2510         {
2511             "question": "Is the wooden chair positioned higher than the blue
2512             table?",
2513             "anchor": "blue table",
2514             "target": "wooden chair",
2515             "relationship": "higher",
2516             "task": "position"
2517         }
2518     ]
2519
2520     [
2521         {
2522             "question": "Is the wooden chair positioned higher than the blue
2523             table?",
2524             "anchor": "blue table",
2525             "target": "wooden chair",
2526             "relationship": "higher",
2527             "task": "position"
2528         }
2529     ]
2530
2531     [
2532         {
2533             "question": "Is the wooden chair positioned higher than the blue
2534             table?",
2535             "anchor": "blue table",
2536             "target": "wooden chair",
2537             "relationship": "higher",
2538             "task": "position"
2539         }
2540     ]
2541
2542     [
2543         {
2544             "question": "Is the wooden chair positioned higher than the blue
2545             table?",
2546             "anchor": "blue table",
2547             "target": "wooden chair",
2548             "relationship": "higher",
2549             "task": "position"
2550         }
2551     ]
2552
2553     [
2554         {
2555             "question": "Is the wooden chair positioned higher than the blue
2556             table?",
2557             "anchor": "blue table",
2558             "target": "wooden chair",
2559             "relationship": "higher",
2560             "task": "position"
2561         }
2562     ]
2563
2564     [
2565         {
2566             "question": "Is the wooden chair positioned higher than the blue
2567             table?",
2568             "anchor": "blue table",
2569             "target": "wooden chair",
2570             "relationship": "higher",
2571             "task": "position"
2572         }
2573     ]
2574
2575     [
2576         {
2577             "question": "Is the wooden chair positioned higher than the blue
2578             table?",
2579             "anchor": "blue table",
2580             "target": "wooden chair",
2581             "relationship": "higher",
2582             "task": "position"
2583         }
2584     ]
2585
2586     [
2587         {
2588             "question": "Is the wooden chair positioned higher than the blue
2589             table?",
2590             "anchor": "blue table",
2591             "target": "wooden chair",
2592             "relationship": "higher",
2593             "task": "position"
2594         }
2595     ]
2596
2597     [
2598         {
2599             "question": "Is the wooden chair positioned higher than the blue
2600             table?",
2601             "anchor": "blue table",
2602             "target": "wooden chair",
2603             "relationship": "higher",
2604             "task": "position"
2605         }
2606     ]
2607
2608     [
2609         {
2610             "question": "Is the wooden chair positioned higher than the blue
2611             table?",
2612             "anchor": "blue table",
2613             "target": "wooden chair",
2614             "relationship": "higher",
2615             "task": "position"
2616         }
2617     ]
2618
2619     [
2620         {
2621             "question": "Is the wooden chair positioned higher than the blue
2622             table?",
2623             "anchor": "blue table",
2624             "target": "wooden chair",
2625             "relationship": "higher",
2626             "task": "position"
2627         }
2628     ]
2629
2630     [
2631         {
2632             "question": "Is the wooden chair positioned higher than the blue
2633             table?",
2634             "anchor": "blue table",
2635             "target": "wooden chair",
2636             "relationship": "higher",
2637             "task": "position"
2638         }
2639     ]
2640
2641     [
2642         {
2643             "question": "Is the wooden chair positioned higher than the blue
2644             table?",
2645             "anchor": "blue table",
2646             "target": "wooden chair",
2647             "relationship": "higher",
2648             "task": "position"
2649         }
2650     ]
2651
2652     [
2653         {
2654             "question": "Is the wooden chair positioned higher than the blue
2655             table?",
2656             "anchor": "blue table",
2657             "target": "wooden chair",
2658             "relationship": "higher",
2659             "task": "position"
2660         }
2661     ]
2662
2663     [
2664         {
2665             "question": "Is the wooden chair positioned higher than the blue
2666             table?",
2667             "anchor": "blue table",
2668             "target": "wooden chair",
