Imaginative Perception Tokens Enhance Spatial Reasoning in Multimodal Language Models

Anonymous CVPR submission

Paper ID *****

Abstract

Vision-language models (VLMs) excel at many tasks, yet continue to struggle with spatial reasoning—problems where the key information is not directly observable in the input. Many spatial questions require imaginative perception*: simulating an unseen viewpoint, tracing a trajectory through an occluded space, or integrating partial views into a coherent spatial map. Humans naturally support this kind of reasoning through imagination. Prior work has introduced intermediate visual representations (e.g., visual thoughts, depth, or box tokens), but these intermediates often refine structure already visible rather than predicting the missing spatial structure implied by the evidence. We introduce* ***Imaginative Perception Tokens (IPT)****, intermediate perceptual representations that externalize what a VLM would perceive under an alternative spatial configuration while remaining consistent with the observed input. To study this capability, we formulate three tasks that require imaginative perception:* ***Perspective Taking (PET)****,* ***Path Tracing (PT)****, and* ***Multiview Counting (MVC)****. For each task, we construct datasets of ∼20K examples spanning simulated and real-world settings, paired with ground-truth intermediate imaginations, final answers, and curated evaluation benchmarks. Using the unified VLM BAGEL [12] as our backbone, IPT supervision improves spatial reasoning across several settings and often outperforms textual chain-of-thought training, even when no image is generated at inference time. For example, on MVC, IPT improves accuracy by 3.4% and achieves performance competitive with strong closed-source models on Path Tracing. We also find that mixed training with IPT and label-only data can further improve performance. In contrast, textual chain-of-thought can be detrimental on these tasks, substantially degrading performance in some cases, highlighting a modality mismatch when forcing spatial computation through language. Overall, IPT provides a principled supervision signal for reasoning over unobserved structure, yielding stronger spatial generalization and a more interpretable intermediate aligned with the underlying geometry of the task.*

1. Introduction

Spatial reasoning still still remains a persistent challenge for vision-language models (VLMs) [1, 8, 11]. Many spatial questions require reasoning about how objects relate within a three-dimensional environment, how these relationships change under viewpoint transformations, or how information from multiple partial observations should be integrated into a coherent scene representation [17, 41] for vision-language models (VLMs) [1, 8, 11]. While current models can often recognize objects and attributes, they frequently struggle when reasoning requires manipulating spatial structure, such as predicting how a scene would appear from another viewpoint [21, 23] or aggregating information across multiple views [37].

A key reason for this difficulty is that many spatial reasoning problems cannot be solved by analyzing the input alone. Instead, they require constructing a spatial representation that is not directly observed. Humans naturally address such problems through imagination: when asked what lies to the left after moving to a new position, or how many objects exist in a room seen from several viewpoints, we mentally simulate the scene from unseen perspectives or integrate partial observations into a unified spatial map [34, 41, 42]. In other words, spatial reasoning often depends on imagining missing spatial structure that proceed despite incomplete observations.

Existing approaches provide only partial solutions. Recent work teaches models to generate intermediate visual thoughts alongside language [14, 16, 20], while others introduce structured perceptual intermediates, such as depth maps or bounding boxes represented as tokens [2, 27, 38]. Although these methods demonstrate that intermediate visual representations can support reasoning, they primarily operate over information already present in the input observation, refining visible structures or extracting perceptual attributes. However, as discussed above, many spatial rea-

