DEPTHEVIDENCE: UNIFYING METRIC DEPTH PREDICTION AND GEOMETRIC REASONING IN MULTIMODAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial reasoning with metric constraints requires linking objects to geometric measurements and preserving their numerical content during language reasoning. We present DepthEvidence, a 4B model that uses its own dense metric predictions as object-grounded evidence for language generation. A camera-conditioned decoder predicts full-resolution metric depth using multi-scale visual features and high-resolution RGB refinement. A dense-to-language interface converts predicted depths and decoder features into object-aligned continuous geometry tokens anchored to object identifiers. Geometric supervision encourages metric information to remain recoverable before and after language-context interaction, while instruction tuning supports object measurement and compositional reasoning. We introduce a Depth-VQA benchmark evaluating object-depth queries, relative comparisons, and decisions combining spatial and numerical constraints. Across nine datasets, DepthEvidence achieves the highest average dense δ_1 among evaluated methods, competitive with specialized estimators. It also leads the evaluated methods in instance-level metric depth estimation and overall accuracy on both relative and metric reasoning tracks, while broadly preserving general VQA performance and improving spatial understanding relative to the base model.

1 INTRODUCTION

Answering spatial questions with metric constraints requires connecting object semantics to geometric measurements. Knowing which object is closer establishes a relative relation, but selecting objects within a depth interval requires absolute scale and the correct association between each object and its measured depth. Consider the question in Figure 1(a): *which referenced objects to the right of a wooden cabinet have an average camera depth between 1.30 m and 1.33 m?* The answer must satisfy both the spatial filter and the numerical interval. Dense metric estimators provide pixel-level measurements (Yin et al., 2023; Piccinelli et al., 2024), while spatially aware multimodal large language models (MLLMs) connect objects with language (Chen et al., 2024a; Cheng et al., 2024). This question requires bringing those capabilities together at the level of individual objects.

Recent depth-aware VLMs have substantially advanced geometric prediction. DepthLM (Cai et al., 2026) demonstrates accurate pixel-level metric estimation through language modeling, using visual prompting and intrinsic-conditioned augmentation without a dedicated depth head. DepthVLM (Yu et al., 2026) extends this capability to full-resolution dense prediction through a lightweight head and joint depth–text training, while retaining general multimodal capability and improving spatial reasoning. G^2VLM also couples geometric reconstruction with spatial reasoning (Hu et al., 2026). Building on these advances, we ask: *how can a VLM use its own dense metric predictions as object-grounded evidence for language reasoning?* We focus on explicitly linking the predicted metric field to the referenced objects and preserving its numerical content as the language model processes the question.

Using predicted geometry as language evidence requires both object alignment and metric information preservation. Measurements must come from the appropriate image regions and remain associated with the correct references as the model processes multiple objects. Their numerical content must also survive projection into language tokens and interaction with the question. Answer supervision targets

