From Macro to Micro: Benchmarking Microscopic Spatial Intelligence on Molecules via Vision-Language Models

Zongzhao Li[1,2,3*], Xiangzhe Kong[4,5*], Jiahui Su[6], Zongyang Ma[7], Mingze Li[1,2,3], Songyou Li[1,2,3], Yuelin Zhang[1,2,3], Yu Rong[8,9], Tingyang Xu[8,9], Deli Zhao[8,9], and Wenbing Huang[1,2,3,10†]

[1] Gaoling School of Artificial Intelligence, Renmin University of China
lizongzhao2023@ruc.edu.cn
[2] Beijing Key Laboratory of Research on Large Models and Intelligent Governance
[3] Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE
[4] Dept. of Comp. Sci. & Tech., Tsinghua University
[5] Institute for AI Industry Research (AIR), Tsinghua University
[6] SKL-ESPC & SEPKL-AERM, College of Environmental Sciences and Engineering, Peking University
[7] MAIS, Institute of Automation, Chinese Academy of Sciences
[8] DAMO Academy, Alibaba Group, Hangzhou, China
[9] Hupan Lab, Hangzhou, China
[10] Beijing Academy of Artificial Intelligence {hwenbing@126.com}

Abstract. This paper introduces the concept of Microscopic Spatial Intelligence (MiSI), the capability to perceive and reason about the spatial relationships of invisible microscopic entities, which is fundamental to scientific discovery. To assess the potential of Vision-Language Models (VLMs) in this domain, we propose a systematic benchmark framework MiSI-Bench. This framework features over 163,000 question-answer pairs and 587,975 images derived from approximately 4,000 molecular structures, covering nine complementary tasks that evaluate abilities ranging from elementary spatial transformations to complex relational identifications. Experimental results reveal that current state-of-the-art VLMs perform significantly below human level on this benchmark. However, a fine-tuned 7B model demonstrates substantial potential, even surpassing humans in spatial transformation tasks, while its poor performance in scientifically-grounded tasks like hydrogen bond recognition underscores the necessity of integrating explicit domain knowledge for progress toward scientific AGI. The datasets are available at https://huggingface.co/datasets/zongzhao/MiSI-bench.

Keywords: Microscopic Spatial Intelligence · Vision-Language Models · Spatial Reasoning

* Equal contribution. † Corresponding author.

1 Introduction

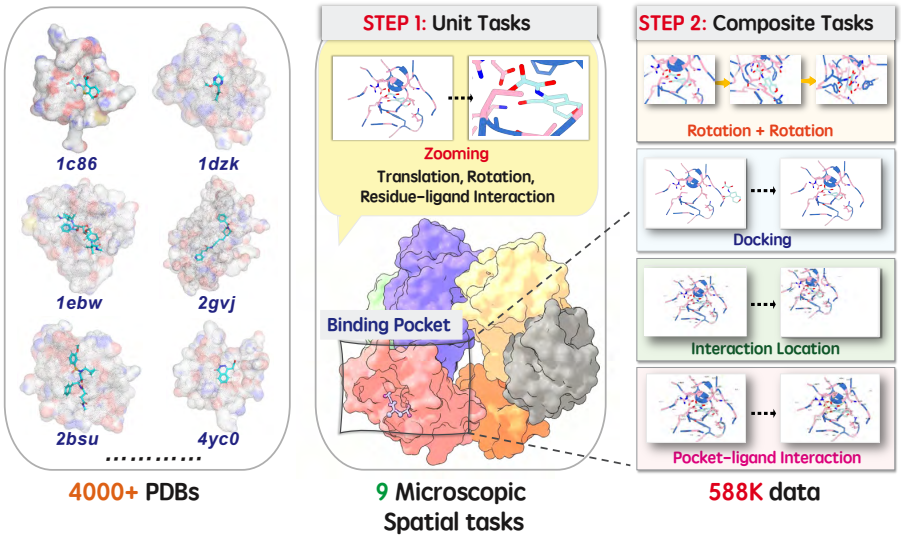

Fig. 1: Overview of our `MiSI-Bench`. Our dataset is derived from around 4,000 PDBs and comprises 9 distinct tasks.

Spatial intelligence [12, 53], a critical component of advanced artificial intelligence, empowers systems to perceive, interpret, and interact with the three-dimensional world. Such capability necessitates a profound comprehension of geometries, spatial relationships, and even fundamental physical rules. Contemporary efforts address this by utilizing Vision-Language Models (VLMs) [6, 28, 29], which jointly process visual and textual data to learn spatial properties and relationships from complex scenes [32, 54, 56]. This reasoning capacity regarding object layouts, occlusions, and perspectives establishes a vital foundation for embodied interaction in real-world environments [57].

Yet, beyond this visible macroscopic realm lies the microscopic world, composed of invisible particles (e.g., atoms and molecules) that constitute matter [17], where spatial reasoning takes on a profoundly different form. In molecular sciences, experts routinely visualize and manipulate microscopic entities such as proteins and drugs using software tools (*e.g.*, PyMOL [16], ChimeraX [41]) to explore geometric complementarity, analyze interactions, and design new functional molecules. This process relies on a specialized form of spatial reasoning: the ability to reconstruct three-dimensional structures from two-dimensional projections and to infer physical relationships (*e.g.*, hydrogen bonds). In this paper, we refer to this capability as **Microscopic Spatial Intelligence (MiSI)**, the cognitive foundation underlying human expertise and discovery in scientific fields such as structural biology, drug discovery, and material design.

The success of Large Language Models (LLMs) in general-purpose tasks has spurred their exploration in scientific discovery [9, 37, 46], motivating the use of VLMs to analyze scientific data. VLMs are uniquely suited for this role, as they can process both visual and textual modalities within a unified architecture. Traditional domain-specific software (e.g., VinaDock [50]) methods excel at numerical optimization but lack the semantic domain knowledge and human-like inference required to interpret scientific insights encoded in literature. On the contrary, VLMs can perceive structural patterns while grounding their interpretations in scientific concepts. This cross-modal capability enables a more human-like, context-aware reasoning about molecular structures by seamlessly linking them with textual semantics. However, shifting from human-scale daily objects to atom-level invisible entities, MiSI requires exceptional scientific expertise to perceive *spatial transformations* and reason over *relational identifications* such as atomic interactions. It remains unclear whether the VLMs are ready for tackling the challenges in the microscopic scientific fields.

