Despa: Resolving Spatial Collapse in VLMs via Depth-Grounded Geometry

Yujing Lou[1,2], Pingyi Chen[2,3], Shen Cao[2], Lubin Fan[2,†], Yue Wu[2], Lizhuang Ma[1], Jieping Ye[2]
[1]Shanghai Jiao Tong University [2]Alibaba Token Hub, Alibaba Group
[3]Zhejiang University

Abstract

Vision-Language Models (VLMs) excel at open-world 2D perception but struggle with precise metric spatial reasoning, a key requirement for applications like autonomous driving, navigation, and embodied intelligence. We identify this as a structural limitation: images are 2D projections of the 3D world, inherently suffering from projective ambiguity, while VLM components favor semantic understanding and rely on 2D positional bias, leading to Spatial Collapse. To address this, we propose **Despa**, a **De**pth-grounded **spa**tial VLMs that injects depth-based geometric information via Geometric Positional Embedding (GPE) and Depth-aware Rotational Positional Embedding (DoPE). To expose masked limitations in current benchmarks, we construct SpaDataset (1.7M QA pairs) and SpaBench (2,051 QAs), to our knowledge the largest fine-grained training corpus and benchmark suite for metric spatial reasoning. They cover three difficulty levels and ten sub-tasks across diverse settings. The fine-tuned Despa-4B model consistently outperforms general-purpose, closed-source, and specialized VLMs on SpaBench by 27.9% over Qwen3-VL-235B-A22B and 31.1% over GPT-5.2, generalizes to MSMU (66.3%), RefSpatial-U (42.4%), and the multi-frame VSI-Bench (66.0%), and consistently lifts LLaVA-1.5, Qwen2.5-VL, and Qwen3-VL backbones, achieving the performance with only a 4B model.

1 Introduction

Vision-Language Models (VLMs) have advanced rapidly in open-world 2D perception and language grounding [1–10], and their capabilities are evolving from 2D image understanding toward 3D spatial intelligence. When applied to tasks that demand metric quantities, e.g., judging whether a robotic arm can reach a target, whether a vehicle can pass through a narrow passage, or estimating relative distances during navigation, a natural question arises: *how far are current VLMs from precise metric spatial reasoning?* Recent evaluations consistently show that even the most advanced VLMs fall short on such tasks [11–14]. A common hypothesis is that scaling visual-instruction data will eventually close this gap. We argue, however, that the deficit is not primarily a data problem: it is rooted in a structural Spatial Collapse of standard VLM architectures, in which both ingredients through which geometry can enter self-attention are 2D in nature.

To achieve precise metric spatial reasoning, VLMs should capture accurate geometries from the inputs. However, the input images/videos themselves does not provide an explicit 3D coordinate system. Instead, current VLM architectures prioritize semantic invariance over geometric fidelity. Specifically, a self-attention block captures geometry from visual context through exactly two ingredients: the visual token features and the rotary position embeddings that modulate their inner products. In current general VLMs both are 2D, which decomposes spatial collapse into two facets. (1) Token-side

[†]Corresponding author.

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

collapse. Standard vision encoders such as CLIP and SigLIP are trained for semantic alignment under explicit scale and translation invariance [15, 16], so visual tokens preserve strong semantics while compressing and entangling metric quantities such as absolute scale and exact depth. The issue compounds in the multi-frame case: the same physical point seen from two viewpoints projects to different patches with no shared geometric anchors to align them consistently. (2) Position-side collapse. Vanilla RoPE is designed to distinguish sequential positional shifts. Multimodal RoPE allocates every channel to one of time-height-width (t, h, w), so the relative phase between two visual tokens depends on $(\Delta t, \Delta h, \Delta w)$ alone. Two patches that sit at nearly the same image position but on opposite sides of an occlusion boundary, with arbitrarily large depth gap Δz, are still treated as one neighborhood by self-attention. Distinguishing them requires depth as a third axis that the rotary spectrum can actually read.

Rather than forcing VLMs to implicitly rediscover depth from 2D supervision, we propose Despa, a depth-grounded extension of standard VLMs that re-enters metric 3D geometry at exactly these two loci, with minimal disturbance to the rest of the pretrained backbone: (1) Geometric Positional Embedding (GPE) attaches a metric anchor to every visual token and fuses it into the token features at the multimodal interface. The anchor is the per-patch depth in the single-image case (Depth Positional Embedding, DPE) and the per-pixel world coordinate, jointly estimated over the entire clip, in the video case (World-Coord Positional Embedding, WPE), so multi-view consistency is enforced at the input rather than left to attention to recover. (2) Depth-aware Rotational Positional Embedding (DoPE) integrates depth into the rotary positional encoding that self-attention uses, so the relative phase between two visual tokens reflects their 3D offset including the depth gap, restoring the model's ability to discriminate tokens that lie close in the image plane but far apart in depth. In practice we realize this with a lightweight channel reallocation that overwrites only a small subset of low-frequency RoPE channels with depth, so that the pretrained (t, h, w) priors are kept largely intact.

Spatial collapse exists especially on tasks that strictly require metric quantities. Existing spatial benchmarks are dominated by qualitative questions, e.g., "above/below", that strong semantic priors can answer without geometric grounding, which masks the underlying deficit. To expose and stress-test metric reasoning, we construct SpaBench, a benchmark of 2,051 evaluation questions spanning three difficulty levels and ten fine-grained metric sub-tasks, grouped into Fundamental Metric Perception, Relational Metric Reasoning, and Interactive Metric Inference. Its companion training corpus SpaDataset provides 1.7M QA pairs derived from nearly 5M physical metric annotations across indoor, outdoor, synthetic, and real-world scenes, simultaneously achieving large scale and high metric density.

Extensive experiments validate Despa. Fine-tuned on SpaDataset, Despa-4B reaches 61.8% on SpaBench, outperforming the open-source Qwen3-VL-235B-A22B[17] by 27.9 points, the closed-source GPT-5.2[18] by 31.1 points, and the spatial specialist SenseNova-SI-1.1[19] by 36.0 points, with the largest gains on depth-sensitive tasks (Route, Point Distance, Point Depth) where every other baseline collapses. Trained on the corresponding splits, Despa-4B further reaches 66.3% on MSMU[20], 42.4% on RefSpatial-U[21], and 66.0% on the multi-frame VSI-Bench[22], beating domain specialists (SD-VLM, RoboRefer, Cambrian-S) and 235B-class generalists, and consistently lifts LLaVA-1.5 (1D RoPE)[5], Qwen2.5-VL (chunked M-RoPE)[8], and Qwen3-VL (interleaved M-RoPE)[17] backbones on SpaBench, confirming that depth grounding at both loci is a generalizable solution rather than a backbone-specific trick. Our contributions are summarized as follows:

- We identify Spatial Collapse as the structural cause of metric reasoning failures in current VLMs, and decompose it into token-side collapse and position-side collapse.
- We propose Despa, an extension of standard VLMs that injects metric 3D geometry through Geometric Positional Embedding (GPE) and Depth-aware Rotational Positional Embedding (DoPE), making self-attention 3D-aware with minimal disturbance to the pretrained priors.
- We construct SpaDataset and SpaBench, to our knowledge the largest and most metric-dense training corpus and benchmark suite for fine-grained metric spatial reasoning, designed to expose deficits that qualitative benchmarks hide.
- Across multiple benchmarks and three RoPE families, Despa consistently surpasses general-purpose, closed-source, and spatial-specialist baselines, achieving state-of-the-art performance with a 4B backbone.

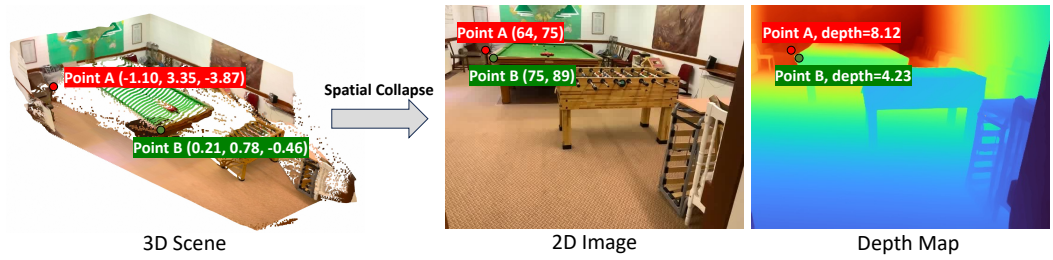

Figure 1: **Spatial Collapse**. Two adjacent points A and B in an image may have huge 3D spatial positional divergence (e.g., world coordinates and depth).

2 Related Work

Spatial VLMs. Despite significant progress in visual reasoning, VLMs are fundamentally constrained in 3D spatial cognition. To bridge this gap, a prominent line of research has explored direct 3D representations by integrating point cloud encoders with large language models [23–28]. Another paradigm leverages video sequences as the primary input modality to acquire spatial understanding through temporal correspondence and multi-view geometry [22, 29–31]. Complementary to complex modalities, an emerging paradigm focuses on depth-enhanced 2D perception. They treat depth maps as additional visual inputs to strengthen their vision embeddings [32, 13, 33, 20]. Recent works demonstrate that spatial intelligence can be substantially scaled through systematic data curation and multi-task fine-tuning [34, 35, 19].