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

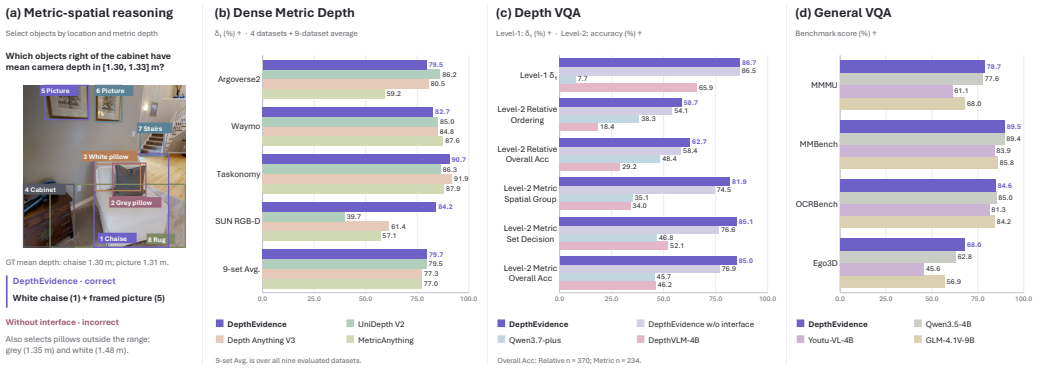

Figure 1: Overview of DepthEvidence's capabilities. (a) A metric-spatial question requires selecting objects by image location and mean camera depth; the dense-to-language interface rejects out-of-range distractors. (b) Dense metric depth results (δ_1) on four representative datasets and the nine-dataset average, compared with specialized estimators. (c) Level 1 metric estimation and Level 2 relative and metric reasoning, compared with general-purpose and depth-aware VLMs as well as the no-interface variant. (d) General VQA retention on MMMU, MMBench, and OCRBench, and spatial understanding on Ego3D. All scores are percentages and higher is better.

the final response without directly constraining the metric content of intermediate geometry tokens. This motivates explicit geometric supervision to keep object-level depth information recoverable before and after interaction with the language context.

We present DepthEvidence, a 4B model that turns its own dense predictions into object-aligned evidence for language generation. Given an image and camera intrinsics, a camera-conditioned decoder combines multi-scale visual features with high-resolution RGB refinement to produce metric depth and preserve local structure. The dense-to-language interface pools predicted depths and decoder features within each referenced region, then encodes them as object-aligned continuous geometry tokens anchored to object identifiers. Auxiliary depth objectives supervise metric information in these tokens before and after language-context interaction, while a ranking objective supervises object ordering. The tokens enter answer generation through cached continuation, avoiding repeated encoding of the image and question. In the cabinet example, this interface supplies the candidate objects' geometry for spatial filtering and metric-range selection.

To evaluate how geometric prediction translates into language answers, we organize depth understanding into dense metric depth prediction (Level 0), instance-level metric depth estimation (Level 1), and compositional depth reasoning (Level 2). Instance-level instruction tuning and the interface support both measurement and reasoning. Our Depth-VQA benchmark evaluates whether models can turn object-level depth measurements into compositional answers, covering relative comparisons and decisions that jointly depend on spatial relations and metric gaps, ranges, or thresholds. DepthEvidence achieves the highest average dense δ_1 across nine datasets among the evaluated methods and the highest overall accuracy on both reasoning tracks, while general VQA performance remains broadly stable relative to the base model (Figure 1).

Our contributions are:

- An object-aligned dense-to-language interface with geometric supervision before and after contextualization, enabling language generation to condition on the model's own metric predictions.
- DepthEvidence, a unified 4B model combining camera-conditioned dense prediction with instance-level depth instruction tuning, together with a Depth-VQA benchmark that distinguishes object measurement, relative reasoning, and metric-constrained reasoning.
- DepthEvidence leads the evaluated methods in nine-dataset average dense δ_1, instance-level metric estimation, and overall compositional reasoning, while retaining general VQA performance and improving spatial understanding over the base model.

