

Think with 3D: Geometric Imagination Grounded Spatial Reasoning from Limited Views

Zhangquan Chen^{1*} Manyuan Zhang^{2†} Xinlei Yu³ Xufang Luo Mingze Sun¹ Zihao Pan²
Yan Feng² Peng Pei² Xunliang Cai² Ruqi Huang^{1‡}

1. Tsinghua Shenzhen International Graduate School, Tsinghua University
2. Meituan 3. National University of Singapore

Abstract

Though recent advances in vision–language models (VLMs) have achieved remarkable progress across a wide range of multimodal tasks, understanding 3D spatial relationships from limited views remains a significant challenge. Previous reasoning methods typically rely on pure text (e.g., topological cognitive maps) or on 2D visual cues. However, their limited representational capacity hinders performance in specific tasks that require 3D spatial imagination. To address this limitation, we propose 3DThinker, a framework that can effectively exploit the rich geometric information embedded within images while reasoning, like humans do. Our framework is the first to enable 3D mentalizing during reasoning without any 3D prior input, and it does not rely on explicitly labeled 3D data for training. Specifically, our training consists of two stages. First, we perform supervised training to align the 3D latent generated by VLM while reasoning with that of a 3D foundation model (e.g., VGGT). Then, we optimize the entire reasoning trajectory solely based on outcome signals, thereby refining the underlying 3D mentalizing. Extensive experiments across multiple benchmarks show that 3DThinker consistently outperforms strong baselines and offers a new perspective toward unifying 3D representations into multimodal reasoning. Our code will be available at <https://github.com/zhangquanchen/3DThinker>.

1. Introduction

Spatial understanding is a critical capability for machines to interact with the real 3D world (e.g., embodied AI, autonomous driving) [61, 69, 73, 83]. These systems typically rely on ego-centric, multi-view observations, typically pro-

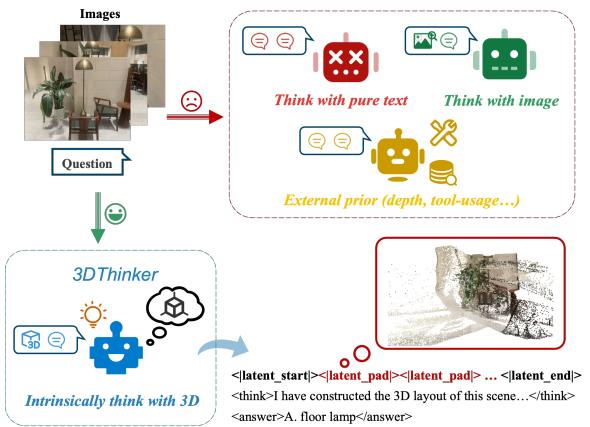


Figure 1. Illustration of our 3DThinker. Existing methods typically perform reasoning based solely on pure text or 2D visual cues, without fully exploiting the rich spatial and geometric information inherent in images. Other methods attempt to enhance the input by introducing auxiliary modalities (e.g., depth maps or coordinates), yet these often depend on additional annotations or external tools. In contrast, our framework enables VLMs to intrinsically form 3D mental representations during reasoning, thereby improving their spatial understanding.

vided by multiple cameras simultaneously capturing limited views of their surroundings. These views are not interchangeable or purely visual; they inherently carry spatial semantics tied to the machine’s frame of reference [24]. Consequently, imagining the full scene and performing reasoning based on a few limited views presents an essential problem for spatial intelligence [79]. Although recent VLMs are pretrained on large-scale image–text corpora, their performance on such spatial reasoning tasks remains notably limited [8, 15, 36, 74]. The core bottleneck lies in their *inability to extract 3D geometry embedded within images* and their *restricted capacity for spatial imagination*.

Recent advances have attempted to enhance the spatial reasoning capabilities of VLMs [13, 24, 34, 39, 40, 45, 68]. As illustrated in Fig. 1, existing methods can be broadly

*The work was conducted during the internship of Zhangquan Chen (czq23@mails.tsinghua.edu.cn) at Meituan.

†Project leader

‡Corresponding author: ruqihuang@sz.tsinghua.edu.cn

divided into two categories. The first category performs reasoning with pure text [7, 14, 39, 40, 45] or 2D visual cues [13, 17, 71], whose representational capacity for complex spatial layouts is inherently limited. To mitigate this limitation, methods such as MindCube [80] train models to generate cognitive maps of 3D layouts; however, they rely on bird’s-eye-vie (BEV) annotations to construct these maps. Ego3D [24] further employs external models—GroundingDINO [38] for referring expression comprehension (REC) and DepthAnythingv2 [76] for depth estimation, to automatically generate cognitive maps. Yet, constrained by the performance of these models, such methods often fail on low-resolution or uncurated images. The second category incorporates auxiliary modalities as additional inputs (e.g., point clouds, camera parameters) [9, 32]. However, these settings restrict the model’s applicability in real-world scenarios where only monocular images are available. Moreover, several recent methods invoke external encoder or tool-usage to obtain prior information (e.g., encoded 3D tokens [21], depth maps [5, 13, 40]). Importantly, these techniques do not constitute an intrinsic capability of the model and introduce additional inference overhead.

These challenges motivate the need for a new method that: G1) **3D-imaginable**: can directly learn 3D geometry from limited 2D images; G2) **Annotation-free**: does not rely on densely annotated data; and G3) **Intrinsic**: requires no external priors or auxiliary models during inference.

The most relevant mental model, Mirage [79], leverages ground-truth image embeddings for supervised training, facilitating the continuation of a multimodal trajectory without the need for pixel-level image generation. However, the training of [79] is heavily reliant on ground-truth image supervision and remains constrained to the “thinking with image” paradigm, which prevents its effectiveness on (G1) and (G2). Nevertheless, it provides a valuable inspiration, prompting us to *introduce a new novel framework, 3DThinker, which enables thinking with 3D mentaling*. Unlike prior works that depend on external priors or complex training data construction, our method intrinsically integrates 3D representations into the VLMs, enabling unified reasoning and 3D latent generation within the model. For (G1), our framework enables the model to generate geometric representations from images during the reasoning process. Regarding (G2), we directly project the 3D latent to align with a 3D foundation model, thereby circumventing the need for raw 3D data construction. Consequently, our model can inherently “think with 3D” without relying on any prior or auxiliary geometry encoder, corresponding to (G3). Simultaneously, since our method allows for the recovery of 3D representations(e.g., point clouds) from 3D latents via the projector, it significantly enhances the interpretability of the large reasoning model.

Specifically, we first construct a batch of Chain-of-

Thought (CoT) data that incorporates 3D special tokens. Our training framework then proceeds in two main stages. In the first stage, we perform supervised learning, where features from the 3D foundation model (e.g., VGGT [59]) are distilled into the native reasoning process of the VLM. To enable the model to think with a 3D mentaling while maintaining textual coherence, we employ both 3D latent alignment loss and the cross-entropy loss. In the second stage, we employ reinforcement learning, optimizing the tokens across the entire sampling trajectory based solely on outcome-driven signals, while preserving the alignment of the 3D latent. That is, we refine 3D mentaling within the trajectory using only outcome as the optimization signal.

Our contributions can be summarized as follows.

