From Views to Worlds: Active Exploration over 3D Worlds for Vision-Language Models

Qijian Tian[1], **Jiayu Ying**[2], **Ke Fan**[1], **Lizhuang Ma**[1], **Xin Tan**[2*]

[1]Shanghai Jiao Tong University, Shanghai, China
[2]East China Normal University, Shanghai, China

tianqijian@sjtu.edu.cn, 51295901118@stu.ecnu.edu.cn
slipperyfrank@sjtu.edu.cn, ma-lz@cs.sjtu.edu.cn, xtan@cs.ecnu.edu.cn

https://fangzhou2000.github.io/projects/v2w

Abstract

Vision-language models (VLMs) have achieved remarkable progress in 2D visual understanding tasks, yet the passive view-centric perception paradigm limits their capacity for 3D spatial reasoning. Such reasoning often requires actively acquiring spatial evidence beyond the currently visible views, such as camera poses and cross-view spatial relationships, to infer 3D relationships that are not explicitly represented in 2D images. However, existing methods either lack active exploration or merely provide additional views during inference, without explicitly acquiring such spatial evidence beyond 2D images. In this paper, we propose View-to-World (V2W), an active exploration framework that enables VLMs to actively explore 3D worlds and lifts VLMs from passive view-centric perception to active world-centric spatial reasoning. Given a spatial reasoning task with multi-view images, V2W first reconstructs an explicit 3D world with geometry, camera poses, and language-grounded semantics, and exposes it to VLMs through visual and linguistic interfaces. Through multi-turn interaction with these interfaces, VLMs iteratively acquire spatial evidence from the constructed 3D world, enabling active world-centric spatial reasoning. V2W can be integrated with existing VLMs in a training-free manner and can be further enhanced by optimizing the exploration policy via reinforcement learning. Experiments on spatial mental modeling benchmarks demonstrate that V2W substantially improves VLM baselines in the training-free setting and achieves state-of-the-art performance with the learned exploration policy, highlighting the importance of active exploration for spatial intelligence.

1 Introduction

Recent advances in vision-language models (VLMs) have demonstrated remarkable capabilities in 2D visual understanding tasks, including image captioning [4, 14], visual question answering [15, 34, 77], and visual grounding [2, 3, 44]. Building on their strong image-level perception and reasoning capabilities, recent works [8, 13] have begun to explore the potential of VLMs for 3D spatial reasoning. However, 2D images are inherently partial observations of the 3D world from specific viewpoints, making view-centric perception fundamentally limited for understanding 3D space. This limitation stems from two key factors. First, critical spatial evidence may be occluded or absent from the currently visible views, which requires active viewpoint changes beyond fixed-view inputs. Second, even with more viewpoints, 2D images alone fail to explicitly represent the cross-view 3D

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

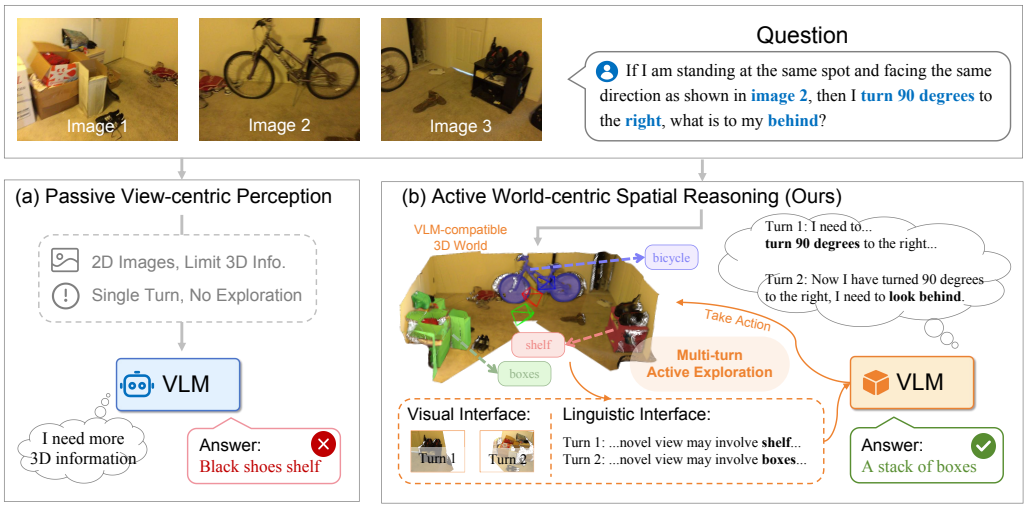

Figure 1: Comparison of passive view-centric perception and active world-centric spatial reasoning

information, such as relative camera poses, cross-view entity correspondences, and object-level 3D spatial relations. Addressing these limitations requires models to actively acquire evidence over a coherent 3D world representation that integrates geometric, visual, and semantic information.

However, most existing methods still rely on passively perceiving 2D inputs for spatial reasoning. Some methods attempt to incorporate geometric priors into VLMs, including encoding 3D geometry into the representation space [19, 24, 45, 65, 85, 88], fusing implicit 3D information at the feature level [17, 33, 80, 84], and introducing 3D priors through distillation [6, 25, 60] or latent reasoning [12, 31, 72]. Although these methods provide VLMs with additional 3D priors, they mainly improve the internal representation capability rather than provide VLMs with an explicitly interactive and actively explorable 3D world, and therefore still belong to the passive perception from fixed-view input. More recent methods [5, 70, 83] have further explored active perception by providing VLMs with additional 2D observations. Nevertheless, these methods remain largely view-centric, as additional observations are still presented as 2D views and lack cross-view 3D information needed for spatial reasoning. Such 3D information is not explicitly represented in 2D views, but is grounded in the underlying 3D world. We refer to reasoning over such information as world-centric spatial reasoning. Therefore, advancing VLMs toward 3D spatial reasoning requires moving from passive view-centric perception to active world-centric spatial reasoning, as shown in Fig. 1.

In this paper, we propose View-to-World (V2W), a model-agnostic framework that enables VLMs to actively explore over an explicit 3D world representation. Given multi-view images, V2W first recovers scene geometry and camera poses with off-the-shelf 3D foundation models and incorporates language-grounded semantics to build a geometrically consistent, semantically aligned, and language-queryable 3D world. We denote it as a VLM-compatible 3D world, since the geometry, camera poses, and semantic evidence are exposed to VLMs through images and natural language rather than raw 3D representations. We design two interfaces to achieve the exposure: a visual interface that provides 2D observations from newly explored viewpoints and a linguistic interface that converts geometric and semantic information from the 3D world into explicit natural-language cues. V2W enables VLMs to actively explore the constructed 3D world through visual and linguistic interfaces and progressively acquire and integrate spatial evidence via multi-turn interaction, thereby performing spatial reasoning in an active world-centric manner. Experiments on spatial mental modeling benchmarks show that V2W improves existing VLMs in a training-free manner, revealing their inherent capability for active exploration. However, we also find that smaller VLMs often struggle to perform effective active exploration, exhibiting insufficient or uninformative exploration. To this end, we further propose to train the exploration policy through reinforcement learning, thereby improving the active exploration capability. With the learned exploration policy, V2W achieves state-of-the-art performance, substantially outperforming both baseline methods and proprietary commercial models (e.g., surpassing GPT-4o by 48.5 percentage points on MindCube-Tiny), demonstrating the

importance of active exploration capability for spatial reasoning. Our contributions are summarized as follows:

- We construct a VLM-compatible 3D world representation that is geometrically consistent, semantically aligned, and language-queryable, providing a 3D environment that lifts VLMs from view-centric perception to world-centric spatial reasoning.
- We propose V2W, an active exploration framework that enables VLMs to interact with the constructed 3D world through visual and linguistic interfaces and progressively acquire world-grounded evidence via multi-turn exploration.
- We introduce reinforcement learning to train the exploration policy of V2W, enabling more effective evidence acquisition from the constructed 3D world.
- Experiments on spatial mental modeling benchmarks demonstrate that V2W substantially improves VLM baselines in the training-free setting and achieves state-of-the-art performance with the learned exploration policy.