To bridge the gap, we propose `MiSI-Bench`, a systematic framework for training and evaluating microscopic spatial intelligence in VLMs. As illustrated in Fig. 1, `MiSI-Bench` contains 163,514 question-answer pairs and 587,975 images over a diverse set of 9 complementary tasks, constructed from around 4,000 molecular structures [52]. We visualize these three-dimensional microscopic objects as two-dimensional orthographic projections, just like how human experts interpret them. We then disentangle the intelligence for *spatial transformations* and *relational identifications* into four elementary operations: translation, rotation, zooming, interaction. Subsequently, we design four unit tasks to evaluate these fundamental abilities independently, and further design five composite tasks which integrate multiple elementary operations to test the models' high-order reasoning ability.

Experimental results show that current advanced VLMs (e.g., o3 [40], Claude Sonnet4.5 [5]) perform well below human level on our benchmark. While human evaluators excel in some tasks, they struggle with continuous spatial transformation and 3D reconstruction. Remarkably, after SFT on our dataset, a 7B model outperforms all leading VLMs and even surpasses humans in spatial transformation tasks, revealing VLMs' untapped potential for spatial reasoning. However, its poor performance in biologically-grounded tasks like hydrogen bond recognition highlights the need for injecting explicit scientific knowledge during pre-training to progress toward AGI.

2 Related Work

Macroscopic Spatial Intelligence In recent years, researchers have developed a variety of datasets and benchmarks with distinct focuses to evaluate the spatial intelligence of VLMs [21,44,59]. For instance, VIS-Bench [53] and MuriBench [51] emphasize the model's ability to associate and reason across video/multi-images; LEGO-Puzzles [47] examines multi-step spatial reasoning in a synthetic block-building environment. However, MiSI is distinct from macro-spatial tasks in three

aspects: 1) Unlike 3D reconstruction of common objects, MiSI requires mapping fine-grained visual elements to hierarchical scientific entities (e.g., atoms and residues); 2) Tasks like "binding" require translating abstract scientific concepts into precise geometric constraints, a mapping largely absent in general vision; 3) Unlike macro-tasks that tolerate spatial imprecision, MiSI necessitates a much tighter, expert-level coupling between visual perception and domain-specific language. Thus, MiSI represents a unique dimension where spatial reasoning is inextricably tied to scientific logic. Therefore, we propose `MiSI-Bench` to draw attention to this direction and to establish a reliable benchmark for evaluating models' micro-level spatial intelligence.

Three-Dimensional Molecular Understanding Conventional modeling of three-dimensional molecules usually relies on Cartesian coordinates as model inputs. Starting from physical force fields [1, 45], models have evolved from 3D convolutional neural networks [42] to graph-based relational modeling, where graph neural networks [8, 33, 34, 58] have been widely used to capture structured dependencies [19, 36]. This line has further advanced toward equivariant graph neural networks [24, 30, 31, 48] and transformers [23, 27] with the rise of geometric deep learning [10]. However, these approaches primarily operate within geometric coordinate space only, while MLLMs offer a complementary, human-like perspective which can learn to reason about three-dimensional molecular geometry through visual abstractions and natural language grounding, unlocking a new mode of molecular understanding that bridges microscopic visual perception and textual knowledge.

3 Definition of Major Concepts

Study Scope To explore whether current VLMs possess the ability to understand 3D molecular structures, which we term **MiSI** as above, we begin by defining the representations adopted for molecular structures, and then introduce the fundamental perceptual expertise humans rely on to comprehend the micro world. These elements together serve as the preliminaries for our benchmark design.

Our physical world is organized across multiple hierarchical levels, from macroscopic organisms to organs, cells, and to molecules such as proteins, and DNAs [17]. Thanks to the continuous scientific exploration, people now understand the phenomena observed in the macroscopic world are the results of microscopic particles. And advances in imaging technologies, such as cryo-electron microscopy, further allow us to visualize these particles at near-atomic resolution [7]. In this work, we shift the focus from the macroscopic world to its microscopic foundation, investigating how VLMs perceive and reason about 3D molecular structures composed of atoms.

Orthographic Projection of Molecules Throughout human history, people have sought to represent the three-dimensional world on two-dimensional media. One classical approach employs perpendicular rays of light to generate *orthographic*