- We are the first to introduce the “think with 3D mentaling” framework, which operates without dependence on densely labeled training data (e.g., cognitive maps).
- We propose a two-stage training framework (shown in Fig. 2), progressing from feature alignment to learning intrinsic geometry awareness from outcome-based signals, thus enabling 3D mentaling without any external prior.
- *3DThinker* overcomes the lack of interpretability in latent reasoning. Specifically, *3DThinker* enables the recovery of 3D representations from the latent space via a projector during the reasoning process.
- Extensive experiments across multiple benchmarks demonstrate that *3DThinker* consistently outperforms strong baselines. Furthermore, our results indicate that the effectiveness of *3DThinker* generalizes well across different base VLMs, highlighting its broad applicability.

2. Related Work

2.1. Multimodal Reasoning

Large language models (LLMs) have experienced rapid development, and demonstrated strong performance across a wide range of tasks [11, 16, 35, 70, 82, 85, 92]. Building on these advances, recent works have highlighted that in-context learning, including intermediate rationales, can significantly enhance the performance of LLMs [22, 37, 49, 63, 64, 81, 93]. Current reasoning methods can be categorized into three types: pure-text, visual, and latent reasoning.

Pure-text reasoning: [3, 10, 27, 29, 53, 56] elicit textual step-by-step reasoning inspired by [25]. They typically rely on textual descriptions, which can limit the reasoning capabilities when dealing with visual evidences that cannot be adequately described using pure textual language.

Visual reasoning: to solve the problem mentioned above, some methods integrate visual evidences directly into the reasoning trajectory, whether in multi-hop or continuous modes. Some intrinsic multi-hop methods [12, 39, 51, 62], first generate detailed visual cues within the model itself (e.g., bounding boxes, coordinates, or masks), and then the

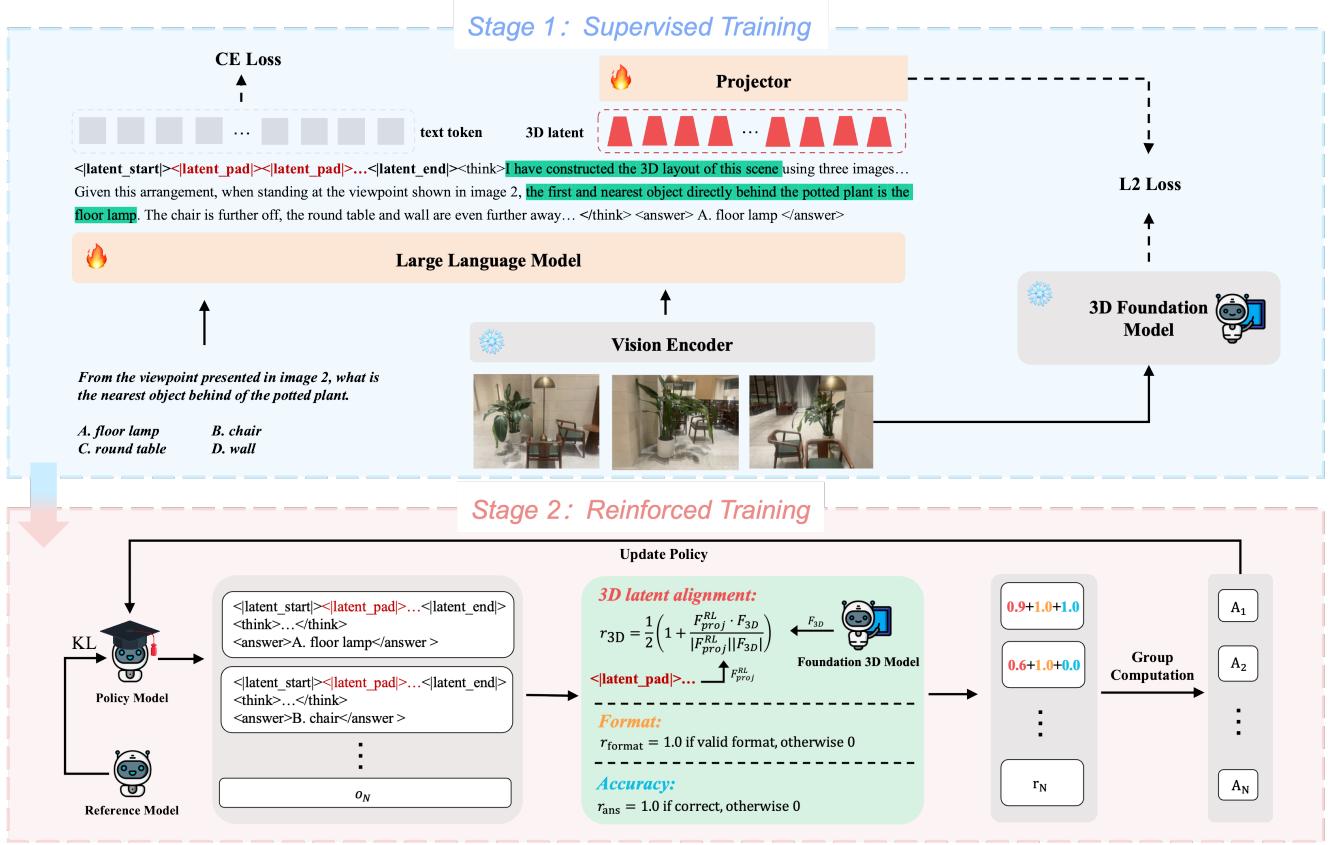


Figure 2. The schematic illustration of our *3DThinker*, a framework that enables thinking with 3D mentalizing. (1) Stage 1: *3DThinker* is first trained under supervision using our constructed CoT data (see Sec. 3.1), aligning the generated 3D latents with the feature space of 3D foundation model. This alignment allows the model to leverage suitable 3D spatial mentalizing while reasoning. (2) Stage 2: After supervised training, we further optimize the entire trajectory using only outcome signals, while maintaining the alignment of the 3D latents.

further reasoning is conducted based on these cues. Other extrinsic tool-usage methods [44, 54, 67, 91], enhance the “think with image” capability by dynamically invoking external image tools. On the other hand, continuous methods like GRIT [20] and SIFThinker [13] generate continuous visual reasoning to enable iterative corrections during the single-step reasoning process. **Latent reasoning:** some studies have shown that incorporating intermediate hidden representations into LLMs can effectively enhance model capabilities [4, 18, 55, 77]. [26] replaces CoT tokens with continuous latent embeddings, allowing unconstrained reasoning in the latent space to tackle complex tasks. More recently, Mirage [79] and LVR [31] utilize special visual tokens alongside ordinary text during reasoning. They explore visual information within the model by implicitly supervising the generation of image latent, thereby enabling reasoning with 2D visual latent.

While prior works primarily focus on enhancing reasoning ability in textual or 2D spaces, our method takes a different perspective: *we treat latent tokens as a bridge for the model to think with 3D at a mental-level, aligning more*

closely with human cognition.