2 Related Work

2.1 Vision-Language Models for Spatial Reasoning

Spatial reasoning is essential for intelligent perception and interaction. It allows agents to represent, organize, and transform objects and their spatial relationships in 3D environments [22, 27, 28, 75]. However, recent studies show that existing VLMs still struggle with spatial reasoning [1, 9, 10, 23, 48, 68, 69, 74, 78, 79, 81]. Prior efforts have attempted to improve the capability of VLMs mainly in two ways. The first line of work injects 3D priors into VLM representations. These methods encode 3D geometry into the representation space [19, 24, 45, 65, 85, 88], fuse implicit 3D information at the feature level [17, 33, 80, 84], or introduce 3D knowledge through distillation [6, 25, 60] or latent reasoning [12, 31, 72]. While effective, they often require large-scale aligned data and costly training, and they do not provide explicit geometric constraints. The second line of work provides 3D information externally in the VLM context [11, 16, 20, 38, 83], which offers a more direct and model-agnostic way to improve 3D spatial awareness. However, a key challenge is how to help VLMs understand and use this external 3D information effectively. In this work, we also provide 3D information externally, but we construct a VLM-compatible 3D world that supports active exploration, allowing VLMs to make better use of 3D information.

2.2 3D Foundation Models and Scene Understanding

Recent feed-forward reconstruction models [30, 35, 57, 61–63] have made rapid progress in 3D modeling. These models can recover 3D geometry from multi-view images and accurately estimate camera parameters and other 3D information. Their strong generalization ability further positions them as 3D foundation models. In parallel, visual foundation models, such as the SAM series [7, 29, 47], have demonstrated strong capabilities in scene parsing and understanding. Recent studies [32, 56, 58, 59, 87] further integrate 3D and visual foundation models to support semantic 3D scene understanding. While the above work mainly focuses on 3D representation and perception-oriented tasks, our work investigates how VLMs can benefit from such foundation models to improve spatial reasoning capability.

2.3 Vision-Language Models with Active Interaction

Recent VLMs have advanced from direct reasoning over given inputs to interactive reasoning [37, 71, 73]. Representative methods allow models to interact with external resources during inference through multi-turn reasoning, including tool-augmented VLMs [50, 53, 66, 67, 86] and visual programming methods [21, 38, 39, 54]. Such interaction reflects the active exploration capability of VLMs [52, 64, 89]. This capability is particularly important for spatial reasoning, since understanding 3D space often requires integrating observations from different viewpoints and reasoning about regions beyond the current field of view. Recent works have introduced active exploration into spatial reasoning through novel-view rendering [83] or generation [5, 70]. However, these methods mainly provide additional 2D views to VLMs and therefore remain view-centric, without explicitly exposing 3D information beyond 2D images. Another work [38] performs active exploration through code

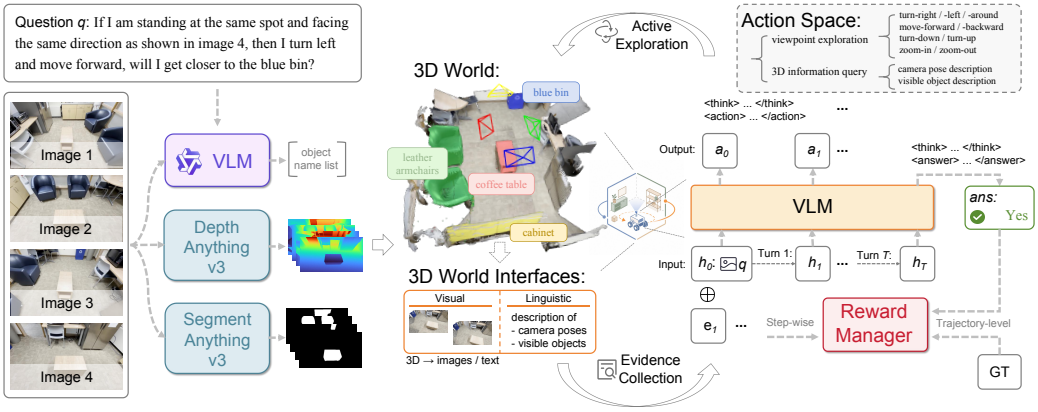

Figure 2: Overview of the View-to-World (V2W) framework. V2W constructs a VLM-compatible 3D world from multi-view images. The 3D information is exposed through visual and linguistic interfaces, enabling the VLM to actively explore the world via multi-turn interaction, and the exploration policy can be further optimized via reinforcement learning.

generation, but this paradigm restricts multi-turn exploration and relies heavily on the code-generation ability of VLMs. In contrast, V2W constructs a VLM-compatible 3D world with visual and linguistic interfaces, enabling VLMs to gather spatial evidence through multi-turn interaction and shift from passive view-centric perception to active world-centric exploration.

3 Methods

3.1 Overview

Given multi-view images $\mathcal{I} = \{I_i\}_{i=1}^{N}$ and a spatial reasoning question q, V2W aims to produce the answer by enabling a VLM to actively explore an explicit 3D world. As shown in Fig. 2, V2W first constructs a VLM-compatible 3D world $\mathcal{W}$ by reconstructing the 3D scene and lifting question-relevant object semantics into 3D space (Sec. 3.2). Since VLMs only take images and text as input, $\mathcal{W}$ is exposed through two interfaces: a visual interface that renders novel-view images, and a linguistic interface that provides textual cues about visible objects and relative camera poses. During inference, V2W formulates active exploration as a multi-turn interaction via function call (Sec. 3.3). We further train the exploration policy with supervised fine-tuning and reinforcement learning (Sec. 3.4), which strengthens the VLM's ability to actively explore the 3D world.

3.2 VLM-compatible 3D World Construction

Active exploration requires an explicit 3D world that serves as an interactive environment for existing VLMs. Since VLMs typically take images and natural language as inputs, raw 3D representations such as point clouds, meshes, or 3D Gaussian Splatting cannot be directly used. Meanwhile, spatial reasoning often requires both geometry and language-grounded semantics. Therefore, we construct a VLM-compatible 3D world that integrates scene geometry, camera poses, and object-level semantics, and exposes them to VLMs through visual and linguistic interfaces. These interfaces convert explicit 3D information into rendered observations and natural-language cues, allowing VLMs to acquire spatial evidence without modifying their architectures.

Geometric Reconstruction and Semantic Lifting. We first use an off-the-shelf 3D foundation model Φ_{3D} [35] to estimate the geometry and camera parameters from the input images $\mathcal{I} = \{I_i\}_{i=1}^{N}$,

$$\{\mathbf{D}_i, \mathbf{E}_i, \mathbf{K}_i\}_{i=1}^{N} = \Phi_{\text{3D}}(\mathcal{I}), \tag{1}$$

where $\mathbf{D}_i \in \mathbb{R}^{H \times W}$ is the predicted depth map, $\mathbf{E}_i \in \text{SE}(3)$ is the camera-to-world pose, and $\mathbf{K}_i$ is the camera intrinsic. We then obtain the point map by back-projecting each pixel into 3D space:

$$\mathbf{P}_i(\mathbf{u}) = \pi^{-1}(\mathbf{u}, \mathbf{D}_i(\mathbf{u}); \mathbf{K}_i, \mathbf{E}_i) = \mathbf{R}_i \left(\mathbf{D}_i(\mathbf{u})\mathbf{K}_i^{-1}\tilde{\mathbf{u}}\right) + \mathbf{t}_i, \tag{2}$$

where $\mathbf{u} = (u, v)$ is a pixel coordinate, $\tilde{\mathbf{u}} = (u, v, 1)^\top$ is its homogeneous form, and $\mathbf{E}_i = [\mathbf{R}_i|\mathbf{t}_i]$. The resulting point map $\mathbf{P}_i \in \mathbb{R}^{H\times W\times 3}$ assigns each pixel in I_i a 3D coordinate in the global coordinate system. The point maps provide the scene geometry, while the camera poses establish the spatial relationships among input views. To incorporate semantics, we use a lightweight VLM, Qwen2.5-VL-7B [3], to extract question-relevant object names from the input images $\mathcal{I}$ and the question q, and then use these objects as text prompts for SAM 3 [7] to obtain their 2D masks in the input views. The masks are lifted to 3D by associating masked pixels with their corresponding 3D points in the point maps. For an object o, its 3D semantic point set is defined as

$$\mathcal{X}_o = \{\mathbf{P}_i(\mathbf{u}) \mid \mathbf{M}_{i,o}(\mathbf{u}) = 1,\ i = 1, \ldots, N\}, \tag{3}$$

where $\mathbf{M}_{i,o}$ is the mask of object o in view I_i, and $\mathbf{u}$ is a pixel location. This step lifts language-grounded object semantics onto the reconstructed geometry, yielding a 3D representation with object-level semantic information. We additionally obtain 3D floor points through the same mask-lifting process, and align the reconstructed 3D world with the real scene by estimating the gravity direction from the floor plane using RANSAC [18]. This allows the subsequent action space to be defined on the horizontal plane with intuitive directions such as forward, backward, left, and right.