Positional Encoding Architectures. Vanilla RoPE [36] processes multi-modal inputs as 1D sequences. Multimodal RoPE (M-RoPE) [8, 17, 37] extends rotary embedding into temporal, height, and width dimensions, yet fusing the temporal axis into high-frequency channels impairs long-range dependencies. Subsequent work addresses this via frequency reallocation, VideoRoPE [38] with low-frequency temporal allocation, VRoPE [39] with balanced spatio-temporal encoding, and long-range modeling such as Hyperbolic RoPE (HoPE) [40]. Nevertheless, these methods still lack genuine 3D spatial awareness. LieRE [41] generalizes RoPE to n-dimensional inputs via Lie group rotations, yet depth remains implicitly obscured in planar projections. Depth Positional Encoding (DPE) [20] is therefore proposed to compensate for this geometric loss with explicit depth cues.

3 Preliminaries

A decoder self-attention block consumes a token sequence $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ formed by concatenating visual and text tokens, and produces queries, keys, and values via

$$\mathbf{Q} = f_{rope}(\mathbf{W}_Q\,\mathbf{X};\,\mathbf{p}), \quad \mathbf{K} = f_{rope}(\mathbf{W}_K\,\mathbf{X};\,\mathbf{p}), \quad \mathbf{V} = \mathbf{W}_V\,\mathbf{X}, \tag{1}$$

which then feed standard scaled dot-product attention. Here $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_N)$ collects the RoPE position descriptors. For simplicity, here use $f_{rope}(\mathbf{x}_i, p_i)$ as the operation on each token, which is obtained by rotating each pair in a 2D subspace:

$$f_{rope}(\mathbf{x}_i, p_i) = \mathbf{R}_{p_i}\mathbf{x}_i, \tag{2}$$

where $\mathbf{R}_{p_i} \in \mathbb{R}^{D\times D}$ is a block-diagonal rotation matrix formulated as $\mathbf{R}_{p_i} = \mathrm{diag}(\mathbf{r}_{p_i}^{(0)}, \ldots, \mathbf{r}_{p_i}^{(D/2-1)})$. Each 2×2 block $\mathbf{r}_{p_i}^{(j)}$ corresponds to the j-th frequency band:

$$\mathbf{r}_{p_i}^{(j)} = \begin{pmatrix} \cos p_i\theta_j & -\sin p_i\theta_j \\ \sin p_i\theta_j & \cos p_i\theta_j \end{pmatrix}, \tag{3}$$

where $\theta_j = b^{-2j/D}$ is the rotation frequency. Eq. (1) thus exposes exactly two loci at which geometry can enter the block: (i) the token features $\mathbf{X}$ that define $\mathbf{Q}, \mathbf{K}, \mathbf{V}$, and (ii) the channel-allocation rule inside f_{rope} that maps each $\mathbf{p}_i$ to per-channel rotation angles.

Visual Tokenization. A vision encoder E_v followed by an MLP projector turns image $\mathbf{I}$ into flattened visual tokens $\{\mathbf{v}_i\}_{i=1}^{N_v}$ with $\mathbf{v}_i \in \mathbb{R}^D$, each summarizing a 2D patch on the $h\times w$ grid. These are then concatenated with text tokens to form the input sequence $\mathbf{X}$ of Eq. (1).

Self-Attention Discriminates Positions via RoPE. Self-attention is permutation-invariant on its own, and the position-awareness of Eq. (1) comes entirely from f_{rope}. Given a scalar position p and a projected query/key $\mathbf{x} \in \mathbb{R}^D$, f_{rope} groups $\mathbf{x}$ into $D/2$ pairs and rotates the j-th pair by an angle $p\theta_j$ with $\theta_j = b^{-2j/D}$ and $b = 10{,}000$. So query–key inner products depend on their positional difference. Letting $p_i^{(j)}$ denote the scalar position fed to channel j for token i, the per-channel relative phase between two tokens at positions p_m and p_n is $\Delta\phi_j(p_m, p_n) = (p_n^{(j)} - p_m^{(j)})\,\theta_j$. Existing positional encodings differ only in the channel-allocation rule that defines $p_i^{(j)}$. Multimodal RoPE (M-RoPE) promotes each visual token's position to a 3-tuple $\mathbf{p}_i = (t_i, h_i, w_i)$ on the temporal–height–width lattice and partitions the channel index set $\mathcal{I} = \{0, \ldots, D/2 - 1\}$ into three disjoint subsets $\mathcal{I} = \mathcal{I}_t \cup \mathcal{I}_h \cup \mathcal{I}_w$, so that

$$p_i^{(j)} = \begin{cases} t_i, & j \in \mathcal{I}_t, \\ h_i, & j \in \mathcal{I}_h, \\ w_i, & j \in \mathcal{I}_w, \end{cases} \tag{4}$$

making $\Delta\phi_j$ between two visual tokens a function of $(\Delta t, \Delta h, \Delta w)$ alone. In practice, chunked M-RoPE [8] assigns three contiguous channel chunks to $t/h/w$, while interleaved M-RoPE [38, 42, 37, 17] alternates the axes across the spectrum so every axis sees both high-/low-frequency channels.

4 Analysis: Spatial Collapse in General VLMs

Eq. (1) exposes two geometric entry points: the token feature $\mathbf{X}$ and the multimodal positions $\mathbf{p}$. In current VLMs both are 2D in nature. For $i \in \mathcal{V}$ where $\mathcal{V}$ is the index set of visual tokens, $\mathbf{x}_i$ is a flat patch embedding of 2D semantics, and $\mathbf{p}_i = (t_i, h_i, w_i)$ encodes only image-plane and sequential coordinates. We refer to this dual absence as **Spatial Collapse** and dissect it into a token-side and a position-side collapse.

Token-side Collapse. As shown in Fig. 1, the pinhole projection from a 3D point (x, y, z) to a pixel (u, v) already discards one geometric degree of freedom. Standard vision encoders, e.g., CLIP [15] and SigLIP [16], trained for semantic alignment under explicit scale and translation invariance, they produce embeddings $\mathbf{v}_i$ that preserve strong semantics while compressing metric quantities (absolute scale, exact depth). Whenever a scene departs from data priors picked up during spatial QA tuning, the next-token distribution stays semantically sharp but geometrically ambiguous, surfacing as hallucination on precise spatial reasoning. The pathology is amplified for multi-frame inputs: the same physical point seen from two viewpoints projects to different patches with different latent geometries, with no shared coordinate to align them across frames. **Insight 1: vision tokens are dominated by semantics and their metric geometry is collapsed.**

Position-side Collapse. Even granting a token feature aware of its 3D position, the geometry of pairwise attention remains effectively planar. Under the M-RoPE allocation of Eq. (4), $\Delta\phi_j$ between two visual tokens is a function of $(\Delta t, \Delta h, \Delta w)$ alone, leaving no channel for an out-of-plane axis such as depth. In Fig. 1, an object-boundary patch and a background patch are adjacent in the image plane (small $\Delta h, \Delta w$) but separated by a large depth gap. M-RoPE encodes them as one neighborhood, and attention is structurally biased to mix them regardless of Δz. The model is forced to fight its own positional prior, recovering 3D separation purely from semantic context—an objective that is data-hungry and brittle under heavy occlusion or unusual viewpoints. **Insight 2: the visual positional manifoldis intrinsically 2D in space, and self-attention is structurally biased regardless of the input tokens' 3D positions.**

5 Method

Following the two loci of Eq. (1), we propose **Despa**, a **De**pth-grounded **spa**tial VLM. Despa is mainly composed of two modifications: (i) Geometric Positional Embedding (GPE), attaches a metric geometric anchor to every visual token and fuses it into the LLM's input visual tokens. (ii) Depth-aware Rotational Positional Embedding (DoPE), make the self-attention depth-aware for distinguishing 3D positional divergence.

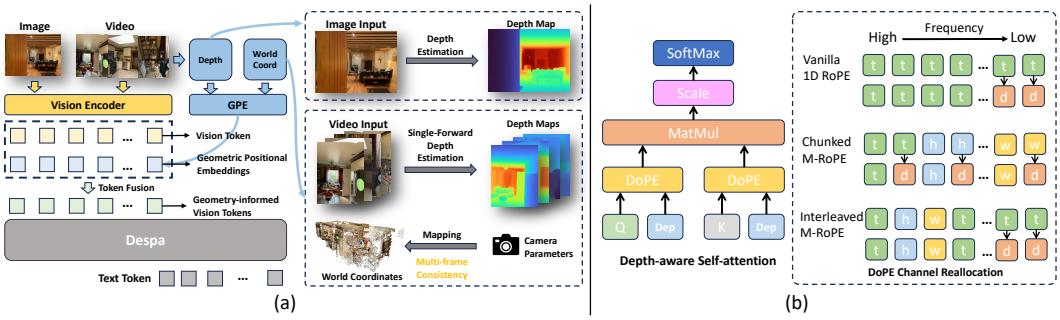

Figure 2: (a) Despa architecture with GPE. (b) Depth-aware self-attention via DoPE and the depth channel reallocation strategies.