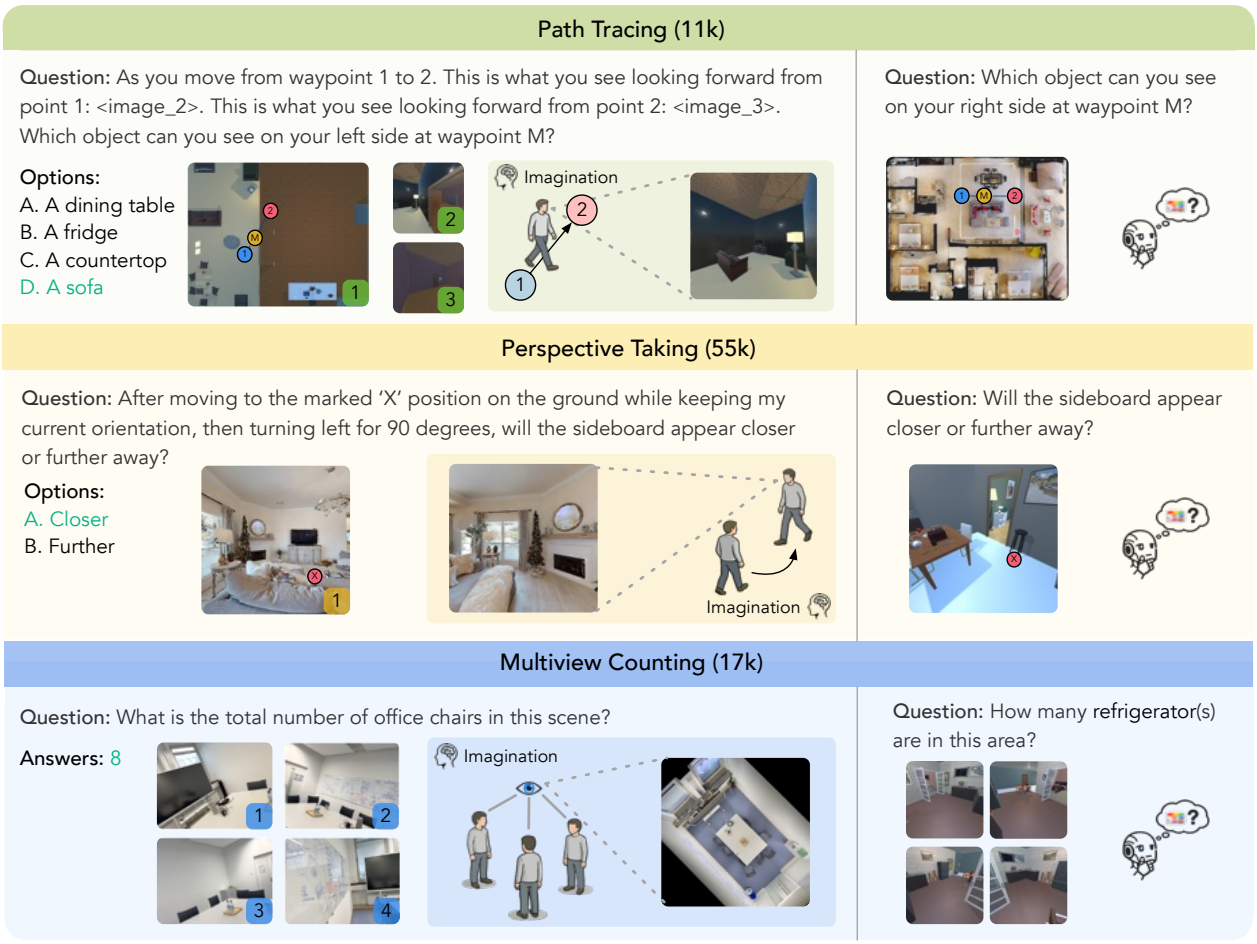

Figure 1. **Overview of the three spatial imagination tasks.** The left columns show training examples with ground-truth imaginative perception; the right columns show evaluation examples.

soning problems arise precisely because the required spatial information is not directly observable, and therefore requires imagination.

To address this gap, we propose **Imaginative Perceptual Tokens** for VLMs. When VLMs are trained with them, they enable intermediate reasoning steps that represent novel spatial views. Unlike standard perceptual intermediates that describe structures visible in the input, imaginative representations correspond to what the model would perceive if it were observing the input from a different spatial configuration, such as from an unseen viewpoint or after integrating multiple partial observations into one. At the same time, they are not unconstrained imagination: the predicted percept must remain consistent with the observed scene. These tokens externalize the model's prediction of what would be perceived given incomplete spatial evidence.

To study this capability, we propose three spatial reasoning tasks that fundamentally require imaginative perception. (1) Perspective Taking requires predicting how a scene would appear from a new viewpoint given a single first-person observation ("If you move to the marked position and turn left, will the chair appear on your left or right?"); (2) Path Tracing requires inferring what an agent would see along a navigation path based on a top-down view ("If you walk along the marked path, which object will you see on your side?"); Finally, (3) Multiview Counting requires integrating multiple partial observations into a top-down view to determine the number of objects present in the scene. These tasks would be made easy when correctly predicting what would be perceived in a different spatial configuration. For each task we construct a dataset of approximately 20k examples each drawn from both real-world and synthetic simulated environments, with ground-truth intermediate spatial imaginations paired with final answers. Each dataset is accompanied by a human-filtered benchmark for evaluation. Together these constitute the first datasets designed explicitly to train and evaluate visually-grounded intermediate spatial reasoning in models.

Empirically, we find that training with imaginative perceptual supervision can improve performance on these spatial reasoning tasks compared to answer-only supervision, and often compares favorably to textual chain-of-thought approaches. These improvements can persist even when the model does not explicitly generate intermediate images at inference time, suggesting that such supervision may help models develop stronger internal spatial representations. At the same time, we observe that the benefits vary across tasks and settings, indicating that imagination quality and task structure both play important roles.

Overall, our results suggest that supervising models with intermediate perceptual predictions offers a useful direction for improving spatial reasoning, particularly in settings where the required structure is not directly observable from the input.

2. Related Works

Evaluation of VLMs' spatial reasoning. A growing body of benchmarks has established that spatial reasoning remains a persistent weakness of modern vision-language models. Early datasets target brittleness in basic spatial predicates: SpatialSense [35] reduces language priors through adversarial crowdsourcing, while VSR [22] scales relation types in a caption-verification format, and What'sUp [17] uses minimal-pair testing to reveal systematic failures on left/right and above/below distinctions. More recent work shifts from 2D relations to viewpoint and 3D structure. 3DSRBench [23] shows that models fail under modest changes in perspective, depth, and occlusion, and ViewSpatial-Bench [21] identifies a "perspective gap": models often succeed in camera-centered views but break when asked to adopt human-centered viewpoints.