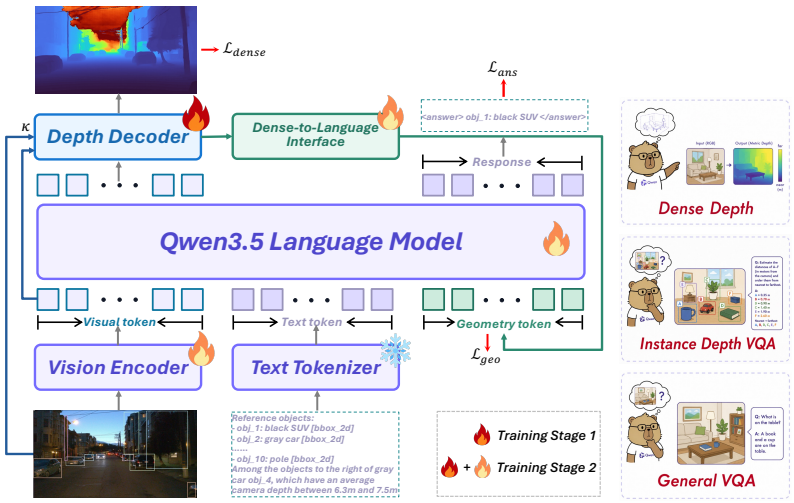

Figure 2: Overview of DepthEvidence. The dense decoder supplies metric predictions and features to the dense-to-language interface, which converts them into object-aligned geometry tokens for answer generation. This path connects Level-0 dense prediction to Level-1 instance measurement and Level-2 compositional reasoning within a shared VLM.

2 Related Work

Monocular Metric Depth Estimation. Monocular depth estimation has progressed from multi-scale convolutional models with scale-invariant supervision (Eigen et al., 2014) to transformer decoders using multi-level visual tokens (Ranftl et al., 2021). Depth Anything and Depth Anything V2 scale transferable representations with unlabeled or pseudo-labeled data (Yang et al., 2024a;b), whereas Metric3D and UniDepth address camera variation through canonicalization or camera-conditioned features (Yin et al., 2023; Piccinelli et al., 2024). These methods primarily treat geometry as a dense output; DepthEvidence additionally uses metric predictions as object-grounded evidence for measurement and reasoning.

Geometry-Aware Vision–Language Models. Prior work uses geometry either to support spatial reasoning or as an explicit model output. SpatialVLM constructs metric spatial question–answer data through automatic 3D annotation (Chen et al., 2024a), and SpatialRGPT combines 3D scene-graph regions with a depth-aware visual plugin (Cheng et al., 2024); G^{2}VLM, Vid-LLM, and SpatialStack more directly couple reconstruction, 3D tokens, or geometry-encoder features with language reasoning (Hu et al., 2026; Chen et al., 2026; Zhang et al., 2026). AURORA and Perceptio emit discrete depth-related tokens before text (Bigverdi et al., 2025; Li et al., 2026), DepthLM predicts metric depth at prompted pixels (Cai et al., 2026), and DepthVLM jointly outputs full-resolution depth and text (Yu et al., 2026). However, explicit or parallel depth prediction does not necessarily condition answers on the predicted metric field. DepthEvidence bridges prediction and reasoning by converting its single-image dense outputs into object-aligned continuous geometry tokens, supervised before and after language-context interaction, for object-level measurement and relative or metric compositional reasoning.

3 Method

3.1 Framework Overview

As illustrated in Figure 2, DepthEvidence conditions object-level answers on geometric evidence predicted from the same image. A camera-conditioned dense decoder supplies a metric depth map and spatial features. The dense-to-language interface extracts evidence for objects specified by boxes or referring prompts and binds their geometry tokens to explicit object identifiers. Auxiliary objectives supervise metric-depth prediction from these tokens before and after language-context interaction. Instance-level instruction tuning trains the VLM to use this context for measurement and compositional reasoning.

3.2 LEVEL-0 DENSE METRIC DEPTH PREDICTION

Camera-conditioned decoder. The Level-0 branch follows a coarse-to-fine DPT design (Ranftl et al., 2021), fusing intermediate vision-encoder features with the spatially reshaped final visual hidden states. To recover metric scale, it converts the supplied camera intrinsics $\boldsymbol{\kappa}$ into a compact descriptor containing normalized calibration, field-of-view, and aspect-ratio information. An MLP maps this descriptor to FiLM parameters (Perez et al., 2018) that condition every decoder scale. A lightweight high-resolution RGB pathway supplies finer appearance cues during upsampling, and the resulting full-resolution representation predicts both the metric depth map and the dense features used by the dense-to-language interface. This numerical camera encoding is specific to the dense branch; Depth-VQA instead receives the raw intrinsic values in its instruction. Appendix A.1.1 provides the feature dimensions, camera parameterization, and decoder details.