5.1 Complementing Visual Tokens with Geometric Context

In this section, we introduce the geometric positional encoding to encode a modality-dependent metric anchor $\mathbf{c}_i$ for each visual token. The anchor captures its physical out-of-plane location. GPE utilizes a parameter-free sinusoidal map ϕ for positional encoding and fuse the result into the visual token by a single residual addition at the multimodal interface:

$$\tilde{\mathbf{v}}_i \;=\; \mathbf{v}_i + \phi(\mathbf{c}_i), \qquad \phi(\mathbf{c}_i)[2j] = \sin\!\left(\frac{c_i^{\alpha(j)}}{b^{2j/D}}\right), \quad \phi(\mathbf{c}_i)[2j{+}1] = \cos\!\left(\frac{c_i^{\alpha(j)}}{b^{2j/D}}\right), \tag{5}$$

for $j = 0, \ldots, D/2-1$, with base $b = 10{,}000$ matching the RoPE frequency basis $\theta_j = b^{-2j/D}$. Text tokens are untouched ($\tilde{\mathbf{x}}_i = \mathbf{x}_i$ for $i \notin \mathcal{V}$). Geometries enters the residual stream once at this interface and propagates through subsequent decoder layers. The two ingredients of Eq. (5), the dimensionality of $\mathbf{c}_i$ and the axis-binding map $\alpha(j)$, are dictated by the modality.

Depth Positional Embedding (DPE) for Single Images. Given a single image $\mathbf{I}$ and an estimated depth map $\mathbf{Z}$, we align it with the patch grid by adaptive average pooling, take the per-patch scalar z_i as the 1D anchor $\mathbf{c}_i = z_i \in \mathbb{R}$, and bind every channel to that single axis ($\alpha(j) \equiv 0$, with $c_i^0 = z_i$).

World-Coord Positional Embedding (WPE) for Videos. For videos, a per-frame, camera-frame depth is no longer comparable across views, so we lift depth into a frame-shared world frame. To make this lift well defined, a single feed-forward pass of a depth estimation model over the entire clip jointly produces (i) per-frame depths $\{\mathbf{Z}_t\}$ that are consistent across views and (ii) a unified set of intrinsics $\{\mathbf{K}_t\}$ and extrinsics $\{(\mathbf{R}_t, \mathbf{t}_t)\}$ in a common world frame. It avoids the per-frame scale and pose drift of independent monocular estimates. Per-pixel depth is then lifted via pinhole back-projection. For a pixel (u_i, v_i) with depth z_i, the world coordinates is

$$\mathbf{c}_i^{\text{world}} \;=\; \mathbf{R}_{t_i}\, \pi^{-1}(u_i, v_i, z_i;\, \mathbf{K}_{t_i}) \;+\; \mathbf{t}_{t_i}, \tag{6}$$

and average-pooled to the patch grid, yielding a per-token 3D anchor $\mathbf{c}_i = (X_i, Y_i, Z_i) \in \mathbb{R}^3$ in metric meters. WPE make multi-frames consistent: along each viewing ray the world coordinate is determined by z_i up to a fixed rigid transform, while identical physical points across frames now share an identical anchor by construction. To encode this 3D anchor through Eq. (5), we share the full frequency spectrum across all three world axes by binding $\alpha(j) = j \bmod 3$ (with $0 \to X$, $1 \to Y$, $2 \to Z$), so every axis sees both high and low-frequency channels in the spirit of interleaved M-RoPE.

5.2 Depth-Aware Self-Attention via RoPE Reallocation

GPE enriches token features with explicit coordinates, but pairwise attention between visual tokens are still 2D. Two adjacent visual tokens in the image plane but far apart in depth are still treated as one neighborhood by only encoding (t, h, w). To compensate this collapse on the relation side, **DoPE** (**D**epth-aware **Ro**tational **P**ositional **E**mbedding) extends the position of visual tokens to (t, h, w, z), and picks a subset $\mathcal{I}_z$ of the RoPE frequency channels and rewrites the channel-allocation rule.

Depth Channel Reallocation. A channel in $\mathcal{I}_z$ has its pretrained signal overwritten by z_i. We set two criteria for depth channel reallocation to identify low-cost channels: (i) low frequency, which encodes

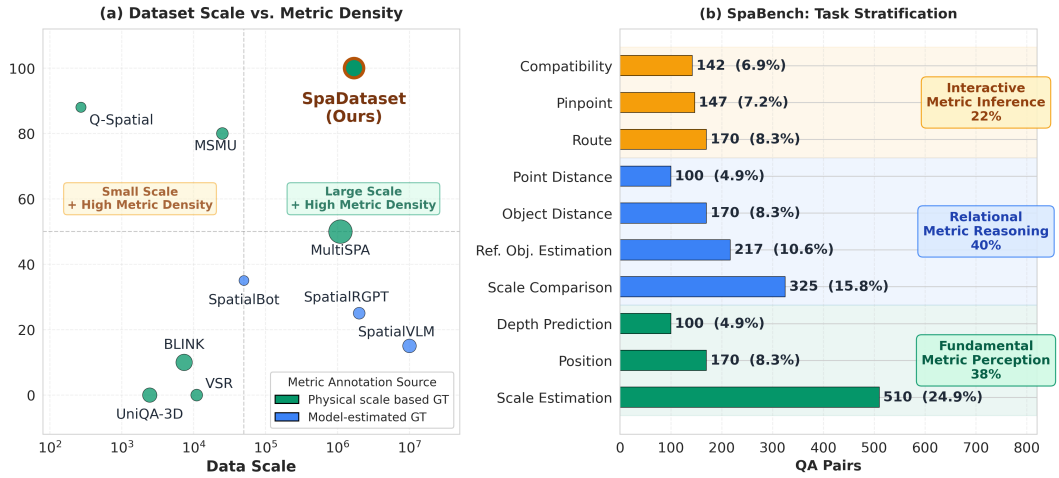

Figure 3: (a) Comparison of various spatial datasets; (b) distribution of tasks in SpaBench across ten fine-grained subcategories, grouped by the three functional tiers.

only long-range global structure [43] and contributes least to local visual discrimination; (ii) weak axis, when a channel is bound to a specific axis of (t, h, w) whose pretrained role for visual tokens is itself weakly informative. The two criteria specialize cleanly to the three RoPE families used by modern VLMs. Vanilla 1D RoPE: a single linearly-ordered axis, where (ii) is vacuous. We take $\mathcal{I}_z$ as the last k channels (smallest θ_j). Chunked M-RoPE: channels form three contiguous $T/H/W$ chunks, so we fall back to (i) and take the lowest-frequency channel of each chunk. Interleaved M-RoPE: the H, W axes saturate before the low-frequency tail, so the tail is purely temporal. For a single image t is constant and carries no signal, so this pure-T tail satisfies both criteria and leaves H, W entirely untouched. The reallocation strategies are visualized in Fig. 2 (b).

The DoPE Operator. Every Despa decoder block replaces $f_{rope}(\cdot; \mathbf{p})$ in Eq. (1) with $f'_{rope}(\cdot; \mathbf{p}, \mathbf{z})$, where $\mathbf{z} = (z_1, \ldots, z_N)$ collects the scalar depth coordinates supplied to DoPE. With channel $\mathcal{I}_z$ fixed as above, DoPE replaces the per-channel rotation by

$$\mathbf{r}_i'^{(j)} = \mathbf{r}^{(j)}_{\tilde{z}_i} \text{ if } i \in \mathcal{V} \text{ and } j \in \mathcal{I}_z, \quad \text{else } \mathbf{r}^{(j)}_{p_i^{(j)}}. \tag{7}$$

$\tilde{z}_i^{(j)} = \ln z_i$. The rotation frequencies for $\mathcal{I}_z$ are geometrically spaced over $[2\pi/10, 2\pi]$. Text tokens ($i \notin \mathcal{V}$) and channels outside $\mathcal{I}_z$ pass through the original RoPE unchanged, so the pretrained (t, h, w) priors survive on the dominant majority of the spectrum.

6 Data Construction for Precise Metric Spatial Reasoning

Existing spatial reasoning datasets primarily focus on qualitative spatial reasoning tasks, which can often be solved using semantic information alone, without relying on visual cues. To address this limitation, we propose a more challenging spatial dataset, aiming to evaluate quantitative spatial abilities, which demand a higher level of spatial reasoning and remain largely underexplored. We propose metric density to measure metric-related spatial annotations per interaction. As shown in Fig. 3(a), existing datasets polarize: small-scale benchmarks (e.g., Q-Spatial[44]) offer high metric density but low volume, while large-scale ones (e.g., SpatialVLM[12], SpatialRGPT[32]) focus on qualitative tasks. Our SpaDataset and SpaBench fill this gap, uniquely occupying the upper-right quadrant, achieving both large scale and high metric density.

Data Sources and Composition. To ensure broad domain coverage and robust generalization, we extract images and physical property annotations from five heterogeneous sources spanning indoor, outdoor, synthetic, and real-world scenes: Hypersim [45], nuScenes [46] , Objectron [47], and LingBot-Depth [48]. Per-source descriptions are provided in Appendix A.1.

SpaDataset focuses on quantitative metric spatial reasoning with physical numerical outputs. We organize metric-related tasks into three functional tiers: *Fundamental Metric Perception*, *Relational*

Table 1: Results on SpaBench.