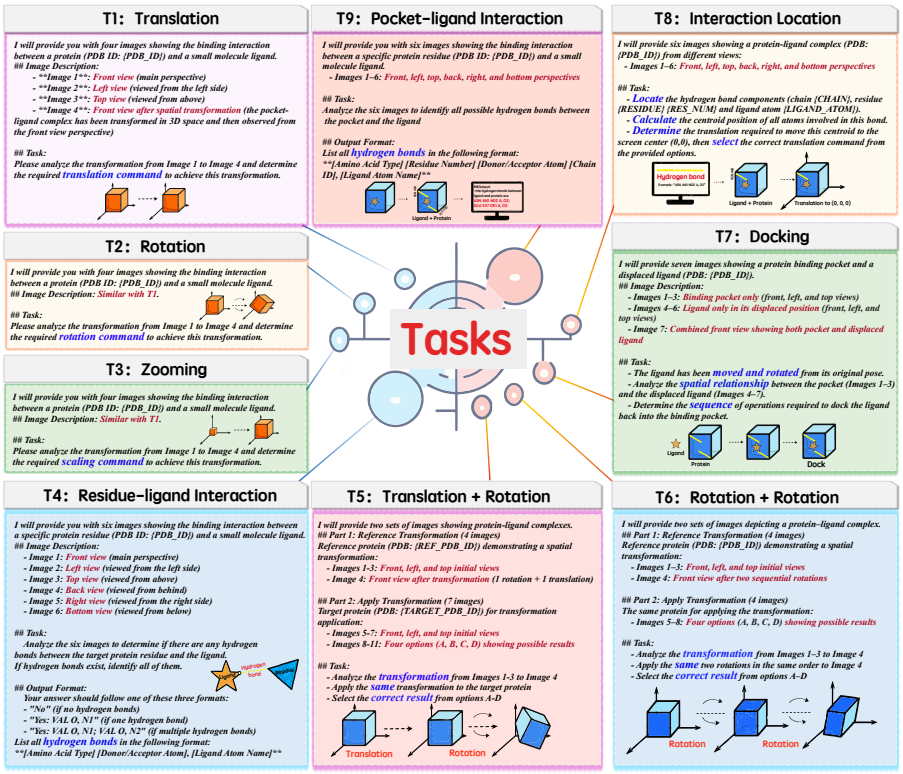

Fig. 2: A brief demonstration of the `MiSI-Bench` dataset. The examples have been simplified for clarity; for complete examples, please refer to Appendix C.

projections, and typically canonical views, such as the front, top, and left views, are employed to reconstruct the full 3D structure of an object [11]. Following this convention, we adopt orthographic views as 2D representations of microscopic 3D molecular structures.

Taxonomy of Human Expertise Understanding microscopi molecules requires both spatial reasoning and domain expertise. Experts rely on fundamental *spatial transformation* abilities, such as translation, rotation, and zooming, to establish a more complete panorama of molecular structures [13, 26]. Beyond geometric manipulation, they use domain knowledge to identify interaction patterns such as hydrogen bonds, which reveal

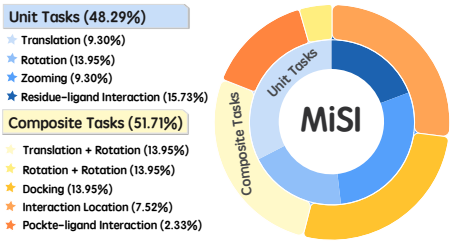

Fig. 3: The statistics for all tasks.

the underlying physical principles of molecular organization [2]. We refer to this process as *relational identification*. In this work, we summarize the expert skills with the above-mentioned four elementary microspace operations, namely `translation`, `rotation`, `zooming`, and `interaction`, then design unit tasks to independently evaluate each capability. We further introduce combinatorial tasks that require integrating multiple operations, enabling a more comprehensive assessment of VLMs in microspace understanding.

4 MiSI-Bench

We construct our `MiSI-Bench` for evaluating VLMs using the refined PDBbind dataset [52], a widely adopted benchmark for structure-based drug discovery [22, 25, 49]. Each entity in PDBbind dataset corresponds to a complex composed of a protein and a ligand. We visualize all complexes in ChimeraX [41] to generate orthographic projections images at a resolution of 1160×803 as model inputs. After removing complexes with visualization issues, the final dataset contains 3,503 protein–ligand complexes for Supervised FineTuning (SFT) and 490 for testing, all with experimentally solved crystal structures. The benchmark encompasses nine tasks, including four unit tasks involving single elementary operations and five composite tasks involving combinations of multiple operations. For each task, the QA pairs are generated using fixed templates. The problem templates and a brief illustration of all tasks are shown in Fig. 2. The overall pipeline for constructing `MiSI-Bench` is detailed in supplementary materials. Our benchmark contains a total of 150,597 Question Answering (QA) pairs for training and 12,917 for testing, summing up to 538,015 and 49,960 images in the train and the test set, respectively. The statistics for all tasks are presented in Fig. 3. Two formats of questions are used to evaluate model performance.

Cloze Questions require the model to complete partially specified instructions by filling in missing actions or parameters. These tasks assess the ability of the models to identify the correct operations with precise attributes (e.g., axes, distances, or angles).

Multiple-Choice Questions present several candidate options, among which the model should identify the correct answer while rejecting distracting decoys. These tasks evaluate whether the models have discriminative understanding of spatial configurations and their ability to reason about the consequences of microspace operations.

4.1 Unit Task

We first establish four unit tasks to evaluate the spatial understanding of elementary microspace operations, where each task isolates one essential ability involved in manipulating or interpreting microscopic 3D molecular structures. For translation, rotation, and zooming tasks, the three orthographic projections (*i.e.*, the top, the front, and the left side views) of the initial complex and the

front view of the complex after the operation are given to the models. For residue-ligand interaction task, since overlapping atom names might interfere with the performance, we give all six orthographic projections to the models.

Translation (*Cloze*) In this task, the molecular complex is translated along one of the axes parallel to the visualization plane (*i.e.*, the x or y axis) by a random distance between -4 and 4 angstrom (Å). The model must infer both the direction and the magnitude of motion, completing a prompt of the form: `move _x_ _3_`. To avoid being too harsh on numerical precisions, the translation range is discretized into 1.0 Å bins. For SFT, two samples per complex are generated along each axis, yielding 14,012 training samples, and one per axis for evaluation, totaling 980 test samples.

Rotation (*Cloze*) The complex is rotated along one of the three coordinate axes (x, y, or z) by a random angle uniformly drawn from $[-90°, 90°]$. Models must determine both the rotation axis and the degree of rotation, filling in the prompt: `roll _x_ _15_`. Rotation angles are discretized into 15° bins. Each complex results in two samples per axis for SFT (21,018 training samples) and one per axis for testing (1,470 samples).

Zooming (*Cloze*) To simulate zooming operations, the complex is moved along the axis perpendicular to the visualization plane (*i.e.*, the z axis) by a random depth between 40 and 60 Å. This range corresponds to the magnification levels most suitable for visualizing the pocket–ligand interactions near the center of the view (See the distribution figure in Appendix A for details). The model fills in prompts like: `move z _50_`, where depth values are discretized into 1.0 Å bins. Four samples per complex are created for SFT (14,012 training samples) and two per complex for testing (980 samples).

Residue-Ligand Interaction (*Cloze*) Given a residue and the ligand, models must first identify whether the residue interacts with the ligand (`Yes` or `No`), and output all atom pairs participating in the interaction: `ARG NH2, O22; ARG N, O23`. For this proof-of-concept benchmark, we focus on hydrogen bonds as interactions of interest, with detailed geometric configurations provided in the Appendix A. The dataset includes 11,572 positive and 12,125 negative samples for SFT, and 1,499 positive and 1,603 negative samples for evaluation.