Benchmarks have also expanded to multi-image and video settings where maintaining a consistent spatial state is essential. VSI-Bench [34] tests whether models can build a persistent mental map from videos, while MMSI-Bench [37] reports large human–model gaps on cross-view scene reconstruction. MindCube [41] is closely aligned with our motivation, targeting spatial mental modeling from limited views, including perspective-taking and "what-if" scene dynamics. Counting Stacked Objects [13] studies 3D object counting under heavy occlusion across multiple views, directly analogous to our Multiview Counting setting. Finally, benchmark design work emphasizes that shortcuts remain pervasive: Brown *et al*. [3] construct VSI-Bench-Debiased by iteratively pruning samples solvable via priors, reinforcing the need for evaluations where success requires genuine spatial computation.

Collectively, these benchmarks diagnose *where* VLMs fail spatially, but they typically evaluate *discriminative* understanding—reading off a relation from an observed view—rather than *constructive* spatial imagination. Our work is complementary: we isolate *imaginative perception* as a standardized intermediate substrate, and pair each task with a ground-truth intermediate spatial imagination rather than only a final answer label.

Intermediate representations for spatial reasoning. Chain-of-thought prompting [31] can improve multi-step reasoning, but serializing viewpoint transformations, occlusions, and geometric constraints into language is often awkward and error-prone. This motivates intermediate representations in modalities better aligned with spatial computation. One direction externalizes reasoning into explicit visual buffers: Visual Sketchpad [16] equips models with drawing actions for iterative refinement, and MVoT [20] introduces visualization-of-thought traces that help on dynamic spatial tasks where text CoT struggles. ThinkMorph [14] studies interleaved text–image reasoning traces, and OpenAI describes o3/o4-mini as using chains-of-thought that include simple image transformations during reasoning [25]. A complementary line introduces latent visual scratchpads: Mirage [38] frames latent tokens as "machine mental imagery," and Mull-Tokens [27] generalizes to modality-agnostic latent thinking tokens.

Our work differs in what the intermediate is meant to represent. Many prior approaches treat intermediate images or latents as optional visualizations of *visible* structure. We instead target *imaginative perception*: predicting what would be perceived under an unobserved spatial configuration (e.g., a rotated viewpoint or a top-down path state), a representation constrained by the input but not present in it. This framing provides a principled criterion for when intermediate visual thoughts are necessary and a controlled way to supervise them.

Unified multimodal models for interleaved understanding and generation. Producing imaginative perceptual intermediates within a single model requires the ability to both understand and generate images. Unified decoder-only architectures treat image tokens as first-class sequence elements, enabling arbitrary text–image interleaving. Chameleon [5] is an early example, while Show-o2 [33] and Janus [7] offer alternative unified designs that balance understanding and generation. We build on BAGEL [12], a unified model pretrained on large interleaved corpora that exhibits strong spatial capabilities, making it a natural substrate for producing intermediate spatial imaginations. Crucially, however, a unified architecture alone does not guarantee that intermediate images are *used* in a way that supports reasoning; our work provides task constructions and supervision that make imaginative perception the relevant computational substrate.

3. Spatial Imagination: Tasks and Datasets

We introduce three spatial reasoning tasks that require constructing a missing spatial representation from incomplete

Table 1. **Dataset and benchmark statistics.** All experiments use the AI2-THOR subset for training; additional data sources are released for future research. †: human-verified subset.

Task		
Perspective Taking	Source (samples)	AI2-THOR (20,531) + Habitat (19,998) + Real images (15,000)
	Train	55,529
	Eval	AI2-THOR† (238), Habitat† (300)
	IPT Format	Novel-viewpoint image
Path Tracing	Source (samples)	AI2-THOR (11,204)
	Train	11,204
	Eval	AI2-THOR† (329), Real† (332)
	IPT Format	Sideview image
Multiview Counting	Source (samples)	AI2-THOR (17,079) + MessyTable (1,880) + ScanNet (540)
	Train	19,499
	Eval	AI2-THOR† (260)
	IPT Format	Top-down BEV map

inputs (single-view, partial-view, or map inputs). For each task, we build a 10k–50k training set with paired *ground-truth spatial imaginations* (task-specific intermediate visual supervision) and final answers, and we release a human-filtered benchmark for controlled evaluation. All datasets will be released publicly; for consistency across tasks, we train our models on the AI2-THOR [19] subset of each training set. Table 1 and Fig 1 summarize the training data and evaluation benchmarks.

3.1. Perspective Taking

Given a first-person view of an indoor scene with target positions marked, the model must answer a spatial question (*e.g.*, "After moving to 'X' and turning left 90°, will the {object} be on your left or right?") about the scene from the new viewpoint. Since the target view is never provided, the model must mentally simulate the spatial transformation rather than read off the answer directly.