Model Name	Scale Est.	Ref. Obj. Est.	Abs. Dist.	Pinpoint	Scale Comp.	Position Point	Compat.	Route	Point Depth	Point Dist.	Average
Gemini-3-Pro-Preview	18.2	26.7	10.6	2.7	73.5	30.0	66.2	15.3	21.0	24.0	28.8
GPT-5.2	23.7	24.4	11.2	6.8	78.1	**51.8**	68.3	9.4	17.0	16.0	30.7
InternVL3.5-241B-A28B	23.3	17.5	11.2	19.1	70.8	34.1	63.4	17.1	17.0	14.0	28.7
InternVL3.5-8B	18.0	14.3	15.3	26.5	67.4	32.9	62.0	20.6	18.0	12.0	28.7
Qwen3.6-Plus	30.2	41.5	17.6	0.7	76.0	33.5	70.4	8.8	21.0	23.0	32.3
Qwen3-VL-235B-A22B	30.4	45.6	18.2	1.4	74.5	32.4	69.7	22.4	29.0	15.0	33.9
Qwen3-VL-8B	27.6	26.3	24.7	0.0	69.9	27.1	69.0	8.2	22.0	13.0	28.8
Qwen3-VL-4B	27.6	12.9	18.2	0.0	69.5	30.0	67.6	14.7	8.0	10.0	25.9
SD-VLM	36.2	34.9	20.1	3.5	65.5	14.5	21.1	4.7	0.0	3.7	20.4
Cambrian-S	15.3	15.2	8.8	0.0	63.7	29.4	50.0	5.3	15.0	3.0	20.6
HY-Embodied-0.5	22.8	19.8	14.1	5.9	68.3	37.6	53.5	16.5	15.0	13.0	26.6
SenseNova-SI-1.1-Qwen3-VL-8B	29.4	20.3	12.4	2.0	69.5	29.4	62.7	10.0	13.0	9.0	25.8
Direct SFT (Qwen3-VL-4B)	71.6	61.8	42.9	36.1	81.5	47.1	74.7	18.8	63.0	64.0	56.1
Despa-4B	**74.3**	**62.2**	**47.6**	**38.8**	**83.1**	47.6	**76.1**	**42.9**	**70.0**	**75.0**	**61.8**

Metric Reasoning, and *Interactive Metric Inference*, which are further decomposed into ten fine-grained sub-categories visualized as the SpaBench distribution in Fig. 3(b). Per-category definitions are provided in Appendix A.2.

Data Generation. Our data generation pipeline proceeds in three stages: (i) extracting and unifying 3D bounding boxes from heterogeneous source formats into a common 3D camera coordinate system, together with a curated cross-source semantic vocabulary; (ii) filtering out objects whose 2D projections are heavily occluded ($\geq 50\%$) so that the remaining instances provide sufficient visual evidence for metric reasoning; and (iii) instantiating a rich set of task-specific templates with dynamic placeholders for object labels, spatial relations, and metric values. Implementation details of each stage are provided in Appendix A.3, and the full QA template pool is listed in Appendix B.

Metric Density. Unlike existing datasets that typically query a single spatial property per question, our templates are designed so that one QA pair often supervises multiple physical dimensions at once (e.g., predicting an object's length, width, and height jointly under a single *Scale Estimation* prompt), yielding dense quantitative supervision across the dataset.

SpaDataset Statistics. SpaDataset contains 1.7M QA pairs, and the total number of physical metric annotations is 5M, ensuring that models are trained on dense, fine-grained quantitative supervision.

SpaBench Statistics. We propose SpaBench, which contains 2,051 QA pairs with physical metric annotations and is the only benchmark that covers all ten fine-grained metric sub-categories of SpaDataset (full per-benchmark comparison shown in Appendix A.5). As shown in Fig. 3(b), the benchmark achieves balanced stratification across three functional tiers, enabling fine-grained diagnosis of model capabilities from basic metric perception to complex spatial reasoning. For evaluation, LLMs extract the predicted numerical value. A prediction is considered correct if it lies within a predefined threshold of the ground-truth metric.

7 Experiments

Models and Training. We instantiate Despa on Qwen3-VL-4B-Instruct as the default backbone and apply it to LLaVA-1.5-7B (1D RoPE), Qwen2.5-VL-7B (chunked M-RoPE), and Qwen3-VL-8B (interleaved M-RoPE) to cover all RoPE variants. Per-token depth and world coordinates required by GPE are precomputed offline. All parameters are fine-tuned on A100 GPUs (24 GPU hours for the 4B model), with per-GPU batch size 1 and a learning rate decayed from $1e{-5}$ to $1e{-6}$. We utilize Depth Anything 3 (DA3) [49] as an offline depth estimator, obtaining the metric depth for each frame in a single forward pass per input video.

Baseline. We compare with (i) proprietary VLMs Gemini-3-Pro-Preview [3] and GPT-5.2 [18], (ii) open-source generalists from the Qwen3-VL [17] and InternVL3.5 [50] families, and (iii) spatial specialists including SD-VLM [20], Cambrian-S [34], HY-Embodied-0.5 [35], SenseNova-SI-1.1 [19], and RoboRefer [21]. To isolate the contribution of GPE/DoPE from training data, we also report Direct SFT: fine-tuning the Qwen3-VL-4B backbone on each benchmark's training set using standard interfaces of Eq. (1) without depth injection or rotary modulation.

Table 2: Generalization across spatial datasets.

Model Name	MSMU	RefSpatial-U	VSI-Bench
Gemini-3-Pro-Preview	43.4	**45.3**	52.5
GPT-5.2	34.9	10.8	54.5
InternVL3.5-241B-A28B	34.1	11.7	**69.5**
InternVL3.5-8B	34.7	3.0	56.1
Qwen3-VL-235B-A22B	43.9	11.7	62.7
Qwen3-VL-8B	43.3	23.4	57.9
SD-VLM-7B	56.3	-	-
RoboRefer8B-SFT	-	37.7	-
Cambrian-S-7B	-	-	62.9
Direct SFT (Qwen3-VL-4B)	62.6	31.2	62.4
Despa-4B	**66.3**	42.4	66.0

Table 3: Performance across architectures and scales.

Method	SpaBench	VSI-Bench
Direct SFT (LLaVA-1.5-7B)	33.2	-
Despa (LLaVA-1.5-7B)	35.5 (+2.3)	-
Direct SFT (Qwen2.5-VL-7B)	51.7	52.9
Despa (Qwen2.5-VL-7B)	53.6 (+1.9)	54.7 (+1.8)
Direct SFT (Qwen3-VL-8B)	54.9	61.5
Despa (Qwen3-VL-8B)	57.9 (+3.0)	64.3 (+2.8)
Direct SFT (Qwen3-VL-4B)	56.1	62.4
Despa (Qwen3-VL-4B)	**61.8 (+5.7)**	**66.0 (+3.6)**

Table 4: Transfer to other spatial benchmarks.

Model Name	MMSI Bench	MindCube Tiny	Omni Spatial
Gemini-3-Pro-Preview	**45.2**	70.9	**69.1**
GPT-5.2	41.3	60.4	57.4
InternVL3.5-8B	29.0	40.2	47.4
Qwen3-VL-8B	31.1	29.4	47.0
Cambrian-S-7B	27.1	37.9	41.9
SenseNova-SI-1.1-Qwen3-VL-8B	38.1	**73.8**	43.0
Despa-4B	38.4	73.6	43.1

Table 5: Ablation study.

Dataset	GPE	DoPE	SpaBench	VSI-Bench
✗	✗	✗	24.3	52.6
✓	✗	✗	56.1	62.4
✓	✓	✗	58.4	64.9
✓	✗	✓	58.8	63.9
✓	✓	✓	**61.8**	**66.0**

7.1 Results on SpaBench

Despa establishes a new state of the art with a 4B backbone. Table 1 reports per-task accuracies on the ten metric categories of SpaBench. Despa reaches an average of 61.8%, surpassing the best open-source generalist Qwen3-VL-235B-A22B (33.9%) by 27.9 points and the strongest closed-source model GPT-5.2 (30.7%) by 31.1 points despite using $\sim$60$\times$ fewer parameters. Spatial specialists such as SD-VLM (20.4%) and SenseNova-SI-1.1 (25.8%), which inject 2D spatial supervision but leave the token and rotary interfaces untouched, also fall far below. The gap is therefore not closed by either parameter scaling or task-specific fine-tuning, supporting our claim in Sec. 4 that the bottleneck is structural rather than a matter of capacity.

Gains over Direct SFT concentrate on metric- and depth-sensitive tasks. Sharing identical training data and backbone, Despa improves the average by 5.7 points, but with a strongly task-dependent pattern. The largest gains appear precisely where reasoning depends on out-of-plane geometry: Route +24.1, Point Distance +11.0, Point Depth +7.0, and Absolute Distance +4.7, tasks on which every other baseline collapses to single digits. In contrast, tasks largely solvable from 2D layout (Reference Object, Position Point, Scale Comparison, Compatibility) show only modest gains ($\leq$1.6), since Direct SFT already saturates these from planar cues. This decomposition matches our analysis: GPE attaches an explicit metric anchor to mitigate token-side collapse (Insight 1), while DoPE encodes Δz in the rotary spectrum to restore depth-aware attention (Insight 2). Pinpoint additionally illustrates that off-the-shelf generalists score near chance ($\leq$6.8) regardless of scale, and Despa adds another +2.7 over Direct SFT, showing that GPE/DoPE is complementary to data-side supervision rather than a substitute.

Comparison to methods using geometry-oriented vision encoders. We further train two representative approaches, Spatial-MLLM [30] and VG-LLM [51], which augment VLMs with geometry-oriented vision encoders (VGGT [52]), on the same SpaDataset. They reach only 49.5% and 50.8%, respectively, below Despa 61.8%.