4.2 Composite Tasks

We further design five composite tasks that require models to understand and reason on multiple microspace operations, the capability of which are commonly required for human experts during molecular structural analysis.

Spatial Transformation Reasoning This category evaluates the ability of the models to reason about sequential spatial transformations and generalize

them across different molecular complexes. Denote two complexes as c_1 and c_2, and two spatial transformations as f_1 and f_2. The models are given the three orthographic projections of both c_1 and c_2, along with the front view of $f_2(f_1(c_1))$, which is the result of applying f_1 followed by f_2 to c_1. The task is to identify the correct front view of $f_2(f_1(c_2))$ among four candidate images, where the other three are decoys. The decoys are constructed by exerting perturbation on the ground-truth transformations through one of three schemes: 1) Altering the magnitude (translation distance or rotation angle) of both f_1 and f_2; 2) Flipping the sign of f_1 (e.g., clockwise to counterclockwise) and adjusting the magnitude of f_2; 3) Changing the axis of f_1 and modifying the magnitude of f_2.

Translation-Rotation Movement (*Multiple-Choice*) In this task, f_1 is sampled from translational operations and f_2 from rotational operations. For each complex in the dataset, we pair it with another random complex and construct six and three distinct transformation combinations for SFT and testing, respectively, yielding a total of 21,018 and 1,470 questions for SFT and testing.

Rotation-Rotation Movement (Multiple-Choice) Both f_1 and f_2 are sampled from rotational operations, constrained to different axes to prevent trivial correlations. The scale matches the previous task, with 21,018 questions for SFT and 1,470 for testing.

Local Relational Reasoning This category evaluates whether the models are capable of interpreting fine-grained, domain-specific spatial relations within molecular complexes, such as hydrogen bonds between atom pairs, and then reasoning about how to manipulate the visualization to highlight specific interactions.

Interaction Location (*Multiple-Choice*) The model is provided with six orthographic projections of a molecular complex, along with a specified atom pair representing a hydrogen bond (e.g., `ARG 45 NH2 A, O1B`, denoting the interaction between the `NH2` atom of residue `ARG45` on chain `A` and the `O1B` atom on the ligand). The objective is to distinguish the correct transformation that repositions the corresponding interaction to the center of the visualization from three other decoys. The decoy transformations are generated by perturbing the sign and magnitude of the ground-truth translation parameters.

Global Relational Reasoning This series of tasks considers the overall spatial relations of the entire complex instead of single localized ones, which is inherently more difficult than previous tasks, as they test the ability of the models to reason high-order combinations of spatial operations.

Ligand Docking (*Cloze*) This task emulates the molecular docking process to evaluate whether the model can infer the complementary binding configuration and corresponding geometric transformations. The model is provided with three

orthographic views of the ligand alone, the pocket alone, and one undocked complex view obtained by translating and rotating the ligand away from its pocket. Rotation angles are uniformly sampled from $[-90°, 90°]$, while translation distances are adaptively determined for each complex to minimize spatial overlap between the displaced ligand and pocket (Specific details can be found in Appendix A). The model must predict the sequence of transformations required to recover the native docking conformation, such as `roll y 45, move x -12`. Each complex generates six training samples (21,018 in total) and three test samples (1,470 in total).

Pocket-Ligand Interaction (*Cloze*) This task extends the Residue–Ligand Interaction task to the entire binding pocket, requiring the model to reason about global intermolecular contact patterns. Given six orthographic projections of the full complex, the models are required to output all hydrogen-bond interactions between the ligand and the pocket in a structured format like `ARG 221 NH2 A, O22; ARG 221 N A, O23; ARG 221 NE A, O22`. Each interaction is expressed as a tuple specifying the residue type, residue index, interacting atom in the residue, chain identity, and interacting atom in the ligand, with semicolons separating multiple entries.

5 Experiments

We first introduce the experimental setup in Sec. 5.1, and then report the main results of all compared models on our benchmark in Sec. 5.2. Finally, we conduct factor analysis and case study in Sec. 5.3.

5.1 Setup

Benchmark Subsets Due to the expensive cost of closed-source models (especially reasoning models), we create `MiSI-Bench(tiny)` by randomly sampling 50 question-answer pairs from each task in our dataset. This tiny subset is then used for affordable evaluation. This tiny subset will be used for evaluating the performance of all closed-source models and open-source mixture-of-experts (MoE) models, ensuring an intuitive and fair comparison. All models are evaluated under few-shot settings to provide them with necessary scientific prior knowledge.

Metrics For *Multiple-Choice Questions* and Zooming task, we follow the convention to adopt Accuracy (ACC) as the main metric [20, 55], which is calculated as the proportion of answers exactly matching the ground truth. For *Cloze Questions*, where answers might involve continuous numerical values and multiple entries, we employ a weighted composite score, as inspired by previous literature [18, 35], to reflect the degree of correctness beyond exact matching. For tasks involving spatial transformations (*i.e.*, `move` and `roll`), when the model predicts the correct axis, we will further assign scores based on the predicted values. Specifically, the score is determined by the normalized absolute error

Table 1: Evaluation on `MiSI-Bench`. Bold indicates the best result among all models. Since in the Trans-Rot. and Rot-Rot. tasks, the predictions of almost all models are close to random guessing, we exclude these two rows when calculating the average scores.