Sub-categories. Questions span two spatial relation types across six balanced sub-categories. *Distance change* asks whether a target object becomes closer or further after the viewpoint shift: (1) *closer* and (2) *further*. *Relative position* asks whether the object falls to the left or right in the new view, defined by the object's lateral position before and after the transformation: (3) *left→left*; (4) *left→right*; (5) *right→left*; (6) *right→right*. Overall accuracy is the unweighted mean across all six sub-categories so that each spatial relationship contributes equally, preventing models from gaming the metric by over-predicting common cases.

Imaginative perception target. A novel-viewpoint rendering of the scene from the target position, directly supervised against ground-truth renders from the 3D scene.

Data. Synthetic data is generated from AI2-THOR [19] and Habitat [24, 26, 29] by sampling source/target camera pairs, rendering first-person views, and annotating the source view with a red "X" marking the target. Questions cover two relation types (distance change, relative position) across six balanced sub-categories. A *mixed* training data variant additionally incorporates real-world examples from the Visual Spatial Tuning dataset [36] (camera motion subset) as a synthetic-to-real bridge. The base training set contains 20,531 AI2-THOR examples; the mixed variant totals 55,529. We evaluate on held-out human-verified AI2-THOR (238) and Habitat (300) benchmarks. Full subcategory breakdowns and data generation details are provided in the Appendix.

3.2. Path Tracing

Given a top-down map with a marked path $1 \rightarrow 2$, a midpoint M_1, and egocentric forward views at waypoints 1 and 2, the model must identify which object is visible on a queried side at M_1. Neither the top-down map nor the endpoint views reveal first-person visibility at the midpoint, requiring the model to imagine what the agent would see from ground level.

We evaluate under three input settings of increasing spatial cues: *Path* (map only), *PathArr* (map + query direction arrow), and *EgoDir* (map + egocentric endpoint views).

Imaginative perception target. A sideview image — a first-person rendering from M_1 — that externalizes the 3D visibility reasoning the top-down input cannot support. Ground-truth sideviews are rendered directly from the simulator at M_1.

Data. Synthetic data is generated from AI2-THOR [19] and ProcTHOR [10], sampling feasible two-waypoint paths balanced across room types and distance bins. Questions are template-generated with four answer choices and quality-filtered via TIFA-style verification [15] using GPT-4.1 majority voting; samples answerable from endpoint views alone are removed to ensure genuine imagination is required. The synthetic training set contains 11,204 examples. A real-world benchmark of 332 human-verified questions is constructed from Matterport3D [6] top-down views and evaluated on Path and PathArr settings only. Full filtering criteria and real-world annotation pipeline details are in the Appendix.

3.3. Multiview Counting

Given several first-person frames of the same environment, the model must select the correct count of a queried object (*e.g.*, "How many chairs are in this area?"). Since no single view reveals the full layout, and the same object often appears across multiple frames, the model must construct a unified spatial representation that resolves both occlusions and cross-view duplicates.

Imaginative perception target. A top-down bird's-eye view (BEV) map aggregating all input views, making deduplication explicit by mapping each object to a single spatial location. Ground-truth BEV maps are rendered from an overhead camera in the 3D scene.

Data. Synthetic examples are generated via multi-camera and rotation trajectory types. Real-world data is sourced from MessyTable [4] (fixed multi-camera rig; overhead image as ground-truth BEV) and ScanNet++ [40] (point-cloud

BEV maps converted to photorealistic overhead images via Qwen Edit [32]). Questions are four-choice MCQ with distractors sampled near the true count. The base training set contains 17,079 synthetic examples; the mixed variant totals 19,499. We evaluate on a human-verified benchmark of 260 samples. Details on trajectory types, BEV rendering, and distractor sampling are in the Appendix.

4. Method: Imaginative Perception Tokens

The core of our approach is to enable Multimodal Language Models (MLLMs) to externalize spatial reasoning through **Imaginative Perception Tokens**. Unlike standard textual chain-of-thought or methods outsourcing visual imagination with an external visual generation model, our method requires the model to generate a visual representation of a *non-observed* spatial configuration—such as an unseen viewpoint or an integrated top-down map—as a functional prerequisite for answering a spatial query.

4.1. Problem Formalization

Given an input context $\mathcal{C}$ consisting of one or more observed images $\mathcal{I}_{obs} = \{I_1, \ldots, I_k\}$ and a spatial language query Q, the goal is to predict the correct answer A. We decompose this into a two-stage generative process. First, the model generates **imaginative perception tokens** $\hat{I}_{imag}$, representing the implied spatial structure requested by the task (e.g., the view from a new coordinate): $P(\hat{I}_{imag}|\mathcal{I}_{obs}, Q)$ Second, the conditioned on this imaginative perception tokens $\hat{I}_{imag}$, the model produces the final answer: $P(A|\mathcal{I}_{obs}, Q, \hat{I}_{imag})$.

4.2. Architecture

We implement this approach using BAGEL [12], a unified decoder-only transformer that natively supports interleaved multimodal understanding and generation. BAGEL employs a Mixture-of-Transformer-Experts (MoT) design: the model utilizes two transformer experts, one optimized for multimodal understanding and another for generation. Both operate on the same token sequence through shared self-attention at every layer. Images are represented via two distinct paths. *Understanding tokens* (U) are extracted via a SigLIP2 [30] ViT encoder to capture semantic content, while *Generation tokens* (G) are latent representations from a FLUX VAE used for high-fidelity synthesis. Because all tokens (text, U, and G) coexist in a single shared context window, the model maintains lossless interaction between understanding and generation modules.