7.2 Generalization Analysis

Generalization on Different Datasets. To verify that Despa generalizes across spatial training distributions rather than being tied to SpaDataset, we fine-tune the same Qwen3-VL-4B backbone separately on three datasets and evaluate on their matching benchmarks. The results are shown in Table 2. On MSMU [20], Despa-4B reaches 66.3%, surpassing general-purpose baselines like Qwen3-VL-235B (43.9%) and Gemini-3-Pro-Preview (43.4%), as well as the spatial specialist

Table 6: Channel reallocation of DoPE.

Strategy	SpaBench	VSI-Bench
Full Channel	57.7	57.5
DoPE on (h, w)	<u>59.6</u>	60.8
Interleaved	56.7	<u>64.6</u>
Tail Channels	**61.8**	**66.0**

Table 7: Robustness analysis of depth noise.

Noise Level	0	0.05	0.1	0.5	no depth
SpaBench	**61.8**	<u>61.7</u>	60.3	58.9	56.1

SD-VLM-7B (56.3%). On RefSpatial-U [21], Direct SFT reaches only 31.2% while Despa attains 42.4%, exceeding RoboRefer8B-SFT (37.7%) and trailing only Gemini-3-Pro-Preview (45.3%). On multi-frame VSI-Bench [22], Despa reaches 66.0%, outperforming Cambrian-S-7B (62.9%) and Qwen3-VL-235B (62.7%), trailing only InternVL3.5-241B (69.5%). The results confirm that the dual-loci correction of GPE/DoPE can lift every spatial training distribution we tried, including the multi-frame regime.

Generalization on Different Architectures. To verify Despa as a generic design, we instantiate it on three RoPE-based architectures: LLaVA-1.5-7B (1D RoPE), Qwen2.5-VL-7B (chunked M-RoPE), and Qwen3-VL-4B/8B (interleaved M-RoPE). Table 3 shows consistent gains across four backbones (+2.3, +1.9, +3.0, +5.7 on SpaBench; +1.8, +2.8, +3.6 on VSI-Bench), confirming compatibility with any pretrained frequency layout. The largest gain is on the smallest backbone: Despa-4B outperforms its 8B counterpart (61.8 vs. 57.9) despite similar Direct SFT baselines (54.9 vs. 56.1). This supports Insight 1: larger backbones have stronger semantic priors that compress metric quantities, reducing plasticity for GPE, confirming spatial collapse is architectural, not capacity-limited.

7.3 Transfer to Other Spatial Benchmarks

We evaluate Despa-4B trained with SpaDataset and other general spatial datasets including SenseNova-SI[19] in a transfer setting on three other spatial benchmarks: MMSI-Bench (multi-image), MindCube-Tiny (viewpoint/mental rotation), and OmniSpatial (broad QA). Table 4 shows that Despa-4B remains competitive. On MindCube-Tiny, it achieves 73.6, matching the 8B specialist SenseNova-SI-1.1 (73.8) and outperforming closed-source models like Gemini-3-Pro-Preview (70.9). On MMSI-Bench, Despa-4B (38.4) surpasses Qwen3-VL-8B (31.1) and InternVL3.5-8B (29.0), matching the 8B SenseNova specialist (38.1) at half the parameters. On OmniSpatial, Despa-4B (43.1) edges past SenseNova (43.0) and Cambrian-S-7B (41.9), but trails large generalists and closed-source models. The dual-locus correction lifts a 4B model to the level of 8B spatial specialists on these benchmarks, with the residual gap concentrated on qualitative reasoning rather than metric or viewpoint understanding.

7.4 Ablation Study and Analysis

Component Ablation: GPE and DoPE. The factorial ablation in Table 5 disentangles GPE and DoPE. On top of Direct SFT (56.1 / 62.4), GPE alone yields +2.3 / +2.5, and DoPE alone yields +2.7 / +1.5. GPE benefits multi-frame tasks (VSI-Bench) requiring world-coordinate consistency, while DoPE excels on metric-dense single-image tasks (SpaBench). Combining both pushes performance to 61.8 / 66.0 (+5.7 / +3.6). The super-additivity on SpaBench confirms the dual-loci design: GPE alone enriches content but keeps attention 2D; DoPE alone enables 3D attention but on entangled tokens. Only together do GPE's depth anchors reach the right neighbors via DoPE's depth-aware attention.

WPE Design for Multi-view Consistency. DPE and WPE are the two instances of GPE, differing only in whether the anchor lives in the camera frame or in a shared world frame. On VSI-Bench, replacing WPE with DPE, WPE reaches 66.0% versus 63.0% for DPE, a 3.0-point gap that quantifies the cost of view-dependent anchors. By lifting depth into a frame-shared anchor via Eq. (6), WPE enforces multi-view consistency at the input rather than leaving it as a learning burden on attention.

DoPE Channel Allocation. Table 6 stress-tests the two design criteria for $\mathcal{I}_z$ in Sec. 5.2: a channel is a low-cost reallocation target only when it (i) lies at the low-frequency end of $\mathcal{I}$ and (ii) is bound to an axis whose pretrained role is weakly informative for visual tokens. Full Channel (replacing all channels with depth) violates both and yields the worst score (57.7 / 57.5). Interleaved (round-robin subset) violates only (i) and collapses on SpaBench (56.7) where pixel-level cues matter. DoPE on

Table 8: Depth probes on visual tokens.

Input	Spearman↑	AbsRel↓	δ<1.25↑
Visual tokens	0.597	0.303	0.585
+ GPE	**0.936**	**0.074**	**0.947**

Table 9: Depth probes on attention weights.

Setting	Spearman↑	AbsRel↓	δ<1.25↑
Vanilla	0.528	0.328	0.560
GPE only	0.574	0.305	0.584
DoPE only	<u>0.723</u>	<u>0.236</u>	<u>0.691</u>
GPE+DoPE	**0.729**	**0.229**	**0.707**

(h, w) respects (i) but violates (ii), stalling at 59.6 / 60.8. Our default Tail Channels strategy is the only one to satisfy both criteria, reaching the best 61.8 / 66.0.

Robustness to Depth Noise. Table 7 probes whether degraded depth from a weaker estimator would erase Despa's gains by injecting multiplicative Gaussian noise of relative scale $\{0, 0.05, 0.1, 0.5\}$ at inference time. Performance degrades smoothly: 5% noise causes a 0.1-point drop, 10% causes 1.5 points, and even 50% (far beyond typical monocular errors) retains 58.9, still 2.8 points above Direct SFT. In the case of no depth (anchor set to a constant), Despa recovers exactly the 56.1 of Direct SFT, indicating that GPE/DoPE introduces no penalty when its input is uninformative. Despa can therefore be deployed with off-the-shelf depth estimators rather than requiring depth ground truth.

7.5 Mechanistic Evidence for Spatial Collapse

We conduct two stage-specific diagnostics using ridge probes to predict per-token metric depth, measured by Spearman correlation, AbsRel, and δ<1.25. The probes are trained on 10K SpaDataset images and evaluated on a disjoint set of 1K images.

We predict per-token metric depth from the original visual tokens versus the geometry-informed tokens after GPE. Table 8 shows Spearman increasing from 0.597 to 0.936. GPE directly restores metric-depth information in the token representation, confirming its role against token-side collapse.

We probe depth from answer-to-visual attention weights, concatenating all 36 layers $\times$ 32 heads into a 1152-dim feature per token. As attention depends jointly on GPE and DoPE, Table 9 reports their factorial comparison. GPE alone improves Spearman only marginally whereas adding DoPE yields a much larger gain. Metric-depth structure in attention is thus dominated by the rotary injection, supporting DoPE's role against position-side collapse.

8 Conclusion

In this paper, we identify that VLMs' metric spatial reasoning deficits stem from spatial collapse, comprising token-side and position-side collapse. To address this, we propose Despa, a lightweight extension that injects metric 3D information via Geometric Positional Embedding (GPE) and Depth-aware Rotational Positional Embedding (DoPE), enabling 3D-aware self-attention with minimal disruption to pretrained knowledge. We also introduce SpaDataset and SpaBench, the largest dense fine-grained training corpus and benchmark for metric spatial reasoning in VLMs. Despa consistently outperforms general-purpose, closed-source, and spatial-specialized baselines across multiple benchmarks and RoPE-based architectures, achieving SOTA on a 4B backbone. Our findings underscore spatial collapse as a key bottleneck and point toward robust spatial intelligence in VLMs.

Limitations. High-quality physical metric annotations remain scarce, and our image GPE still rely on modality-specific anchors rather than a unified geometric interface. In future work, we will unify GPE and further scale up metric data to push the limits of spatial intelligence.

Acknowledgments

This work was supported by Alibaba Postdoc Program. We would like to thank the Qwen Team at Alibaba Token Hub (ATH), Alibaba Group, for providing the computational resources and foundation models (Qwen) used in this research.

References

[1] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421, 2023.

[2] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024.

[3] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. NuerIPS, 2023.

[5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[6] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024.

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966, 2023.

[8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

[9] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479, 2025.

[10] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025.

[11] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In EMNLP, 2023.

[12] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In CVPR, June 2024.

[13] Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. arXiv preprint arXiv:2406.13642, 2024.

[14] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. arXiv preprint arXiv:2503.13111, 2025.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021.

[16] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF international conference on computer vision, pages 11975–11986, 2023.

[17] 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, 2025.

[18] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[19] 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, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026.

