# How and What to Imagine? Visual Thinking in Unified Multimodal Models for Cross-View Spatial Reasoning

**Qian Yang**[1,2], **Ankur Sikarwar**[*1,2], **Huy Le**[*1,2], **Le Zhang**[1,2], **Zhuan Shi**[1,3], **Perouz Taslakian**[1,3,4], **Aishwarya Agrawal**[1,2,5]
[1] Mila - Québec AI Institute [2] Université de Montréal [3] McGill University
[4] ServiceNow AI Research [5] Canada CIFAR AI Chair
{qian.yang, aishwarya.agrawal}@mila.quebec

## Abstract

Cross-view spatial reasoning remains a weak spot for vision-language models (VLMs): they reason in language and discard the fine-grained geometry the task requires. Thinking with images aims to fix this by generating an intermediate *thinking-image*, but recent work shows the visual evidence in these traces is largely ignored. We therefore ask **how** to make visual thinking matter, and **what kind** of visual thinking works best. We ask these questions for unified multimodal models (UMMs) that natively support interleaved image–text generation. For the **how**, we propose **View Dropout (VDrop)**, a training-time intervention that hides parts of one input view from the answer span while leaving it visible to the thinking-image tokens. This incentivizes the model to make use of the thinking-image when answering, rather than answering based on the input views only. With the thinking-image now being used in answer prediction, we ask **which kind** of visual thinking works best. We frame this as a **Learnability–Informativeness (L–I) tradeoff** and compare three thinking-image variants: top-down, panoramic, and point-matching renderings. Trained on synthetic scenes and evaluated on five real-world out-of-domain benchmarks, panoramic visual thinking with VDrop is the only configuration that is simultaneously informative and learnable, and achieves the best out-of-domain generalization.[1]

## 1 Introduction

*Cross-view spatial reasoning* requires inferring scene layout, object placement, and geometry from images taken at different viewpoints. It underlies a range of vision-language model (VLM) applications, from embodied agents navigating a room (Wang et al., 2025; Han et al., 2025) to video VLMs integrating temporally distant frames (Wu et al., 2026), all of which reduce to the same problem: maintaining a consistent scene representation across viewpoints that share only partial visual content. We study this capability in its basic form: given two partially overlapping views and a question, a VLM must reason across views to answer correctly, the format adopted by recent multi-view benchmarks (Yang et al., 2026a; Jia et al., 2026; Wang et al., 2026). Despite strong single-image performance, the strongest VLMs perform marginally above chance at cross-view spatial reasoning (Yang et al., 2026a; Jia et al., 2026). We argue this stems from a representational mismatch: cross-view reasoning is inherently *visual*, yet VLMs reason only through language, verbalizing observations into intermediates that discard the fine-grained geometry the task demands. Humans, in contrast, reason spatially in the visual domain by mentally constructing internal layouts (Tversky, 2003; Levinson, 2003; Garrod and Anderson, 1987), suggesting that letting models *think visually*, by using visual intermediates as part of the reasoning chain, is key to closing this gap. Existing think-with-image approaches realize this by generating or invoking intermediate visual representations, such as 3D reconstructions, depth maps, or predicted camera trajectories (Yang et al., 2026b; Zhang et al., 2026b; Chen et al., 2026). Yet, recent work questions whether these intermediates do real perceptual work: under controlled interventions, model predictions barely change when the visual content of the intermediate is altered, indicating that visual evidence is largely ignored (Liu et al., 2025).

Our experiments corroborate that visual evidence is under-used: with standard supervised fine-tuning (SFT), dropping the generated thinking-image at inference barely changes accuracy (Figure 3, "Visual Thinking w/o VDrop"). SFT teaches the model to *generate* a plausible thinking-image but not to *use* it when answering: the thinking-image becomes a decorative by-product of training,

[*] Equal Contribution.
[1] Code is available at https://github.com/MyLittleChange/VDrop.



*Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 32092–32111
October 24-29, 2026 ©2026 Association for Computational Linguistics

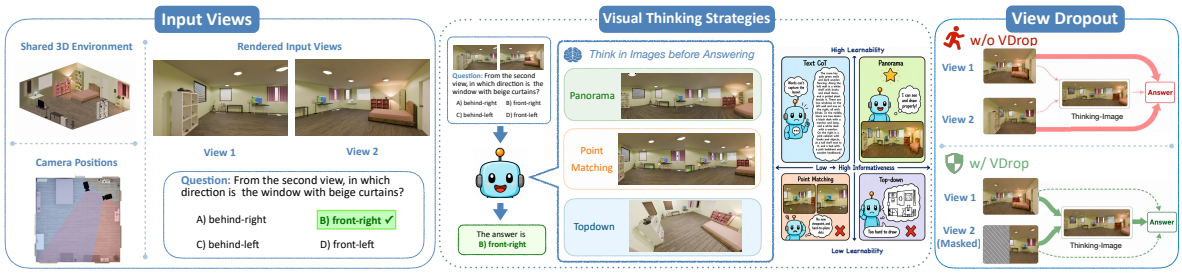



Figure 1: **Visual thinking for cross-view spatial reasoning.** Given two input views and a cross-view spatial question (**left**), a UMM can generate one of three intermediate thinking-image types (**middle**) before answering: panorama, point matching, or top-down. **Right:** without View Dropout, the answer pathway takes a shortcut through the input views, leaving the generated thinking-image unused; with View Dropout, part of one input view is masked, forcing the answer to route through the thinking-image and making visual thinking causally load-bearing.

present in form but not in function. This under-use motivates our first research question: (1) how to make visual thinking matter during learning. Once the thinking-image is genuinely used, a second question arises: (2) which kind of thinking-image is most effective for cross-view spatial reasoning, among natural candidates such as panoramic views, top-down layouts, and point-matching overlays that explicitly connect the two views. To study these questions, we use unified multimodal models (UMMs), which natively generate the thinking-image, enabling end-to-end learning of visual thinking and controlled comparison across intermediate representations within a single model.

To make visual thinking matter during learning, we propose **View Dropout (VDrop)** (Figure 1 Right), a training-time intervention that masks part of one input view from the answer span, so the only remaining path for that spatial evidence runs through the generated thinking-image. VDrop requires no architectural change and is agnostic to which thinking-image is generated. Across every thinking-image variant we test, it consistently improves cross-view spatial reasoning. With VDrop in place, candidate thinking-image representations become meaningfully comparable, and we ask which works best. The different thinking-image variants trade off along two axes that prior work has not cleanly separated: **informativeness** (how much spatial structure the thinking-image variant unveils) and **learnability** (how reliably the UMM can produce that variant). We formalize this as a Learnability–Informativeness (*L–I*) tradeoff: a thinking-image type benefits spatial reasoning only if it is both spatially informative *and* learnable from data alone, and neither axis is sufficient on its own. We instantiate this study with synthetic scenes from Infinigen Indoors, testing on BAGEL (Deng et al., 2025) and SenseNova-U1 (Diao et al., 2026), and evaluating on one in-domain synthetic benchmark and five real-world out-of-domain benchmarks. Experiments show that visual thinking improves cross-view spatial reasoning both in- and out-of-domain, and that VDrop makes the generated thinking-image causally load-bearing, consistently improving OOD performance on both UMMs. Trained on only 8K synthetic samples, our best configuration achieves a 6.7-point OOD gain over vanilla BAGEL and surpasses all prior methods we compare against, including methods trained on at least 3× more data. Once visual thinking is forced to matter, the choice of representation also matters: top-down views, though informative, are not directly learnable by current UMMs, leaving panoramic visual thinking as the only candidate that scores high on both *L–I* axes and the only one that consistently beats prior methods on OOD generalization.

Our contributions are as follows:

- **Method.** We identify under-use as a pervasive failure mode of visual thinking and propose **View Dropout**, a training-time intervention that requires no architectural change, is agnostic to the thinking-image type, and consistently improves OOD cross-view spatial reasoning across all three thinking-image variants and across two architecturally distinct UMMs.
- **Framework.** We frame the choice of visual thinking as a *Learnability–Informativeness tradeoff*, disentangling two axes that prior work has conflated: a representation may fail either because it does not encode sufficient information or because it is difficult to learn.
- **Empirical analysis.** On one in-domain synthetic benchmark and five real-world OOD benchmarks, we show that the thinking-image becomes causally

used only after VDrop training, and that panoramic visual thinking is the most informative and learnable representation. With only 8K training samples, it outperforms prior BAGEL-based visual-thinking methods trained on at least 3× more data.

## 2 Related Work

**Cross-View Spatial Reasoning.** Cross-view spatial reasoning, the task of reasoning about object positions, distances, depth ordering, and viewpoint relationships across multiple views, has emerged as a documented weakness of current VLMs (Yang et al., 2026a; Jia et al., 2026; Li et al., 2026b; Fu et al., 2024; Zhang et al., 2026a), with even strong open-source VLMs scoring only marginally above chance. Existing remedies primarily target the language pathway: spatial instruction tuning (Chen et al., 2024; Cai et al., 2026) and curated reasoning-trace fine-tuning train VLMs to verbalise spatial structure into text. Other approaches add architectural components, injecting spatial priors via depth or 3D-aware encoders (Thai et al., 2025), but still rely on language for the reasoning itself. Across both lines, the reasoning remains *linguistic*: the scene is verbalised into text, discarding the fine-grained geometry the task requires. Our work investigates a different axis, whether models can be trained to reason *visually* by generating intermediate visual representations of the scene.