While BAGEL's standard generation tokens are typically used for open-ended text-to-image generation or editing, we repurpose this generative capacity for **spatial reasoning**. In our framework, the generation target is not a stylistic output but a precise **view imagination**—a visually grounded intermediate that represents the unobserved 3D structure of the scene.

4.3. Training and Inference

Training Objective. We optimize the framework using a multi-task loss $\mathcal{L}_{total} = \lambda_{fm}\mathcal{L}_{fm} + \lambda_{lm}\mathcal{L}_{lm}$. The model is trained to jointly produce the imaginative perception and the final answer:

1. **Flow-Matching Loss ($\mathcal{L}_{fm}$):** For the imaginative intermediate, BAGEL adopts the **Rectified Flow** method. The model learns to predict the velocity field v_t required to transform Gaussian noise into the target latent G_{gt} representing the unobserved view, conditioned on the preceding context $\mathcal{C}$:

$$\mathcal{L}_{fm} = \mathbb{E}_{t,G_0,\mathcal{C}}\left[\|v_t(G_t|\mathcal{C}) - (G_{gt} - G_0)\|^2\right] \quad (1)$$

2. **Language Modeling Loss ($\mathcal{L}_{lm}$):** We minimize the negative log-likelihood of the final VQA answer tokens A, conditioned on the observed context and the ground-truth imaginative tokens:

$$\mathcal{L}_{lm} = -\sum_{i=1}^{|A|} \log P(a_i|\mathcal{C}, U_{gt}, G_{gt}, a_{<i}) \quad (2)$$

Inference. At inference time, the model operates in one of two modes depending on the task and configuration. In the **text-only** mode, the model produces only a textual answer without generating any visual intermediate $A \sim P(A \mid \mathcal{C})$, serving as a baseline. In the **imagination** mode, the model first performs iterative denoising over VAE tokens to produce the imaginative latent: $\hat{G}_{imag} = \int_0^1 v_t(G_t \mid \mathcal{C})\, dt$ The decoded image $\hat{I}_{imag}$ is immediately re-encoded and appended to the context as both ViT understanding tokens and VAE generation tokens: $\mathcal{C}' = \left[\mathcal{C}, \text{ViT}(\hat{I}_{imag}), \text{VAE}(\hat{I}_{imag})\right]$ The model then attends to its own imagination to predict the final answer $A \sim P(A \mid \mathcal{C}')$.

5. Experiments

We evaluate *imaginative perception tokens* on the three spatial reasoning tasks introduced in Sec. 3: **Perspective Taking (PET)**, **Path Tracing (PT)**, and **Multiview Counting (MVC)**. To enable controlled comparisons, we train all task-specific models on the AI2-THOR subset of each dataset. We additionally report transfer to cross-environment benchmarks (Habitat), real-world images, and external datasets. All tasks use multiple-choice evaluation with balanced answer distributions.

PT is evaluated under three input variants that provide increasing spatial cues: **EgoDir** (egocentric direction only), **Path** (top-down path overlay), and **PathArr** (path with directional arrows) and average accuracy reported. Unless

Table 2. **Main results.** Accuracy (%) on AI2-THOR (in-domain) and different-environment (out-of-domain) benchmarks. PT reports the average across input settings (EgoDir/Path/PathArr for AI2-THOR; Real/Real+Arr for different environments). Text CoT generates a textual chain-of-thought before answering. IPT (Imaginative Perception Token) generates an intermediate image before answering. For our models, accuracy reports the **maximum** between answer-only and free-generation inference. Best per group in **bold**.

	AI2-THOR			Different Env.	
Model	PET	PT	MVC	PET	PT
VQA Models					
GPT-5	**79.8**	**60.2**	53.5	**69.3**	**80.9**
GPT-5.2	45.5	32.9	44.2	54.0	63.0
Gemini 2.5 Flash	51.0	41.5	30.8	66.3	71.4
Gemini 3 Flash	55.0	42.3	**56.9**	51.3	83.2
InternVL3.5-8B	51.5	35.8	44.6	47.7	47.4
Qwen2.5-VL-7B	50.7	37.3	38.8	54.3	44.8
Qwen3-VL-8B	52.0	35.9	43.8	46.7	64.1
Unified Models					
Janus-Pro-7B	51.8	33.5	33.1	44.7	35.3
Chameleon 7B	34.3	16.3	5.4	47.3	24.5
Ours (fine-tuned BAGEL)					
Bagel (base)	40.3	29.9	35.4	62.7	42.7
Bagel (label-only)	97.5	65.7	63.9	82.0	54.7
+ Text CoT	83.1	49.7	62.3	70.3	52.2
+ IPT	96.8	49.0	**67.3**	87.0	57.5
+ Mixed Training	**97.8**	**66.7**	62.3	**87.7**	**58.6**

otherwise stated, we report accuracy (%) and use the same prompt formatting across baselines and our models.