2.2. Spatial Understanding

Spatial understanding encompasses skills such as 3D imagination and spatial cognition, which are essential for perceiving and manipulating spatial relationships in both 2D and 3D environments [6, 19, 42, 43, 58, 72, 84, 86, 87]. Recently, much efforts have been dedicated to evaluating the spatial understanding ability of VLMs [24, 30, 41, 48, 80, 89, 90]. Additionally, several methods have been proposed to enhance spatial understanding. For example, [5, 17, 40, 45, 46] equip LLM with additional multiview, depth or point cloud inputs, essentially serving as input enhancement. Furthermore, 3DRS [28] introduces a teacher model for 3D supervision to achieve explicit spatial representation alignment; however, this method requires input that includes the 3D coordinates corresponding to each pixel. Moreover, VLM-3R [21] employs implicit 3D tokens from a pre-trained model (e.g., CUT3R [60]) to achieve spatial awareness by incorporating prior information, necessitating inference with extensively 3D foundation model. Re-

cently, methods like MindCube [80] and Ego3D-VLM [24] have facilitated spatial understanding by constructing textual cognitive maps.

Despite these advancements, existing methods often rely on input enhancement or constructed cognitive maps, necessitating complex data collection and annotation. However, *3DThinker enables 3D mentalizing directly from multi views by learning 3D latent distilled from 3D foundation models, thereby facilitating spatial reasoning without relying on densely annotated data.*

3. Methodology

Human cognition is inherently rooted in the comprehension of 3D environments. Inspired from the cognitive mechanism of mental imagery, we propose *3DThinker*, a framework that enables VLMs to imagine 3D scenes during reasoning processes. In contrast to existing methods that reason with pure text or 2D visual cues, our framework integrates 3D representations into the interleaved multimodal trajectories. Specifically, *3DThinker* generates compact latent embeddings that serve as 3D tokens, closely emulating the mental 3D scenes that humans intuitively imagine in spatial reasoning. As illustrated in Fig. 2, *3DThinker* first aligns the VLM-generated 3D latent with the 3D foundation model, followed by reinforced training to optimize the trajectory. In this section, we will explain how we achieve this from three aspects: data generation, supervised training (stage 1), and reinforcement training (stage 2).

3.1. Data Generation

Due to the fact that VLMs naturally only generate textual tokens, they require additional supervised training to learn how to produce interleaved reasoning patterns that incorporate 3D information. Therefore, we synthesize specific training corpora based on the 10K training data from Mind-Cube dataset [80]. Given an image set from different views $\mathcal{I} = \{I_1, I_2, \dots, I_n\}$, a question Q , and the ground truth response R , we employ a high-level model (i.e., GPT-4o) M to complete the reasoning chain. Specifically, we prompt the model M to generate step-by-step reasoning that contains placeholders (3D special tokens), where these tokens represent imagined 3D scenes in the mind. Denote the response o as:

$$o = M(Q, \mathcal{I}, R). \quad (1)$$

Here, o represents the step-by-step reasoning process with embedded 3D placeholders, whose last layer hidden states are required to be consistent with features extracted from the 3D foundation model during supervised training. By prompting the large-scale reasoning VLM with various inputs, we are able to collect a training dataset $\mathcal{D} = \{(Q^{(i)}, \mathcal{I}^{(i)}, R^{(i)}, o^{(i)})\}$, where each $o^{(i)}$ contains interleaved text and 3D placeholders.

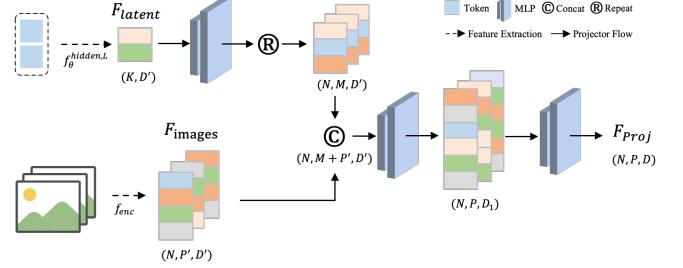


Figure 3. Illustration of our projector, which transforms VLM-generated 3D latent into the feature space of VGGT.

3.2. Supervision for 3D Grounded Reasoning

To teach the model reasoning with 3D, a naive solution is to explicitly align its outputs with 3D representations (e.g., point cloud). However, this often depends on labor-intensive data annotation and requires the model to have explicit 3D generation capabilities, which can be quite challenging. Instead, we introduce the 3D foundation model (i.e., VGGT [59]) during training, and distill its features to the 3D special token generated within the VLM reasoning process, thereby facilitating effective *3D-aware reasoning without the need for exhaustive manual labeling*.

Specifically, for each training example $(Q, \mathcal{I}, R, o) \in \mathcal{D}$, the reasoning trajectory o can be decomposed into three sequential components through concatenation operations:

$$o = o_{\text{pre}} \oplus t_{\text{3D}} \oplus o_{\text{post}}, \quad (2)$$

where $t_{\text{3D}} = \{t_1, \dots, t_k\}$ represents the token sequence of human-like 3D mental imagery. The salient vectors $F_{\text{latent}} = \{h_1, \dots, h_k\}$, which operationalize the 3D cognitive tokens t_{3D} , are extracted from the last layer hidden states of VLM $f_\theta(\cdot)$ with parameter θ . These salient vectors are recursively generated conditioned on the preceding context:

$$h_i = \begin{cases} f_\theta^{\text{hidden},L}(Q, \mathcal{I}, o_{\text{pre}}), & i = 1, \\ f_\theta^{\text{hidden},L}(Q, \mathcal{I}, o_{\text{pre}}, t_{1:i-1}), & i \geq 2. \end{cases} \quad (3)$$

Concurrently, we can obtain patch-level visual features $F_{\text{images}} = f_{\text{enc}}(\mathcal{I})$ from the image encoder, and acquire the geometry features $F_{\text{3D}} = f_{\text{vggt}}(I)$ through the last layer of VGGT aggregator. To ensure dimensional consistency between the generated 3D latent features and the predicted geometry features, we employ the projector as illustrated in Fig. 3 to transform F_{latent} into a compatible feature space:

$$F_{\text{proj}} = \text{Projector}(F_{\text{latent}}, F_{\text{images}}). \quad (4)$$

Our objective is to achieve optimal alignment between the projected 3D features derived from the VLM and the corresponding VGGT features. To this end, we formulate

the 3D alignment as the Frobenius loss:

$$\mathcal{L}_{3D} = \|F_{\text{proj}} - F_{3D}\|_F^2. \quad (5)$$

On the other hand, to ensure textual coherence while introducing 3D tokens, we employ cross-entropy loss to optimize the prediction of surrounding textual tokens. Specifically, the prediction of i -th textual tokens before is t_{3D} conditioned on both the preceding response tokens and the original input sequence.

$$\mathcal{L}_{\text{text}}^{\text{pre}} = \sum_{i=1}^{|o_{\text{pre}}|} \ell_{\text{CE}}(o_{\text{pre},i}, f_{\theta}(Q, \mathcal{I}, o_{\text{pre},<i})). \quad (6)$$

In contrast, textual tokens positioned after t_{3D} incorporates k textual 3D special tokens.

$$\mathcal{L}_{\text{text}}^{\text{post}} = \sum_{i=1}^{|o_{\text{post}}|} \ell_{\text{CE}}(o_{\text{post},i}, f_{\theta}(Q, \mathcal{I}, o_{\text{pre}}, t_{3D}, o_{\text{post},<i})). \quad (7)$$

Finally, the textual loss is formulated as follows:

$$\mathcal{L}_{\text{text}} = \mathcal{L}_{\text{text}}^{\text{pre}} + \mathcal{L}_{\text{text}}^{\text{post}}. \quad (8)$$

The overall training objective incorporates both 3D alignment and textual losses, thereby enabling the model to seamlessly incorporate 3D imagining into its textual reasoning process. Here, λ_{3D} and λ_{text} serve as hyperparameters that balances coefficients.

$$\mathcal{L}_{\text{total}} = \lambda_{3D} \mathcal{L}_{3D} + \lambda_{\text{text}} \mathcal{L}_{\text{text}}. \quad (9)$$

3.3. Reinforced Spatial Mentaling

At the supervised training stage, our primary objective is to enable the model to perform textual reasoning while simultaneously generating formatted 3D tokens. Additionally, we pre-train the projector to achieve effective alignment of 3D latents. During the reinforced training stage, we expect to use *only outcome signals* to optimize the sampling trajectories and refine the imagined mental 3D representations as well. Specifically, we employ outcome-based group-relative policy optimisation (GRPO) [52], while VGGT features are utilized to further optimize the 3D visual token generated by the model. Notably, the projector remains frozen in this stage. We formalize the RL framework as follows.