**Think with Images.** Recent works show that models can "think in images" by sketching annotations on inputs or composing reasoning chains from generated images (Hu et al., 2024; Cheng et al., 2026; Xu et al., 2026). A parallel line introduces 3D-derived intermediates, such as depth maps, Gaussian splats, or 3D reconstructions (Zhang et al., 2026b; Chen et al., 2026), while others predict camera trajectories to mentally simulate unseen viewpoints (Yang et al., 2026b; Yu et al., 2026). However, the visual content in these intermediates is often ignored by the answer pathway (Liu et al., 2025), questioning whether the generated image is doing real perceptual work. Our work takes this critique as a starting point and asks two questions prior work leaves open: *how* to train models so that the thinking-image is causally used, and *which kind* of thinking-image is most effective once it is. Our fix can be seen as shortcut removal (Geirhos et al., 2020): VDrop blocks the input-view shortcut during training, while keeping the masked evidence reachable through the model's own generated thinking-image. Closest to ours, Markovian training forces the answer to route through a textual CoT (Viteri et al., 2026); VDrop enforces the same dependence on a visual intermediate.

**Unified Multimodal Models.** Unified multimodal models (UMMs) extend the single-encoder/decoder paradigm of standard VLMs to support *interleaved* image–text generation within one architecture (Xie et al., 2025; Wu et al., 2025; Chen et al., 2025; Deng et al., 2025; Liu et al., 2026b,a; Diao et al., 2026). Designs differ in how they reconcile the conflicting representational needs of understanding and generation: Janus (Wu et al., 2025; Chen et al., 2025) decouples visual encoding into two specialised pathways feeding a shared transformer; BAGEL (Deng et al., 2025) couples a multimodal understanding encoder with a diffusion-based image generator via a unified token interface; SenseNova-U1 (Diao et al., 2026) instead removes pretrained visual encoders altogether, consuming raw pixels directly and generating images via pixel flow matching rather than a VAE-based diffusion head. Recent work already uses UMMs as backbones for interleaved visual reasoning, training them to generate intermediate images during decoding (Gu et al., 2026; Li et al., 2026a). This native generation capability makes UMMs a natural testbed for our study: a single model can produce and reason over thinking-images end-to-end, enabling controlled comparison across thinking-image types without external tools. Following ThinkMorph (Gu et al., 2026), we conduct our experiments on BAGEL and further validate cross-architecture transfer on SenseNova-U1.

## 3 Method

### 3.1 Unified Multimodal Models

Given two input views $V_1, V_2$ and a textual question $q$, a UMM generates an output sequence $\mathbf{o} = (o_1, o_2, \ldots, o_T)$ where each $o_t$ is either a text token or an image token drawn from a shared vocabulary space. This allows the model to produce an intermediate visual representation $I_{\text{vt}}$ (the *thinking-image*; the subscript denotes "visual thinking") as part of its reasoning before producing the final textual answer $a$. We refer to the full process as *visual thinking*, and to a single sequence $(V_1, V_2, q) \to I_{\text{vt}} \to a$ as a *visual-thinking trace*. Given a dataset of such traces $\mathcal{D} = \{(V_1^{(i)}, V_2^{(i)}, q^{(i)}, I_{\text{vt}}^{(i)}, a^{(i)})\}_{i=1}^{N}$, supervised fine-tuning trains the UMM to generate the interleaved sequence end-to-end: first the thinking-image $I_{\text{vt}}$ conditioned on the inputs, then the answer $a$ conditioned on the inputs and the

generated thinking-image.

### 3.2 View Dropout

**Motivation.** Standard SFT supervises both the thinking-image and the answer, but does not enforce that the answer tokens *depend* on $I_{\text{vt}}$ when reasoning. Thus, the model can successfully minimise the thinking-image generation loss and the answer loss without actually making use of the thinking-image while answering, leaving the generated thinking-image as a decorative side-product. Recent analyses (Liu et al., 2025) report that predictions remain nearly unchanged under visual intervention, indicating that the visual evidence in the thinking-image is largely ignored.

**Method overview.** To force the thinking-image to be a load-bearing component of reasoning, we introduce *View Dropout (VDrop)*, a training-time intervention that hides a randomly selected contiguous region of one input view[2] from the answer tokens, while leaving the thinking-image tokens fully visible (Figure 2). Under standard SFT, the layout information needed to answer is fully available across the two input views, so the model is not compelled to rely on the thinking-image. VDrop removes this shortcut: with part of one view hidden from the answer tokens, the complete layout is recoverable only from the thinking-image, so the answer pathway must attend to it.

**Attention mask construction.** Let $M$ denote the standard per-sample attention mask, with $M_{qk} = 0$ permitting attention from query $q$ to key $k$ and $M_{qk} = -\infty$ blocking it, and let $Q_a$ denote the query positions of the answer span. We sample a primary view $v \in \{V_1, V_2\}$ uniformly and a contiguous subset $D_v$ of its patch-token positions, and edit the mask so the answer cannot read $D_v$: $M_{qk} \leftarrow -\infty$ for all $q \in Q_a$ and $k \in D_v$. For thinking-image queries, the mask entries to $V_1$ and $V_2$ are unchanged, so $I_{\text{vt}}$ generation continues to attend fully to both input views (see Figure 2).

**Region selection.** We sample $D_v$ as a contiguous axis-aligned rectangle of patch positions on the chosen view $v$. By hiding a coherent chunk of the scene rather than scattered patches, the answer pathway cannot interpolate the missing region from nearby patches and must recover it from the thinking-image. In BAGEL, each input view is encoded into two parallel token streams: ViT tokens, which carry semantic content for understanding, and VAE tokens, which carry pixel-level detail for generation. A masked region must therefore be hidden in *both* streams; masking only one would let the answer recover the region through the other. We thus mask the ViT and VAE tokens covering $D_v$ jointly. $D_v$ covers a fixed fraction $\rho$ of patch positions on view $v$.[3]

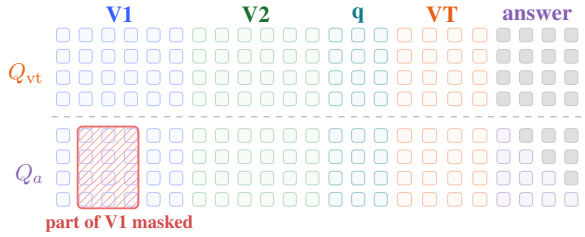



Figure 2: **VDrop attention mask.** Answer queries $Q_a$ cannot attend to the masked region (red hatched), while thinking-image queries $Q_{\text{vt}}$ retain full access to all.

**Training curriculum.** Applying the mask from the first step collapses learning: the model is asked to route evidence through $I_{\text{vt}}$ before SFT has shaped what $I_{\text{vt}}$ should encode. We therefore anneal the masking probability $p_{\text{mask}}(s)$ over training steps $s$: $p_{\text{mask}} = 0$ for $s < s_w$ (warmup), $p_{\text{mask}} = (s - s_w)/s_a$ for $s_w \le s < s_w + s_a$ (linear anneal), and $p_{\text{mask}} = 1$ thereafter. Warmup lets the model first learn to generate and use $I_{\text{vt}}$ alongside the full input; annealing then introduces VDrop pressure gradually, so a working $I_{\text{vt}} \to a$ route is in place by the time the input views are fully masked.[4]

**Compatibility.** VDrop modifies only the answer-side attention mask and leaves the SFT objective unchanged. The thinking-image is generated from the full input views and supervised toward its ground-truth render, as under standard SFT; VDrop changes only whether the answer is *forced to use* the thinking-image, not how it is generated. It is compatible with any thinking-image strategy.

### 3.3 Visual Thinking Strategies

To study which type of thinking-image is most effective for cross-view spatial reasoning, we consider three visual-thinking variants, each capturing a distinct strategy for bridging the two input views (Figure 1). **Panoramic view**: a wide-angle rendering from the observer's pose, *reconstructs the full*

[2]We mask only one of the two views; masking both removes too much spatial evidence and performs worse in our ablations (Appx. C.2).

[3]We set $\rho = 0.5$ in all main experiments. We sweep over $\rho$ and compare contiguous-region masking against random-patch masking in Appx. C.2; contiguous-region masking at $\rho = 0.5$ gives the best OOD accuracy.

[4]We use $s_{\text{w}} = 500$ and $s_{\text{a}} = 1500$; Appx. C.2 ablates these values.

| Type | Task description | Example question |
|---|---|---|
| Anchor | Identify an object visible in both views. | "Which object appears in both views?" |
| Counting | Count total instances of a given object across views. | "How many chairs are in the scene?" |
| Relative Distance | Identify the closest or farthest object from a reference. | "Which object is closest to the desk?" |
| Relative Direction | Identify the direction of an object relative to a reference. | "Which side of the sofa is the lamp on?" |

Table 1: The four cross-view question types in our 8K Infinigen training set.

*scene* so that $V_1$ and $V_2$ become sub-regions of one unified visual field. **Top-down view**: a high-angle rendering from a top corner of the room, *lifts to a shared external frame* that exposes the global layout while still revealing object sides and depth. **Point matching**: the two input views shown side by side with coloured markers on corresponding objects, *makes cross-view correspondences explicit* without changing the camera frame.

Together, these variants span a natural design space: panorama unifies the views into one scene, top-down reprojects them into a shared external frame, and point matching annotates the views in place with cross-view identity. We deliberately avoid intermediates that require auxiliary modules, such as depth maps or 3D reconstructions, to isolate the contribution of the thinking-image itself rather than the supervision signal of an external tool.

### 3.4 Training Data: Infinigen Indoors

To obtain clean training signal for each visual-thinking strategy, we construct training data from Infinigen Indoors (Raistrick et al., 2024), whose procedural 3D annotations yield unambiguous ground-truth answers and ground-truth thinking-images. Each scene provides two egocentric views with overlapping fields of view, along with the corresponding top-down, panoramic, and point-matching renderings used as ground-truth $I_{\text{vt}}$. Following the COSMIC benchmark (Sikarwar et al., 2026), we construct four cross-view question types (Table 1) that require integrating spatial information from both views. The full training set contains 7,921 QA pairs across 1,584 unique scenes; per-type descriptions, scene-generation details, and trace construction are in Appx. B.