Visual and Linguistic Interfaces. The constructed 3D representation contains both geometry and object-level semantics, whereas existing VLMs typically take only images and text as input. To bridge this modality gap, we expose the 3D world through two VLM-compatible interfaces: the visual and linguistic interfaces. The visual interface returns rendered images from queried camera poses, while the linguistic interface returns textual cues derived from object-level semantics and spatial relations.

Given a queried camera pose, the visual interface renders the corresponding novel-view image as the visual observation for the VLM. To support object-level textual cues in the linguistic interface, we internally render object masks in the queried view. Specifically, for a target pose $\mathbf{E}_\tau$, the mask of object o is obtained by projecting its associated 3D points into the target view:

$$\mathcal{P}_o = \{\mathbf{P}_i(\mathbf{u}) \mid \mathbf{M}_{i,o}(\mathbf{u}) = 1, i = 1, \ldots, N\}, \qquad \hat{\mathbf{M}}_{\tau,o} = \Pi_\tau(\mathcal{P}_o). \tag{4}$$

where $\mathcal{P}_o$ denotes the set of 3D points lifted from the masks of object o across all input views, and $\Pi_\tau(\cdot)$ projects these points into the target view to obtain the internal mask $\hat{\mathbf{M}}_{\tau,o}$. The projected mask is used to identify visible objects, which are then returned as textual cues by the linguistic interface. The linguistic interface also describes spatial relations between camera poses. Given two camera poses and the gravity direction, we project the camera centers onto the horizontal plane to describe their relative translation as forward, backward, left, right, or diagonal, and compare their orientations to describe the relative rotation as turning left or right. Together, the visual and linguistic interfaces provide the VLM with image observations and textual spatial cues through its native input modalities.

3.3 Active Exploration via Function Calling

The visual and linguistic interfaces of the constructed VLM-compatible 3D world are formulated as callable functions, enabling V2W to perform active exploration as a structured tool-calling process. Each function corresponds to a predefined exploration action. During inference, the VLM calls a function with specific arguments, receives the returned images or text from the visual or linguistic interface, and then decides whether to continue exploration or produce the final answer.

Exploration Action Space. V2W defines an exploration action space with two types of actions: viewpoint exploration and 3D information query, as shown in Fig. 2. Viewpoint exploration actions change the current camera pose to acquire additional visual evidence from novel views, including horizontal rotations around the gravity direction, i.e., turn left, turn right, and turn around; translations on the horizontal plane, i.e., move forward and move backward; and view-centered adjustments, i.e., zoom in, zoom out, turn up, and turn down. After each viewpoint exploration action, the 3D world returns a rendered novel-view image through the visual interface. 3D information query actions provide textual spatial cues from the 3D world through the linguistic interface. These cues include visible object descriptions, which summarize the object categories observed from the explored viewpoint, and relative camera pose descriptions, which return the relative translation and rotation between two existing views in the gravity-aligned coordinate system.

Multi-turn Interaction. During inference, V2W formulates exploration as a multi-turn interaction between the VLM and the constructed 3D world. Let $\mathcal{W}$ denote the VLM-compatible 3D world, h_t denote the cumulative interaction history up to turn t, including the original question, previous function calls, and the accumulated evidence, and π_θ denote the VLM policy parameterized by θ, which conditions on the interaction history to generate exploration actions and the final answer. Starting from $h_0 = (\mathcal{I}, q)$, the VLM iteratively performs an exploration action, receives world-grounded evidence, and updates the interaction context:

$$a_t = \pi_\theta^{\text{act}}(h_t), \qquad e_{t+1} = \mathcal{W}(a_t), \qquad h_{t+1} = h_t \oplus (a_t, e_{t+1}), \tag{5}$$

where a_t denotes the exploration action selected by the VLM, π_θ^{act} denotes using the policy to generate the next exploration action, e_{t+1} denotes the returned visual or linguistic evidence, and $\oplus$ denotes appending the new action and evidence to the context. This recursive process allows the VLM to progressively accumulate world-grounded spatial evidence. When sufficient evidence has been collected at turn T, the VLM produces the final answer based on the interaction history:

$$\text{ans} = \pi_\theta^{\text{ans}}(h_T). \tag{6}$$

where π_θ^{ans} denotes using the same policy to generate the final answer from interaction history h_T.

3.4 Exploration Policy Learning

Although V2W can be applied to existing VLMs without training, smaller VLMs often struggle with active exploration. They may terminate prematurely, repeat uninformative actions, or select actions weakly related to the question. This suggests that effective exploration also depends on the VLM's exploration policy. We therefore optimize this policy using supervised fine-tuning (SFT) and reinforcement learning (RL).

Supervised Fine-tuning. To learn an effective exploration policy, we first perform supervised fine-tuning on demonstration trajectories before reinforcement learning. This stage teaches the model the function-calling format, valid exploration actions, and how to reason over accumulated evidence. We use Qwen3-VL-235B [2] as the teacher model to generate trajectories. Instead of generating a full trajectory at once, the teacher interacts with the VLM-compatible 3D world turn by turn. At each turn, it predicts the next reasoning step and function call based on the current interaction history, receives the returned visual or linguistic evidence, and appends the new step to the trajectory. After the interaction ends, we concatenate the reasoning steps, function calls, returned evidence, and final answer into a complete trajectory. Each trajectory follows the format: `<think>` reasoning `</think>` `<action>` function call `</action>` $\cdots$ `<think>` reasoning `</think>` `<answer>` final answer `</answer>`. Here, each `<think>` segment contains intermediate reasoning, each `<action>` segment specifies a function call with its arguments, and the final `<answer>` segment gives the answer. We filter out invalid trajectories with malformed function calls, failed executions, or incorrect answers, and fine-tune the VLM on the filtered data using a standard next-token prediction objective.

Reinforcement Learning. After supervised fine-tuning, we optimize the exploration policy with reinforcement learning. We adopt Group Relative Policy Optimization (GRPO) [49] as the policy optimization algorithm. For reward design, we first introduce three trajectory-level rewards, including a format reward that enforces valid function calls and answer formats, an answer reward that evaluates whether the final prediction is correct, and an exploration reward that checks whether the trajectory contains valid exploration actions.

Since trajectory-level rewards provide only coarse supervision, we further introduce step-wise rewards to guide intermediate exploration decisions. For each training sample, we use Qwen3-VL-235B [2] to pre-generate question-conditioned preferred exploration actions. During training, the policy receives a step-wise matching reward when its selected action is consistent with the preferred action at each interaction step. We also apply a step-wise action penalty for failed function calls and uninformative actions, such as actions that lead to entirely empty new views from the visual interface. The overall reward for a trajectory τ is defined as

$$R(\tau) = \underbrace{r_{\text{fmt}} + r_{\text{ans}} + r_{\text{explore}}}_{\text{trajectory-level}} + \underbrace{\lambda r_{\text{match}} - \beta p_{\text{act}}}_{\text{step-wise}} \tag{7}$$

where r_{match} denotes the step-wise action matching reward, p_{act} denotes the step-wise action penalty, and λ and β are weighting coefficients.