2669             "relationship": "higher",
2670             "task": "position"
2671         }
2672     ]
2673
2674     [
2675         {
2676             "question": "Is the wooden chair positioned higher than the blue
2677             table?",
2678             "anchor": "blue table",
2679             "target": "wooden chair",
2680             "relationship": "higher",
2681             "task": "position"
2682         }
2683     ]
2684
2685     [
2686         {
2687             "question": "Is the wooden chair positioned higher than the blue
2688             table?",
2689             "anchor": "blue table",
2690             "target": "wooden chair",
2691             "relationship": "higher",
2692             "task": "position"
2693         }
2694     ]
2695
2696     [
2697         {
2698             "question": "Is the wooden chair positioned higher than the blue
2699             table?",
2700             "anchor": "blue table",
2701             "target": "wooden chair",
2702             "relationship": "higher",
2703             "task": "position"
2704         }
2705     ]
2706
2707     [
2708         {
2709             "question": "Is the wooden chair positioned higher than the blue
2710             table?",
2711             "anchor": "blue table",
2712             "target": "wooden chair",
2713             "relationship": "higher",
2714             "task": "position"
2715         }
2716     ]
2717
2718     [
2719         {
2720             "question": "Is the wooden chair positioned higher than the blue
2721             table?",
2722             "anchor": "blue table",
2723             "target": "wooden chair",
2724             "relationship": "higher",
2725             "task": "position"
2726         }
2727     ]
2728
2729     [
2730         {
2731             "question": "Is the wooden chair positioned higher than the blue
2732             table?",
2733             "anchor": "blue table",
2734             "target": "wooden chair",
2735             "relationship": "higher",
2736             "task": "position"
2737         }
2738     ]
2739
2740     [
2741         {
2742             "question": "Is the wooden chair positioned higher than the blue
2743             table?",
2744             "anchor": "blue table",
2745             "target": "wooden chair",
2746             "relationship": "higher",
2747             "task": "position"
2748         }
2749     ]
2750
2751     [
2752         {
2753             "question": "Is the wooden chair positioned higher than the blue
2754             table?",
2755             "anchor": "blue table",
2756             "target": "wooden chair",
2757             "relationship": "higher",
2758             "task": "position"
2759         }
2760     ]
2761
2762     [
2763         {
2764             "question": "Is the wooden chair positioned higher than the blue
2765             table?",
2766             "anchor": "blue table",
2767             "target": "wooden chair",
2768             "relationship": "higher",
2769             "task": "position"
2770         }
2771     ]
2772
2773     [
2774         {
2775             "question": "Is the wooden chair positioned higher than the blue
2776             table?",
2777             "anchor": "blue table",
2778             "target": "wooden chair",
2779             "relationship": "higher",
2780             "task": "position"
2781         }
2782     ]
2783
2784     [
2785         {
2786             "question": "Is the wooden chair positioned higher than the blue
2787             table?",
2788             "anchor": "blue table",
2789             "target": "wooden chair",
2790             "relationship": "higher",
2791             "task": "position"
2792         }
2793     ]
2794
2795     [
2796         {
2797             "question": "Is the wooden chair positioned higher than the blue
2798             table?",
2799             "anchor": "blue table",
2800             "target": "wooden chair",
2801             "relationship": "higher",
2802             "task": "position"
2803         }
2804     ]
2805
2806     [
2807         {
2808             "question": "Is the wooden chair positioned higher than the blue
2809             table?",
2810             "anchor": "blue table",
2811             "target": "wooden chair",
2812             "relationship": "higher",
2813             "task": "position"
2814         }
2815     ]
2816
2817     [
2818         {
2819             "question": "Is the wooden chair positioned higher than the blue
2820             table?",
2821             "anchor": "blue table",
2822             "target": "wooden chair",
2823             "relationship": "higher",
2824             "task": "position"
2825         }
2826     ]
2827
2828     [
2829         {
2830             "question": "Is the wooden chair positioned higher than the blue
```