[20] Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. Sd-vlm: Spatial measuring and understanding with depth-encoded vision-language models, 2025.

[21] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. arXiv preprint arXiv:2506.04308, 2025.

[22] Jihan Yang, Shusheng Yang, Anjali 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, 2024.

[23] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. NeurIPS, 2023.

[24] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In CVPR, 2024.

[25] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In ICML, 2024.

[26] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual reasoning. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025.

[27] Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. Layoutvlm: Differentiable optimization of 3d layout via vision-language models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 29469–29478, 2025.

[28] Yongsen Mao, Junhao Zhong, Chuan Fang, Jia Zheng, Rui Tang, Hao Zhu, Ping Tan, and Zihan Zhou. Spatiallm: Training large language models for structured indoor modeling. In Advances in Neural Information Processing Systems, 2025.

[29] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding, 2024.

[30] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747, 2025.

[31] Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Eric Xin Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. In Proceedings of the IEEE/CVF international conference on computer vision, pages 8600–8612, 2025.

[32] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In NeurIPS, 2024.

[33] Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning, 2025.

[34] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video. arXiv preprint arXiv:2511.04670, 2025.

[35] Tencent Robotics X and HY Vision Team. Hy-embodied-0.5: Embodied foundation models for real-world agents. arXiv preprint arXiv:2604.07430, 2026.

[36] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021.

[37] Jie Huang, Xuejing Liu, Sibo Song, Ruibing Hou, Hong Chang, Junyang Lin, and Shuai Bai. Revisiting multimodal positional encoding in vision-language models, 2025.

[38] Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? arXiv preprint arXiv:2502.05173, 2025.

[39] Zikang Liu, Longteng Guo, Yepeng Tang, Tongtian Yue, Junxian Cai, Kai Ma, Qingbin Liu, Xi Chen, and Jing Liu. VRoPE: Rotary position embedding for video large language models. In EMNLP, pages 14460–14472, 2025.

[40] Chang Dai, Hongyu Shan, Mingyang Song, and Di Liang. Hope: Hyperbolic rotary positional encoding for stable long-range dependency modeling in large language models, 2025.

[41] Sophie Ostmeier et al. LieRE: Lie rotational positional encodings. In ICML, 2025.

[42] Haoran Li, Yingjie Qin, Baoyuan Ou, Lai Xu, and Ruiwen Xu. Hope: Hybrid of position embedding for long context vision-language models. In The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.

[43] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568:127063, 2024.

[44] Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. In EMNLP, 2024.

[45] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In International Conference on Computer Vision (ICCV) 2021, 2021.

[46] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving, 2020.

[47] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.

[48] Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, et al. Masked depth modeling for spatial perception. arXiv preprint arXiv:2601.17895, 2026.

[49] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647, 2025.

[50] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265, 2025.

[51] Duo Zheng, Yanyang Li, Liwei Wang, et al. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. Advances in neural information processing systems, 38:20560–20586, 2026.

[52] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5294–5306. IEEE, 2025.

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

[54] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. Transactions of the Association for Computational Linguistics, 11:635–651, 2023.

[55] Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J. Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. In CVPR, 2026.

[56] Yiming Zuo, Karhan Kayan, Maggie Wang, Kevin Jeon, Jia Deng, and Thomas L Griffiths. Towards foundation models for 3d vision: How close are we? International Conference on 3D Vision (3DV), 2025.

[57] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In European conference on computer vision, pages 216–233. Springer, 2024.

[58] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9556–9567, 2024.

[59] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences, 67(12):220102, 2024.

[60] Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. arXiv preprint arXiv:2605.25979, 2026.

Despa: Resolving Spatial Collapse in VLMs via Depth-Grounded Geometry

Appendix

Table of Contents

A SpaDataset Construction Details **16**

A.1 Data Sources . . . 16

A.2 Task Taxonomy . . . 16

A.3 Details of SpaDataset/SpaBench Generation . . . 17

A.4 Metric Density . . . 17

A.5 Comparison with Existing Spatial Benchmarks . . . 18

B QA Templates **18**

B.1 Scale Estimation . . . 19

B.2 Position . . . 19

B.3 Depth Prediction . . . 19

B.4 Object Distance . . . 20

B.5 Point Distance . . . 20

B.6 Scale Comparison . . . 20

B.7 Reference Object Estimation . . . 21

B.8 Pinpoint . . . 21

B.9 Compatibility . . . 22

B.10 Route . . . 22

C Example Results on SpaBench **22**

D Example Results on RefSpatial **22**

E Orther Experiments **22**

E.1 Sensitivity to the Choice of Depth Estimator . . . 22

E.2 Performance on General Benchmarks . . . 24

A SpaDataset Construction Details

This appendix complements Section 6 by providing the implementation details of SpaDataset that are deferred from the main text. We describe (i) the five heterogeneous data sources, (ii) the full task taxonomy with per-category definitions, (iii) the annotation extraction and unification pipeline, (iv) the object filtering criterion, and (v) the QA generation procedure together with the high metric-density design. The full QA template pool used during generation is listed separately in Appendix B.

A.1 Data Sources

SpaDataset is constructed from five heterogeneous sources that together span indoor, outdoor, synthetic, and real-world domains and provide complementary forms of physical ground truth:

- **Hypersim** [45]: photorealistic synthetic indoor scenes rendered from artist-built 3D meshes. The mesh-derived ground truth allows millimeter-level annotation accuracy and is free of sensor noise.
- **nuScenes** [46]: real-world outdoor autonomous-driving scenarios with 3D bounding boxes calibrated to LiDAR returns, providing reliable metric depth at long range and outdoors.
- **Objectron** [47]: real-world object-centric videos with consistent-scale 3D bounding boxes around everyday objects, useful for fine-grained per-object metric estimation.
- **LingBot-Depth** [48]: large-scale RGB-D captures from diverse indoor and outdoor scenes that primarily supply dense depth supervision rather than object-level boxes.

This combination intentionally crosses two axes, indoor/outdoor and synthetic/real, so that models trained on SpaDataset are exposed to both clean ground truth and noisy in-the-wild geometry within a single training distribution.

A.2 Task Taxonomy

We decompose metric spatial reasoning into three functional tiers, which together comprise ten fine-grained sub-categories. The grouping reflects the cognitive load of each task, from single-object property recognition to multi-object reasoning to action-oriented inference. The same taxonomy is used as the column structure of Table 10 and as the proportion stratification in Fig. 3(b).

Fundamental Metric Perception. Single-object physical-property recognition that does not require cross-object comparison.

- *Scale Estimation*: predict an object's metric dimensions (length, width, height).
- *Position*: predict the 2D image-plane coordinates of an object, or look up an object given coordinates.
- *Depth*: predict the metric depth of a point or an object in the camera frame.

Relational Metric Reasoning. Quantification of metric relations between two or more objects.

- *Object Distance*: metric distance between two objects.
- *Point Distance*: metric distance between two annotated 2D points back-projected to 3D.
- *Scale Comparison*: which of multiple objects is taller, lower, larger, or smaller.
- *Reference Object Estimation*: estimate one object's dimension given another object's dimension as a metric reference, requiring scale propagation across the scene.

Interactive Metric Inference. Action planning and feasibility prediction grounded in precise metric estimation.

- *Route*: produce a directional move (front/back, left/right) and metric distance to navigate from one object to another.
- *Pinpoint*: produce a 2D point with a specified metric offset relative to an anchor object.
- *Compatibility*: decide whether an agent of a given physical width can fit between two objects.

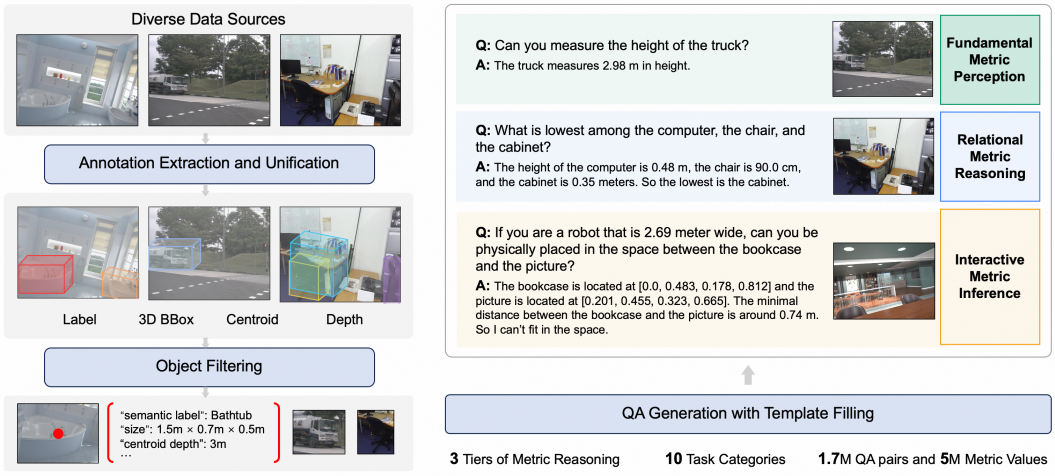

Figure 4: The data generation pipeline of SpaDataset which contains hierarchical and abundant metric reasoning tasks for spatial intelligence.

A.3 Details of SpaDataset/SpaBench Generation

A visual illustration of the data generation pipeline is shown in Fig. 4.