4 Experiments

4.1 Experimental Settings

Benchmarks. Spatial mental modeling refers to internal representations of unseen space that support reasoning about layout, perspective, and motion [27, 28]. It is closely aligned with our goal of evaluating whether VLMs can perform human-like spatial reasoning from limited observations. We therefore evaluate our framework on two related benchmarks: (1) **MindCube** contains 21,154 questions over 3,268 images and focuses on inferring spatial layouts from limited observations, particularly through what-if questions that require mental simulation. Its training set contains 10K samples, and the commonly used MindCube-Tiny evaluation set includes 1,050 questions, comprising 600 Among, 200 Rotation, and 250 Around questions. Following prior work [5, 12, 38, 74], we primarily evaluate our method on MindCube-Tiny, and the overall accuracy is weighted by the number of questions in each category. (2) **SAT** similarly evaluates human-like spatial reasoning from sparse visual observations across diverse categories, including Egocentric Movement, Object Movement, Allocentric Perspective, Goal Aiming, and Action Consequence. It also includes questions that require mental simulation. Its training set contains 175K questions across 22K procedurally generated scenes, and the SAT-val evaluation set contains 4,000 questions. We evaluate our method on SAT-val, and the overall accuracy is averaged equally across the five categories, following the original SAT protocol [48].

Baselines. Our framework can be applied to existing VLMs in a training-free manner and also supports further optimization of the VLM exploration policy via RL. We therefore evaluate our framework under both training-free and training-based experimental settings. The baselines are grouped into four categories: (1) generalist open-source VLMs, including Qwen series: Qwen2.5-VL-7B-Instruct [3], Qwen3-VL-8B-Instruct [2], Qwen3-VL-235B-A22B-Instruct [2], and the latest Qwen3.5-9B [46]; (2) proprietary commercial models, e.g. GPT-4o [41] and GPT-4.1 series [43]; (3) training-free spatial reasoning methods, e.g. pySpatial [38]; and (4) specialized spatial reasoning VLMs, e.g. 3DThinker [12] and SpatialDreamer [5], representing training-based methods.

Implementation Details. We use off-the-shelf foundation models, including 3D foundation model Depth Anything 3 [35] and the segmentation model SAM 3 [7], to construct the VLM-compatible 3D world. During exploration, we set the maximum number of interaction turns to 4. For the training-free setting, all evaluations are conducted on a single H200 GPU. Since our framework is model-agnostic, we adapt V2W with different base models to align with the underlying VLMs used by different baselines. For the training-based setting, we train our method separately on the training set of each benchmark using 8 H200 GPUs. To ensure consistency with the main comparison methods [5, 12], we also adopt Qwen2.5-VL-7B-Instruct [3] as the base model. In the SFT stage, we fine-tune all model parameters for one epoch using a learning rate of 1.0×10^{-5}, a cosine learning-rate schedule with a 0.1 warmup ratio, and a batch size of 64. In the RL stage, we implement the exploration policy training with the verl [51] framework, using a learning rate of 1.0×10^{-6} and a batch size of 128.

4.2 Quantitative Results

We evaluate V2W under both training-free and training-based settings. First, in the training-free setting, we assess whether V2W can serve as a plug-and-play framework to directly enhance the spatial reasoning ability of existing VLMs without additional training. Second, in the training-based setting, we further investigate whether a learned exploration policy can lead to stronger spatial reasoning capabilities. Specifically, we evaluate training-free methods on MindCube-Tiny, and evaluate training-based methods on both MindCube-Tiny and SAT-val.

Training-free Evaluation As shown in Table 1, V2W substantially improves the overall performance of different base VLMs on MindCube-Tiny without additional training. V2W improves the overall accuracy of Qwen2.5-VL-7B, Qwen3-VL-8B, Qwen3-VL-235B, and GPT-4.1-mini by 9.3, 19.6, 17.3, and 18.4 percentage points, demonstrating its ability to enhance existing VLMs across different model families and scales. To verify that these gains do not merely arise from eliciting explicit textual reasoning, we further compare V2W with chain-of-thought (CoT) prompting, which encourages the model to perform free-form textual reasoning before producing the final answer. V2W also substantially outperforms CoT under the same base models, suggesting that the improvements

Table 1: Quantitative results on MindCube-Tiny under training-free and training-based settings.

Method	Setting	**Overall**	Among	Rotation	Around
Qwen2.5-VL-7B [3]	Training-free	34.5	32.7	35.0	38.4
Qwen2.5-VL-7B [3] w CoT		29.1	27.7	31.5	30.8
V2W (Qwen2.5-VL-7B)		43.8	43.7	35.0	51.2
Qwen3-VL-8B [2]		30.6	29.2	31.5	33.2
Qwen3-VL-8B [2] w CoT		38.9	33.7	40.0	50.4
V2W (Qwen3-VL-8B)		50.2	44.8	38.5	72.4
Qwen3-VL-235B [2]		42.3	38.8	42.5	50.4
Qwen3-VL-235B [2] w CoT		45.4	41.0	49.5	52.8
V2W (Qwen3-VL-235B)		59.6	60.2	49.5	66.4
Qwen3.5-9B [46]		33.7	26.5	33.5	51.2
Qwen3.5-9B [46] w CoT		45.9	38.7	49.0	60.8
V2W (Qwen3.5-9B)		55.4	47.7	65.0	66.4
GPT-4o [41]		42.3	43.0	35.0	46.4
GPT-4.1 [43]		44.7	45.3	35.5	50.4
GPT-4.1-mini [43]		43.3	45.0	36.0	44.8
ViperGPT [54]		37.0	41.0	20.5	40.4
VADAR [39]		40.8	40.7	33.5	46.8
pySpatial (GPT-4.1-mini) [38]		58.2	62.0	37.5	65.6
V2W (GPT-4.1-mini)		**61.7**	61.3	47.0	74.4
SpaceMantis [8]	Training-based	26.3	20.5	35.0	33.2
Spatial-MLLM [65]		33.4	30.5	39.0	36.0
Space-Qwen [8]		36.9	36.3	32.5	41.6
RoboBrain [26]		44.0	45.2	36.0	47.6
3DThinker (Qwen2.5-VL-7B) [12]		76.0	83.0	55.0	76.0
SpatialDreamer (Qwen2.5-VL-7B) [5]†		84.9	78.2	87.5	93.5
V2W (Qwen2.5-VL-7B)		**90.8**	91.3	87.5	92.0

† Results are taken from the corresponding original paper.

are primarily driven by active exploration over explicit 3D world representations. In addition, V2W outperforms other training-free spatial reasoning methods, such as pySpatial, a visual-programming-based approach. Compared with visual programming pipelines, V2W enables multi-turn active exploration via lightweight function calls, thereby avoiding potential failures arising from code generation and execution.

Training-based Evaluation Although training-free V2W already improves the overall performance of existing VLMs, its gains remain limited on more challenging question types such as Rotation, which often involve complex what-if hypotheses and require multi-step exploration to identify critical spatial evidence. This suggests that existing VLMs still lack effective exploration strategies for complex spatial reasoning. To address this limitation, we further optimize the exploration policy via reinforcement learning, enabling VLMs to better conduct multi-step exploration and aggregate spatial evidence. As shown in Table 1, training-based V2W achieves 90.8% overall accuracy on MindCube-Tiny, outperforming existing specialized spatial reasoning models. Table 2 further shows that V2W also surpasses competing methods on SAT-val, demonstrating that RL-based exploration policy training effectively enhances the active exploration capability of VLMs.

4.3 Qualitative Analysis

Fig. 3 presents the spatial reasoning results of training-based V2W on MindCube-Tiny. The first case illustrates how V2W actively explores the 3D world and collects spatial evidence through both visual and linguistic interfaces over multiple interaction turns. The second case demonstrates that, through our linguistic interface, V2W can actively acquire 3D information that is not explicitly available in the original 2D images, thereby facilitating more effective spatial reasoning. More qualitative results and analyses are provided in the appendix.

Table 2: Quantitative results on SAT-val under the training-based setting.