Methods	Rank	Avg.	*Translation*	*Rotation*	*Zooming*	*Res-Lig Inter Pos.*	*Res-Lig Inter Neg.*	*Trans-Rot.*	*Rot-Rot.*	*Docking*	*Inter Location*	*Poc-Lig Inter.*
			Unit Task					Composite Task				
Reasoning Models												
Human Level	-	81.18	100.00	70.18	30.00	100.00	100.00	32.00	26.00	74.54	92.00	82.78
GPT-5-mini	9	27.71	47.71	30.55	4.00	29.33	34.00	28.00	22.44	27.24	47.82	1.01
O4-mini	8	28.55	39.71	36.36	2.08	12.67	76.00	40.00	20.00	31.51	30.00	0.00
O3	3	33.65	52.29	43.82	2.00	18.67	**94.00**	22.00	22.44	20.69	36.00	1.71
Gemini-2.5-pro	6	29.94	50.15	38.88	0.00	28.67	52.00	30.61	21.62	30.38	38.00	1.44
Gemini-2.5-flash-lite	11	16.00	36.29	22.55	4.00	6.67	0.00	30.00	25.00	32.25	26.00	0.25
Claude Opus4	4	33.13	57.43	24.73	6.00	33.67	74.00	12.00	26.00	34.39	34.00	0.77
Claude Sonnet4.5	2	34.37	45.71	44.18	6.00	22.33	84.00	28.00	26.00	34.12	38.00	0.60
Qwen3-vl-235b-a22b-thinking	10	23.34	46.36	25.21	6.00	17.03	25.00	20.40	22.00	29.32	38.00	0.00
General Models												
GPT-4.1	7	29.20	29.71	37.45	2.00	7.33	80.00	33.33	29.26	32.90	36.00	0.2
Claude Sonnet3.5	5	31.23	47.14	37.11	10.00	27.50	70.00	18.00	32.00	27.52	28.00	2.55
Our Model												
Qwen2.5VL-7B-SFT	1	**62.96**	**99.84**	**99.71**	**27.14**	**63.46**	89.52	**88.44**	**89.59**	**57.79**	**88.37**	**10.72**

between the predicted ($\hat{d}$) and ground-truth (d) magnitudes (*i.e.*, distances and angles), $|\hat{d} - d|$. For composite tasks involving multiple transformations, such as Ligand Docking, each component operation (*e.g.*, `move` and `roll`) contributes equally to the total score, summing up to 1.0. In this task, since the axis for the `move` operation is fixed to x, scores are assigned only when the sign of the predicted ($\hat{d}$) matches the sign of the ground-truth (d). For Residue–Ligand Interaction and Pocket–Ligand Interaction tasks, we compute the ratio of correctly predicted interactions among all provided outputs. To penalize cheating behaviors where models output all the correct interactions along with hallucinations of irrelevant ones, such cases are assigned with a score of 0.5. Furthermore, if the number of hydrogen bonds in the model's response exceeds twice the number in the ground-truth, we consider the model to be attempting to score through exhaustive enumeration, in which case the model receives a score of 0. Furthermore, we also provide the results for exact matching for *Cloze Questions* in Appendix B.

Benchmark Models We comprehensively evaluate ten VLMs spanning four major model families, including nine closed-source and one open-source representatives. From Open AI, we include GPT-5-mini (w/ mixed modes of reasoning) [39], o4-mini (w/ reasoning) [40], o3 (w/ reasoning) [40], and GPT-4.1 (w/o reasoning) [38]. From Anthropic's Claude series, we test Claude 4.5 Sonnet (w/

Table 2: Ablation study on modality dependence and instruction robustness for microscopic spatial intelligence tasks. For convenience, we denote all tasks in the `MiSI-Bench` as T1–T9.

	Tasks								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
Diff-template	99.7	99.6	**29.4**	76.9	**89.2**	87.8	53.9	87.4	10.2
Random image	33.5	16.1	5.0	0.0	26.0	27.4	13.9	25.2	0.0
Qwen2.5VL-7B-SFT	**99.8**	**99.7**	27.1	**78.1**	88.4	**89.5**	**57.8**	**88.3**	**10.7**

reasoning) [5], Claude 4 Opus (w/ reasoning) [4], and Claude 3.5 Sonnet (w/o reasoning) [3]. From Google's Gemini series, we include Gemini-2.5-pro (w/ reasoning) [15] and Gemini-2.5-flash-lite (w/ reasoning) [14]. For open-source models, our preliminary experiments show that most models below 32B parameters achieve very low performance on `MiSI-Bench`. Therefore, we select the strong Qwen3-vl-235b-a22b-thinking (w/ reasoning) [43] from larger open-source models as a representative baseline. Furthermore, we propose Qwen2.5VL-7B-SFT, which is finetuned on the training split of our benchmark to investigate how to better trigger the *Microscopic Spatial Intelligence* in VLMs.

Human-Level Performance We estimate the performance of humans by recruiting PhD candidates in biology to complete the questions for residue-ligand interaction, ligand docking, and pocket-ligand interaction, which requires more domain expertise than the other tasks. For the rest of tasks, we employ PhD candidates in broader field of science, technology, engineering, and mathematics to propose answers. Each participant is required to answer questions in `MiSI-Bench(tiny)` independently, whose responses are later assessed with the same metrics as the models.

5.2 Main Results

Human Level Performance. The evaluation results are shown in Tab. 1. Human evaluators perform well in most unit tasks, demonstrating strong 3D spatial modeling abilities and the potential to integrate biological knowledge with spatial reasoning for basic interactive tasks. However, their performance decline significantly in complex spatial reasoning tasks. They manage small-angle rotations by tracking key atomic changes, while large-scale rotations—requiring maintained spatial continuity and multi-atom tracking—increase cognitive load and impaired axis and angle judgment. Zooming tasks prove even more challenging, as their judgments rely on overall intuition regarding boundary shifts and atomic density changes, lacking clear reference points and resulting in greater estimation errors.

In composite tasks, consecutive spatial operations (e.g., Trans-Rot., Rot-Rot.) lead to error accumulation and frequent reference frame shifts, significantly

degrading human performance. Such tasks impose high demands on working memory and the stability of spatial mental simulation, forming a bottleneck in human performance. In Docking task, performance is the poorest due to the need for both sequential spatial transformations and biological knowledge to determine hydrogen bond formation and optimal docking positions. For Poc-Lig Inter. task, the main challenge lies in integrating multiple 2D views to reconstruct the 3D conformation of occluded residues before identifying hydrogen bonds, making it highly demanding. In contrast, Inter Location. task are simpler: they involve no rotational operations—only distance perception—and provide hydrogen bond information upfront, eliminating the need for specialized knowledge.