```

1728     },
1729     {
1730         "question": "Does the red bicycle locate more to the left than
1731         the man in a floral shirt?",
1732         "anchor": "man in a floral shirt",
1733         "target": "red bicycle",
1734         "relationship": "more left",
1735         "task": "position"
1736     }
1737 ],
1738 "modified_questions": [
1739     {
1740         "question": "Is the wooden chair farther from the observer than
1741         the blue table?",
1742         "anchor": "blue table",
1743         "target": "wooden chair",
1744         "relationship": "farther",
1745         "task": "position"
1746     },
1747     {
1748         "question": "Is the red bicycle located at a lower elevation than
1749         the man in a floral shirt?",
1750         "anchor": "man in a floral shirt",
1751         "target": "red bicycle",
1752         "relationship": "lower",
1753         "task": "position"
1754     }
1755 ]
1756 """
1757 messages.append({"role": "user", "content": '\n'.join(query)})

```

Listing 12: Prompts for design questions of task *Size Comparison*

```

1756
1757 messages = [{"role": "system", "content": """
1758     I will give you an image and a description about the image. You should
1759     design 2 questions regarding size judgments around top-level objects
1760     in the image (top-level means it's not a part of another object). The
1761     question should involve two object (anchor, target) and a type of
1762     relationship:
1763
1764     - *The anchor and target object* should be randomly chosen from the top
1765     -level objects. You possibly need to add the attributions about
1766     appearance, behavior, posture, position in the image, etc. to the
1767     description of the anchor and target objects so that they can be
1768     distinguished from others.
1769     - *The relationship* should be randomly selected from: larger, smaller,
1770     taller, shorter, wider, narrower.
1771     - Only design questions about top-level objects. Ignore those not in
1772     top-level objects list. Ignore environment objects such as water, sky
1773     , grass, cloud, etc.
1774
1775     After that, generate 2 more questions based on the designed questions
1776     by choose another relationship and keep other parts unchanged.
1777
1778     Please respond in JSON format. All content should be in English. Here
1779     is an example of output:
1780
1781     {
1782         "questions": [
1783             {
1784                 "question": "Is the green vase taller than the brown table?",
1785                 "anchor": "brown table",
1786                 "target": "green vase",
1787                 "relationship": "taller",
1788                 "task": "size"
1789             }
1790         ]
1791     }
1792 """
1793 messages.append({"role": "user", "content": '\n'.join(query)})

```

```

1782     },
1783     {
1784         "question": "Is the plate with food on it narrower than the white
1785         box in the middle?",
1786         "anchor": "white box in the middle",
1787         "target": "plate with food on it",
1788         "relationship": "narrower",
1789         "task": "size"
1790     }
1791 ],
1792 "modified_questions": [
1793     {
1794         "question": "Is the green vase smaller than the brown table?",
1795         "anchor": "brown table",
1796         "target": "green vase",
1797         "relationship": "smaller",
1798         "task": "size"
1799     },
1800     {
1801         "question": "Is the plate with food on it larger than the white
1802         box in the middle?",
1803         "anchor": "white box in the middle",
1804         "target": "plate with food on it",
1805         "relationship": "larger",
1806         "task": "size"
1807     }
1808 ]
1809 """
1810 ]
1811
1812 messages.append({"role": "user", "content": '\n'.join(query)})

```

Listing 13: Prompts for design questions of task *Existence Estimation* and *Object Counting*

```

1810 messages = [{"role": "system", "content": f"""
1811 I will give you an image and a description about the image. You should
1812 design 2 questions regarding existence judgments and 2 questions
1813 regarding counting around top-level objects in the image (top-level
1814 means it's not a part of another object). The conditions in the
1815 question need to involve an anchor object and a type of relationship:
1816
1817 - *The anchor object* should be randomly chosen from the top-level
1818 objects. If multiple objects in the image are similar to anchor
1819 object, you need to add the attributions about appearance, behavior,
1820 posture, position in the image, etc. to the description of anchor
1821 object so that it can be distinguished from others.
1822 - *The relationship* should randomly selected from: more to the left,
1823 more to the right, closer (to the observer), farther (to the observer
1824 ), higher, lower, larger, smaller, taller, shorter, wider, narrower.
1825 - Only design questions about top-level objects. Ignore those not in
1826 top-level objects list. Ignore environment objects such as water, sky
1827 , grass, cloud, etc.
1828
1829 After that, generate 4 more questions based on the designed questions
1830 by choose another type of relationship and keep other parts unchanged
1831 .
1832
1833 Please respond in JSON format. All content should be in English. Here
1834 is an example of output:
1835 {
1836     "questions": [
1837         {
1838             "question": "Are there chairs wider than the blue table?",
1839             "anchor": "blue table",
1840             "target": "chair",
1841             "relationship": "wider",
1842         }
1843     ]
1844 }

```

```

1836         "task": "existence"
1837     },
1838     {
1839         "question": "Is there a bicycle located more to the left than the
1840         man in a floral shirt?",  

1841         "anchor": "man in a floral shirt",
1842         "target": "bicycle",
1843         "relationship": "more left",
1844         "task": "existence"
1845     },
1846     {
1847         "question": "How many green vases are positioned higher than the
1848         middle wooden table?",  