## 4 Experiments

### 4.1 Experimental Setup

**Model and baselines.** Our main experiments fine-tune *BAGEL* (Deng et al., 2025), a SOTA open-source UMM built on a Mixture-of-Transformers architecture (14B total parameters, 7B active per token); for cross-architecture validation (§4.2) we additionally fine-tune *SenseNova-U1* (Diao et al., 2026), an encoder-free 8B UMM natively trained for interleaved vision–language generation. We compare four categories of baselines against our visual-thinking variants. First, *vanilla BAGEL* without fine-tuning, measuring the gain attributable to visual-thinking SFT. Second, two non-visual-thinking baselines fine-tuned on the same 8K synthetic Infinigen data as our visual-thinking variants: *No-Think*, which answers directly from $V_1, V_2$ and the question, and *Text CoT*, which produces a textual CoT instead of an image, annotated by prompting a strong off-the-shelf VLM with the input views and the ground-truth answer (details in Appx. B.1). These isolate the contribution of generating an intermediate *image* from fine-tuning alone or a textual intermediate. Third, two BAGEL-based visual-thinking methods that fine-tune BAGEL without VDrop: *ThinkMorph* (Gu et al., 2026), which continues training BAGEL on ∼24K interleaved reasoning traces, and *BAGEL-Zebra-CoT* (Li et al., 2026a), which fine-tunes BAGEL on 182K interleaved text-image reasoning traces. Fourth, *Qwen3-VL* (Bai et al., 2025), a strong understanding-only VLM, to contextualise the gap between standard VLMs and visual-thinking-trained UMMs. All baselines use the same multiple-choice prompt and answer-extraction protocol.

**Training hyperparameters.** We apply LoRA fine-tuning on BAGEL with rank 32 and alpha 64, training for 7,000 steps on 4×H100 GPUs. We use the Adam optimizer with a learning rate of $1 \times 10^{-5}$ and a cosine-decay schedule, and weight the cross-entropy and MSE losses equally (1.0 each). The maximum context length is 35,000 tokens for text-only training (No-Think and Text CoT) and 20,000 tokens for visual-thinking training (panoramic, top-down, and point-matching, each with and without VDrop). For SenseNova-U1, whose framework lacks LoRA support, we instead unfreeze the top six LLM layers, training for 490 steps with learning rate $5 \times 10^{-5}$ and a 14,336-token context.

**VDrop hyperparameters.** View Dropout is controlled by three hyperparameters: the warmup length $s_{\text{w}}$, the anneal length $s_{\text{a}}$, and the drop fraction $\rho = |D_v|/|V_v|$ that determines the proportion of the chosen view's patch tokens hidden from the answer span. Unless stated otherwise, we use $s_{\text{w}} = 500$, $s_{\text{a}} = 1500$, and $\rho = 0.5$, with the *contiguous region* masking strategy as the default

| | Model | VDrop | COSMIC (in-domain) | | | | Avg. | MMSI | MindCube | OmniSpatial | | STARE | BLINK | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Anchor | Count. | Rel-Dist. | Rel-Dir. | ID | Overall | Overall | CL | PT | Persp. | MultiView | OOD |
| **Understanding VLMs** | Qwen3-VL-4B | ✗ | 62.8 | 44.4 | 40.4 | 22.4 | 42.5 | 27.4 | 29.0 | 28.6 | 40.1 | 29.2 | 48.9 | 33.8 |
| | Qwen3-VL-8B | ✗ | 64.8 | 54.4 | 40.8 | 26.8 | 46.7 | 28.0 | 34.4 | 25.4 | **45.5** | 31.6 | 55.6 | 37.0 |
| **BAGEL-based prior work** | BAGEL | ✗ | 18.0 | 42.1 | 24.4 | 21.2 | 26.4 | 26.9 | 31.7 | 31.1 | 38.9 | 28.0 | 45.1 | 33.3 |
| | BAGEL-Zebra-CoT (182K) | ✗ | 8.8 | 24.4 | 28.4 | 24.8 | 21.6 | 23.2 | 21.7 | 29.0 | 43.0 | 28.4 | 24.8 | 26.8 |
| | ThinkMorph (24K) | ✗ | 49.6 | 43.6 | 32.0 | 30.0 | 38.8 | 26.5 | 39.2 | 33.7 | 44.6 | 28.8 | 52.6 | 37.2 |
| **BAGEL w/ 8K non-visual** | No-Think | ✗ | 86.8 | 82.4 | 67.6 | 85.6 | 80.6 | 27.4 | **41.1** | 30.2 | 44.0 | 24.8 | 45.1 | 35.1 |
| | Text CoT | ✗ | 56.0 | 60.4 | 44.4 | 40.8 | 50.4 | 24.6 | 25.3 | 29.4 | 35.8 | 26.0 | 54.9 | 32.7 |
| **BAGEL w/ 8K visual thinking** | Panoramic | ✗ | 93.6 | **83.6** | 76.4 | 87.2 | 85.2 | 24.9 | 36.9 | 32.9 | 43.9 | 32.4 | 55.6 | 37.6 |
| | | ✓ | 89.2 | 78.8 | 74.8 | 93.2 | 84.0 | 26.0 | 34.1 | **38.9** | 45.3 | **35.6** | **62.4** | **40.0** (+2.4) |
| | Point Matching | ✗ | 92.8 | 81.2 | 73.2 | **94.4** | **85.4** | 27.8 | 35.2 | 31.0 | 41.7 | 24.0 | 52.6 | 35.2 |
| | | ✓ | 92.8 | 78.8 | **78.0** | 91.6 | 85.3 | 28.3 | 34.1 | 33.7 | 43.3 | 34.4 | 45.1 | 36.1 (+0.9) |
| | Top-down | ✗ | 93.6 | 83.2 | 68.4 | 92.0 | 84.3 | 28.8 | 35.2 | 31.4 | 39.8 | 28.0 | 58.7 | 37.3 |
| | | ✓ | **94.4** | 76.8 | 72.0 | 89.2 | 83.1 | **32.0** | 36.5 | 32.5 | 44.2 | 26.0 | 57.1 | 38.0 (+0.7) |

Table 2: **Comparison with baselines and VDrop ablation across thinking-image types.** The two blue **Avg.** columns report the in-domain mean over the four COSMIC subtasks and the out-of-domain mean over five real-world benchmarks. Per column, the **best** and second-best accuracy are highlighted.

selection rule. The primary view $v$ is sampled uniformly from $V_1$ and $V_2$ at each training step. On SenseNova-U1, we keep the same masking strategy and $\rho$, and scale the curriculum to its shorter run ($s_w = 50$, $s_a = 100$), so both models train under full masking for $\approx$70% of their steps.

**Evaluation Benchmarks.** We evaluate on one ID benchmark, **COSMIC** (Sikarwar et al., 2026), built on Infinigen scenes from our training domain, and five real-world OOD benchmarks covering diverse cross-view spatial reasoning skills. **MMSI-Bench** (Yang et al., 2026a) poses expert-authored questions requiring spatial reasoning across multiple images of a scene (overall split). **MindCube** (Wang et al., 2026) tests building a coherent spatial mental model from partial, incrementally revealed views (MindCube-Tiny, 1,050 samples). **OmniSpatial** (Jia et al., 2026) covers higher-order relational reasoning and non-egocentric viewpoints (Complex Logic and Perspective Taking subsets, averaged). **STARE-Perspective** (Li et al., 2026b) requires reasoning about object relations from a viewpoint other than the camera's. **BLINK-MultiView** (Fu et al., 2024) tests integrating evidence across multiple images of one scene. Per-benchmark details are in Appx. A.

### 4.2 How to Imagine? Does View Dropout Make Visual Thinking Matter?

We now test whether VDrop converts the generated thinking-image into a load-bearing component of reasoning, as designed in §3.2.

**Impact of view dropout and visual thinking.** Table 2 contrasts three visual-thinking strategies trained under standard SFT (§3.1) against the same strategies trained with VDrop.[5] **(1)** *VDrop consistently improves OOD cross-view spatial reasoning across all three thinking-image types:* Adding VDrop raises the OOD average for every strategy: panoramic 37.6 $\rightarrow$ 40.0 (+2.4; significant under a paired bootstrap, $p = 0.018$, Appx. C.1), top-down 37.3 $\rightarrow$ 38.0 (+0.7), and point matching 35.2 $\rightarrow$ 36.1 (+0.9). The effect is positive across all three, with panoramic benefiting most. With VDrop, panoramic visual thinking reaches an OOD average of 40.0%, a 6.7-point gain over vanilla BAGEL (33.3%) and above Qwen3-VL-8B (37.0%), a strong general-purpose VLM. **(2)** *With only 8K training samples, VDrop outperforms prior methods trained on far more data:* Panoramic visual thinking with VDrop (40.0% OOD, 8K samples) outperforms prior BAGEL-based visual-thinking methods despite their much larger fine-tuning sets: ThinkMorph (37.2%, 3$\times$ the samples) and BAGEL-Zebra-CoT (26.8%, 23$\times$ the samples). Since these methods fine-tune the same BAGEL backbone, the results show that having the right intermediate representation and the training recipe (VDrop) matters more than the scale of fine-tuning data. **(3)** *Visual intermediates outperform non-visual baselines on both in- and out-of-domain:* All three visual-thinking variants with VDrop beat both non-visual baselines on the ID and OOD averages. The OOD gains are substantial: relative to No-Think (35.1%), panoramic gains +4.9 points ($p < 0.001$, paired bootstrap; Appx. C.1), top-down +2.9, and point matching +1.0, with larger gains over Text CoT (+7.3, +5.3, +3.4, respectively). On ID, the margin is smaller, as fine-tuning BAGEL on our 8K set already reaches 80.6% without visual thinking (No-Think) and visual thinking lifts this only to 83–85%: the ID task is largely