Advancing VLMs Performance. As evidenced by the results in the table, all advanced models perform sub-optimally across tasks in `MiSI-Bench(tiny)`. Overall, the models exhibit better performance in distance-related tasks compared to rotation-related ones. For instance, in tasks such as "Translation" versus "Rotation", and "Interaction Location" versus "Rotation–Rotation Movement," the models consistently achieve higher scores in the former. This may stem from the fact that most current VLMs are primarily trained on two-dimensional data, making distance—as a two-dimensional attribute—more readily adaptable for the models. Furthermore, in the "Residue–Ligand Interaction–Pos" and "Pocket–Ligand Interaction" tasks, the performance gap between the models and human-level performance is most pronounced, highlighting their still-insufficient knowledge reserves in specialized domains such as biology. In contrast, in the "Residue–Ligand Interaction–Neg" task, most models perform relatively well, likely because the greater spatial distance between residues and ligands in negative samples allow the models to make correct judgments based solely on spatial proximity.

SFT Model Performance. Experimental results demonstrate that after fine-tuning on the `MiSI-Bench` dataset, model performance improves significantly, surpassing mainstream VLMs across all tasks and exceeding human-level performance in complex spatial tasks such as Rotation. Notably, in the Rot–Rot. task, where human performance approaches random guessing, the model maintains approximately 90% accuracy, indicating that advanced VLMs possess potential for 3D spatial cognition. Previous underperformance of advancing VLMs may have stemmed from domain adaptation barriers: although models have generic spatial understanding, they lack visual priors for specialized structures such as proteins, hindering knowledge transfer. Appropriate fine-tuning can establish cross-domain mappings and unlock their spatial reasoning capabilities. However, in tasks such as Res/Poc–Lig Inter., which rely on domain-specific knowledge, models still lag behind humans, suggesting that the absence of domain priors in foundational training remains a bottleneck. Future work should focus on further exploring the spatial potential of models and investigating how to effectively integrate explicit knowledge from scientific fields such as structural biology.

5.3 Analysis

Instruction Generalization and Robustness Analysis. To verify whether the model genuinely comprehends the underlying microscopic spatial structures, rather than merely relying on pattern matching from the fixed text templates in the benchmark, we further evaluate its generalization capabilities when faced with free-form and less-constrained instructions.

Specifically, we employ an advanced large language model (GPT-5.2) to comprehensively paraphrase the original instruction templates across all 9 tasks. While strictly preserving the core semantics, the paraphrased instructions introduce richer syntactic structures and diverse expressions, thereby constructing an evaluation set that more closely resembles real-world interaction scenarios.

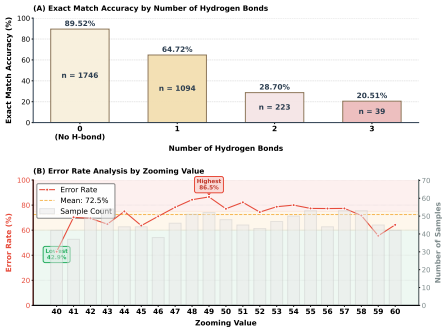

Fig. 4: Factor Analysis of `MiSI-Bench`.

Subsequently, we re-evaluate our supervised fine-tuned (SFT) model on these newly generated prompts. As shown in Row 1 of Tab. 2, the model's performance on the paraphrased instructions remains highly consistent with the evaluation results under the original templates (Row 3 of Tab. 2). This negligible performance fluctuation provides compelling evidence that our model does not achieve high scores by memorizing or over-fitting specific textual shortcuts, but rather has developed profound, highly generalizable microscopic spatial intelligence and reasoning capabilities.

Ablation Study on Input Modalities and Language Priors. To verify whether the model over-relies on language priors (e.g., exploiting shortcuts through the distribution of discrete values like rotation angles or translation distances) rather than genuinely understanding microscopic visual content, we conduct ablation studies on input modalities. Specifically, we investigate the impact of "image-only" and "text-only" settings on model performance. 1) Infeasibility of the Image-only Setting: In our task formulation, entirely removing text instructions is infeasible. This is because, in the evaluation of microscopic spatial intelligence, many distinct tasks may share identical initial visual states. Without text instructions to specify the exact task requirements, the model is fundamentally unable to determine the required operation. Therefore, text instructions are an indispensable prerequisite for defining task objectives. 2) Performance Verification in the Text-only Setting: To isolate and evaluate the independent contribution of text instructions, we design a "text-only" ablation study. In this setting, we retain all original text instructions but replace the input molecular images with random images. As shown in Row 2 of Tab. 2, when effective visual input is severed, the model's performance suffers a catastrophic collapse,

plummeting to near-random guessing levels, with some tasks even dropping to zero.

Robustness to View Selection and Depth Cues. To investigate the model's sensitivity to the choice and number of orthographic views, and to understand how additional depth cues affect performance, we conduct an ablation study across different levels of task complexity in microscopic spatial intelligence.

For fundamental unit tasks, such as Rotation, reducing the visual input from three orthogonal views to a single view resulted in only a marginal performance drop (from 99.71 to 97.51). This indicates that for basic spatial operations, the model possesses robust monocular depth perception, effectively inferring 3D transformations from limited 2D cues without a heavy reliance on multi-view inputs.

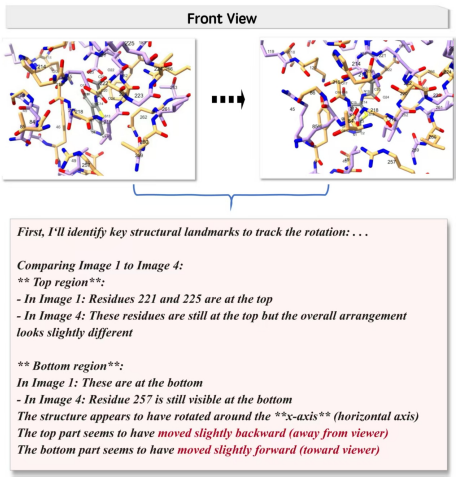

Fig. 5: Case study of the Rotation task.