5.1. Setup

Baselines. We compare against two groups of models, evaluated zero-shot with task-specific prompts. *VQA models* include GPT-5, GPT-5.2, Gemini 2.5 Flash, Gemini 3 Flash, InternVL3.5-8B, Qwen2.5-VL-7B, and Qwen3-VL-8B. *Unified models* that support both understanding and generation include Janus-Pro-7B and Chameleon 7B.

Our model variants. We fine-tune BAGEL [12] under several configurations to isolate the contribution of imagination supervision. Each fine-tuned model is task-specific and trained on a single task using AI2-THOR data only (unless noted otherwise):

- **Bagel (base)**: pretrained model with no task-specific fine-tuning.
- **Bagel (label-only)**: fine-tuned with answer supervision only, with no intermediate thought.
- **+ Text CoT**: trained to generate a textual chain-of-thought describing the imagined spatial configuration before answering. Training CoTs are generated by GPT-5.1 using simulator ground-truth scene metadata.
- **+ IPT**: trained to generate an intermediate image (the imaginative perception token) before answering.
- **+ Mixed Training**: trained on a mixture of IPT examples (image-generation targets) and label-only examples (answer supervision only).

Training details. We fine-tune BAGEL-7B-MoT with AdamW (lr 1×10^{-5}, 2,000 warmup steps) on 8 GPUs using FSDP bf16, following BAGEL [12] and ThinkMorph [14]. For multi-image inputs, each image is resized to 512×512. Unless noted, IPTs use Latent-64 resolution.

5.2. Main results

Table 2 reports results on our benchmarks.

Spatial reasoning remains difficult for current VLM and unified models.

Among the zero-shot baselines, GPT-5 is the strongest across nearly all settings, yet still trails our best fine-tuned variants on multiple in-distribution tasks. Smaller open VLM models (InternVL3.5-8B, Qwen2.5-VL-7B, Qwen3-VL-8B) hover near chance on PET (50–52%) and struggle on PT, indicating that these tasks are not solvable through superficial cues. Unified models perform worse overall: Chameleon 7B drops to 34.3% on PET and 5.4% on MVC, suggesting that current unified designs often trade away understanding robustness in exchange for generation capability.

Answer supervision alone yields large gains and transfers across environments.

Bagel (label-only) substantially improves over Bagel (base) across all tasks, rising from 40.3% to 97.5% on AI2-THOR PET, from 29.9% to 65.7% on PT, and from 35.4%

Table 3. **Ablation on latent size.** Accuracy (%) with w/ Thought inference mode at different imagination resolutions. Best per column in **bold**.

Latent Size	Resolution	PET		MVC
		AI2-THOR	Habitat	AI2-THOR
Latent-4	64×64	87.4	73.3	53.5
Latent-16	256×256	95.3	81.0	56.2
Latent-32	512×512	95.0	**87.0**	58.9
Latent-64	1024×1024	**96.8**	83.3	**63.1**

Table 4. **Ablation on thought modality and inference mode.** Accuracy (%) on AI2-THOR benchmarks (PT uses EgoDir variant). We compare Text CoT vs. IPT training and vary inference mode: generate thought then answer (w/ text/image), answer directly (answer-only), or condition on ground-truth (w/ GT image).

Training	Inference	PET	PT	MVC
Text CoT	w/ text	83.1	53.1	61.5
Text CoT	answer-only	78.3	55.8	62.3
IPT	w/ image	96.8	50.4	63.1
IPT	answer-only	96.8	61.1	62.3
IPT	w/ GT image	96.7	**86.7**	**67.3**

to 63.9% on MVC. These improvements transfer: label-only reaches 82.0% on Habitat PET, showing that spatial reasoning can be learned in simulation and generalized to new environments.

Imagination supervision helps most when language is a poor interface.

On MVC, IPT achieves the best accuracy (67.3%), outperforming label-only (63.9%) and Text CoT (62.3%). On different-environment PET (Habitat), IPT reaches 87.0% (vs. 82.0% for label-only), and Mixed Training improves further to 87.7%. On PT, Mixed Training achieves the best results on both synthetic (66.7%) and real (58.6%) benchmarks, outperforming label-only (65.7% / 54.7%) and all baselines. IPT also improves real-world PT transfer (57.5%) over label-only (54.7%) and Text CoT (52.2%). Notably, IPT models are evaluated in *answer-only* mode: the model does not generate an image at inference, yet the imagination targets during training strengthen internal spatial representations that transfer across environments.

Text CoT underperforms label-only and IPT.

Text CoT typically falls behind label-only (e.g., PET 83.1% vs. 97.5%, PT 49.7% vs. 65.7%) and also behind IPT (e.g., MVC 62.3% vs. 67.3%, PET 83.1% vs. 96.8%). Compared to label-only, the Text CoT objective forces the model to allocate capacity to generating long spatial descriptions during fine-tuning, which competes with answer prediction. Compared to IPT, the gap reflects a modality mismatch: viewpoint changes, occlusions, and cross-view correspondences are difficult to serialize into natural language, and the resulting textual traces introduce noise rather than useful structure. IPT represents these relationships directly in the visual modality where they are naturally expressed.