1849         "anchor": "middle wooden table",
1850         "target": "green vase",
1851         "relationship": "higher",
1852         "task": "count"
1853     },
1854     {
1855         "question": "How many plates are larger than the white box in the
1856         middle?",  

1857         "anchor": "white box in the middle",
1858         "target": "plate",
1859         "relationship": "larger",
1860         "task": "count"
1861     }
1862 ],
1863 "modified_questions": [
1864     {
1865         "question": "Are there chairs which have lower elevation than the
1866         blue table?",  

1867         "anchor": "blue table",
1868         "target": "chair",
1869         "relationship": "lower",
1870         "task": "existence"
1871     },
1872     {
1873         "question": "Is there a bicycle closer to the observer than the
1874         man in a floral shirt?",  

1875         "anchor": "man in a floral shirt",
1876         "target": "bicycle",
1877         "relationship": "closer",
1878         "task": "existence"
1879     },
1880     {
1881         "question": "How many green vases are wider than the middle
1882         wooden table?",  

1883         "anchor": "middle wooden table",
1884         "target": "green vase",
1885         "relationship": "wider",
1886         "task": "count"
1887     },
1888     {
1889         "question": "How many plates are shorter than the white box in
1890         the middle?",  

1891         "anchor": "white box in the middle",
1892         "target": "plate",
1893         "relationship": "shorter",
1894         "task": "count"
1895     }
1896 ],
1897 }
1898 """
1899 ]
1900 messages.append({"role": "user", "content": '\n'.join(query)})
```

## 1890 E TRAINING DETAILS

1891  
 1892 We train InternVL-Spatial-8B using LoRA(Hu et al., 2022) with approximately 291K general train-  
 1893 ing samples from InternVL2.5 (Chen et al., 2024c) and 2M samples from InternSpatial, counted  
 1894 with repetition. The training is conducted on 16 A100 GPUs for approximately 14 hours. We report  
 1895 the models and training hyperparameters of InternVL-Spatial-8B in Table 8.  
 1896

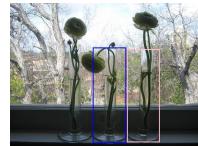
1897  
 1898 **Table 8: Training settings and hyperparameters for InternVL-Spatial-8B models.** Key config-  
 1899 urations for InternVL-Spatial-8B, including model architectures and training parameters.

1900 <b>InternVL-Spatial-8B</b>		
1901 Model	1902 ViT	1903 InternViT-300M
1904	1905 LLM	1906 Internlm2.5-7b-chat
1907 Training Hyperparameters	1908 Tile Resolution	1909 448
	1910 Lora Rank	1911 16
	1912 Packed Batch Size	1913 64
	1914 Optimizer	1915 AdamW
	1916 Learning Rate	1917 2.00E-05
	1918 Warmup Ratio	1919 0.03
	1920 LR Scheduler	1921 Cosine
	1922 Weight Decay	1923 0.05
	1924 ViT Drop Path	1925 0.1
	1926 Image Tile Threshold	1927 40
	1928 Context Length	1929 12.8K
	1930 Epochs	1931 1

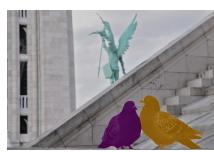
## 1932 F VISUALIZATION OF INTERNSPATIAL



1933 **Image Source:** COCO  
 1934 **Task type:** Positional Comparison  
 1935 **Instruction Format:** Origin Image + With <box>  
 1936 **Question:**  
 1937 **Answer:**  
 1938 Can it be said that <ref>region</ref><ref><box>[251.0, 467.0, 557.0, 744.0]<box></box></ref></ref> is elevated over <ref>region</ref><ref><box>[508.0, 411.0, 698.0, 741.0]<box></box></ref></ref>?  
 1939 Yes, <ref>region</ref><ref><box>[251.0, 467.0, 557.0, 744.0]<box></box></ref></ref> is situated above <ref>region</ref><ref><box>[508.0, 411.0, 698.0, 741.0]<box></box></ref></ref>.



1940 **Image Source:** COCO  
 1941 **Task type:** Positional Comparison  
 1942 **Instruction Format:** Image with boxes + Natural Language  
 1943 **Question:**  
 1944 <image>  
 1945 Is the home decor in the blue box at a greater height compared to the home decor in the pink box?  
 1946 **Answer:**  
 1947 The proximity in the vertical aspect makes it difficult to ascertain.



1948 **Image Source:** COCO  