Method	**Overall**	EgoM	ObjM	EgoAct	GoalAim	Pers
GPT-4V [40]	44.8	54.7	32.7	52.0	50.5	34.2
GPT-4o [41]	63.1	64.7	86.8	51.9	68.7	43.4
GPT-4.1 [43]	68.4	75.3	89.0	57.8	78.3	41.5
o1 [42]	74.0	78.0	85.9	65.4	86.0	54.6
RoboPoint-13B [76]	53.3	50.2	69.4	48.8	72.6	25.5
LLaVA-1.5-13B [36]	51.1	46.6	73.8	49.7	45.6	39.9
LLaVA-Video-7B [82]	57.4	56.4	82.7	48.0	52.9	47.1
SAT (LLaVA-1.5-13B) [48]	87.7	61.7	90.2	91.4	96.8	98.5
SAT (LLaVA-Video-7B) [48]	78.0	79.6	80.4	85.3	56.4	88.4
SpatialDreamer (Qwen2.5-VL-7B) [5]	93.0	95.4	93.1	91.8	94.9	89.8
V2W (Qwen2.5-VL-7B)	**95.5**	99.8	90.3	90.5	97.8	99.2

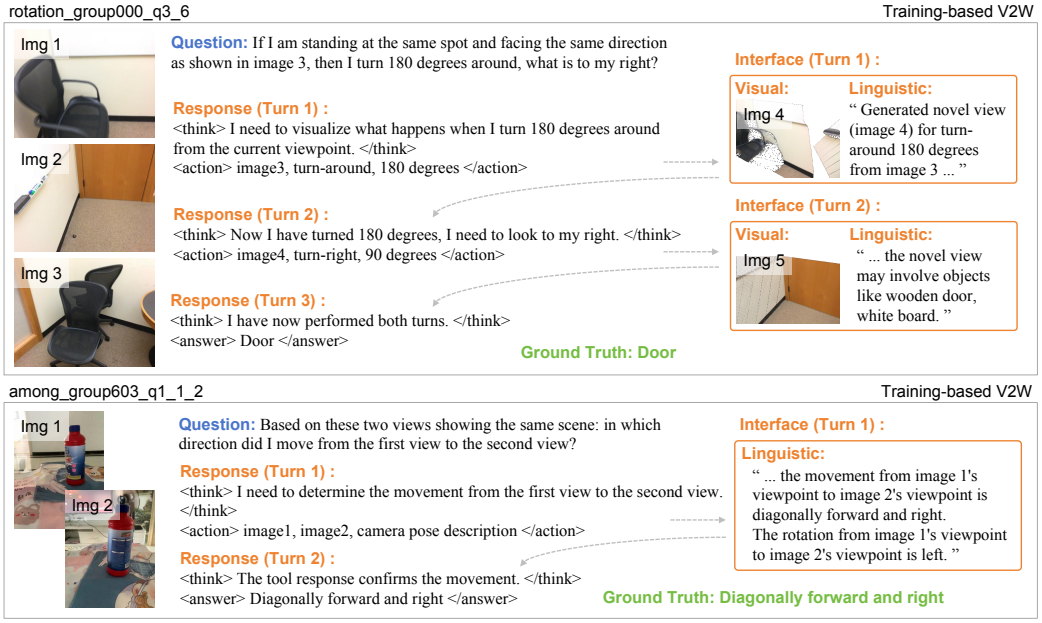

Figure 3: Qualitative results on MindCube-Tiny.

4.4 Ablation Study

In Table 3, we conduct ablation studies on MindCube-Tiny with Qwen2.5-VL-7B [3] as the base model (row 0). The results identify two factors underlying the performance gains. First, the world interface improves performance across different training stages. In the training-free setting, introducing both visual and linguistic interfaces improves the overall accuracy from 34.5% to 43.8% (rows 0-3). In the training-based setting, the full interface consistently improves overall accuracy across different training configurations, yielding gains of 6.1 points in the SFT-only configuration (rows 4-5), 14.4 points in the RL-only configuration (rows 6-8), and 16.0 points in the two-stage SFT-then-RL configuration (rows 9-12). These results highlight the importance of active world-centric spatial reasoning enabled by the world interface.

Second, RL more effectively elicits active exploration than SFT. Without the world interface, RL-only achieves 69.5% overall accuracy, outperforming SFT-only at 65.2% (rows 6 vs. 4). With the full interface, RL-only further reaches 83.9%, substantially higher than SFT-only at 71.3% (rows 8 vs. 5). Nevertheless, the best performance is achieved by SFT-then-RL with the full world interface, reaching 90.8% overall accuracy (row 12). This suggests that SFT provides a useful imitation-learning initialization, while RL further optimizes the exploration policy, allowing the model to better exploit the world interface for active world-centric spatial reasoning. We further ablate the step-wise reward (SR) in RL (row 11), demonstrating the effectiveness of fine-grained reward signals.

Table 3: Ablation studies on MindCube-Tiny.

#	Setting	Training Stage		World Interface		MindCube-Tiny			
		SFT	RL	Visual	Linguistic	Overall	Among	Rotation	Around
0	Training-free					34.5	32.7	35.0	38.4
1				✓		38.5	36.7	34.0	46.4
2					✓	41.4	40.2	35.0	49.6
3				✓	✓	43.8	43.7	35.0	51.2
4	Training-based	✓				65.2	66.8	57.0	68.0
5		✓		✓	✓	71.3	77.8	43.0	78.4
6			✓			69.5	73.7	58.5	68.4
7			✓	✓		78.8	81.7	75.0	74.8
8			✓	✓	✓	83.9	84.5	74.0	90.4
9		✓	✓			74.8	76.8	69.0	74.4
10		✓	✓	✓		83.9	85.2	74.5	88.4
11		✓	w/o SR	✓	✓	89.0	90.0	83.5	91.2
12		✓	✓	✓	✓	90.8	91.3	87.5	92.0

More ablation studies across different base models and different 3D foundation models used for 3D world construction, as well as efficiency analyses, are provided in the appendix.

5 Conclusion

In this work, we propose View-to-World (V2W), an active exploration framework that enables VLMs to explore explicit 3D worlds for spatial reasoning. Unlike existing VLMs that passively perceive fixed-view inputs in a view-centric manner, V2W allows models to actively acquire and reason over 3D information from the underlying world, a paradigm that we refer to as active world-centric spatial reasoning. V2W constructs a VLM-compatible 3D world from sparse input views and exposes its geometry, camera poses, and language-grounded semantics through visual and linguistic interfaces, enabling VLMs to progressively gather world-grounded evidence through multi-turn interaction. Experiments show that V2W improves VLMs in a training-free setting, validating the effectiveness of active exploration over explicit 3D worlds. By further optimizing the exploration policy with reinforcement learning, V2W learns more effective exploration strategies and achieves state-of-the-art performance on spatial mental modeling benchmarks. These results highlight active world-centric spatial reasoning as a promising direction for advancing the spatial intelligence of VLMs.

Acknowledgments and Disclosure of Funding

This work was supported by the Shanghai Committee of Science and Technology (Grant Nos. 25511103300, 25511104302, 25511102700), the National Natural Science Foundation of China (Grant Nos. 62302167, U23A20343, W2521174, 72192821, 62472282), the Young Elite Scientists Sponsorship Program by CAST YESS20240780, and the Fundamental Research Funds for the Central Universities (project number: YG2023QNA35).

References

[1] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19129–19139, 2022.

[2] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] D. Bucciarelli, N. Moratelli, M. Cornia, L. Baraldi, and R. Cucchiara. Personalizing multimodal large language models for image captioning: an experimental analysis. In *European Conference on Computer Vision (ECCV) Workshops*, pages 351–368, 2024.

[5] M. Cao, X. Li, X. Liu, I. Reid, and X. Liang. Spatialdreamer: Incentivizing spatial reasoning via active mental imagery. *arXiv preprint arXiv:2512.07733*, 2025.

[6] M. Cao, H. Lin, H. Li, H. Tang, R. Xu, D. An, X. Liu, I. Reid, and X. Liang. Seeing through imagination: Learning scene geometry via implicit spatial world modeling. *arXiv preprint arXiv:2512.01821*, 2025.

[7] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

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

[9] D. Z. Chen, A. X. Chang, and M. Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision (ECCV)*, pages 202–221, 2020.

[10] Z. Chen, A. Gholami, M. Nießner, and A. X. Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2021.

[11] Z. Chen, X. Lu, Z. Zheng, P. Li, L. He, Y. Zhou, J. Shao, B. Zhuang, and L. Sheng. Geometrically-constrained agent for spatial reasoning. *arXiv preprint arXiv:2511.22659*, 2025.

[12] Z. Chen, M. Zhang, X. Yu, X. Luo, M. Sun, Z. Pan, Y. Feng, P. Pei, X. Cai, and R. Huang. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*, 2025.

[13] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:135062–135093, 2024.