Annotation Extraction and Unification. The five sources expose 3D geometry in incompatible formats: Hypersim[45] provides mesh vertices, nuScenes[46] provides LiDAR-derived cuboids , Objectron[47] provides per-frame object-centric boxes, and LingBot-Depth[48] provides dense depth maps. We extract from each source a canonical 3D representation comprising the object centroid coordinates and physical dimensions in metric units, expressed in a common 3D camera coordinate system. This step normalizes the box conventions (axis-aligned vs. oriented, world frame vs. camera frame) and converts every measurement into meters, so that downstream QA generation operates on a single homogeneous annotation interface regardless of source.

We further curate the semantic vocabulary across sources to resolve label inconsistencies: objects annotated with divergent category names (e.g., semantic variations for common household items) are mapped to a unified taxonomy. We then filter and retain only object categories that have a sufficient instance count across the merged dataset, which both ensures statistical reliability for fine-grained evaluation and removes long-tail labels with ambiguous semantics.

Object Filtering. Even after annotation unification, not all object instances are suitable QA targets. Heavily occluded objects expose only fragmented visual evidence, which makes any metric prediction depend more on prior knowledge than on the input image. Before generating QA pairs, we therefore exclude objects whose projected 2D regions are severely occluded, retaining only instances with less than 50% of their pixels occluded by other objects or scene structures. The threshold is chosen empirically as the point at which the remaining objects still provide sufficient visual evidence for metric reasoning while keeping the dataset size large enough for training. This filter protects the model from learning to hallucinate metric values from cues that do not exist in the visible image.

QA Generation. For each retained object (or object pair/triple), we instantiate a rich set of task-specific templates that differ both in questioning style (declarative, interrogative, comparative) and in reasoning complexity (single-property vs. multi-property answers). Each template carries dynamic placeholders for object labels, spatial relations, and metric values, which are filled in from the unified annotations described above. The full template pool, organized by sub-category, is listed in Appendix B.

A.4 Metric Density

A central design choice is to make each QA pair as metric-dense as possible. Unlike existing datasets that typically query a single spatial property per question, our templates often require estimating

Table 10: Comparison of metric spatial reasoning tasks across benchmarks. ✓ indicates full support with precise physical-scale GT; ✗ indicates unsupported or purely qualitative reasoning. SpaBench is the sole benchmark covering all ten fine-grained metric tasks.

		Fundamental Perception			**Relational Reasoning**				**Interactive Inference**			**Nums**
Benchmark	**Volume**	Scale Est.	Position	Depth	Obj.Dist	Pt.Dist	ScaleCmp	Ref.Obj	Route	Pinpoint	Compat.	
MultiSPA [55]	7,800	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	4
BLINK [53]	3,807	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0
UniQA-3D [56]	500	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	1
SpatialVLM [12]	546	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗	3
MSMU [20]	987	✓	✓	✗	✓	✗	✓	✓	✗	✗	✗	5
SpatialRGPT [32]	1,406	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗	3
VSR [54]	2,195	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0
Q-Spatial [44]	271	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	2
SpatialBot [13]	120	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	3
SpaBench (Ours)	**2,051**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**10**

multiple physical dimensions within a single interaction. For instance, a *Scale Estimation* template may demand complete dimensional prediction:

> *Q: What is the size of [object A]?*
> *A: The size of [object A] is [Length] × [Width] × [Height].*

This single QA pair yields three independent metric values. Aggregated across SpaDataset's 1.7M QA pairs, this design produces nearly 5M physical metric annotations, providing the dense quantitative supervision that we observe in Section 7 to be critical for closing the spatial-collapse gap.

A.5 Comparison with Existing Spatial Benchmarks

To position SpaBench against prior work, Table 10 compares its task coverage against nine existing spatial reasoning benchmarks across the ten metric sub-categories defined above. ✓ indicates full support with precise physical-scale ground truth, while ✗ indicates either no support or purely qualitative reasoning without metric values. Two observations follow.

First, **SpaBench is the only benchmark that supports all ten sub-categories simultaneously**; among prior benchmarks the highest task coverage is five (MSMU[20]), and several widely used benchmarks such as BLINK[53] and VSR[54] support none of the ten precise-metric tasks because they are formulated as qualitative spatial reasoning. Second, **the structural gap is concentrated on the Interactive Metric Inference tier**: none of the nine prior benchmarks supports any of Route, Pinpoint, or Compatibility, all of which require composing precise metric estimation with action-oriented planning. This task-coverage gap mirrors the metric-density gap visualized in Fig. 3(a) and motivates SpaBench as a unified evaluation suite for fine-grained quantitative spatial reasoning.

B QA Templates

Inspired by MSMU [20], we design and expand spatial QA templates to construct a more comprehensive dataset. We provide several samples of each category in the Spa Dataset, which are demonstrated as below:

B.1 Scale Estimation

```
# Scale Estimation
size_template_questions = [ "What is the size of [A]?", "How big is [A]?",
     "Can you provide the size measurement of [A] ?", ]
size_template_answers = [ "The size of [A] is [Length] x [Width] x [
    Height]. ", "[A] is with the length of [Length], width of [Width],
    and height of [Height]."]

height_template_questions = [ "What is the height of [A]?", "How tall is
    [A]?", "Can you measure the height of [A] ?", ]
height_template_answers = [ "The height of [A] is [Height].", "[A] is
    with the height of [Height].","[A] measures [Height] in height."]

width_template_questions = [ "What is the width of [A]?", "Determine the
    width of [A].", "Can you measure the width of [A] ?", ]
width_template_answers = [ "The width  of [A] is [Width].", "[A] is with
    the width of [Width].","[A] measures [Width] in width."]
```

B.2 Position

```
# Position
position1_template_questions = [ "What object is located at ([x],[y])?",
    "What can you find at ([x],[y])?", "What object does the position ([x
    ],[y]) belong to?", ]
position1_template_answers = [ "It is [A].", "That is [A].",'[A].']
position2_template_questions = [ "What is the coordinate of [A] ?", ]
position2_template_answers = [ "([x],[y])." ,"It is located at ([x],[y])
    in the image."]
```

B.3 Depth Prediction

```
# Grounding
position1_template_questions = [ "What object is located at ([x],[y])?",
    "What can you find at ([x],[y])?", "What object does the position ([x
    ],[y]) belong to?", ]
position1_template_answers = [ "It is [A].", "That is [A].",'[A].']
position2_template_questions = [ "What is the coordinate of [A] ?", ]
position2_template_answers = [ "([x],[y])." ,"It is located at ([x],[y])
    in the image."]
```

B.4 Object Distance

```
# Object Distance
object_distance_template_questions = [ "What is the distance between [A]
    and [B]?", "How far away is [A] from [B]?", "Can you provide the
    distance measurement between [A] and [B]?", ]
object_distance_template_answers = [ "[A] and [B] are [X] apart.", "A
    distance of [X] exists between [A] and [B].", "[A] and [B] are [X]
    apart from each other.","The distance is [X]." ]
```

B.5 Point Distance

```
# Marker-based
point1_distance_template_questions = [ "Calculate the distance between
    the two green points in meters. Output numerical value only."]
point1_distance_template_answers =  ["[X]." ]

# Position-based
point2_distance_template_questions = [ "Compute the spatial physical
    distances for the following 2 pairs of points [[[x1, y1], [x2, y2]],
    [[x3, y3], [x4, y4]]]. Output numerical value only."]
point2_distance_template_answers =  ["[Dist1, Dist2]." ]
```

B.6 Scale Comparison

```
# Scale Comparison
taller_template_questions = ["Between [A] and [B], which one is taller/
    lower?","Which one is taller/lower, [A] or [B]? "]
taller_template_answers = ["The height of [A] is [Height A] and [B] is [
    Height B], so [X] is taller/lower."]

tallest_template_questions = ["What is tallest/lowest among [A], [B], and
     [C]?"]
tallest_template_answers = ["The  height of [A] is [Height A], height of
    [B] is [Height B], and height of [C] is [Height C], so the  tallest
    is [X]."]

larger_template_questions = ["Between [A] and [B], which one is larger/
    smaller?","Which one is larger/smaller, [A] or [B]? "]
larger_template_answers = ["The size of [A] is [Length A] x [Width A] x [
    Height A] and [B] is [Length B] x [Width B] x [Height B], so [X] is
    larger/smaller."]
```

B.7 Reference Object Estimation

```
# two objects
refer1_template_questions = [ "The height of [A] is [Height A], can you
    measure the height of [B]?"]
refer1_template_answers =  ["Since the height of [A] is [Height A], i
    think [B] is [Height B] in height.", ]

refer2_template_questions = ["The width of [A] is [Width A], can you
    measure the width of [B]?"]
refer2_template_answers =  ["Since the width of [A] is [Width A], i think
     the width of [B] is [Width B]"]

refer3_template_questions = ["The height of [A] is [Height A], can you
    measure the size of [B]?"]
refer3_template_answers =  ["Since the height of [A] is [Height A], i
    think the size of [B] is [Length B] x [Width B] x [Height B]."]

# three objects
refer4_three_template_questions=["The height of [A] is [Height A], what
    is the height of [B] and [C] ?"]
refer4_three_template_answers=["Since the height of [A] is [Height A], i
    think the height of [B] is [Height B] and the height of [C] is [
    Height C]."]

refer5_three_template_questions=["The distance between [A] and [B] is [
    dis A2B], what is the distance between [B] and [C] ?"]
refer5_three_template_answers=["Since the distance between [A] and [B] is
     [dis A2B], i think the distance between [B] and [C] is [dis B2C]."]
```