 1949 **Task type:** Size Comparison  
 1950 **Instruction Format:** Image with masks + According to Image  
 1951 **Question:**  
 1952 <image>  
 1953 Is the animal with the orange mask less broad than the animal with the purple mask?  
 1954 **Answer:**  
 1955 Not according to the information, the animal with the orange mask is not less broad than the animal with the purple mask.



1956 **Image Source:** Visual Genome  
 1957 **Task type:** Position Comparison  
 1958 **Instruction Format:** Origin Image + Natural Language  
 1959 **Question:**  
 1960 <image>  
 1961 Comparing balcony on side of building and the front end of the blue car, which one is more to the right?  
 1962 **Answer:**  
 1963 Among the given options, the front end of the blue car is the one that is closer to the right.

1944

1945

1946

1947

1948

1949

1950

1951

1952

1953

1954

1955

1956

1957

1958



Image Source: Visual Genome

Task type: Position Comparison

Instruction Format: Image with `bbboxes` + With `<ref><box>`

Question:

&lt;image&gt; Comparing &lt;ref&gt;region&lt;/ref&gt;&lt;box&gt;[367.0, 520.0, 441.0, 735.0]&lt;box&gt; and &lt;ref&gt;region&lt;/ref&gt;&lt;box&gt;[436.0, 490.0, 545.0, 734.0]&lt;box&gt;, which is more behind?

Answer:

&lt;ref&gt;region&lt;/ref&gt;&lt;box&gt;[367.0, 520.0, 441.0, 735.0]&lt;box&gt; is clearly more behind than the other.



Image Source: Visual Genome

Task type: Position Comparison

Instruction Format: Image with numbers + With `<ref><box>`

Question:

&lt;image&gt; Between &lt;ref&gt;blue and clear sky&lt;/ref&gt;&lt;box&gt;[820.0, 3.0, 125.0, 205.0]&lt;box&gt; and &lt;ref&gt;decorative back roof on the building&lt;/ref&gt;&lt;box&gt;[125.0, 5.0, 977.0, 145.0]&lt;box&gt;, which object is closer to the observer?

Answer:

&lt;ref&gt;decorative black roof on the building&lt;/ref&gt;&lt;box&gt;[125.0, 5.0, 977.0, 145.0]&lt;box&gt; is in front.

1952

1953

1954

1955

1956

1957



Image Source: AS-1B

Task type: Size Comparison

Instruction Format: Image with `bbboxes` + With `<ref><box>`

Question:

&lt;image&gt; Which has a greater height, the vase in the yellow box or the tapestry in the orange box?

Answer:

The height of the tapestry in the orange box surpasses the other.



Image Source: AS-1B

Task type: Size Comparison

Instruction Format: Image with numbers + According to Image

Question:

&lt;image&gt; Which has a greater width, the island with a circled '1' drawn on it or the surfer with a circled '2' drawn on it?

Answer:

The width of the island with a circled '1' drawn on it is greater.

1963

1964

1965

1966

1967



Image Source: AS-1B

Task type: Size Comparison

Instruction Format: Image with `maskss` + With `<box>`

Question:

&lt;image&gt; Is &lt;ref&gt;region&lt;/ref&gt;&lt;box&gt;[295.0, 89.0, 475.0, 374.0]&lt;box&gt; larger than &lt;ref&gt;region&lt;/ref&gt;&lt;box&gt;[510.0, 282.0, 983.0, 495.0]&lt;box&gt;?

Answer:

&lt;image&gt; &lt;ref&gt;region&lt;/ref&gt;&lt;box&gt;[510.0, 282.0, 983.0, 495.0]&lt;box&gt; is larger than &lt;ref&gt;region&lt;/ref&gt;&lt;box&gt;[295.0, 89.0, 475.0, 374.0]&lt;box&gt;.



Image Source: COCO

Task type: Existence Estimation

Instruction Format: Origin Image + With `<box>`

Question:

&lt;image&gt; Are there any people closer to the observer than &lt;ref&gt;region&lt;/ref&gt;&lt;box&gt;[282.0, 331.0, 615.0, 558.0]&lt;box&gt;?

Answer:

No.

1974

1975

1976

1977

1978



Image Source: COCO

Task type: Object Counting

Instruction Format: Origin Image + Natural Language

Question:

&lt;image&gt; How many cars are positioned lower than the small house with a red roof?

Answer:

2.



Image Source: Cityscapes

Task type: Size Comparison

Instruction Format: Origin Image + With `<box>`

Question:

&lt;image&gt; Is the car with a circled '1' drawn on it smaller than the car with a circled '2' drawn on it, isn't it?

Answer:

Definitely, the car with a circled '1' drawn on it is smaller than the car with a circled '2' drawn on it.