[5] OmniSpatial reports two splits (CL and PT); we average them into a single score that contributes equally with the other four benchmarks to the out-of-domain mean.

| Setting | COSMIC (in-domain) | | | | Avg. | MMSI | MindCube | OmniSpatial | | STARE | BLINK | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Anchor | Count. | Rel-Dist. | Rel-Dir. | ID | Overall | Overall | CL | PT | Persp. | MultiView | OOD |
| SenseNova-U1 (no FT) | 68.8 | 31.2 | 44.8 | 30.8 | 43.9 | 29.0 | **44.8** | 27.0 | 33.5 | **33.2** | 38.4 | 35.1 |
| Panoramic SFT | **72.8** | **48.4** | 48.8 | 50.0 | **55.0** | 28.6 | 40.6 | 26.2 | 42.4 | 30.0 | 37.6 | 34.2 |
| + VDrop | 66.8 | 45.6 | **49.2** | **52.4** | 53.5 | **31.0** | 42.3 | **29.4** | **45.1** | 31.6 | **41.4** | **36.7** (+2.5) |

Table 3: **Cross-architecture validation on SenseNova-U1.** Panoramic visual thinking trained with standard SFT vs. with VDrop. The two blue **Avg.** columns report the in-domain mean over the four COSMIC subtasks and the out-of-domain mean over five real-world benchmarks. Best per column in bold.

| Setting | MMSI | MindCube | OmniSpatial | | STARE | Avg. |
|---|---|---|---|---|---|---|
| | Overall | Overall | CL | PT | Persp. | OOD |
| SenseNova-U1 (no FT) | 29.0 | **44.8** | 27.0 | 33.5 | **33.2** | 34.3 |
| No-Think | **32.0** | 42.2 | 26.2 | **50.5** | 27.6 | 35.0 |
| Panoramic SFT | 28.6 | 40.6 | 26.2 | 42.4 | 30.0 | 33.4 |
| + VDrop | 31.0 | 42.3 | **29.4** | 45.1 | 31.6 | **35.5** |

Table 4: **No-Think control on SenseNova-U1 (OOD, BLINK excluded).** Best per column in bold.

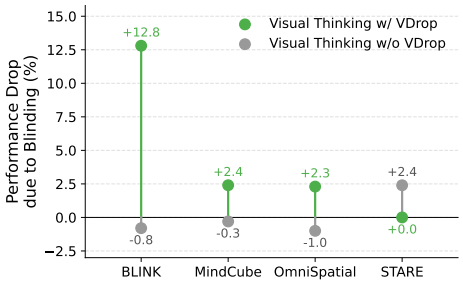



Figure 3: **Generate-then-blind probe across 4 OOD benchmarks.** Accuracy drop when the generated thinking-image is blinded at answer time; a larger drop means more dependence on the thinking-image. VDrop-trained models show larger drops on three benchmarks.

solved by fine-tuning alone, so we treat the OOD average, far from saturated, as the primary metric. Notably, Text CoT performs *worse* than No-Think: a textual CoT hurts rather than helps. We attribute this to inconsistent textual supervision, as the same 3D scene admits many valid descriptions, making the generated traces often imprecise (Appx. B.1). The VDrop hyperparameter ablation is in Appx. C.2.

**Does VDrop transfer to other UMM architectures?** On SenseNova-U1, fine-tuned on the same spatial training data and evaluated on the same benchmarks, the headline finding mirrors BAGEL: VDrop again delivers the best OOD generalization (Table 3). However, Panoramic SFT alone brings no OOD gain here (34.2% vs. 35.1% for the untrained model): SenseNova-U1 starts from a stronger base, and, as the underuse thesis predicts, SFT teaches the model to *draw* the panorama but not to *read* it. VDrop closes this gap, raising the OOD average to 36.7% (+2.5 over Panoramic SFT), with ID gains concentrated on the two relational subtypes that most require cross-view reasoning (Rel-Dir and Rel-Dist). VDrop also outperforms a *No-Think* control trained on the same data with the thinking-image removed (35.5 vs. 35.0 on the OOD average excluding BLINK; Table 4).[6]

**Generate-then-Blind: Is the impact of thinking-image causal?** Performance gains alone do not imply that the answer *causally* depends on the generated thinking-image: the model could produce a decorative image whose only effect is regularisation. To test whether the thinking-image causally affects answer prediction, we introduce a *generate-then-blind* intervention at inference. For each test item, the model first generates the thinking-image as in normal inference; before answer decoding begins, we then mask the answer span's attention to the thinking-image tokens. The answer span still attends to $V_1$, $V_2$, and the question, but can no longer read the thinking-image: the model still "thinks" by producing the image, yet answers without consulting it. A model that genuinely uses the thinking-image should lose accuracy under blinding; one that ignores it should be unaffected.

We apply this probe to two panoramic-trained variants, standard visual-thinking SFT and SFT with VDrop (Figure 3; full setup in Appx. C.4). The VDrop variant shows substantial accuracy drops under blinding on most OOD benchmarks while standard SFT is largely invariant, confirming that VDrop makes the thinking-image load-bearing rather than decorative. A per-category breakdown on MMSI (Figure 5 in Appx. C.4) shows VDrop's blinding effect concentrates on questions answerable by visually aligning the two views, confirming its causal dependence on the thinking-image appears where visual reasoning helps.

**Mechanism check: Does VDrop increase answer-token attention to the thinking-image?** We measure how the answer tokens divide their attention among the visual inputs during answer generation. At each decoder layer, we compute how

[6] BLINK is excluded because the No-Think checkpoint collapses to a constant predictor there; see Appx. C.3 for detailed analysis.

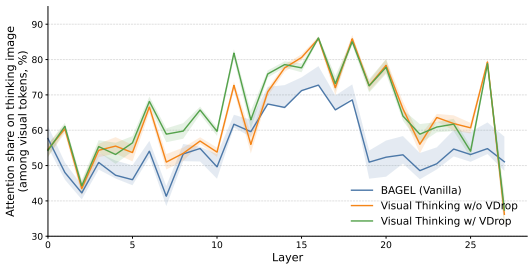



Figure 4: **Mean answer-token attention on thinking-image tokens across decoder layers (BLINK).** VDrop-trained model places more attention on the generated thinking-image than standard SFT model, especially in early and mid layers. Shaded bands show the interquartile range across samples (middle 50%).

much the answer attends to the generated thinking-image as a fraction of its total attention to all visual content, i.e. the thinking-image plus the two input views $V_1, V_2$. A higher fraction means the answer relies more on the thinking-image than on the input views (full setup in Appx. C.5).[7] Averaging across all BLINK samples and decoder layers (Figure 4), the thinking-image's share of the answer span's visual attention rises across the three models: $55.3\%$ for vanilla BAGEL, $63.6\%$ for standard SFT, and $65.2\%$ with VDrop. The gap widens in early and mid decoder layers, where VDrop attends $3.9$ percentage points more to the thinking-image than standard SFT.[8] We observe the same early-to-mid-layer pattern on STARE (Appx. C.5).

***Finding 1.*** Visual thinking improves cross-view spatial reasoning: generating an intermediate thinking-image outperforms non-visual baselines on both ID and OOD benchmarks.

***Finding 2.*** VDrop consistently improves cross-view spatial reasoning across all three thinking-image types and across two architecturally distinct UMMs (BAGEL and SenseNova U1), increasing the answer span's attention to the thinking-image and making it a load-bearing component of reasoning.

### 4.3 What to Imagine? A Learnability – Informativeness Tradeoff

With VDrop ensuring the thinking-image is genuinely used, we now ask which thinking-image type yields the largest gains. We formalise the end-to-end benefit of a type $T$ along two axes.

• **Informativeness** $I(T)$: how much a *perfect* instance of $T$ would reduce the model's reasoning burden on the target task. A panorama exposes the joint scene directly and a top-down view exposes the global layout, whereas point matching only marks correspondences within the original views and adds no new viewpoint.

• **Learnability** $L(T)$: how reliably a UMM can be trained to produce faithful instances of $T$ from the available data. A panorama is a single coherent rendering the model can generate directly, whereas a faithful top-down view requires synthesising an unseen viewpoint, and precise point markers must be placed exactly on small target objects.

We hypothesise that the benefit of visual thinking is bounded by these two axes: a thinking-image type helps only if it is both informative *and* learnable, with neither alone sufficient. This choice of axes is grounded in imitation learning: visual-thinking SFT trains the model to reproduce a target thinking-image, i.e., behavioural cloning of the target (Stiennon et al., 2020; Ouyang et al., 2022), whose theory bounds the cloned policy by two independent error terms: *(i)* the value of the imitated target, since cloning is at best as good as the target it imitates (Rajaraman et al., 2020), and *(ii)* the realizability of that target under the learner, since a target the model class cannot represent leaves an irreducible misspecification error (Espinosa Dice et al., 2025). These correspond exactly to $I(T)$ and $L(T)$ (details in Appx. C.6). We measure each axis in isolation on a paired subset of the COSMIC test set[9] by controlling *which* thinking-image the answerer sees: ground-truth renders isolate the target's value from generation quality, while generated images expose how much of that value the model realises; any gap between the two measurements is the model failing to realise the target.

**Informativeness: oracle measurement.** We feed *ground-truth* renders of each thinking-image type as a third image to two off-the-shelf VLMs, Qwen3-VL-32B and Qwen3-VL-235B (Table 5). The informativeness ranking, panorama $>$ top-down $>$ point matching, is consistent across both VLMs. *Panorama* delivers the largest uplift, concentrated on the relative-distance and relative-direction sub-

[7] Vanilla BAGEL does not reliably emit a thinking-image; we force one by injecting the image-generation token, giving all three models a comparable thinking-image span.

[8] The three shares average over all decoder layers; the 3.9-point figure averages over the first 14 layers, where the effect concentrates.