For each question-images pair (Q, \mathcal{I}) , the reinforcement learning (RL) framework generates a set of candidate completions $\{o_1, \dots, o_N\}$ from the current policy $\pi_{\theta_{\text{old}}}$, and subsequently updates the policy to π_{θ} by maximizing the following objective:

$$\mathcal{J}(\theta) = \frac{1}{N} \sum_{i=1}^N \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \left[\text{clip}(r_{i,t}, 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t}, r_{i,t} \hat{A}_{i,t} \right] - \beta \mathbb{D}_{\text{KL}}[\pi_{\theta} || \pi_{\text{ref}}] \right\}, \quad (10)$$

where $r_{i,t} = \frac{\pi_{\theta}(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}$ denotes the likelihood ratio between the updated and old policies at step t . ϵ, β are hyperparameters, and $\mathbb{D}_{\text{KL}}[\pi_{\theta} || \pi_{\text{ref}}]$ represents the KL divergence [50] between the current policy model and the fixed reference model.

The group-normalized advantage, denoted as $\hat{A}_{i,t}$, is calculated by the task-specific reward $r_{i,t}$.

$$\hat{A}_{i,t} = \frac{r_{i,t} - \text{mean}\{r_{1,t}, \dots, r_{N,t}\}}{\text{std}\{r_{i,t}, \dots, r_{N,t}\} + \delta}. \quad (11)$$

Next, we will introduce several specifically designed rewards to achieve reinforced spatial mentaling.

Reward for 3D visual token. After the supervised training, the model has begun to exhibit the ability of 3D mentaling during thinking process. To further optimize the 3D visual token in the reasoning process, we can extract the last layer hidden state of 3D special token t_{3D} (i.e., <|latent_start|><|latent_pad|> ... <|latent_end|>) in each trajectory o and perform optimization. Specifically, the projected features F_{proj}^{RL} are computed based on Eq. 4 at step t , with VGGT features F_{3D} serving as constraints during the RL stage. That is, the cosine similarity between the VGGT features and the projected features is calculated to serve as the reward r_{3D} .

$$r_{3D} = \frac{1}{2} (1 + \frac{F_{\text{proj}}^{RL} \cdot F_{3D}}{\|F_{\text{proj}}^{RL}\| \|F_{3D}\|}), \quad (12)$$

Reward for outcome-based optimization. We expect to optimize the entire trajectory using only the outcome-based signals, without relying on explicit annotations of intermediate processes. Thus, we design corresponding rewards for both format (r_{format}) and final answer (r_{ans}). **(1) Format reward:** the model’s output should adhere to the format: ...<|latent_start|><|latent_pad|>...<|latent_end|>...<think>...</think><answer>...</answer>. A reward of 1.0 is assigned to responses that strictly comply with this format. **(2) Answer reward:** we also provide the 0/1 binary reward by comparing the generated answer with the ground truth option. This outcome-based reward is evenly distributed across each token in the trajectory, including 3D visual tokens.

So, the task reward $r_{i,t}$ is a composite signal comprising the sum of three components: r_{3D} , r_{format} and r_{ans} .

4. Experiments

Evaluation metric. For multiple-choice questions, we use Accuracy, which is calculated based on exact matches between the model’s predictions and the ground truth. For numerical-answer questions, we use Mean Relative Accuracy (MRA) introduced by [75], a metric that measures the

Table 1. Accuracy comparison of generalist VLMs and our method (*3DThinker*) on MindCube-Tiny and Ego3D-Bench, with our training conducted on stage 1 (S1) and on both stage 1 and stage 2 (S1 + S2). The best results achieved based on different VLMs are **bolded**. The overall/average results of each model are highlighted in **blue**, with the best results among all models highlighted in **red**.