1986

1987

1988

1989

1990



Image Source: Cityscapes

Task type: Size Comparison

Instruction Format: Image with `bbboxes` + According to Image

Question:

&lt;image&gt; Which is larger, the bicycle in the yellow box or the car in the pink box?

Answer:

the car in the pink box is the larger one.



Image Source: Cityscapes

Task type: Positional Comparison

Instruction Format: Image with `maskss` + According to Image

Question:

&lt;image&gt; Can you determine if the car with the red mask is closer or farther from the observer compared to the car with the green mask?

Answer:

the car with the red mask is in front of the car with the green mask.

1994

1995

1996

1997

## 1998 G VISUALIZATION OF RESULTS ON INTERNSPATIAL-BENCH

1999

2000

2001

2002

2003

2004

2005

2006

2007

2008

2009

2010

2011

2012

2013

2014

2015

2016

2017

2018

2019

2020

2021

2022

2023

2024

2025

2026

2027

2028

2029

2030

2031

2032

2033

2034

2035

2036

2037

2038

2039

2040

2041

2042

2043

2044

2045

2046

2047

2048

2049

2050

2051



Question: <image> Where is <ref>the woman in the red coat</ref> positioned in relation to <ref>the sign</ref>? Answer with the option's letter from the given choices directly. (A) To the right side of <ref>the sign</ref> (B) In front of <ref>the sign</ref> (C) Behind <ref>the sign</ref> (D) To the left side of <ref>the sign</ref>

GT: A

InternVL-8B Pred: (C) Behind the sign

InternVL-Spatial-8B Pred: (A) To the right side of the sign

Task: positional



Question: <image> Are there any vehicles more to the right than <ref>the silver car in the left lane</ref>? Answer with a single word or option's letter.

GT: Yes

InternVL-8B Pred: No

InternVL-Spatial-8B Pred: Yes

Task: existence



Question: <image> Is <ref>region</ref><box>[599, 390, 654, 472]</box> positioned farther from the observer than <ref>region</ref><box>[918, 378, 1000, 538]</box>? Answer with a single word or option's letter.

GT: Yes

InternVL-8B Pred: No

InternVL-Spatial-8B Pred: Yes

Task: positional

2052

2053

2054

2055

2056

2057

2058

2059

2060

2061

2062

2063

2064

2065

2066

2067

2068

2069

2070

2071

2072

2073

2074

2075

2076

2077

2078

2079

2080

2081

2082

2083

2084

2085

2086

2087

2088

2089

2090

2091

2092

2093

2094

2095

2096

2097

2098

2099

2100

2101

2102

2103

2104

2105



Question: <image> Is <ref>the blue parking sign on the right side of the road</ref> shorter than <ref>the white building on the left side</ref>? Answer with a single word or option's letter.

GT: Yes

InternVL-8B Pred: No

InternVL-Spatial-8B Pred: Yes

Task: size



Question: <image> Which is above, <ref>region</ref><box>[210, 179, 453, 566]</box> or <ref>region</ref><box>[209, 0, 506, 516]</box>?

GT: Positioned higher is <ref>region</ref><box>[209, 0, 506, 516]</box>.

InternVL-8B Pred: region[210, 179, 453, 566] is above region[209, 0, 506, 516].

InternVL-Spatial-8B Pred: region[209, 0, 506, 516] is the one that is above.

Task: above\_choice



Question: <image> Is the picture with the purple mask below the pillow with the pink mask?

GT: Actually, the picture with the purple mask is above the pillow with the pink mask.

InternVL-8B Pred: Yes, the picture with the purple mask is below the pillow with the pink mask.

InternVL-Spatial-8B Pred: No, the picture with the purple mask is not situated beneath the pillow with the pink mask.

Task: below\_predicate



Question: <image> Does the counter with a circled '1' drawn on it come behind the picture with a circled '2' drawn on it?

GT: No, it is not.

InternVL-8B Pred: Yes, the counter with a circled '1' is behind the picture with a circled '2'.

InternVL-Spatial-8B Pred: No, the counter with a circled '1' drawn on it is not behind the picture with a circled '2' drawn on it.

Task: behind\_predicate

2106



Question: <image> Does the books with the yellow mask come in front of the chair with the blue mask?