[9] The $L$–$I$ analysis uses the paired subset of $n$=720 COSMIC test examples for which ground-truth thinking-images of all strategies could be rendered (requiring complete source-mesh data); the subset spans the four subtypes near-uniformly.

| Condition | Anchor | Counting | Rel-Dist | Rel-Dir | Overall |
|---|---|---|---|---|---|
| *Qwen3-VL-32B-Instruct* | | | | | |
| Input views only | 83.33 | 73.22 | 56.32 | 39.66 | 62.78 |
| $T_{\text{pano}}$ ($\Delta$) | −5.95 | +1.64 | **+12.63** | **+25.14** | **+8.61** |
| $T_{\text{topdown}}$ ($\Delta$) | −5.36 | **+3.83** | +5.26 | +13.97 | +4.58 |
| $T_{\text{point mat.}}$ ($\Delta$) | **+2.98** | −10.38 | −0.53 | +2.23 | −1.53 |
| *Qwen3-VL-235B-A22B-Instruct* | | | | | |
| Input views only | 73.21 | 66.12 | 49.47 | 39.11 | 56.67 |
| $T_{\text{pano}}$ ($\Delta$) | **+0.60** | **+9.84** | **+15.26** | **+16.76** | **+10.83** |
| $T_{\text{topdown}}$ ($\Delta$) | −6.54 | −1.09 | +10.00 | +8.38 | +2.92 |
| $T_{\text{point mat.}}$ ($\Delta$) | 0.00 | −9.29 | 0.00 | +6.15 | −0.83 |

Table 5: **Oracle measurement of $I(T)$ on COSMIC**. For each VLM block, the *Input views only* row reports accuracy (%) given just $V_1$ and $V_2$; subsequent rows report uplift $\Delta$ over that baseline when a third image of type $T$ is supplied.

types whose answers require a viewpoint neither input view provides. *Top-down* yields a smaller but consistently positive uplift, exposing global spatial relations through an allocentric view. *Point matching* is flat-to-negative on both: its annotations stay in the original viewpoint and add little geometric structure beyond the input views. Anchor moves little under *any* intermediate: it is already near-saturated from the input views alone (the Qwen3-VL-32B answerer reaches 83.33 on Anchor with just $V_1$ and $V_2$), leaving little headroom for a third image, whereas the other subtypes start far lower and gain much more. This dip is a property of the probe, not of the trained UMM: end-to-end, the panoramic variants still beat direct answering on Anchor (93.6/89.2 vs. 86.8; Table 2).

**Learnability: generation quality.** We assess whether BAGEL can produce faithful instances of each type, using two measurements computed on the *same* set of BAGEL-generated thinking-images, so generation fidelity and functional usefulness are evaluated on identical images. ***Directly***, we compute SigLIP cosine similarity between BAGEL-generated thinking-images and their ground-truth renders. Panorama and top-down generations are the most faithful, with near-identical similarity (0.950 and 0.948), both above point matching (0.928). But SigLIP captures whole-image appearance, not geometric fidelity: a generated top-down view can resemble its target room while misplacing objects or distorting depth (see Appendix C.7 for qualitative examples). We therefore treat it as a necessary but insufficient signal, and turn to a functional measurement. ***Indirectly***, we feed BAGEL's *generated* thinking-image to a frozen VLM (Qwen3-VL-235B) and measure whether it still helps answer cross-view questions (Table 6): a learnable strategy should yield images the VLM

| Condition | Anchor | Counting | Rel-Dist | Rel-Dir | Overall |
|---|---|---|---|---|---|
| Input views only | 73.21 | 66.12 | 49.47 | 39.11 | 56.67 |
| $T_{\text{pano\_gen}}$ ($\Delta$) | −13.69 | **+3.28** | +2.63 | **+6.70** | 0.00 |
| $T_{\text{td\_gen}}$ ($\Delta$) | −13.10 | −2.19 | **+8.42** | +2.79 | −0.69 |
| $T_{\text{point mat.\_gen}}$ ($\Delta$) | −0.60 | −2.19 | −3.68 | −0.56 | −1.81 |

Table 6: **Generated-image measurement of $L(T)$ on COSMIC.** The *Input views only* row reports accuracy (%) given $V_1$ and $V_2$ on Qwen3-VL-235B-A22B-Instruct; subsequent rows report the uplift $\Delta$ from supplying a thinking-image of type $T$, generated by BAGEL after VDrop training, as a third image.

can use as spatial evidence. Generated panoramas remain net-positive on all three spatial subtypes (Counting, Rel-Dist, Rel-Dir), showing BAGEL transfers their spatial *structure* even from generated images; the lone regression is on Anchor, which benefits little even at the oracle level. Generated top-down images help on two spatial subtypes but regress on the others, leaving overall accuracy below baseline, so BAGEL captures the top-down layout only partially. Point matching is flat-to-negative throughout, consistent with its low oracle informativeness rather than a generation failure.

***Finding 3.*** On the L–I framework, top-down is informative but only partially learnable, while point matching is limited mainly by low informativeness and only moderately learnable, as its fine markers are hard to place precisely. Panorama alone is strong on both axes.

## 5 Conclusion

We studied how to make visual thinking matter for cross-view spatial reasoning in UMMs, and which kind of visual thinking is most effective. We identified *under-use* as a pervasive failure of current visual-thinking pipelines and proposed **View Dropout (VDrop)**, a training-time intervention that forces the generated thinking-image to become a load-bearing component of reasoning. With visual thinking made load-bearing, we framed the choice of intermediate as a *Learnability–Informativeness tradeoff* and identified panoramic visual thinking with VDrop as the only configuration that is simultaneously informative, learnable, and used. Trained on only 8K samples, it achieves the best out-of-distribution generalization, outperforming prior methods on the same backbone trained on at least 3× more data. The bottleneck for visual thinking is not data scale, but a training signal that forces the thinking-image to be used.

## Limitations

First, our study focuses on a single task family, cross-view spatial reasoning, where text-based reasoning demonstrably fails and visual imagination is inherently required; whether VDrop and the Learnability–Informativeness framework extend to other visual-thinking tasks remains to be tested. Second, VDrop makes the thinking-image causally load-bearing, but it does not by itself improve the *quality* of the generated thinking-image: when the generated image is low-fidelity or not genuinely useful for the question, forcing the answer to route through it provides little benefit. VDrop is therefore complementary to, not a substitute for, supervision that improves thinking-image generation itself; combining VDrop with higher-fidelity thinking-image targets is a natural direction for future work.

## Acknowledgements

We thank the Mila IDT team for providing the computational resources for this research, their technical support and for maintaining the Mila compute cluster. We also thank Calcul Québec and Digital Research Alliance of Canada for providing the computational resources for this research. Throughout this project, Aishwarya Agrawal received support from the Canada CIFAR AI Chair award.

## References

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, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, and 10 others. 2026. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. 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.

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

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. 2026. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2624.

Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, and Libo Qin. 2026. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *Advances in Neural Information Processing Systems*, 38:96084–96112.

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. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, Yan Li, Yubo Wang, Zhijie Cao, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Yuwei Niu, Yue Zhu, Bo Liu, and 39 others. 2026. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*.

Nicolas Espinosa Dice, Sanjiban Choudhury, Wen Sun, and Gokul Swamy. 2025. Efficient imitation under misspecification. In *International Conference on Learning Representations*, volume 2025, pages 81466–81504.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. 2026. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In *The Fourteenth International Conference on Learning Representations*.

Leekyeung Han, Hyunji Min, Gyeom Hwangbo, Jonghyun Choi, and Paul Hongsuck Seo. 2025. Dialnav: Multi-turn dialog navigation with a remote guide. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 8514–8523. IEEE.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, XinQiang Yu, Jiawei He, He Wang, and Li Yi. 2026. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *The Fourteenth International Conference on Learning Representations*.

Stephen C Levinson. 2003. *Space in language and cognition: Explorations in cognitive diversity*, volume 5. Cambridge University Press.

Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. 2026a. Zebra-cot: A dataset for interleaved vision-language reasoning. In *The Fourteenth International Conference on Learning Representations*.

Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. 2026b. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. In *The Fourteenth International Conference on Learning Representations*.

Zhiheng Liu, Weiming Ren, Xiaoke Huang, Shoufa Chen, Tianhong Li, Mengzhao Chen, Yatai Ji, Sen He, Jonas Schult, Tao Xiang, Wenhu Chen, Ping Luo, Luke Zettlemoyer, and Yuren Cong. 2026a. Tuna-2: Pixel embeddings beat vision encoders for unified understanding and generation. *arXiv preprint arXiv:2604.24763*.

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, Viktar Atliha, Tony Ng, Xiao Han, Chuyan Zhu, Chenyang Zhang, Ding Liu, Juan-Manuel Perez-Rua, Sen He, Jürgen Schmidhuber, and 6 others. 2026b. Tuna: Taming unified visual representations for native unified multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15740–15751.

Zujing Liu, Junwen Pan, Qi She, Yuan Gao, and Guisong Xia. 2025. On the faithfulness of visual thinking: Measurement and enhancement. *arXiv preprint arXiv:2510.23482*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. 2024. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21783–21794.

Nived Rajaraman, Lin Yang, Jiantao Jiao, and Kannan Ramchandran. 2020. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33:2914–2924.

Ankur Sikarwar, Debangan Mishra, Sudarshan Nikhil, Ponnurangam Kumaraguru, and Aishwarya Agrawal. 2026. Communicating about space: Language-mediated spatial integration across partial views. *arXiv preprint arXiv:2603.27183*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Anh Thai, Songyou Peng, Kyle Genova, Leonidas Guibas, and Thomas Funkhouser. 2025. Splattalk: 3d vqa with gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4712–4721.

Barbara Tversky. 2003. Structures of mental spaces: How people think about space. *Environment and behavior*, 35(1):66–80.