B.8 Pinpoint

```
# Pinpoint
pinpoint_template_questions = [ "Pinpoint several points above [A].", "
    Pinpoint several locations above [A].", "Annotate multiple points
    above [A]."]
pinpoint_template_answers =  ["[A] is located at [bbox1]. The points are
    ([x1],[y1]), ([x2],[y2])."]
```

B.9 Compatibility

```
# Compatibility
compatibility_template_questions = ["If you are a robot that is [X] meter
    wide, can you fit in the space between [A] and [B]?","If you are a
    robot that is [X] meter wide, can you be physically placed in the
    space between [A] and [B]?"]
compatibility_template_answers = ["[A] is located at [bbox1] and [B] is
    located at [bbox2]. The minimal distance between [A] and [B] is
    around [dist] m. So I [can/can't] fit in the space."]
```

B.10 Route

```
# Route
route_template_questions = ["How to move from [A] to [B]?","Navigate from
    [A] to [B].","What is the route from [A] to [B]?"]
route_template_answers = ["Move to the [front/back]-[left/right] for [X].
    "]
```

C Example Results on SpaBench

Figure 5 presents qualitative comparisons between Despa and baseline models (Gemini-3-Pro-Preview[3], GPT-5.2[18], and Direct SFT) on the SpaBench. In terms of indoor scene understanding, Despa generates an accurate prediction about the height of the window. Meanwhile, competing models struggle with substantially larger errors. For point-based localization on outdoor scenes, Despa predicts coordinates [787, 411] (highlighted in green) that precisely align with the target, whereas Direct SFT [784, 405] exhibits minor deviation. It is worth noting that Gemini-3-Pro-Preview and GPT-5.2 demonstrate substantial localization errors. These results underscore Despa's superior spatial capabilities.

D Example Results on RefSpatial

Figure 6 shows the results on RefSpatial-U[21] tasks combinations across different models. The purple region in the image denotes the target mask. It can be observed that our Despa(highlighted in green) is largely on par with Gemini-3-pro[3](highlighted in blue), significantly surpasses Direct SFT(highlighted in red) and GPT-5.2[18](highlighted in yellow) on the multi-step reasoning tasks. As evidenced by the visualizations of Despa and Direct SFT, incorporating depth information significantly enhances the model's understanding of spatial relationships.

E Orther Experiments

E.1 Sensitivity to the Choice of Depth Estimator

We further evaluated sensitivity to the choice of depth estimator. Specifically, we trained Despa using depth maps generated by either DA3 [49] or VGGT [52], and evaluated both models using depth maps from each estimator:

As shown in Table 11, when the training and test estimators are matched, DA3 and VGGT achieve 61.8 and 60.7, respectively, showing only a modest 1.1 point difference between the two estimators. Changing the estimator at test time causes a 0.6–2.1 point decrease, indicating some sensitivity to the train–test depth distribution. Nevertheless, every estimator configuration remains at least 3.6 points above the no-depth setting.

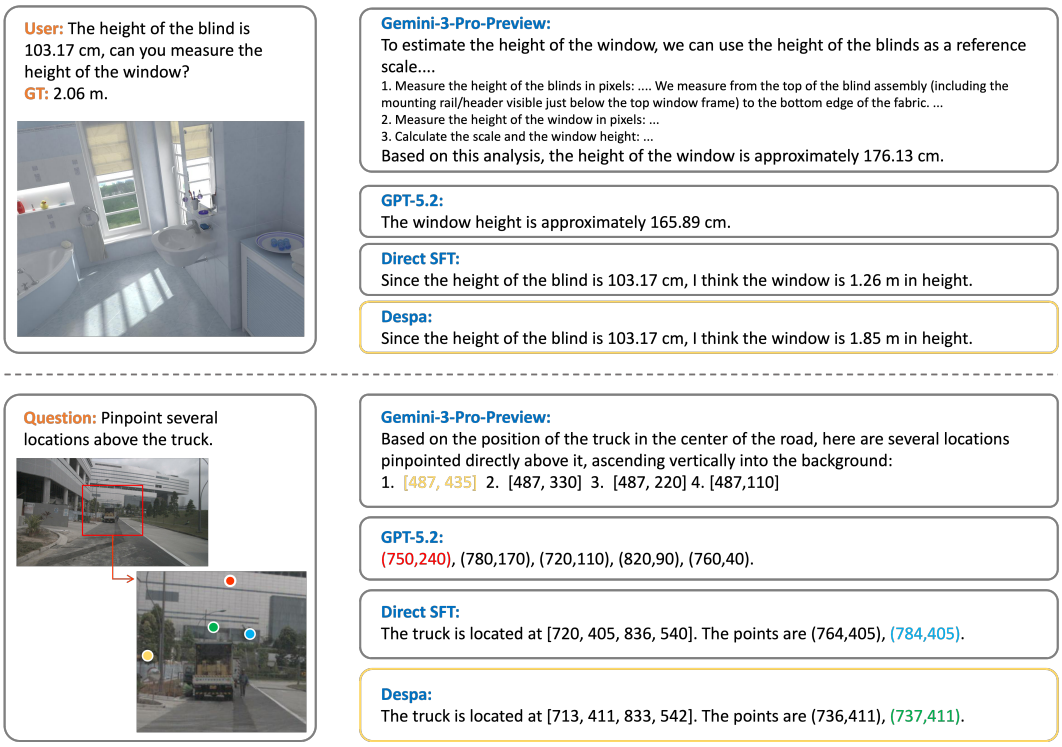

Figure 5: Comparative responses from various models evaluated on SpaBench. Despa demonstrates notable superiority in tasks requiring metric spatial reasoning and precise point-based localization.

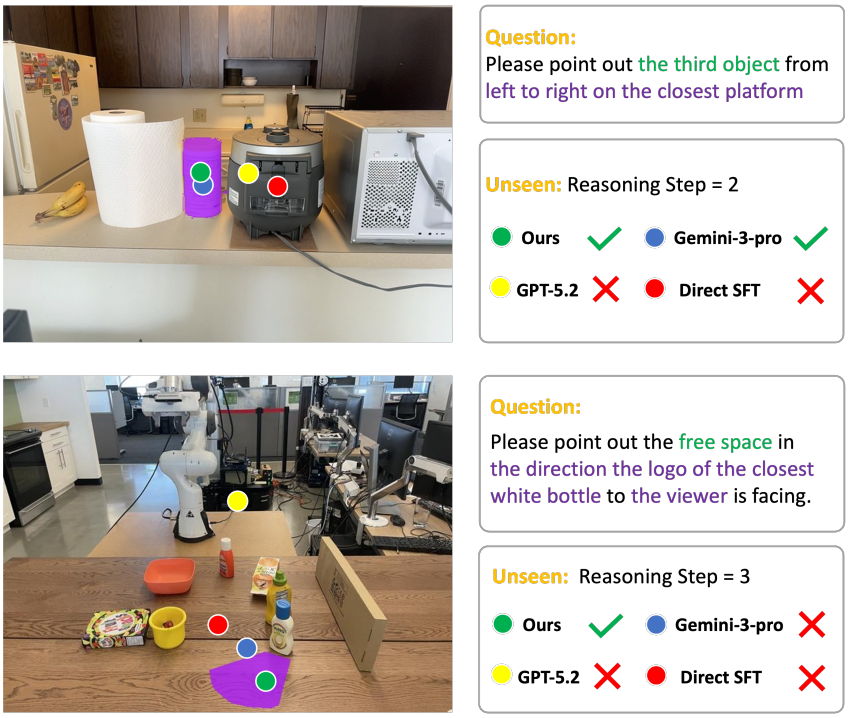

Figure 6: Results of RefSpatial-U Benchmark.Our Despa demonstrates capabilities comparable to Gemini-3-pro.

Table 11: Sensitivity to the choice of depth estimator on SpaBench.

Settings	SpaBench
No Depth	56.1
VGGT-train / VGGT-test	<u>60.7</u>
VGGT-train / DA3-test	60.1
DA3-train / VGGT-test	59.7
DA3-train / DA3-test	**61.8**

E.2 Performance on General Benchmarks

We conduct additional experiments on four general VL benchmarks, namely MMBench [57], BLINK [53], MMMU [58], and OCRBench [59]. For fairness, we sample a subset of LLaVA-OneVision-2 [60] data and mix it into the training set as replay to preserve general capabilities.

Table 12: Performance on standard VL benchmarks.

Model	MMBench	BLINK	MMMU	OCRBench
Baseline	83.4	**65.6**	**56.7**	**875**
Despa	**85.4**	64.7	55.1	853

As shown in Table 12, Despa exhibits only marginal decreases on BLINK, MMMU, and OCRBench. These results do not show the broad or severe degradation characteristic of catastrophic forgetting.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitaions are discussed in the supplementary.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: Complete proofs are shown in the main manuscript.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: It is shown in Section 7.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: Code and data will be open-sourced after acceptance.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: It is included in Section 7.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: No error bars are reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: It is shown in Section 7.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research is conducted in the paper conform.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [No]

    Justification: There is no discussion of broader impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification:

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Non-commercial research and educational purposes of our source data are permitted.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: A new dataset is introduced in this paper. The dataset details are discussed in Section 6. The dataset will be open-sourced after acceptence.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification:

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: LLMs are used to assist in model evaluation.

Guidelines:

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