Method	MindCube-Tiny				Ego3D-Bench								
	Rotation	Among	Around	Overall ↑	Ego Dist.	Obj. Dist.	Loc.	Ego Mot.	Obj. Mot.	Travel Time	Ego Rel.	Obj. Rel.	Avg.↑
<i>Closed-source Models</i>													
gpt-4o-2024-11-20	37.0	44.8	56.4	46.1	33.2	26.5	28.1	78.1	56.7	36.0	60.5	66.0	48.1
gpt-4.1	45.5	44.2	47.2	45.1	51.7	36.2	41.8	82.7	62.6	44.3	65.7	70.2	56.9
glm-4.5v	28.0	43.0	33.2	37.8	49.9	39.6	48.4	88.8	73.4	40.4	57.1	81.9	59.9
gemini-2.5-pro	84.0	39.7	56.8	52.2	58.5	50.2	61.4	92.9	75.5	43.5	72.8	78.6	66.7
claude-sonnet-4	49.5	42.2	12.8	36.6	48.9	36.5	51.6	81.9	55.1	33.6	53.9	69.5	53.9
doubaos-1.6	87.0	35.8	38.0	46.1	55.2	50.8	60.5	89.0	67.3	49.8	71.4	86.0	66.3
o3-2025-04-16	86.5	42.7	66.0	56.6	71.3	59.3	65.6	93.4	80.1	53.5	77.7	83.1	73.0
<i>Qwen2.5-VL Family [2]</i>													
Qwen2.5-VL-3B	37.4	33.3	30.3	33.2	21.5	29.4	28.8	50.3	41.9	30.9	54.1	56.1	39.1
<i>3DThinker-S1</i> _{Qwen2.5-3B}	44.0	64.8	72.4	62.7	36.1	39.4	32.5	54.8	46.2	30.8	64.0	69.7	46.7
<i>3DThinker-S1+S2</i> _{Qwen2.5-3B}	55.5	81.8	75.2	75.2	41.6	46.0	33.1	54.7	53.3	30.8	70.1	76.9	50.8
Qwen2.5-VL-7B	36.5	32.5	38.4	34.7	32.7	31.5	30.5	45.9	44.0	34.5	43.2	66.5	41.1
<i>3DThinker-S1</i> _{Qwen2.5-7B}	43.5	66.3	76.4	64.4	47.9	44.5	36.5	51.9	51.3	39.1	59.1	73.9	50.5
<i>3DThinker-S1+S2</i> _{Qwen2.5-7B}	55.0	83.0	76.0	76.0	54.0	52.3	36.5	52.7	56.6	38.2	66.0	83.1	54.9
Qwen2.5-VL-32B	39.5	34.5	43.6	37.6	45.4	40.7	49.6	75.6	74.1	40.1	54.0	79.0	57.3
<i>3DThinker-S1</i> _{Qwen2.5-32B}	45.0	66.8	77.2	65.1	52.0	51.9	54.8	80.1	79.4	44.3	62.0	83.1	63.5
<i>3DThinker-S1+S2</i> _{Qwen2.5-32B}	56.5	83.2	77.2	76.7	62.2	61.9	54.5	80.2	86.6	43.7	69.9	86.0	68.1
Qwen2.5-VL-72B	40.0	42.5	44.4	42.5	42.4	38.6	54.8	86.8	68.9	38.5	53.3	80.5	58.0
<i>3DThinker-S1</i> _{Qwen2.5-72B}	42.5	68.0	73.6	64.5	49.9	45.9	57.8	85.6	75.6	43.9	58.0	80.8	62.2
<i>3DThinker-S1+S2</i> _{Qwen2.5-72B}	57.0	83.7	77.6	77.1	61.1	59.9	59.7	93.1	84.9	43.7	69.8	87.8	70.0
<i>InternVL3 Family [94]</i>													
InternVL3-8B	37.0	40.3	63.2	45.1	25.8	28.7	29.8	54.1	54.8	36.1	49.9	65.2	43.1
<i>3DThinker-S1</i> _{InternVL3-8B}	43.0	66.8	79.2	65.2	43.8	44.4	32.9	60.6	61.2	46.9	64.1	72.1	53.3
<i>3DThinker-S1+S2</i> _{InternVL3-8B}	55.0	82.5	79.2	76.5	54.6	56.1	36.0	67.2	69.4	46.7	71.0	81.9	60.4
InternVL3-14B	36.0	48.0	55.6	47.5	46.0	35.6	35.9	63.2	65.9	41.6	55.5	70.1	51.7
<i>3DThinker-S1</i> _{InternVL3-14B}	42.0	68.3	77.2	65.4	56.2	49.1	37.3	70.0	71.8	51.1	68.0	77.7	60.2
<i>3DThinker-S1+S2</i> _{InternVL3-14B}	54.5	84.3	77.6	77.0	63.5	59.9	41.3	78.3	80.2	50.0	75.1	84.0	66.5
InternVL3-38B	32.5	48.5	56.0	47.2	35.4	31.0	39.4	66.6	64.9	38.0	61.0	77.3	51.7
<i>3DThinker-S1</i> _{InternVL3-38B}	39.0	68.0	76.8	64.6	44.8	47.0	43.6	73.1	68.6	48.5	71.2	79.1	59.5
<i>3DThinker-S1+S2</i> _{InternVL3-38B}	53.5	85.2	78.0	77.4	54.7	58.1	49.2	86.9	80.4	49.1	79.6	85.9	68.0
InternVL3-78B	38.5	50.5	57.4	49.9	54.6	48.4	50.3	77.7	70.0	44.8	57.0	76.6	59.9
<i>3DThinker-S1</i> _{InternVL3-78B}	43.5	69.0	77.2	66.1	59.8	53.1	52.2	80.1	72.5	53.9	65.1	78.0	64.3
<i>3DThinker-S1+S2</i> _{InternVL3-78B}	57.0	86.2	78.8	78.9	69.9	61.0	61.0	91.9	88.6	54.8	75.3	83.9	73.3
<i>LLaVA-OneVision-1.5 Family [1]</i>													
LLaVA-OneVision-1.5-4B	33.5	38.0	49.2	39.8	39.7	37.1	29.2	51.4	51.8	34.1	52.4	73.5	46.2
<i>3DThinker-S1</i> _{LLaVA-O-1.5-4B}	41.5	59.8	66.0	57.8	40.0	39.1	33.1	51.1	52.6	30.9	58.6	73.8	47.4
<i>3DThinker-S1+S2</i> _{LLaVA-O-1.5-4B}	48.0	67.5	65.2	63.2	40.2	39.9	34.2	51.9	52.3	30.8	61.8	73.8	48.1
LLaVA-OneVision-1.5-8B	34.5	34.7	48.4	37.9	30.3	36.6	34.3	44.9	51.9	36.9	53.4	74.4	45.3
<i>3DThinker-S1</i> _{LLaVA-O-1.5-8B}	43.0	57.8	64.8	56.7	35.1	39.0	36.1	44.9	53.2	31.9	61.0	73.8	46.9
<i>3DThinker-S1+S2</i> _{LLaVA-O-1.5-8B}	49.0	68.2	64.8	63.7	36.5	41.5	37.0	46.2	53.3	32.8	64.9	77.2	48.7

closeness of the model’s predictions to the ground truth values. ”Avg.” denotes the mean value of all subset task.

Hyper-parameters. For *3DThinker*, in the stage 1, we set the MLP depth to 6, with the learning rate of 1e-4, latent size of 12, epoch of 10. In Eqn. 9, the hyper-parameters $\lambda_{3D}, \lambda_{text}$ are uniformly set to 0.1 and 1, respectively. In the stage 2, we set the balancing coefficient of all three reward to 1, with the learning rate of 1e-5, the rollout number of 8. Additional details are provided in the Supp. Mat..

4.1. Benchmarking Generalist VLMs

In this section, we comprehensively investigate different training stages in *3DThinker* across various generalist VLMs. We conduct experiments on MindCube-Tiny [80]

and Ego3D-Bench [24] benchmarks, both of which are designed to evaluate the spatial understanding ability from limited views.

As shown in Tab. 1, *3DThinker*-full achieves consistent improvements over the generalist VLMs across all settings. On MindCube-Tiny, the overall performance gain ranges from **51.8%** to **108.8%**, while on Ego3D-Bench, the improvement spans **18.1%** to **36.9%**. Taking Qwen2.5-VL-3B as an example, *3DThinker* boosts performance on MindCube-Tiny by **88.9%** (62.7 vs. 33.2) after stage 1, and further improves by **19.9%** (75.2 vs. 62.7) after stage 2. Similarly, on Ego3D-Bench, we observe a **19.3%** improvement (46.7 vs. 39.1) after the stage 1 and an additional **8.8%** gain (50.8 vs. 46.7) following the stage 2. Al-

Table 2. The evaluation of various baselines on the VSI-Bench [75], SPBench [34], CV-Bench [57], SPAR-Bench [88], ViewSpatial-Bench [33] and MMSI-Bench [78] datasets. [SI] denotes benchmarks with single image, whereas [MV] refers to multi-view images. The **best**-performing results under each base model are highlighted.