Scott Viteri, Max Lamparth, Peter Chatain, and Clark Barrett. 2026. Markovian transformers for informative language modeling. In *International Conference on Learning Representations*, volume 2026, pages 15823–15843.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. 2026. Mindcube: Spatial mental modeling from limited views. In *The Fourteenth International Conference on Learning Representations*.

Yipu Wang, Yuheng Ji, Yuyang Liu, Enshen Zhou, Ziqiang Yang, Yuxuan Tian, Ziheng Qin, Yue Liu, Huajie Tan, Cheng Chi, Zhiyuan Ma, Daniel Dajun Zeng, and Xiaolong Zheng. 2025. Towards cross-view point correspondence in vision-language models. *Preprint*, arXiv:2512.04686.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. 2025. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. 2026. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *Advances in neural information processing systems*, 38:13569–13597.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2025. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pages 28240–28264.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. 2026. Visual planning: Let's think only with images. In *The Fourteenth International Conference on Learning Representations*.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. 2026a. MMSI-bench: A benchmark for multi-image spatial intelligence. In *The Fourteenth International Conference on Learning Representations*.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. 2026b. Mindjourney: Test-time scaling with world models for spatial reasoning. *Advances in Neural Information Processing Systems*, 38:109855–109885.

Shoubin Yu, Yue Zhang, Zun Wang, Jaehong Yoon, Huaxiu Yao, Mingyu Ding, and Mohit Bansal. 2026. When and how much to imagine: Adaptive test-time scaling with world models for visual spatial reasoning. *arXiv preprint arXiv:2602.08236*.

Le Zhang, Jihan Yang, Soundarya Krishnan, Jimit Majmudar, Xiou Ge, Prasoon Puri, Prathamesh Nandkishor Saraf, Shruti Bhargava, Dhivya Piraviperumal, Yinan Ling, Cindy Pan, Hong Yu, Aishwarya Agrawal, and Bo-Hsiang Tseng. 2026a. From where things are to what they are for: Benchmarking spatial-functional intelligence in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12052–12063.

Zaibin Zhang, Yuhan Wu, Lianjie Jia, Yifan Wang, Zhongbo Zhang, Yijiang Li, Binghao Ran, Fuxi Zhang, Zhuohan Sun, Yizhuang Peng, Zhenfei Yin, Lijun Wang, and Huchuan Lu. 2026b. Think3d: Thinking with space for spatial reasoning. *arXiv preprint arXiv:2601.13029*.

# Appendix

**Appendix Overview**


A Evaluation Benchmarks
B Training Data
C Experiments


## A Evaluation Benchmarks

We evaluate on one in-domain (ID) benchmark and five real-world out-of-domain (OOD) benchmarks covering diverse cross-view spatial reasoning skills. All benchmarks use multiple-choice questions, and we report accuracy. Unless otherwise specified, we parse each model output to extract the predicted answer token and apply exact-match scoring against the ground-truth answer. [10]

**ID Benchmark.** **COSMIC** (Sikarwar et al., 2026) is a cross-view spatial reasoning benchmark built on Infinigen-generated scenes, making it in-domain with respect to our training data. COSMIC covers two levels of spatial reasoning. Object-level tasks include *anchor recognition* (identifying objects shared across views) and *global counting* (aggregating object instances across views while correctly disambiguating shared and unique instances). Relation-level tasks include *relative distance* (inferring which object is closest or farthest from a target distributed across views) and *relative direction* (inferring the egocentric direction of a target object absent from the answerer's view, requiring cross-view perspective transformation). Each subtask contains 250 samples, with no scene overlap between the evaluation set and our training samples.

**OOD Benchmarks.** We evaluate OOD generalisation on five benchmarks covering real-world environments and diverse spatial reasoning skills.

**MMSI-Bench** (Yang et al., 2026a) is a multi-image spatial intelligence benchmark containing 1,000 challenging multiple-choice questions, each meticulously crafted by six 3D-vision researchers and paired with carefully designed distractors and a stepwise reasoning process. MMSI-Bench is highly challenging: the strongest open-source models achieve roughly 30% accuracy and GPT-5 reaches 41.9%, while humans score 97.2%. We report the *Overall* accuracy across all MMSI question types as the headline MMSI score. Following the official recommendation, we use Gemini-3.0-Flash to extract the predicted answer from each model output, and then apply exact-match scoring between the extracted answer and the ground-truth answer for each multiple-choice question.

**MindCube** (Wang et al., 2026) tests whether models can build spatial mental models from partial observations. It evaluates three core spatial settings: *Rotation* (interpreting multiple orthogonal views from a static observation point, requiring holistic understanding despite incremental visibility shifts), *Around* (leveraging occlusion to test object permanence and the ability to convert lateral relations in frontal views into depth cues in side views), and *Among* (maintaining spatial consistency across views captured around a central object, requiring models to deduce overall spatial arrangement when not all elements are simultaneously visible). We evaluate on MindCube-Tiny, a stratified subset of 1,050 samples balanced across the three settings.

**OmniSpatial** (Jia et al., 2026) is a comprehensive spatial reasoning benchmark for VLMs covering a broad set of spatial skills. We evaluate on two subsets relevant to cross-view spatial reasoning: *Complex Logic* (CL), which involves higher-order reasoning about relations, transformations, and geometric structure; and *Perspective Taking* (PT), which probes the ability to reason about a scene from a non-egocentric viewpoint. We report each subset individually in the main results table and an unweighted average of the two as the OmniSpatial headline score.

**STARE-Perspective** (Li et al., 2026b) is the perspective-taking split of the STARE benchmark, which evaluates a model's ability to reason about object positions and relations relative to a non-egocentric viewpoint specified in the question.

**BLINK-MultiView** (Fu et al., 2024) is the multi-view split of BLINK, which tests visual reasoning that requires integrating evidence across multiple images of the same scene rather than from a single image alone.

## B Training Data Details

The synthetic Infinigen source is chosen for two reasons: (i) procedural rendering provides ground-truth thinking-images for every variant in §3.3

[10] The artifacts used in this work, including the base model and associated resources, are released under the Apache-2.0 license. Our use is limited to research on cross-view spatial reasoning and follows the artifacts' stated terms and intended use.

(top-down maps, panoramic stitches, and point-matching overlays) free of real-world rendering noise, and (ii) full access to object-level 3D annotations (positions, bounding boxes, categories) lets us automatically construct cross-view spatial questions whose answers are unambiguous.

**Scene generation.** Infinigen Indoors (Raistrick et al., 2024) generates diverse photorealistic indoor scenes, each populated with a structured layout of furniture, architectural elements, and everyday objects. For every scene we render two egocentric views with deliberately overlapping fields of view: the two camera poses are sampled such that a portion of the scene is co-visible across views, ensuring that shared objects or shared regions anchor the two viewpoints and make a layout connection between them feasible. Alongside each view pair, we render the top-down bird's-eye-views, stitched panoramic views, and point-matching overlays used as ground-truth $I_{\text{vt}}$ for the variants in §3.3.

**Question types.** We automatically generate four types of cross-view spatial questions, each targeting a distinct spatial skill:

- **Anchor.** Given two cross-view images, identify the common object(s). This tests cross-view correspondence: the model must match object identity despite viewpoint-induced appearance changes such as occlusion, scale, and aspect-ratio shifts.
- **Counting.** Given two cross-view images, count the total number of instances of a specified object category. This requires both cross-view correspondence and enumeration: the model must count instances in each view and resolve which are shared to avoid double-counting.
- **Relative Distance.** Given two cross-view images and a set of target objects, determine which object is farthest from a reference object. This requires 3D metric layout recovery, estimating inter-object distances from 2D projections.
- **Relative Direction.** Given two cross-view images, determine the direction of an object visible in one image relative to the viewpoint of the other. This requires reference-frame transformation: localising an object in 3D from one view and projecting its direction into the other view's coordinate frame.

| Question Type | Samples | Unique Scenes |
|---|---|---|
| Anchor | 730 | 455 |
| Counting | 1,191 | 854 |
| Relative Distance | 3,000 | 192 |
| Relative Direction | 3,000 | 893 |
| *Total* | *7,921* | *1,584* |

Table 7: Training data distribution by question type. All training samples are procedurally rendered from Infinigen Indoors (Raistrick et al., 2024); no external real-world data is used.

**Training data distribution.** Table 7 reports the per-type sample and unique-scene counts for the 8K Infinigen training set. We deliberately allocate fewer samples to Anchor and Counting than to the relative-distance and relative-direction types. Anchor recognition is implicitly exercised whenever the model answers any cross-view question, and Counting concerns objects largely visible within the overlapping region of the two views rather than genuine cross-view spatial relations. Both are reflected in the higher zero-shot accuracy of off-the-shelf VLMs on Anchor and Counting than on the relational subtypes (Table 5). We therefore reallocate sample budget toward relative distance and relative direction, the harder subtypes, while retaining sufficient accuracy on Anchor and Counting at the reduced counts.

### B.1 Text Chain-of-Thought Annotation

Each training sample comes with a ground-truth multiple-choice answer derived from the underlying 3D scene, but no natural-language rationale. To produce one, we prompt an off-the-shelf large multimodal model, Qwen3-VL-235B-A22B-Instruct, conditioning it on the two camera images and a category-specific textual prompt. The prompt has three parts: a short instruction block specifying the reasoning shape for that question category, one curated in-context example, and a scene-metadata block synthesised from the QA row (the per-scene list of visible objects with short descriptions, and the two camera poses).

A central design choice is an *oracle–trace separation*. The raw numeric quantities used to construct the question (angles, distances, exact object counts, and the gold answer letter) are passed to the annotator as a tagged private "cheat sheet", so that it lands on the correct option with high proba-

bility; the system prompt, however, forbids citing these quantities in the rationale. The trace must instead argue from visual cues a student model could verify from the images alone: qualitative placements (e.g. mid-centre-right), foreground/background contrasts, direction words, and, for view-dependent categories, shared landmarks together with a relative-camera-pose hint that lets the trace argue which side an off-screen target lies on.