5.3. Ablations

Latent resolution controls imagination quality and downstream accuracy.

Table 3 and Fig. 2 ablate IPT resolution on PET and MVC. At Latent-4 (64×64), imaginations are blurry and lose spatial detail; at Latent-64 (1024×1024), imaginations become sharper and more spatially faithful, preserving object identities and relative positions. Quantitatively, increasing resolution from Latent-4 to Latent-64 improves AI2-THOR PET from 87.4% to 96.8% and MVC from 53.5% to 63.1%. Habitat PET peaks at Latent-32 (87.0%) and drops slightly at Latent-64 (83.3%), suggesting mild overfitting to AI2-THOR appearance statistics at the highest resolution.

Thought modality and inference mode. Table 4 ablates the training signal (Text CoT vs. IPT) and inference mode (generate thought, answer-only, or oracle GT).

IPT training builds stronger spatial representations than Text CoT.

On PT, IPT with answer-only inference (61.1%) outperforms Text CoT with answer-only inference (55.8%) by 5.3 points. On MVC, IPT with image generation (63.1%) outperforms Text CoT with text generation (61.5%).

Imagination supervision is useful, but explicit generation is not required at inference.

For IPT models, answer-only mostly outperforms generating the imagination explicitly: on PT, answer-only reaches 61.1% vs. 50.4% with generation. For Text CoT, generating the chain-of-thought also slightly underperforms answer-only (53.1% vs. 55.8% on PT), though the gap is smaller than for IPT. This asymmetry suggests that producing faithful imaginations is harder than producing text descriptions, and imperfect generations can mislead downstream reasoning. However, training with imagination targets remains valuable: answer-only IPT matches GPT-5 on PT (61.1%).

Ground-truth imaginations reveal headroom.

When given ground-truth imaginations instead of model-generated ones, PT accuracy jumps from 50.4% to 86.7% (+36.3) and MVC rises from 63.1% to 67.3% (+4.2). The large PT gap indicates that imagination quality is the dominant bottleneck for path tracing; for PET, model-generated imaginations nearly match GT (96.8% vs. 96.7%), leaving little room for improvement.

IPT transfers to aligned external benchmarks.

Table 5 evaluates transfer to external benchmarks that test similar spatial capabilities: SAT [28] (perspective-taking subset) and MessyTable [4] (multiview counting). On SAT, Bagel (label-only) improves from 34.9% to 59.1%

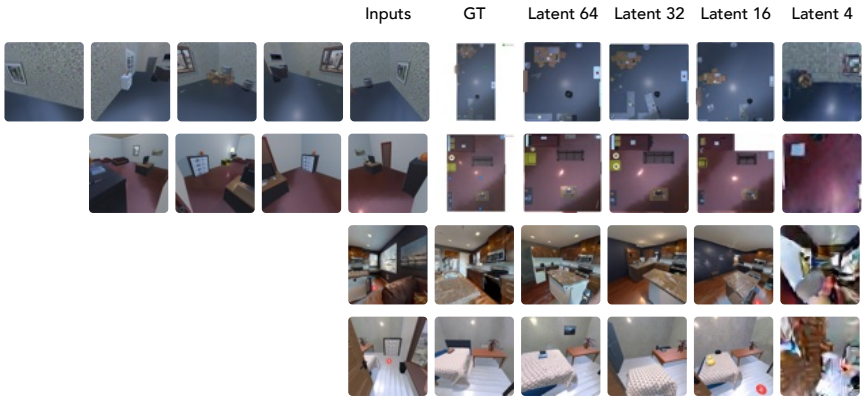

Figure 2. Qualitative examples of model-generated imaginative perception tokens. Top two rows: MVC example showing imagined top-down BEV maps. Bottom: PET examples showing imagined novel viewpoints. From left to right, imagination resolution increases from Latent-4 (64×64) to Latent-64 (1024×1024). Higher resolution produces sharper and more spatially faithful imaginations, preserving object identities and relative positions needed for downstream reasoning.

Table 5. **Transfer to similar external benchmarks.** Accuracy (%). SAT tests perspective taking and MessyTable tests multiview counting, both in domains unseen during training. Best in **bold**.

	PET	MVC
Model	SAT (66)	MessyTable (200)
Bagel (base)	34.9	29.0
Bagel (label-only)	59.1	32.5
+ Text CoT	50.0	30.0
+ IPT	57.6	28.5
+ Mixed Training	**63.6**	**37.0**

over Bagel (base), and Mixed Training further improves to 63.6%. On MessyTable, Mixed Training reaches 37.0%, up from 29.0% for Bagel (base).

Training with our data improves performance on other spatial benchmarks.