[14] K. Cheng, W. Song, J. Fan, Z. Ma, Q. Sun, F. Xu, C. Yan, N. Chen, J. Zhang, and J. Chen. Caparena: Benchmarking and analyzing detailed image captioning in the llm era. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14077–14094, 2025.

[15] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:49250–49267, 2023.

[16] W. Deng, X. Zhang, and M. Qi. Active exploring like a pigeon: Reinforcing spatial reasoning via agentic vision-language models. In *International Conference on Machine Learning (ICML)*, pages 24407–24417, 2026.

[17] Z. Fan, J. Zhang, R. Li, J. Zhang, R. Chen, H. Hu, K. Wang, H. Qu, D. Wang, Z. Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

[18] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6): 381–395, 1981.

[19] X. Gao, Z. Zhang, D. Z. Chen, S. Xu, L. Quan, E. Pérez-Pellitero, and Y. Jang. Map2thought: Explicit 3d spatial reasoning via metric cognitive maps. *arXiv preprint arXiv:2601.11442*, 2026.

[20] Y. Guo, Y. Hou, W. Ma, M. Tang, and M.-H. Yang. Pursuing minimal sufficiency in spatial reasoning. *arXiv preprint arXiv:2510.16688*, 2025.

[21] T. Gupta and A. Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14953–14962, 2023.

[22] D. Harris. Spatial reasoning in context: bridging cognitive and educational perspectives of spatial-mathematics relations. In *Frontiers in Education*, volume 8, page 1302099. Frontiers Media SA, 2023.

[23] J. Huang, B. Jia, Y. Wang, Z. Zhu, X. Linghu, Q. Li, S.-C. Zhu, and S. Huang. Unveiling the mist over 3d vision-language understanding: Object-centric evaluation with chain-of-analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24570–24581, 2025.

[24] J. Huang, X. Ma, X. Linghu, J. He, Q. Li, S.-C. Zhu, Y. Chen, B. Jia, and S. Huang. Leo-vl: Efficient scene representation for scalable 3d vision-language learning. *arXiv preprint arXiv:2506.09935*, 2025.

[25] X. Huang, J. Wu, Q. Xie, and K. Han. 3drs: Mllms need 3d-aware representation supervision for scene understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 38: 67961–67988, 2025.

[26] Y. Ji, H. Tan, J. Shi, X. Hao, Y. Zhang, H. Zhang, P. Wang, M. Zhao, Y. Mu, P. An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1724–1734, 2025.

[27] P. N. Johnson-Laird. Mental models in cognitive science. *Cognitive science*, 4(1):71–115, 1980.

[28] P. N. Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, 1983.

[29] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.

[30] V. Leroy, Y. Cabon, and J. Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision (ECCV)*, pages 71–91, 2024.

[31] B. Li, X. Sun, J. Liu, Z. Wang, J. Wu, X. Yu, H. Chen, E. Barsoum, M. Chen, and Z. Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025.

[32] H. Li, Z. Zou, F. Liu, X. Zhang, F. Hong, Y. Cao, Y. Lan, M. Zhang, G. Yu, D. Zhang, et al. Iggt: Instance-grounded geometry transformer for semantic 3d reconstruction. *arXiv preprint arXiv:2510.22706*, 2025.

[33] H. Li, Q. Cao, T. Tang, K. Xiang, Z. Guo, J. Han, H. Xu, and X. Liang. Thinking with geometry: Active geometry integration for spatial reasoning. *arXiv preprint arXiv:2602.06037*, 2026.

[34] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pages 19730–19742. PMLR, 2023.

[35] H. Lin, S. Chen, J. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[36] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024.

[37] P. Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S.-C. Zhu, and J. Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:43447–43478, 2023.

[38] Z. Luo, C. Zhang, S. Yong, C. Dai, Q. Wang, H. Ran, G. Shi, K. P. Sycara, and Y. Xie. pyspatial: Generating 3d visual programs for zero-shot spatial reasoning. In *International Conference on Learning Representations (ICLR)*, 2026.

[39] D. Marsili, R. Agrawal, Y. Yue, and G. Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 19446–19455, 2025.

[40] OpenAI. GPT-4V(ision) System Card, Sept. 2023. URL `https://openai.com/index/gpt-4v-system-card/`.

[41] OpenAI. GPT-4o System Card, Aug. 2024. URL `https://openai.com/index/gpt-4o-system-card/`.

[42] OpenAI. OpenAI o1 System Card, Dec. 2024. URL `https://openai.com/index/openai-o1-system-card/`.

[43] OpenAI. Introducing GPT-4.1 in the API, Apr. 2025. URL `https://openai.com/index/gpt-4-1/`.

[44] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[45] Z. Qi, Z. Zhang, Y. Fang, J. Wang, and H. Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025.

[46] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

[47] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

[48] A. Ray, J. Duan, E. L. Brown II, R. Tan, D. Bashkirova, R. Hendrix, K. Ehsani, A. Kembhavi, B. A. Plummer, R. Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. In *Conference on Language Modeling (CoLM)*, 2025.

[49] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[50] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:38154–38180, 2023.

[51] G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

[52] Z. Shi, K. Mei, Y. Quan, D. N. Metaxas, and R. Tang. Improving visual reasoning with iterative evidence refinement. *arXiv preprint arXiv:2603.14117*, 2026.

[53] Z. Su, L. Li, M. Song, Y. Hao, Z. Yang, J. Zhang, G. Chen, J. Gu, J. Li, X. Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025.

[54] D. Surís, S. Menon, and C. Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11888–11898, 2023.

[55] X. Tan, B. Liu, Y. Bao, Q. Tian, Z. Gao, X. Wu, Z. Luo, S. Wang, Y. Zhang, X. Wang, et al. Towards safe and trustworthy embodied ai: foundations, status, and prospects. *Shanghai AI Lab., Shanghai, China, Tech. Rep.*, 2025.

[56] Q. Tian, X. Tan, J. Gong, Y. Xie, and L. Ma. Uniforward: Unified 3d scene and semantic field reconstruction via feed-forward gaussian splatting from only sparse-view images. *arXiv preprint arXiv:2506.09378*, 2025.

[57] Q. Tian, X. Tan, Y. Xie, and L. Ma. Drivingforward: Feed-forward 3d gaussian splatting for driving scene reconstruction from flexible surround-view input. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 7374–7382, 2025.

[58] Q. Tian, Z. Wu, X. Wang, L. Ma, and X. Tan. Groupforward: Building referable 3d scenes via instance-grouped feed-forward gaussian splatting. *arXiv preprint arXiv:2608.17535*, 2026.

[59] Q. Tian, J. Ying, X. Wang, Y. Xie, L. Ma, and X. Tan. FLEG: feed-forward language embedded gaussian splatting from any views via compact semantic representation. In *European Conference on Computer Vision (ECCV)*, volume 17016, pages 318–335, 2026.

[60] H. Wang, Y. Zhao, T. Wang, H. Fan, X. Zhang, and Z. Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9275–9286, 2025.

[61] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[62] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024.

[63] Y. Wang, J. Zhou, H. Zhu, W. Chang, Y. Zhou, Z. Li, J. Chen, J. Pang, C. Shen, and T. He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025.

[64] Z. Wang, C. Chen, F. Luo, Y. Dong, Y. Zhang, Y. Xu, X. Wang, P. Li, and Y. Liu. Actiview: Evaluating active perception ability for multimodal large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7605–7633, 2025.

[65] D. Wu, F. Liu, Y.-H. Hung, and Y. Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[66] J. Wu, J. Guan, K. Feng, Q. Liu, S. Wu, L. Wang, W. Wu, and T. Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[67] M. Wu, J. Yang, J. Jiang, M. Li, K. Yan, H. Yu, M. Zhang, C. Zhai, and K. Nahrstedt. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. *arXiv preprint arXiv:2505.19255*, 2025.

[68] P. Xu, S. Wang, Y. Zhu, J. Li, and Y. Zhang. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025.

[69] J. Yang, S. Yang, A. W. Gupta, R. Han, L. Fei-Fei, and S. Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 10632–10643, 2025.

[70] Y. Yang, J. Liu, Z. Zhang, S. Zhou, R. Tan, J. Yang, Y. Du, and C. Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[71] Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

[72] Z. Yang, X. Yu, D. Chen, M. Shen, and C. Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.