We use an open-source model rather than a proprietary one such as Gemini for two reasons. Each of the 8K samples requires conditioning on two input images and a long oracle-information context, so annotating the full set with a proprietary API would be costly; an open-source annotator also keeps the pipeline fully reproducible. Despite the oracle-conditioning protocol, the synthesised text traces are not consistently high-quality: producing a linguistically consistent description of a 3D scene is difficult, since the same spatial configuration admits many equally valid descriptions, yielding surface-form inconsistency across training samples. Visual thinking sidesteps this: a ground-truth panoramic or top-down rendering is a deterministic function of the scene, so the supervision is consistent by construction. This may partially explain why Text CoT underperforms even No-Think in our end-to-end results (§4.2). Designing higher-quality interleaved reasoning supervision that combines visual and textual chains-of-thought is a promising direction for future work.

# C Experiments

## C.1 Confidence Intervals and Significance Tests

We assess the robustness of the cross-strategy ranking in Table 2 with bootstrap resampling over test items.

**Paired significance tests.** For each pair of conditions, we run a *paired* bootstrap over per-item accuracy differences: within each benchmark we resample its questions with replacement and recompute the accuracy difference between the two conditions on the same resampled items, then average the five benchmark-level differences with equal weight, following the OOD protocol of Table 2 (OmniSpatial's CL and PT splits are first combined into a single OmniSpatial score). Both key comparisons are significant: Panoramic + VDrop beats No-Think on the OOD average by $+4.9$ points (95% CI $[+2.5, +7.1]$, $p < 0.001$), and beats Panoramic without VDrop by $+2.4$ points (95% CI $[+0.5, +4.7]$, $p = 0.018$). The pairing removes item-level variance shared between conditions, which is appropriate here because all conditions are evaluated on identical test items.

**Per-condition confidence intervals.** Table 8 reports per-condition 95% percentile-bootstrap confidence intervals (10,000 resamples, resampling items) for every in-domain subtype and OOD benchmark. The VDrop interval on the OOD average, $[38.1, 42.4]$, does not overlap the No-Think interval $[33.0, 37.2]$ and only marginally overlaps the Panoramic-without-VDrop interval $[35.5, 39.9]$; the paired tests above, which account for the shared items, confirm both gaps are significant.

## C.2 VDrop Mask Design Ablation

Having established in §4.2 that VDrop helps across thinking-image types, we ablate which axis of the mask is responsible for the gain. Holding the thinking-image type (Panoramic) and training recipe (LoRA SFT, 8K samples) fixed, we vary three axes of VDrop: the patch-selection *Strategy*, **Region** (a contiguous bounding-box of the chosen view's patches) versus **Random** (an i.i.d. random subset at the same drop ratio); the masked *Scope*, **One view** (a single primary view sampled uniformly from $\{V_1, V_2\}$) versus **Two views** (both input views masked together); and the *drop ratio*. A no-mask reference (Panoramic without VDrop) and a no-warmup ablation (masking from step 0, no anneal) are reported for comparison. Table 9 reports aggregate ID and OOD accuracy for each variant. The default configuration used throughout the paper, **Region** masking on a single view at 50% drop with the warmup–anneal curriculum, achieves the best OOD accuracy (40.0). Four findings stand out. **(i) Region beats Random:** replacing the contiguous region with random patches at the same drop ratio costs 4.4 OOD points, indicating that spatially coherent occlusion is what forces the thinking-image to encode localised structure. **(ii) One view beats two:** masking both views simultaneously costs 3.2 OOD points, suggesting that retaining one full view as an anchor is necessary for the model to learn what content needs routing through $I_{\text{vt}}$. **(iii) An intermediate drop ratio is best:** 50% outperforms both 30% (+7.4 OOD) and 80% (+3.5 OOD); too little masking leaves the input-view shortcut intact, while too much removes

| Condition | COSMIC (in-domain) | | | | MMSI | MindCube | OmniSpatial | | STARE | BLINK | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Anchor | Count. | Rel-Dist. | Rel-Dir. | Overall | Overall | CL | PT | Persp. | MultiView | OOD |
| No-Think | 86.8 [82.4, 90.8] | 82.4 [77.6, 86.8] | 67.6 [61.6, 73.6] | 85.6 [81.2, 89.6] | 27.4 [24.7, 30.2] | 41.1 [38.1, 44.0] | 30.2 [24.6, 36.1] | 44.0 [39.9, 48.1] | 24.8 [19.6, 30.4] | 45.1 [36.8, 53.4] | 35.1 [33.0, 37.2] |
| Text CoT | 56.0 [49.6, 62.0] | 60.4 [54.4, 66.4] | 44.4 [38.4, 50.4] | 40.8 [34.8, 47.2] | 24.6 [22.0, 27.3] | 25.3 [22.8, 28.0] | 29.4 [23.8, 34.9] | 35.8 [31.9, 39.9] | 26.0 [20.8, 31.6] | 54.9 [46.6, 63.2] | 32.7 [30.4, 34.7] |
| Panoramic View | 93.6 [90.4, 96.4] | 83.6 [78.8, 88.0] | 76.4 [71.2, 81.6] | 87.2 [83.2, 91.2] | 24.9 [22.3, 27.6] | 36.9 [33.9, 39.8] | 32.9 [27.4, 38.9] | 43.9 [39.8, 48.0] | 32.4 [26.8, 38.4] | 55.6 [47.4, 63.9] | 37.6 [35.5, 39.9] |
| + VDrop | 89.2 [85.2, 92.8] | 78.8 [73.6, 83.6] | 74.8 [69.2, 80.0] | 93.2 [90.0, 96.0] | 26.0 [23.3, 28.7] | 34.1 [31.2, 37.0] | 38.9 [32.9, 45.2] | 45.3 [41.2, 49.4] | 35.6 [29.6, 41.6] | 62.4 [54.1, 70.7] | **40.0** [38.1, 42.4] |

Table 8: **Per-condition** 95% **bootstrap confidence intervals** (10,000 resamples, resampling items within each benchmark or subtype). The OOD **Avg.** is the equal-weight mean over the five OOD benchmarks; OmniSpatial's CL and PT splits are first combined into a single OmniSpatial score, so the average is over benchmarks, not over the displayed columns. All four conditions fine-tune the same BAGEL backbone on the identical 8K training set and differ only in the thinking strategy.

| Strategy | Scope | Drop $\rho$ % | ID | OOD |
|---|---|---|---|---|
| *No VDrop (reference)* | | | | |
| None | n/a | 0 | 85.2 | 37.6 |
| *+ VDrop (mask variants)* | | | | |
| Region | One view | 30 | 85.2 | 32.6 |
| Region | One view | 50 | 84.0 | **40.0** |
| Region | One view | 80 | **85.9** | 36.5 |
| Region | Two views | 50 | 83.8 | 36.8 |
| Random | One view | 50 | 83.1 | 35.6 |
| *+ VDrop, no warmup* | | | | |
| Region | One view | 30 | 81.9 | 32.2 |
| Region | One view | 50 | 81.6 | 33.7 |

Table 9: **VDrop ablation** on Panoramic visual thinking. **ID** is the mean over the four COSMIC subtasks; **OOD** is the out-of-domain mean defined in §4.1. Bold marks the best per column.

spatial evidence the model needs. **(iv) Warmup is essential:** applying full masking from step 0 without the warmup–anneal schedule drops OOD accuracy by 6.3 points at the same drop ratio, consistent with the curriculum motivation in §3.2: the model must first learn what to put in $I_{\text{vt}}$ before being forced to depend on it.

### C.3 No-Think Ablation on SenseNova-U1

As a further control for the cross-architecture validation (§4.2), we train a *No-Think* SenseNova-U1 checkpoint on the same spatial data with the thinking-image removed from the target sequence (direct answering), using the same partial-unfreeze recipe as the other SenseNova-U1 variants (§4.1). Table 4 compares it against the untrained model, Panoramic SFT, and Panoramic SFT with VDrop on the OOD benchmarks.

We exclude BLINK from this comparison because the No-Think checkpoint's BLINK score does not measure multi-view reasoning: the checkpoint collapses to a constant predictor, answering "B" for every question regardless of its content. The collapse persists under temperature sampling at two seeds, and when we randomly swap the A/B option contents per question, the checkpoint still answers "B" on nearly all items and its accuracy falls to chance. This collapse occurs only for the No-Think checkpoint; no visual-thinking checkpoint exhibits it. On the remaining four benchmarks, VDrop is the strongest setting (35.5), above No-Think (35.0), the untrained model (34.3), and Panoramic SFT (33.4). Moreover, VDrop improves over Panoramic SFT on every individual benchmark.

### C.4 Generate-then-Blind Probe: Setup and Details

The generate-then-blind probe in §4.2 asks a causal question about the role of the generated thinking-image at inference: does the answer pathway genuinely depend on it, or does it attend primarily to the original input views regardless of what the thinking-image contains? We compare two models that differ only in their SFT objective: *(i)* **Standard SFT**, the visual-thinking SFT pipeline with no VDrop, where the thinking-image tokens contribute to the generation loss so the model is trained both to produce the image and to answer conditioned on it; *(ii)* **VDrop SFT**, the same data and recipe with VDrop applied during training, so the answer span must route masked spatial evidence through the generated thinking-image. Both models start from the same BAGEL initialisation and are fine-tuned on the same 8K Infinigen training set (§3.4).

**Intervention procedure.** For each test item we (a) generate the thinking-image autoregressively as in normal inference, letting its tokens enter the KV cache, and (b) before answer decoding begins, set the attention weights from answer queries to thinking-image positions to $-\infty$ before softmax. The answer span therefore attends only to $V_1$, $V_2$, and the question, so the model still "thinks" by producing the image but cannot read it back when answering.