Method	VSI-Bench [75] [MV]	SPBench [34] [SI, MV]	CV-Bench [57] [SI]	SPAR-Bench [88] [SI, MV]	ViewSpatial-Bench [33] [SI, MV]	MMSI-Bench [78] [MV]	Avg. \uparrow
<i>Qwen2.5-VL-3B Based Spatial Models</i>							
Qwen2.5-VL-3B [2]	29.4	38.5	70.6	24.6	35.6	26.5	37.5
Spatial-MLLM-4B [65]	47.3	48.4	73.8	35.1	43.6	31.5	46.6
SpatialLadder-3B [34]	45.7	70.6	73.7	34.4	44.2	29.2	49.6
<i>3DThinker-SI</i> _{Qwen2.5-3B}	53.2	54.8	74.5	52.3	59.5	37.7	55.3
<i>3DThinker-SI+S2</i> _{Qwen2.5-3B}	59.1	60.2	78.4	58.2	64.7	41.9	60.4
<i>Qwen2.5-VL-7B Based Spatial Models</i>							
Qwen2.5-VL-7B [2]	35.8	42.9	73.0	30.2	37.9	26.9	41.1
SpaceR-7B [45]	44.5	54.0	75.3	37.1	45.5	28.8	47.5
VILASR-7B [66]	45.4	53.9	77.1	37.8	46.1	30.2	48.4
Video-R1 [23]	33.4	42.8	69.6	31.5	36.1	29.4	40.5
<i>3DThinker-SI</i> _{Qwen2.5-7B}	57.3	61.5	77.9	56.3	61.7	41.5	59.4
<i>3DThinker-SI+S2</i> _{Qwen2.5-7B}	63.7	68.3	81.1	63.3	68.6	43.3	64.7

though the performance is slightly weaker on certain sub-tasks, e.g., Travel Time, this can be attributed to the need for *richer contextual information to align the normalized 3D representations with the real-world*. Remarkably, our model is trained without any Ego3D-specific data, yet it still achieves promising results on Ego3D-Bench, demonstrating strong cross-dataset generalization. *This highlights that our “think with 3D” framework effectively enhances the model’s generalization capability across diverse spatial understanding scenarios*. It is also worth noting that our best model, *3DThinker-SI+S2*_{Qwen2.5-7B}, outperforms all other models, both open-source and closed-source, including the latest O3 model (**78.9** vs. 56.6 on MindCube-Tiny, **73.3** vs. 73.0 on Ego3D-Bench).

4.2. Comparisons with Baselines

We evaluate our method against several state-of-the-art (SOTA) approaches across a diverse set of categories. Additional details are provided in the Supp. Mat..

Different Spatial Models. As shown in Tab. 2, we categorize the methods into two groups based on the types of base VLMs and then evaluate them across different benchmarks. For the Qwen2.5-VL-3B-based spatial models, *3DThinker* surpasses the recent SOTA, SpatialLadder-3B, by **11.5%** (55.3 vs. 49.6) in stage 1. This improvement is further enhanced to **21.8%** (62.7 vs. 49.6) following stage 2. When using the Qwen2.5-VL-7B model, our method achieves even more remarkable results. *3DThinker* outperforms the SOTA VILASR-7B by **22.7%** (59.4 vs. 48.4) in stage 1, and by **33.7%** (64.7 vs. 48.4) in stage 2. On the other hand, in contrast to methods that exhibit task-specific overfitting (e.g., SpatialLadder-3B on SPBench), *3DThinker* demonstrates consistent improvement across all tasks, highlighting the robust spatial reasoning capability of our method. Additionally, unlike models such as Video-R1, which struggle on single-view tasks (e.g., underperforming the base model on CV-Bench), our method demonstrates strong per-

Table 3. Performance on Ego3D-Bench (Accuracy Avg.) in comparison between *3DThinker* and Ego3D-VLM, employing a series of VLMs with varying parameters. The **best** is highlighted.

Method	Qwen2.5-VL				InternVL3			
	3B	7B	32B	72B	8B	14B	38B	78B
Ego3D-VLM [24]	44.4	54.3	65.5	69.5	60.1	66.1	68.0	71.8
<i>3DThinker</i>	50.8	54.9	68.1	70.0	60.4	66.5	68.0	73.3

Table 4. Results with Qwen2.5-VL-3B on MindCube-Tiny in terms of different training strategies. The **best** is highlighted.

Method	MindCube-Tiny			
	Rotation	Among	Around	Overall \uparrow
raw-QA SFT	34.5	52.5	66.0	52.3
CoT SFT	36.0	54.3	65.2	53.4
Aug-CGMap-FFR-Out-SFT	49.5	52.5	66.4	55.2
Plain-CGMap-FFR-Out-SFT	47.5	62.3	67.6	60.8
<i>3DThinker-SI</i> _{Qwen2.5-3B}	44.0	64.8	72.4	62.7
GRPO	36.5	49.3	64.8	50.6
CoT SFT + GRPO	36.5	55.2	65.6	54.1
Aug-CGMap-FFR-Out-SFT+RL	53.0	76.8	70.0	70.7
Plain-CGMap-FFR-Out-SFT+RL	48.0	79.2	68.4	70.7
<i>3DThinker-SI+S2</i> _{Qwen2.5-3B}	55.5	81.8	75.2	75.2

formance on both single-image and multi-view tasks. This indicates that our 3D mental reasoning framework significantly enhances performance, even in single-image cases.

Different Architectures and Parameter Scales. Tab. 3 compares our method with Ego3D-VLM on Ego3D-Bench across different model series and parameter scales. Although Ego3D-VLM constructs its cognitive map with the aid of external modules—specifically, a referring expression comprehension model (Grounding-DINO-Base [38]) and a depth estimator (Depth-Anything-V2-Metric-Large [76])—our method, which does not rely on any extrinsic priors at inference, still achieves superior performance. In particular, on Qwen2.5-VL-3B, *3DThinker* yields a notable **14.4%** improvement (50.8 vs. 44.4).

4.3. Training Strategies

To further demonstrate the effectiveness of our training paradigm, we compare *3DThinker* against several repre-



Figure 4. The reasoning process for different cases is presented, along with the visualization of the 3D latent representations.

sentative training strategies. Among them, Aug-CGMap-FFR-Out and Plain-CGMap-FFR-Out serve as SOTA baselines introduced in [80]. Specifically, Aug-CGMap-FFR-Out performs reasoning with the augmented cognitive map (camera-view information included), whereas Aug-CGMap-FFR-Out relies solely on plain cognitive maps without augmentation.

Under supervised training, our method surpasses raw-QA SFT, CoT SFT, and even the cognitive-map-based SFT proposed in [80] by a margin of **3.1%** (62.7 vs. 60.8). The relatively smaller improvement observed in the rotation sub-category can be attributed to its requirement for dynamic spatial imagination. *Since our "think with 3D" supervised framework primarily targets static spatial understanding, the RL stage further enhances its dynamic capability by optimizing whole reasoning trajectories.* That is, through outcome-based RL, 3DThinker progressively refines the 3D latents across rollouts, achieving additional gains in both zero-RL and SFT-then-RL settings. Furthermore, 3DThinker achieves a **6.4%** improvement over the cognitive-map-based SFT-then-RL baseline (75.2 vs. 70.7), demonstrating its superior capability in integrating spatial reasoning with reinforcement learning.

4.4. Visualization

We visualize the results of 3D mentalizing in Fig. 4. During inference, we extract the last layer hidden states corresponding to the 3D special tokens. These 3D latents are projected into the VGGT feature space via the projector illustrated in Fig. 3. The projected features are subsequently processed by the DPT [47] of VGGT to generate point clouds. As shown in Fig. 4, the reconstructed mentalizing point clouds roughly depict the underlying scene, *where the clearer regions are typically correlated with prompt-relevant objects.* This observation indicates that the 3D latents effectively encode the mental scene guided by the prompt intent. After reasoning with 3D mentalizing, all three examples yield correct answers. Additional visualizations and analysis are

Table 5. Ablation of different 3D latent size on MindCube-Tiny in terms of $3D\text{Thinker-S1}_{Qwen2.5-3B}$.