Finally, we test whether our training data improves spatial reasoning on tasks with different structures: ScanNet [9] (in-the-wild multiview counting), MindCube [41] (abstract geometric reasoning), and All-Angles-Bench [39] (cross-view matching on EgoHumans [18]). Because IPTs are task-specific by construction (e.g., rotated views for PET, bird's-eye paths for PT), they do not directly transfer to these settings. We therefore fine-tune on AI2-THOR MVC using answer supervision only. Bagel (fine-tuned) consistently improves over Bagel (base) across all three benchmarks (40.5%→52.0% on ScanNet, 39.5%→47.5% on MindCube, 40.0%→50.0% on All-Angles), indicating that our simulator data builds broadly useful spatial representations even when the specific imaginative token target changes.

Table 6. **Does our data help on other spatial tasks?** Accuracy (%) on benchmarks beyond our training task categories. Fine-tuning on our AI2-THOR MVC data consistently improves over Bagel (base), suggesting that the spatial reasoning learned from our datasets transfers broadly. Best per column in **bold**.

Model	ScanNet (200)	MindCube (200)	All-Angles (170)
VQA Models			
GPT-5	58.5	**67.3**	**67.9**
GPT-5.2	48.5	37.5	29.4
Gemini 2.5 Flash	48.0	50.3	37.9
Gemini 3 Flash	62.5	56.5	64.2
InternVL3.5-8B	53.5	42.1	54.8
Qwen2.5-VL-7B	**63.5**	47.8	51.8
Qwen3-VL-8B	62.5	34.5	42.3
Unified Models			
Janus-Pro-7B	39.5	42.0	45.0
Chameleon 7B	5.5	25.4	17.7
Ours (fine-tuned BAGEL)			
Bagel (base)	40.5	39.5	40.0
Bagel (fine-tuned)	**52.0**	**47.5**	**50.0**

6. Conclusion

We introduced Imaginative Perception Tokens (IPTs), intermediate visual representations that externalize spatial reasoning about unobserved structure in multimodal language models, and designed three tasks: Perspective Taking, Path Tracing, and Multiview Counting, with ground-truth intermediate imaginations. Training with imagination supervision consistently outperforms label-only and text chain-of-thought baselines, even without explicit imagination at inference, and ablations confirm that imagination quality directly governs downstream accuracy.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G. Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. *arXiv preprint arXiv:2412.03548*, 2024.

[3] Ellis Brown, Jihan Yang, Shusheng Yang, Rob Fergus, and Saining Xie. Benchmark designers should "train on the test set" to expose exploitable non-visual shortcuts. *arXiv preprint arXiv:2511.04655*, 2025.

[4] Zhongang Cai, Junzhe Zhang, Daxuan Ren, Cunjun Yu, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, and Chen Change Loy. Messytable: Instance association in multiple camera views. In *European Conference on Computer Vision*, pages 1–16. Springer, 2020.

[5] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments, 2017.

[7] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[8] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017.

[10] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation, 2022.

[11] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, 2025.

[12] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[13] Corentin Dumery, Noa Etté, Aoxiang Fan, Ren Li, Jingyi Xu, Hieu Le, and Pascal Fua. Counting stacked objects. In *ICCV*, 2025.

[14] Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. ThinkMorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. *arXiv preprint arXiv:2510.27492*, 2025.

[15] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering, 2023.

[16] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024.

[17] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[18] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An egocentric 3d multi-human benchmark, 2023.

[19] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[20] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025.

[21] Linnan Li, Xiaoyu Chen, Peng Chen, et al. ViewSpatial-Bench: Evaluating multi-perspective spatial under-

standing of vision-language models. *arXiv preprint arXiv:2505.21500*, 2025.

[22] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *arXiv preprint arXiv:2205.00363*, 2022.

[23] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso M. de Melo, and Alan Yuille. 3DSR-Bench: A comprehensive 3D spatial reasoning benchmark. In *ICCV*, 2025.

[24] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[25] OpenAI. Thinking with images. OpenAI Blog, 2025.

[26] Xavi Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Ruslan Partsey, Jimmy Yang, Ruta Desai, Alexander William Clegg, Michal Hlavac, Tiffany Min, Theo Gervet, Vladimír Vondruš, Vincent-Pierre Berges, John Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars and robots, 2023.

[27] Arijit Ray, Ahmed Abdelkader, Chengzhi Mao, Bryan A. Plummer, Kate Saenko, Ranjay Krishna, Leonidas Guibas, and Wen-Sheng Chu. Mull-tokens: Modality-agnostic latent thinking. *arXiv preprint arXiv:2512.10941*, 2025.

[28] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Dynamic spatial aptitude training for multimodal language models, 2025.

[29] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[30] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[32] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025.

[33] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

[34] 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. *arXiv preprint arXiv:2412.14171*, 2025.

[35] Kaiyu Yang, Olga Russakovsky, and Jia Deng. SpatialSense: An adversarially crowdsourced benchmark for spatial relation recognition. In *ICCV*, 2019.

[36] Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025.

[37] Sihan Yang, Runsen Xu, Yiman Xie, et al. MMSI-Bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025.

[38] 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.

[39] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025.

[40] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023.

[41] Baiqiao Yin, Qineng Wang, Pingyue Zhang, et al. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025.

[42] Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, and Manling Li. Theory of space: Can foundation models construct spatial beliefs through active exploration? In *International Conference on Learning Representations (ICLR)*, 2026.