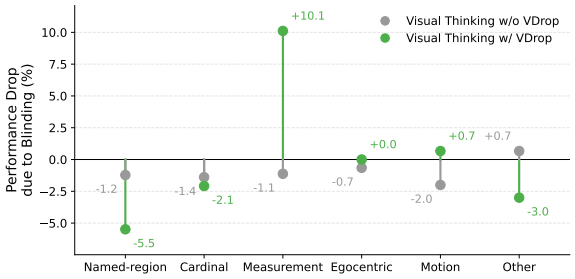



Figure 5: **Generate-then-blind probe on MMSI, by question evidence category.** Accuracy drop when the generated thinking-image is blinded at answer time; a larger positive value means more dependence on the thinking-image. The VDrop-trained model shows a large drop only on Measurement, whose questions are answered by visually aligning the two input views, while standard SFT is unaffected throughout.

**Evaluation.** We compare each blinded run against the same model's unblinded baseline on four OOD benchmarks (Figure 3). In-distribution accuracy on COSMIC is largely saturated by visual-thinking SFT and is therefore an insensitive test of causal dependence; the OOD benchmarks, where models still have substantial headroom, provide a cleaner signal for whether the thinking-image is genuinely load-bearing.

**MMSI breakdown.** MMSI's questions ask for spatially heterogeneous evidence, so a single overall accuracy obscures whether the generated thinking-image helps. We re-partition the 1,000 MMSI questions into six disjoint categories by the kind of evidence each requires. *Measurement* questions need visual comparison ("which is wider", "how many"), exactly the operation enabled by aligning $V_1$ and $V_2$ in one panorama. *Egocentric* questions ask for viewer-relative direction, which a wider-field-of-view panorama makes visible. *Named-region* and *Cardinal* questions depend on symbolic information a panorama cannot encode: room labels ("kitchen") and absolute compass frames ("north of the desk"). *Motion* questions concern camera or object movement, and *Other* covers the remainder, mostly non-spatial counting. Each question is assigned to the first matching category in the order Motion, Named-region, Cardinal, Measurement, Egocentric, Other.

Figure 5 reports the blinding-induced accuracy drop per category. Standard SFT is unaffected in every category (within $\pm 2$ pp of zero, max $|z| = 0.6$[11]), confirming the thinking-image is inert without VDrop. The VDrop-trained model shows a large, significant drop on Measurement ($+10.1$ pp, $z = 2.24$, $p = 0.025$), the category answered by visually aligning the two views; on Named-region and Cardinal, whose answers a panorama cannot encode, blinding has no positive effect. VDrop thus teaches the model to genuinely use the thinking-image, and this reliance surfaces precisely on questions that call for visual reasoning, while standard training leaves it largely ignored.

### C.5 Answer-Token Attention Probe: Setup and Details

The attention probe in §4.2 asks where the answer span's visual attention is directed during answer generation, and whether VDrop measurably increases the share directed to the model's own generated thinking-image.

**Models compared.** We extract attention weights from the answer-generation step across all decoder layers and compare **Standard SFT** (visual-thinking SFT without VDrop) and **VDrop SFT** (visual-thinking SFT with VDrop), both trained on the same data. We additionally include **vanilla BAGEL** as a reference. Vanilla BAGEL is not trained for visual thinking and does not reliably emit a thinking-image before answering; to obtain a comparable thinking-image span, we force generation by injecting the image-generation token into the decoding stream, after which it produces a thinking-image and then an answer. This makes the thinking-image span well-defined for all three models, so the share metric below is computed identically across them. We note that vanilla BAGEL's forced thinking-image is not optimised for the task and serves only as an untrained reference point.

**Probe metric.** To focus on visual evidence specifically, we normalise attention over the named visual spans only: the two input views $V_1$ and $V_2$, plus the generated thinking-image visual tokens $\mathrm{VT}$. For each decoder layer $\ell$ and each evaluation example, the quantity of interest is the per-layer *thinking-image share among visual evidence*:

$$\rho_{\mathrm{vt},\ell} = \frac{\mathrm{attn}_\ell(\mathrm{VT})}{\mathrm{attn}_\ell(V_1) + \mathrm{attn}_\ell(V_2) + \mathrm{attn}_\ell(\mathrm{VT})}$$

[11] We test each category's blinding effect against zero with a paired test over per-question accuracy deltas; $z$ is the effect divided by its standard error, and $|z| \gtrsim 2$ corresponds to $p < 0.05$.

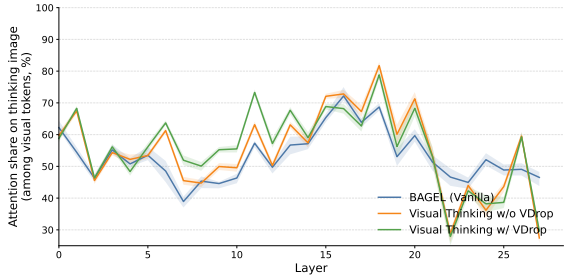



Figure 6: **Mean answer-token attention on thinking-image tokens across decoder layers (STARE).** The VDrop-trained model places more attention on the generated thinking-image than the standard SFT model, especially in early and mid layers, indicating that VDrop shifts the answer pathway toward the thinking-image.

where $\mathrm{attn}_\ell(\cdot)$ is the answer-token attention mass on the named visual span at layer $\ell$, averaged across all answer-token positions and attention heads. By construction $\rho_{\mathrm{vt},\ell} \in [0, 1]$, with higher values indicating that the answer query relies more on the generated thinking-image than on the input views at layer $\ell$. We report the mean of $\rho_{\mathrm{vt},\ell}$ across all evaluation examples.

**STARE-Perspective: per-layer attention share.** Figure 6 reproduces the $\rho_{\mathrm{vt},\ell}$ measurement on 250 STARE-Perspective examples. The pattern matches BLINK in the early and middle layers: averaged over the first 14 layers of the decoder, VDrop places $+3.5$ pp more attention on the thinking-image than standard SFT. Over all layers the gain is smaller ($+0.9$ pp over standard SFT), as the effect is concentrated early, with the two models converging in the late decoder layers. VDrop's increased engagement with the thinking-image is therefore localised to the early-to-middle layers rather than spread uniformly across the network, and this localisation is consistent across BLINK and STARE.

### C.6 Theoretical Grounding of the L–I Framework

This section expands the imitation-learning view of the Learnability–Informativeness framework introduced in §4.3.

**Visual-thinking SFT is behavioural cloning of the target intermediate.** SFT trains the UMM to reproduce a *target* thinking-image from the inputs and then answer conditioned on it, i.e., it clones a demonstrated intermediate rather than discovering one, exactly as instruction tuning clones demonstrated responses (Stiennon et al., 2020; Ouyang et al., 2022). Imitation-learning theory bounds the value of a cloned policy by two *independent* error terms. First, cloning cannot exceed the demonstrator it imitates: the learned policy's value is bounded by the value of the imitated target, up to terms that vanish as imitation becomes exact (Rajaraman et al., 2020). Second, when the target is not realizable by the learner's model class, an irreducible misspecification error remains no matter how much data is used (Espinosa Dice et al., 2025).

**The two bounds are the two axes.** **Informativeness** $I(T)$ instantiates the first term: it measures the value of a *perfect* instance of type $T$ for the downstream answer, which upper-bounds what any model cloning $T$ can gain. We isolate it by feeding *ground-truth* renders to a frozen VLM answerer (Table 5), removing the UMM's generation quality from the measurement entirely. **Learnability** $L(T)$ instantiates the second term: it measures how much of that target the UMM actually realises. We isolate it by feeding the UMM's *generated* images to the same frozen answerer (Table 6); any drop from the ground-truth to the generated measurement is realisation error, not a deficiency of the representation itself.

**The framework is falsifiable and practical.** The imitation-learning view yields a testable prediction: a thinking-image type helps end-to-end only if both terms are favourable, and the two probes, which are cheap to run before any end-to-end training, should predict the end-to-end ranking. Our results confirm this: top-down is informative but only partially learnable, point matching is learnable but uninformative, and only panorama scores on both axes, matching the end-to-end ordering in Table 2. When a strategy underperforms, the probes also pinpoint which axis is the bottleneck, indicating whether to improve the target representation or the model's ability to generate it.

### C.7 Qualitative Analysis

Figure 7 shows the thinking-image our VDrop-trained model generates under each of the three strategies (Panoramic, Point Matching, Top-down) on one example per subtask, together with the question and four options (gold option in green). Below each thinking-image we mark the predicted answer and whether it matches the gold label. The examples illustrate both axes of the L–I analysis (§4.3). On *informativeness*: panoramic and top-

down thinking-images render the scene from a genuinely new viewpoint, directly revealing the joint spatial layout that relative-distance and relative-direction questions require. Point matching instead keeps both input views in their original perspective and overlays correspondence markers; it still helps, by linking objects across the two views, but it conveys cross-view geometry only *indirectly* rather than exposing the full scene directly. On *learnability*: the point-matching examples also reveal a failure specific to this strategy. The correspondence markers are only a few pixels wide, far smaller than the large-area object relationships that panoramic and top-down images convey, and are correspondingly harder to generate reliably. In the Anchor and Relative-Distance examples, the model places a marker on an object in one view but omits the matching marker in the other, breaking the correspondence the strategy depends on. Together this matches the aggregate trend in §4.2, where panoramic visual thinking yields the largest out-of-domain gains.




Figure 7: **Qualitative examples of visual thinking across strategies.** Four samples, one per subtask (Anchor, Counting, Relative Distance, Relative Direction). Each row shows the question and four options (gold option in green), followed by the two input camera views and the generated thinking-image under each strategy (Panoramic, Point Matching, Top-down View). The predicted answer letter and correctness (✓ / ✗) are shown below each thinking-image.