[73] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[74] B. Yin, Q. Wang, P. Zhang, J. Zhang, K. Wang, Z. Wang, J. Zhang, K. Chandrasegaran, H. Liu, R. Krishna, et al. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025.

[75] J. Ying, Q. Tian, R. Xu, X. Zhu, D. Dong, J. Xu, and X. Tan. Exemplar2vqa: A scalable exemplar-driven visual question answering generation framework via multi-agent coding. *arXiv preprint arXiv:2609.37655*, 2026.

[76] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox. Robopoint: A vision-language model for spatial affordance prediction in robotics. In *Conference on Robot Learning (CoRL)*, pages 4005–4020. PMLR, 2025.

[77] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, 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 (CVPR)*, pages 9556–9567, 2024.

[78] J. Zhang, T. Hu, X. Huang, Y. Gong, and D. Zeng. 3dbench: a scalable 3d benchmark and instruction-tuning dataset. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1706–1714, 2024.

[79] J. Zhang, Y. Chen, Y. Xu, Z. Huang, J. Mei, J. Chen, Y. Zhou, Y.-J. Yuan, X. Cai, G. Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2025.

[80] J. Zhang, J. Jiang, H. Li, Y. Chen, K. Jiang, and D. Z. Chen. Gap-mllm: Geometry-aligned pre-training for activating 3d spatial perception in multimodal large language models. *arXiv preprint arXiv:2603.16461*, 2026.

[81] Y. Zhang, Z. Gong, and A. X. Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15225–15236, 2023.

[82] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[83] Z. Zhang, Y. Wu, L. Jia, Y. Wang, Z. Zhang, Y. Li, B. Ran, F. Zhang, Z. Sun, Z. Yin, et al. Think3d: Thinking with space for spatial reasoning. *arXiv preprint arXiv:2601.13029*, 2026.

[84] D. Zheng, S. Huang, Y. Li, and L. Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[85] D. Zheng, S. Huang, and L. Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8995–9006, 2025.

[86] Z. Zheng, M. Yang, J. Hong, C. Zhao, G. Xu, L. Yang, C. Shen, and X. Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

[87] C. Zhou, R. Wang, F. Luo, M. D. Pesé, Z. Fan, Y. Zhong, and S. Huang. Ff3r: Feedforward feature 3d reconstruction from unconstrained views. *arXiv preprint arXiv:2604.09862*, 2026.

[88] C. Zhu, T. Wang, W. Zhang, J. Pang, and X. Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4295–4305, 2025.

[89] M. Zhu, H. Zhong, C. Zhao, Z. Du, Z. Huang, M. Liu, H. Chen, C. Zou, J. Chen, M. Yang, et al. Active-o3: Empowering multimodal large language models with active perception via grpo. *arXiv preprint arXiv:2505.21457*, 2025.

A Appendix

A.1 Ablation Study Across Different Base Models

Table 4: Ablation studies across different base models on MindCube-Tiny.

#	Base Model	World Interface		MindCube-Tiny			
		Visual	Linguistic	Overall	Among	Rotation	Around
0	Qwen2.5-VL-7B [3]			34.5	32.7	35.0	38.4
1		✓		38.5	36.7	34.0	46.4
2			✓	41.4	40.2	35.0	49.6
3		✓	✓	43.8	43.7	35.0	51.2
4	Qwen3-VL-8B [2]			30.6	29.2	31.5	33.2
5		✓		39.0	41.8	34.0	36.0
6			✓	46.8	44.0	36.5	61.6
7		✓	✓	50.2	44.8	38.5	72.4
8	Qwen3-VL-235B [2]			42.3	38.8	42.5	50.4
9		✓		51.0	56.5	41.0	45.6
10			✓	54.8	54.0	42.0	66.8
11		✓	✓	59.6	60.2	49.5	66.4

To verify the effectiveness of our method across different model variants and model scales, we conduct ablation studies on MindCube-Tiny [74] under the training-free setting using Qwen2.5-VL-7B-Instruct [3], Qwen3-VL-8B-Instruct [2], and Qwen3-VL-235B-A22B-Instruct [2] as base models, respectively. As shown in Table 4, both the visual and linguistic world interfaces consistently improve the overall performance across all base models. Specifically, the full V2W improves the overall accuracy from 34.5% to 43.8% on Qwen2.5-VL-7B-Instruct (row 0 vs. 3), from 30.6% to 50.2% on Qwen3-VL-8B-Instruct (row 4 vs. 7), and from 42.3% to 59.6% on Qwen3-VL-235B-A22B-Instruct (row 8 vs. 11). Notably, the Qwen3-VL series exhibits larger improvements than Qwen2.5-VL-7B, which may be attributed to its stronger tool-use capability that enables more effective utilization of the proposed interfaces for active exploration over 3D worlds. The overall performance generally improves with model scale, with Qwen3-VL-235B-A22B-Instruct achieving the best result of 59.6% when both interfaces are enabled (row 11). In addition, the visual and linguistic interfaces exhibit consistent complementarity across all base models, with the full configuration outperforming either individual interface for all base models (row 3 vs. rows 1–2, row 7 vs. rows 5–6, and row 11 vs. rows 9–10). This suggests that the visual and textual modalities provide complementary signals during active exploration, enabling the model to capture richer 3D spatial information and thereby achieve better overall performance.

A.2 Ablation Study Across Different 3D Foundation Models

Table 5: Ablation studies across different 3D foundation models on MindCube-Tiny.

Method	**Overall**	Among	Rotation	Around
V2W(Qwen2.5-VL-7B) w Depth Anything 3 [35]	43.8	43.7	35.0	51.2
V2W(Qwen2.5-VL-7B) w VGGT [61]	42.6	42.2	35.0	49.6
V2W(Qwen3-VL-8B) w Depth Anything 3 [35]	50.2	44.8	38.5	72.4
V2W(Qwen3-VL-8B) w VGGT [61]	50.0	45.8	35.5	71.6

Our V2W leverages an off-the-shelf 3D foundation model during the 3D world construction process, with Depth Anything 3 used by default. To evaluate the robustness of V2W to different 3D foundation models and verify that its effectiveness does not depend on a specific 3D foundation model, we conduct an ablation study on MindCube-Tiny under the training-free setting by adopting VGGT, another commonly used 3D foundation model. We evaluate V2W with both Qwen2.5-VL-7B-Instruct [3] and Qwen3-VL-8B-Instruct [2] as base models, covering VLMs with different exploration capabilities and interaction behaviors during inference. As shown in Table 5, adopting VGGT leads to only a slight overall performance drop compared with the default Depth Anything 3 across both base models. Despite this, V2W with VGGT still substantially outperforms the corresponding base

models reported in Table 1. These results demonstrate that V2W exhibits robustness across different 3D foundation models, suggesting that the effectiveness of our framework does not depend on any particular 3D foundation model.

A.3 Efficiency Analysis

Table 6: Runtime efficiency analysis of the proposed pipeline on MindCube-Tiny.

	3D World Construction	Interface Response		Model Response	Total Time
		Visual	Linguistic		
Avg. Time	4.64s	0.33s	0.02s	1.94s	6.93s / sample

We further analyze the inference efficiency of the training-based V2W (Qwen2.5-VL-7B) on MindCube-Tiny. As shown in Table 6, the overall average inference time is approximately 6.93 seconds per sample. The main computational cost comes from 3D world construction, which takes about 5 seconds, while the interface responses introduce only negligible overhead. The model response itself takes less than 2 seconds, indicating that V2W does not rely on lengthy free-form reasoning during inference. Considering that V2W substantially outperforms models of comparable scale in spatial reasoning performance, such a runtime is acceptable. Moreover, free-form reasoning in larger models may introduce substantial inference overhead. For instance, Qwen3-VL-235B [2] with CoT in Table 1 can require a thinking time comparable to, or even longer than, the total inference time of our entire pipeline, while still achieving lower performance. Rather than relying on lengthy free-form reasoning, V2W learns a more effective active exploration strategy through reinforcement learning, reducing redundant reasoning steps and unnecessary deliberation. Notably, we do not introduce any reward related to reasoning length, since we empirically find that longer reasoning does not necessarily improve the learned exploration strategy or the final answer accuracy. Instead of relying on long reasoning traces, V2W improves performance through more efficient exploration over the constructed 3D world, thereby achieving stronger spatial reasoning ability with acceptable inference efficiency.