Conversely, for composite tasks that require comprehensive global structural analysis, such as Pocket-Ligand Interaction, severe occlusion presents a significant challenge. In these complex scenarios, increasing the number of observation views from 6 to 27 yielded a substantial performance improvement, raising the score from 10.72 to 23.77. These findings demonstrate that while basic microscopic spatial reasoning can be accurately achieved with minimal views, resolving intricate molecular interactions benefits significantly from the supplementary depth cues and the reduction in occlusion provided by a denser set of viewing angles.

Factor Analysis. In this section, we conduct a detailed analysis of the suboptimal performance exhibited by the SFT model on the Res-Lig Inter Pos. and Zooming tasks. Fig. 4(a) and (b) present the prediction accuracy of the model across different statistical intervals for these two tasks. As shown in Fig. 4(a), the prediction accuracy decreases sharply as the number of hydrogen bonds increases, indicating that the model struggles to identify all hydrogen bonds in scenarios with complex hydrogen-bonding interactions. In Fig. 4(b), the prediction error rate curve exhibits an initial increase followed by a decrease. We hypothesize that this pattern may stem from the model's failure to generalize uniformly across the entire scale space. The observed peak likely corresponds to a visually critical scale in molecular structures, where discriminative structural information is minimal, thereby reducing the parsing efficiency of the model's attention mechanism.

Case Study. In this section, we examine Claude Sonnet 4.5, the top-performing model among advancing VLMs, by analyzing its reasoning process in failure cases from the Rotation task. As shown in Fig. 5, the model demonstrates a logically sound approach: it identifies conserved residues across structural changes as anchors and infers the rotation axis and angle accordingly. However, in the key region it localizes, although the model correctly detects spatial rearrangements of residues 221 and 225, it misinterprets the change as "moved slightly backward," leading to an incorrect conclusion. In fact, simply observing the positional shift of residue 221 would suggest a rotation around the y-axis. This case indicates that advancing VLMs still lack adequate spatial reasoning capabilities and require more effective mechanisms to elicit such skills.

6 Conclusion

In this work, we establish Microscopic Spatial Intelligence (MiSI) as a distinct and critical challenge for Vision-Language Models (VLMs), extending beyond macroscopic understanding to the atom-level reasoning essential for scientific discovery. We propose a systematic benchmark framework dubbed as `MiSI-Bench`, for evaluating various advanced VLMs for MiSI. The experiments reveals a significant performance gap between state-of-the-art VLMs and human expertise. Yet, the strong performance of a fine-tuned 7B model underscores the substantial potential of VLMs to master complex spatial transformations, even surpassing humans on complex spatial tasks such as rotation. Ultimately, achieving robust MiSI will require not only scaling model architectures but also the explicit integration of scientific knowledge for real-world scientific applications.

Acknowledgments

This work was jointly supported by the following projects: the National Natural Science Foundation of China (No. 62376276); Beijing Nova Program (No. 20230484278); the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (23XNKJ19); the Beijing Major Science and Technology Project under Contract No. Z251100008125061. This work was supported by Beijing Academy of Artificial Intelligence (BAAI) and Damo Academy through Damo Academy Research Intern Program.

References

1. Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K., et al.: The rosetta all-atom energy function for macromolecular modeling and design. Journal of chemical theory and computation **13**(6), 3031–3048 (2017)
2. Anderson, A.C.: The process of structure-based drug design. Chemistry & biology **10**(9), 787–797 (2003)

3. Anthropic: Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet (2025)
4. Anthropic: Introducing claude 4. https://www.anthropic.com/news/claude-4 (2025)
5. Anthropic: Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5 (2025)
6. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
7. Bai, X.C., McMullan, G., Scheres, S.H.: How cryo-em is revolutionizing structural biology. Trends in biochemical sciences **40**(1), 49–57 (2015)
8. Battaglia, P., Pascanu, R., Lai, M., Jimenez Rezende, D., et al.: Interaction networks for learning about objects, relations and physics. Advances in neural information processing systems **29** (2016)
9. Boiko, D.A., MacKnight, R., Kline, B., Gomes, G.: Autonomous chemical research with large language models. Nature **624**(7992), 570–578 (2023)
10. Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P.: Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine **34**(4), 18–42 (2017)
11. Carlbom, I., Paciorek, J.: Planar geometric projections and viewing transformations. ACM Computing Surveys (CSUR) **10**(4), 465–502 (1978)
12. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
13. Cherezov, V., Rosenbaum, D.M., Hanson, M.A., Rasmussen, S.G., Thian, F.S., Kobilka, T.S., Choi, H.J., Kuhn, P., Weis, W.I., Kobilka, B.K., et al.: High-resolution crystal structure of an engineered human β2-adrenergic g protein–coupled receptor. science **318**(5854), 1258–1265 (2007)
14. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
15. DeepMind, G.: Gemini 2.5: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/ (2025)
16. DeLano, W.L., et al.: Pymol: An open-source molecular graphics tool. CCP4 Newsl. protein crystallogr **40**(1), 82–92 (2002)
17. Eidelman, S., Hayes, K., Olive, K.e., Aguilar-Benitez, M., Amsler, C., Asner, D., Babu, K., Barnett, R., Beringer, J., Burchat, P., et al.: Review of particle physics. Physics letters B **592**(1-4), 1–5 (2004)
18. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
19. Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., Wang, H.: Geometry-enhanced molecular representation learning for property prediction. Nature Machine Intelligence **4**(2), 127–134 (2022)
20. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24108–24118 (2025)

21. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
22. Gao, B., Qiang, B., Tan, H., Jia, Y., Ren, M., Lu, M., Liu, J., Ma, W.Y., Lan, Y.: Drugclip: Contrastive protein-molecule representation learning for virtual screening. Advances in Neural Information Processing Systems **36**, 44595–44614 (2023)
23. Huang, W., Jiao, R., Kong, X., Zhang, L., Yu, Z., Ren, F., Tan, W., Liu, Y.: An equivariant pretrained transformer for unified 3d molecular representation learning (2025)
24. Kong, X., Huang, W., Liu, Y.: Conditional antibody design as 3d equivariant graph translation. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=LFHFQbjxIiP
25. Kong, X., Huang, W., Liu, Y.: Generalist equivariant transformer towards 3d molecular interaction learning. In: International Conference on Machine Learning. pp. 25149–25175. PMLR (2024)
26. Kong, X., Jiao, R., Lin, H., Guo, R., Huang, W., Ma, W.Y., Wang, Z., Liu, Y., Ma, J.: Peptide design through binding interface mimicry with pepmimic. Nature biomedical engineering pp. 1–16 (2025)
27. Kong, X., Zhang, Z., Zhang, Z., Jiao, R., Ma, J., Huang, W., Liu, K., Liu, Y.: Unimomo: Unified generative modeling of 3d molecules for de novo binder design. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=KUN7A7Okb6
28. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
29. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
30. Li, Z., Cen, J., Huang, W., Wang, T., Song, L.: Size-generalizable RNA structure evaluation by exploring hierarchical geometries. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=QaTBHSqmH9
31. Li, Z., Cen, J., Su, B., Xu, T., Rong, Y., Zhao, D., Huang, W.: Large language-geometry model: When LLM meets equivariance. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=Qi9rPcdiTY
32. Li, Z., Ma, Z., Li, M., Li, S., Rong, Y., Xu, T., Zhang, Z., Zhao, D., Huang, W.: Star-r1: Multi-view spatial transformation reasoning by reinforcing multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12041–12051 (2026)
33. Li, Z., Zhu, X., Lei, Z., Zhang, Z.: Deconfounding physical dynamics with global causal relation and confounder transmission for counterfactual prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1536–1545 (2022)
34. Li, Z., Zhu, X., Zhang, X., Zhang, Z., Lei, Z.: Visual commonsense based heterogeneous graph contrastive learning. arXiv preprint arXiv:2311.06553 (2023)
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)

36. Lu, C., Liu, Q., Wang, C., Huang, Z., Lin, P., He, L.: Molecular property prediction: A multilevel quantum interactions modeling perspective. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 1052–1060 (2019)
37. M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A.D., Schwaller, P.: Augmenting large language models with chemistry tools. Nature Machine Intelligence **6**(5), 525–535 (2024)
38. OpenAI: Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/ (2025)
39. OpenAI: Introducing gpt-5. https://openai.com/index/introducing-gpt-5/ (2025)
40. OpenAI: Introducing o3 and o4 mini. https://openai.com/index/introducing-o3-and-o4-mini/ (2025)
41. Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., Ferrin, T.E.: Ucsf chimerax: Structure visualization for researchers, educators, and developers. Protein science **30**(1), 70–82 (2021)
42. Pinheiro, P.O., Jamasb, A.R., Mahmood, O., Sresht, V., Saremi, S.: Structure-based drug design by denoising voxel grids. In: International Conference on Machine Learning. pp. 40795–40812. PMLR (2024)
43. QwenTeam: Qwen3-vl: Sharper vision, deeper thought, broader action. https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef&from=research.latest-advancements-list (2025)
44. Ray, A., Duan, J., Tan, R., Bashkirova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B.A., Krishna, R., Zeng, K.H., et al.: Sat: Spatial aptitude training for multimodal language models. arXiv e-prints pp. arXiv–2412 (2024)
45. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., Serrano, L.: The foldx web server: an online force field. Nucleic acids research **33**(suppl_2), W382–W388 (2005)
46. Swanson, K., Wu, W., Bulaong, N.L., Pak, J.E., Zou, J.: The virtual lab of ai agents designs new sars-cov-2 nanobodies. Nature pp. 1–3 (2025)
47. Tang, K., Gao, J., Zeng, Y., Duan, H., Sun, Y., Xing, Z., Liu, W., Lyu, K., Chen, K.: Lego-puzzles: How good are mllms at multi-step spatial reasoning? arXiv preprint arXiv:2503.19990 (2025)
48. Thölke, P., Fabritiis, G.D.: Equivariant transformers for neural network based molecular potentials. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=zNHzqZ9wrRB
49. Townshend, R.J.L., Vögele, M., Suriana, P.A., Derry, A., Powers, A., Laloudakis, Y., Balachandar, S., Jing, B., Anderson, B.M., Eismann, S., Kondor, R., Altman, R., Dror, R.O.: ATOM3d: Tasks on molecules in three dimensions. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021), https://openreview.net/forum?id=FkDZLpK1Ml2
50. Trott, O., Olson, A.J.: Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry **31**(2), 455–461 (2010)
51. Wang, F., Fu, X., Huang, J.Y., Li, Z., Liu, Q., Liu, X., Ma, M.D., Xu, N., Zhou, W., Zhang, K., et al.: Muirbench: A comprehensive benchmark for robust multi-image understanding. arXiv preprint arXiv:2406.09411 (2024)
52. Wang, R., Fang, X., Lu, Y., Yang, C.Y., Wang, S.: The pdbbind database: methodologies and updates. Journal of medicinal chemistry **48**(12), 4111–4119 (2005)
53. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Pro-

ceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
54. Yin, B., Wang, Q., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., et al.: Spatial mental modeling from limited views. In: Structural Priors for Vision Workshop at ICCV'25 (2025)
55. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
56. Zhang, J., Chen, Y., Zhou, Y., Xu, Y., Huang, Z., Mei, J., Chen, J., Yuan, Y.J., Cai, X., Huang, G., et al.: From flatland to space: Teaching vision-language models to perceive and reason in 3d. arXiv preprint arXiv:2503.22976 (2025)
57. Zhao, B., Wang, Z., Fang, J., Gao, C., Man, F., Cui, J., Wang, X., Chen, X., Li, Y., Zhu, W.: Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 11071–11080 (2025)
58. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. AI open **1**, 57–81 (2020)
59. Zuo, Y., Kayan, K., Wang, M., Jeon, K., Deng, J., Griffiths, T.L.: Towards foundation models for 3d vision: How close are we? In: 2025 International Conference on 3D Vision (3DV). pp. 1285–1296. IEEE (2025)