Latent Size	4	8	12	16	32	64
Accuracy	60.2	60.6	62.7	59.9	25.1	15.5

Table 6. Ablation of different designs including 3D special token position (Token Pos.), projector and rewards in terms of $3D\text{Thinker-S1+S2}_{Qwen2.5-3B}$.

Method	Token Pos.		Projector	Rewards			Full	
	Middle	End		VGGT-to-VLM	w/o r_{format}	w/o r_{ans}		
Accuracy	42.0	74.3	74.1		74.8	64.2	68.3	75.2

provided in the Supp. Mat..

4.5. Ablation Study

Different 3D Latent Size. In Tab. 5, we ablate the effect of different latent sizes on the results. The results indicate that the optimal performance is achieved with the latent size of about 12. *This is because a smaller latent size limits the model's representational capacity, while a larger latent size can compromise the model's natural expressive ability,* leading to repetitive `<|latent_start|>` outputs that fail to yield the final answer.

Different Designs. As shown in Tab. 6, we first conduct an ablation study on the placement of the 3D special tokens. Beyond the approach in Sec. 3.3, where the special tokens is positioned at the beginning (before `<think>`), we also explore placing it between the `<think>` and `</think>`, as well as at the end (after `</answer>`). We observe that placing the 3D tokens in the middle disrupts natural language coherence: *the 3D latent can resemble certain character features, leading to garbled text and premature output termination.* This results in a significant performance drop (75.2 vs. 42.0). In contrast, positioning the 3D tokens at the beginning or end—where it is isolated from natural text—yields significantly better performance.

We also examine two potential projector configurations. The first maps the last layer hidden state of the VLM to the VGGT space (shown in Fig. 3), *allowing the VLM features to be explicitly converted into 3D representations (e.g., point clouds) via the projector.* The alternative compresses VGGT features directly into the VLM space (e.g., via adaptive average pooling), but this approach is unrecoverable to 3D representations. Given the interpretability, visualizability, and better performance (75.2 vs. 74.1) of the first approach, we adopt it as our projector strategy.

Finally, we ablate the three rewards used in stage 2. Among them, the formatting requirement has minimal impact. In contrast, removing 3D alignment leads to a substantial performance drop (75.2 vs. 68.3) due to *the absence of stable constraints on the 3D latent.* The final answer reward is also critical (75.2 vs. 64.2), serving as *the sole ground-truth supervision signal* and guiding optimization of each token across the entire rollout.

5. Conclusion and Limitation

Conclusion. In this paper, we propose *3DThinker*, a framework for VLM to think with 3D spatial mentaling. Unlike recent methods that rely solely on pure text or 2D visual cues for reasoning, *3DThinker* leverages geometric information embedded in images during the reasoning process for the first time. Additionally, our method does not rely on dense annotations or other external priors. To enable thinking with 3D spatial mentaling, we introduce a two-stage training scheme. Stage 1 distills geometric features from a pretrained 3D model to warm up. Stage 2 optimizes the entire reasoning trajectory while maintaining 3D visual alignment based on the outcome signal. Experimental results show that our method outperforms previous methods across various benchmarks, establishing a solid foundation for future exploration.

Limitation & Future Work. (1) Our method recovers 3D mental representations from the last layer hidden state of the special tokens. However, these latents are not autoregressively incorporated into the framework. Thus, developing a unified structure (e.g. unified tokenizer) could be a key area for future improvement. (2) Exploring iterative 3D mentaling within the trajectory may provide additional benefits.

References

- [1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. 6
- [2] Shuai Bai, Kepin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 7
- [3] Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025. 2
- [4] Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. *arXiv preprint arXiv:2406.12775*, 2024. 3
- [5] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024. 2, 3
- [6] Zhongang Cai, Yubo Wang, Qingping Sun, Ruisi Wang, Chenyang Gu, Wanqi Yin, Zhiqian Lin, Zhitao Yang, Chen Wei, Xuanke Shi, et al. Has gpt-5 achieved spatial intelligence? an empirical study. *arXiv preprint arXiv:2508.13142*, 2025. 3
- [7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 2
- [8] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 1
- [9] Hanzhi Chen, Boyang Sun, Anran Zhang, Marc Pollefeys, and Stefan Leutenegger. Vidbot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27661–27672, 2025. 2
- [10] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-V>, 2025. Accessed: 2025-02-02. 2
- [11] Zhangquan Chen, Chunjiang Liu, and Haobin Duan. A three-phases-lora finetuned hybrid llm integrated with strong prior module in the education context. In *International Conference on Artificial Neural Networks*, pages 235–250. Springer, 2024. 2
- [12] Zhangquan Chen, Xufang Luo, and Dongsheng Li. Visrl: Intention-driven visual perception via reinforced reasoning. *arXiv preprint arXiv:2503.07523*, 2025. 2
- [13] Zhangquan Chen, Ruihui Zhao, Chuwei Luo, Mingze Sun, Xinlei Yu, Yangyang Kang, and Ruqi Huang. Sifthinker: Spatially-aware image focus for visual reasoning. *arXiv preprint arXiv:2508.06259*, 2025. 1, 2, 3
- [14] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatial-rgrpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024. 2
- [15] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatial-rgrpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024. 1
- [16] Kexin Chu, Zixu Shen, Dawei Xiang, and Wei Zhang. Safekv: Safe kv-cache sharing in llm serving. In *Machine Learning for Computer Architecture and Systems*, 2025. 2
- [17] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. *arXiv preprint arXiv:2503.13111*, 2025. 2, 3
- [18] Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024. 3
- [19] Tianqi Ding, Dawei Xiang, Pablo Rivas, and Liang Dong. Neural pruning for 3d scene reconstruction: Efficient nerf acceleration. *arXiv preprint arXiv:2504.00950*, 2025. 3
- [20] Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayananaraju, Xinze

- Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*, 2025. 3
- [21] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025. 2, 3
- [22] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:70757–70798, 2023. 2
- [23] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 7
- [24] Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. *arXiv preprint arXiv:2509.06266*, 2025. 1, 2, 3, 4, 6, 7
- [25] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2
- [26] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024. 3
- [27] Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, De-hao Zhang, and Yu Cao. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 2024. 2
- [28] Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. Mllms need 3d-aware representation supervision for scene understanding. *arXiv preprint arXiv:2506.01946*, 2025. 3
- [29] Jiaming Ji, Jiayi Zhou, Hantao Lou, Boyuan Chen, Donghai Hong, Xuyao Wang, Wenqi Chen, Kaile Wang, Rui Pan, Jiahao Li, Mohan Wang, Josef Dai, Tianyi Qiu, Hua Xu, Dong Li, Weipeng Chen, Jun Song, Bo Zheng, and Yaodong Yang. Align anything: Training all-modality models to follow instructions with language feedback. 2024. 2
- [30] Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. *arXiv preprint arXiv:2504.17207*, 2025. 3
- [31] Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muham Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025. 3
- [32] Chengmeng Li, Junjie Wen, Yan Peng, Yaxin Peng, Feifei Feng, and Yichen Zhu. Pointvla: Injecting the 3d world into vision-language-action models. *arXiv preprint arXiv:2503.07511*, 2025. 2
- [33] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025. 7
- [34] Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueling Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025. 1, 7
- [35] Xinjin Li, Yu Ma, Yangchen Huang, Xingqi Wang, Yuzhen Lin, and Chenxi Zhang. Synergized data efficiency and compression (sec) optimization for large language models. In *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, pages 586–591, 2024. 2
- [36] Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. *arXiv preprint arXiv:2409.09788*, 2024. 1
- [37] Junhong Lin, Xinyue Zeng, Jie Zhu, Song Wang, Julian Shun, Jun Wu, and Dawei Zhou. Plan and budget: Effective and efficient test-time scaling on large language model reasoning, 2025. 2
- [38] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 2, 7
- [39] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, et al. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv preprint arXiv:2501.10074*, 2025. 1, 2
- [40] Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. *arXiv preprint arXiv:2505.12448*, 2025. 1, 2, 3
- [41] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Celso M de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024. 3
- [42] Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Weijie Wang, Haoyun Li, Guosheng Zhao, Jie Li, Wenkang Qin, Guan Huang, and Wenjun Mei. Wonderturbo: Generating interactive 3d world in 0.72 seconds. *arXiv preprint arXiv:2504.02261*, 2025. 3
- [43] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Xinze Chen, Guanghong Jia, Guan Huang, and Wenjun Mei. Recondreamer-rl: Enhancing reinforcement learning via diffusion-based scene reconstruction. *arXiv preprint arXiv:2508.08170*, 2025. 3
- [44] OpenAI. Introducing openai o3 and o4-mini, 2025. <https://openai.com/index/introducing-o3-and-o4-mini/>. 3
- [45] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Rein-

- forcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. 1, 2, 3, 7
- [46] Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025. 3
- [47] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 8
- [48] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 3
- [49] Ayushman Sarkar, Mohd Yamani Idris, and Zhenyu Yu. Reasoning in computer vision: Taxonomy, models, tasks, and methodologies. *arXiv preprint arXiv:2508.10523*, 2025. 2
- [50] John Schulman. Approximating kl divergence. *John Schulman's Homepage*, 2020. 5
- [51] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models, 2024. 2
- [52] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 5
- [53] Haozhan Shen, Zilun Zhang, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. <https://github.com/om-ai-lab/VLM-R1>, 2025. Accessed: 2025-02-15. 2
- [54] Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025. 3
- [55] Yu Sun, Yin Li, Ruixiao Sun, Chunhui Liu, Fangming Zhou, Ze Jin, Linjie Wang, Xiang Shen, Zhuolin Hao, and Hongyu Xiong. Audio-enhanced vision-language modeling with latent space broadening for high quality data expansion, 2025. 3
- [56] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. Llamav-o1: Rethinking step-by-step visual reasoning in llms, 2025. 2
- [57] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. 7
- [58] Boyuan Wang, Xinpan Meng, Xiaofeng Wang, Zheng Zhu, Angen Ye, Yang Wang, Zhiqin Yang, Chaojun Ni, Guan Huang, and Xingang Wang. Embodiedreamer: Advancing real2sim2real transfer for policy training via embodied world modeling. *arXiv preprint arXiv:2507.05198*, 2025. 3
- [59] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 4
- [60] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 3
- [61] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024. 1
- [62] XuDong Wang, Shaolun Zhang, Shufan Li, Konstantinos Kallidromitis, Kehan Li, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Segllm: Multi-round reasoning segmentation. *arXiv preprint arXiv:2410.18923*, 2024. 2
- [63] Zixuan Wang, Yu Sun, Hongwei Wang, Baoyu Jing, Xiang Shen, Xin Dong, Zhuolin Hao, Hongyu Xiong, and Yang Song. Reasoning-enhanced domain-adaptive pretraining of multimodal large language models for short video content moderation, 2025. 2
- [64] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2
- [65] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. 7
- [66] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025. 7
- [67] Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. *arXiv preprint arXiv:2505.19255*, 2025. 3
- [68] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:90277–90317, 2025. 1
- [69] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE con-*

- ference on computer vision and pattern recognition, pages 9068–9079, 2018. 1
- [70] Dawei Xiang, Wenyan Xu, Kexin Chu, Zixu Shen, Tianqi Ding, and Wei Zhang. Promptsculptor: Multi-agent based text-to-image prompt optimization. *arXiv preprint arXiv:2509.12446*, 2025. 2
- [71] Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey. *arXiv preprint arXiv:2412.20206*, 2024. 2
- [72] Wenrui Xu, Dalin Lyu, Weihang Wang, Jie Feng, Chen Gao, and Yong Li. Defining and evaluating visual language models’ basic spatial abilities: A perspective from psychometrics. *arXiv preprint arXiv:2502.11859*, 2025. 3
- [73] Tianyi Yan, Dongming Wu, Wencheng Han, Junpeng Jiang, Xia Zhou, Kun Zhan, Cheng-zhong Xu, and Jianbing Shen. Drivingsphere: Building a high-fidelity 4d world for closed-loop simulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27531–27541, 2025. 1
- [74] Tianyi Yan, Junbo Yin, Xianpeng Lang, Ruigang Yang, Cheng-Zhong Xu, and Jianbing Shen. Olidm: Object-aware lidar diffusion models for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9121–9129, 2025. 1
- [75] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 5, 7
- [76] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 2, 7
- [77] Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? *arXiv preprint arXiv:2402.16837*, 2024. 3
- [78] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. 7
- [79] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025. 1, 2, 3
- [80] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025. 2, 3, 4, 6, 8
- [81] Zhenyu Yu, Mohd Yamani Idna Idris, Hua Wang, Pei Wang, Junyi Chen, and Kun Wang. From physics to foundation models: A review of ai-driven quantitative remote sensing inversion. *arXiv preprint arXiv:2507.09081*, 2025. 2
- [82] Zhenyu Yu, Mohd Yamani Idna Idris, Pei Wang, Yuelong Xia, and Yong Xiang. Forgetme: Benchmarking the selective forgetting capabilities of generative models. *Engineering Applications of Artificial Intelligence*, 161:112087, 2025. 2
- [83] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresight-drive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025. 1
- [84] Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. *arXiv preprint arXiv:2509.22548*, 2025. 3
- [85] Yiming Zeng, Wanhai Yu, Zexin Li, Tao Ren, Yu Ma, Jinghan Cao, Xiyan Chen, and Tingting Yu. Bridging the editing gap in llms: Fineedit for precise and targeted text modifications, 2025. 2
- [86] Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*, 2025. 3
- [87] J. Zhang, W. Zhang, C. Tan, X. Li, and Q. Sun. Yolo-ppa based efficient traffic sign detection for cruise control in autonomous driving. *arXiv preprint arXiv:2409.03320*, 2024. 3
- [88] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025. 7
- [89] Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. *arXiv preprint arXiv:2412.12693*, 2024. 3
- [90] Weichen Zhang, Zile Zhou, Zhiheng Zheng, Chen Gao, Jin-qiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. *arXiv preprint arXiv:2503.11094*, 2025. 3
- [91] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deep-eyes: Incentivizing” thinking with images” via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 3
- [92] Pengfei Zhou, Weiqing Min, Chaoran Fu, Ying Jin, Mingyu Huang, Xiangyang Li, Shuhuan Mei, and Shuqiang Jiang. Foodsky: A food-oriented large language model that can pass the chef and dietetic examinations. *Patterns*, 6(5), 2025. 2
- [93] Xiaoling Zhou, Wei Ye, Zheng Lee, Lei Zou, and Shikun Zhang. Valuing training data via causal inference for in-context learning. *IEEE Transactions on Knowledge and Data Engineering*, 2025. 2
- [94] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shen-gloung Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 6