A.4 Additional Qualitative Results

We provide additional qualitative results to further demonstrate the flexibility and effectiveness of V2W. As shown in Fig. 4, training-free V2W can be readily adapted to different base models, such as Qwen2.5-VL-7B [3], Qwen3-VL-8B [2], and GPT-4.1-mini [43], enabling them to actively explore 3D worlds without additional training. However, these models often require relatively long reasoning chains before or after exploration to determine the next action or integrate information obtained through the interface. In contrast, Fig. 5 shows that training-based V2W performs efficient multi-round exploration in the constructed 3D worlds with minimal redundant reasoning. When the query is explicit, it can accurately identify informative exploration actions without unnecessary steps. We attribute this efficient exploration strategy to reinforcement learning, which encourages the model to learn a more concise and goal-oriented exploration policy.

A.5 Prompt for Active Exploration via Function Calling

We provide an example prompt for implementing both training-free and training-based V2W in Fig. 6. Following a tool-usage-like paradigm, V2W enables VLMs to perform multi-turn exploration via function calling.

A.6 Limitations and Future Work

Our framework leverages off-the-shelf foundation models during 3D world construction. Although current 3D foundation models and segmentation models show strong generalization across diverse scenes, they may still fail under challenging conditions, such as low lighting or reflective objects, leading to inaccurate reconstruction or segmentation and potentially misleading subsequent reasoning. One possible direction is to introduce an additional VLM-based verifier to assess the correctness of the constructed 3D world, but this would increase the overall complexity of the framework, which we leave for future work. In addition, due to the available computational budget, we train V2W on a



Figure 4: Additional qualitative results on MindCube-Tiny of training-free V2W with different base models.

focused set of datasets to validate the effectiveness of reinforcement learning. Future work may scale training to broader spatial reasoning data and further study the scaling behavior of active exploration.

A.7 Broader Impacts

This work aims to improve the spatial reasoning capability of vision-language models by enabling more active interaction with 3D environments. We expect the proposed framework to benefit applications such as embodied AI, robotics, and assistive spatial understanding, while its use should follow standard privacy and safety considerations when applied to real-world visual data [55].

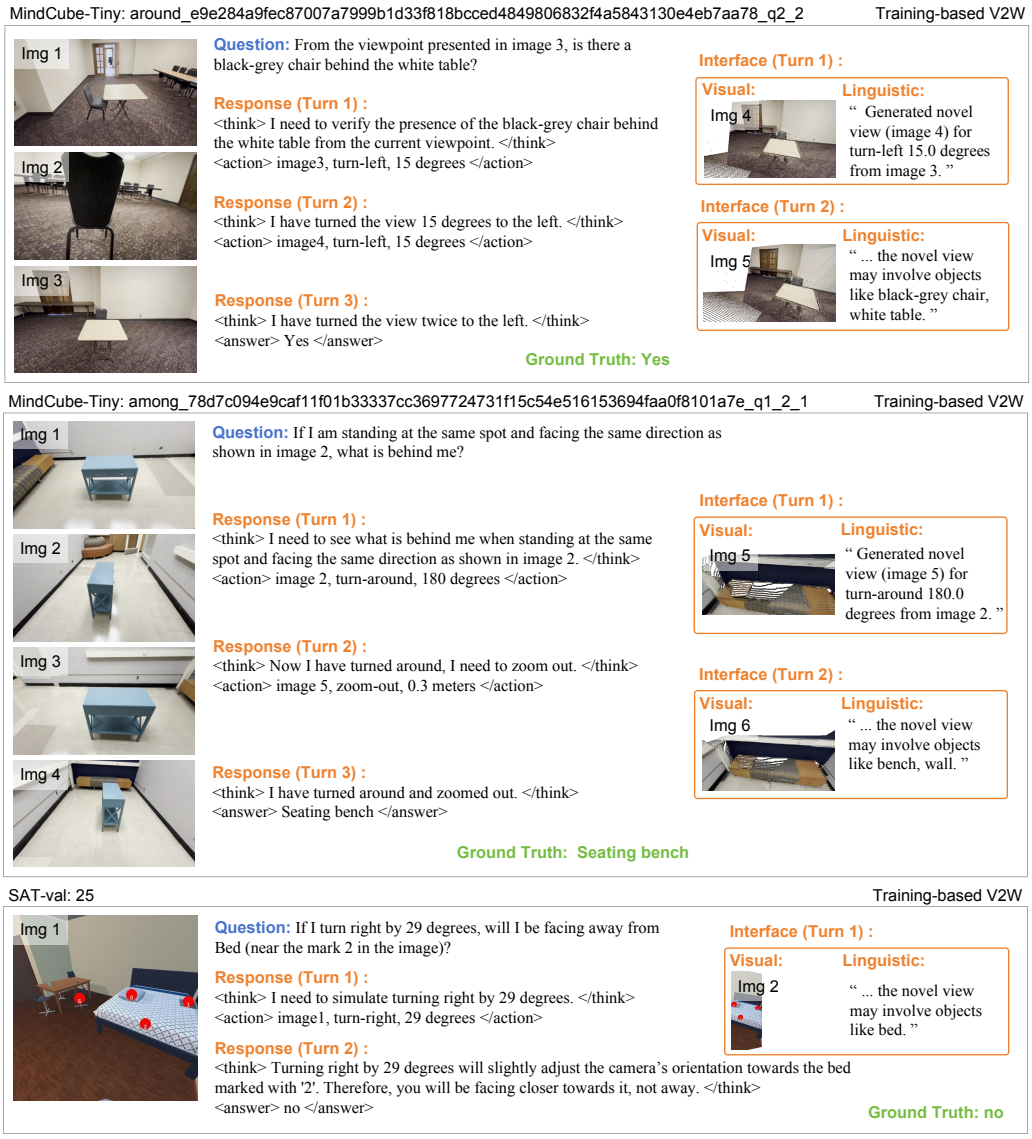

Figure 5: Additional qualitative results on MindCube-Tiny and SAT-val of training-based V2W.

A.7.1 The Use of Large Language Models (LLMs)

In this paper, LLMs were used only as an auxiliary tool for language polishing, including improvements to grammar, clarity, and readability. They were not involved in ideation, data analysis, experimental design, or interpretation of the results.

System Prompt:

```
"You are a spatial reasoning assistant."

## Tools: ...
```

Instruction Following:

```
"## Task\n"
"Reason about the spatial relationship of the 3D scene and answer the
question based on the images.\n\n"
"## Output Format\n"
"On each turn, first output your reasoning inside a <think>...</think>
block, then output exactly one of the following:\n"
"- <tool_call>...</tool_call> — when you need more information from a
tool.\n"
"- <answer>...</answer> — when you have enough information to answer the
question (output option letter followed by the full option text, e.g.
<answer>A. floor lamp</answer>).\n\n"
"## Constraints\n"
"- Never output both <tool_call>...</tool_call> and <answer>...</answer>
in the same turn.\n"
"- Use at most 3 tool calls per question.\n"
"- Output <answer>...</answer> exactly once, as your last output.\n\n"
```

Figure 6: Example prompt of training-free and training-based V2W.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the contribution of studying the active exploration of VLMs and proposing a novel framework, as well as the scope of this paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in the appendix, as well as the analysis of failure cases.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation details in the Section of Experiments provide the details to reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: This paper uses public data, and the code will be released upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the training and test details in the Section of Experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our work proposes a new framework for enhancing the active exploration capability of existing VLMs, which in turn improves their spatial reasoning ability. We validate the effectiveness of the framework on public spatial reasoning benchmarks and report results across different base models and datasets. However, due to the substantial experimental cost of VLM training and evaluation, and following common practice in recent VLM research, we do not report multi-seed results, error bars, or formal statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the GPU resources needed in the Implementation Details of Experiments Section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper complies fully with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts of this work in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All these assets used in this paper are correctly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our method is based on vision-language models (VLMs), and the experiments use existing VLMs as base models for both training-free evaluation and post-training. In addition, we use Qwen3-VL-235B for training data processing. These uses are part of the research methodology and experimental pipeline, and are described in the paper. We also used LLMs for translation, grammar checking, and writing assistance only; such usage did not affect the core methodology, scientific rigor, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.