GT: No, it is not.

InternVL-8B Pred: Yes, the books with the yellow mask are placed in front of the chair with the blue mask.

InternVL-Spatial-8B Pred: No, the books with the yellow mask is not in front of the chair with the blue mask.

Task: front\_predicate

2107

2108

2109

2110

2111

2112

2113

2114

2115

2116

2117

2118

2119

2120

2121

2122

2123

2124

2125

2126

2127

2128

2129

2130

2131

2132

2133

2134



Question: <image> Which is more to the left, the coffee maker in the red box or the box in the yellow box?

GT: From the viewer's perspective, the box in the yellow box appears more on the left side.

InternVL-8B Pred: The coffee maker in the red box is more to the left than the box in the yellow box.

InternVL-Spatial-8B Pred: the box in the yellow box is positioned further to the left.

Task: left\_choice

2135

2136

2137

2138

2139

2140

2141

2142

2143

2144

2145

2146

2147



Question: <image> Can you confirm if the pillow with a circled '1' drawn on it is positioned to the right of the books with a circled '2' drawn on it?

GT: No, the pillow with a circled '1' drawn on it is not to the right of the books with a circled '2' drawn on it.

InternVL-8B Pred: Yes, the pillow with a circled '1' is positioned to the right of the books with a circled '2'.

InternVL-Spatial-8B Pred: the pillow with a circled '1' drawn on it is no to the right of the books with a circled '2' drawn on it, actually.

Task: right\_predicate

2148

2149

2150

2151

2152

2153

2154

2155

2156

2157

2158

2159



Question: <image> Does the toaster with a circled '1' drawn on it have a larger size compared to the counter with a circled '2' drawn on it?

GT: Incorrect, the toaster with a circled '1' drawn on it is not larger than the counter with a circled '2' drawn on it.

InternVL-8B Pred: Yes, the toaster with a circled '1' drawn on it is larger in size compared to the counter with a circled '2' drawn on it.

InternVL-Spatial-8B Pred: Not according to the information, the toaster with a circled '1' drawn on it is not larger than the counter with a circled '2' drawn on it.

Task: big\_predicate

2160

2161

2162

2163

2164

2165

2166

2167

2168

2169

2170

2171

2172

2173

2174

2175

2176

2177



Question: <image> Does the barrier in the purple box have a greater height compared to the truck in the pink box?

GT: In fact, the barrier in the purple box may be shorter than the truck in the pink box.

InternVL-8B Pred: Yes, the barrier in the purple box appears to have a greater height compared to the truck in the pink box.

InternVL-Spatial-8B Pred: Sorry, but the barrier in the purple box does not exceed the truck in the pink box in height.

Task: tall\_predicate



Question: <image> Can you confirm if the painting in the purple box is thinner than the potted plant in the blue box?

GT: In fact, the painting in the purple box might be wider than the potted plant in the blue box.

InternVL-8B Pred: Yes, the painting in the purple box is thinner than the potted plant in the blue box.

InternVL-Spatial-8B Pred: No, the painting in the purple box is not narrower than the potted plant in the blue box.

Task: thin\_predicate

2206

2207

2208

2209

2210

2211

2212

2213

2214 **H THE USE OF LARGE LANGUAGE MODELS (LLMs)**  
22152216 We used large language models (LLMs) as assistive tools during the preparation of this work. Spec-  
2217 ifically, LLMs were employed for language polishing, LaTeX code editing, and debugging of prompts  
2218 in the dataset construction process. The authors take full responsibility for the content of the paper.  
22192220  
2221  
2222  
2223  
2224  
2225  
2226  
2227  
2228  
2229  
2230  
2231  
2232  
2233  
2234  
2235  
2236  
2237  
2238  
2239  
2240  
2241  
2242  
2243  
2244  
2245  
2246  
2247  
2248  
2249  
2250  
2251  
2252  
2253  
2254  
2255  
2256  
2257  
2258  
2259  
2260  
2261  
2262  
2263  
2264  
2265  
2266  
2267