High-resolution RGB refinement. Spatially compressed visual tokens can lose thin structures and sharp depth discontinuities. The RGB pathway therefore injects high-resolution features through learned residual gates, preserving the transformer representation as the main path while restoring local boundary cues. The refined spatial features also provide local appearance and boundary cues for object-level pooling in the interface. Architectural details are given in Appendix A.1.2.

Training objective. We train the dense branch with

$$\mathcal{L}_{\text{dense}} = 0.8\mathcal{L}_{\text{abs}} + 0.2\mathcal{L}_{\text{silog}} + 0.1\mathcal{L}_{\text{der}} + 0.03\mathcal{L}_{\text{lap}} + \mathcal{L}_{\text{ssi}}. \tag{1}$$

The first two terms supervise metric depth, the derivative and Laplacian terms preserve multi-scale structure, and the edge-guided stable local scale-and-shift invariant term $\mathcal{L}_{\text{ssi}}$ emphasizes object boundaries. Appendix A.1.3 provides the detailed definitions and implementation.

3.3 INSTANCE-LEVEL DEPTH INSTRUCTION TUNING

Instance-level measurement and reasoning tasks. We extend dense prediction with two object-grounded instruction levels. Level 1 asks for a specified object's center, mean, median, nearest, or farthest metric depth. Level 2 combines measurements across objects: relative tasks cover near–far and front–behind comparisons, closest/farthest selection, and subset or full depth ordering; metric tasks cover numerical mean-depth gaps, thresholds, ranges, and margin-based decisions. Compositional questions first resolve image-space references or filter candidates, then apply relative comparisons or metric constraints. These instructions teach object-level measurement and its use in spatially and numerically constrained answers.

Unified object grounding. We support two ways to specify an object. A box prompt provides its category and bounding box, with coordinates normalized to a 0–1000 image coordinate system. A referring prompt instead identifies it through a natural-language description enclosed by object-reference tokens. For multi-object questions, each instance receives a stable identifier such as `obj_1`, and the prompt lists its identifier, category, and grounding cue before the question. This common serialization supports both grounding modes and disambiguates repeated object categories without assuming any interface-specific processing.

Structured supervision. Each instruction supplies the raw camera intrinsics as `<focal>` f_x f_y c_x c_y `</focal>`, independently of the dense decoder's numerical camera encoding. The VLM reads this block together with the image and question. We train both instruction levels with the next-token cross-entropy loss $\mathcal{L}_{\text{ans}}$ on the final answer. The answer is a metric depth value for Level 1, or a numerical depth gap, an object or set selection, an ordering, or a discrete comparison or constraint decision for Level 2. Margin-based decisions include an *ambiguous* answer when the closest object does not satisfy the required depth margin. Appendix A.2 provides the detailed targets, prompt serialization, and output formats.

3.4 DENSE-TO-LANGUAGE INTERFACE

Object-aligned dense geometry. For both Level 1 and Level 2 instructions, the interface receives a metric depth map $\widehat{\mathbf{D}}$ and a full-resolution feature map $\mathbf{F}^d$ from the dense decoder. For a box prompt, we use the supplied region directly; for a referring prompt, a lightweight grounding head predicts its

box. For object i, pooling $\mathbf{F}^d$ yields a regional feature $\mathbf{r}_i$, while summarizing the predicted depths yields metric statistics $\mathbf{s}_i$. Together with the normalized box $\mathbf{b}_i$, these form an object descriptor that associates local geometry with the referenced instance. Appendix A.3.1 gives the pooling statistics and grounding-head implementation.

Continuous geometry tokens. A two-layer projector P_g maps each object descriptor to $K = 2$ continuous tokens,

$$\mathbf{g}_{i,k} = P_g([\mathbf{r}_i; \mathbf{b}_i; \mathbf{s}_i])_k + \mathbf{p}_i + \mathbf{t}_k, \qquad k \in \{1, 2\}, \tag{2}$$

where $\mathbf{p}_i$ is a learned object-index embedding and $\mathbf{t}_k$ distinguishes the two token slots. We organize the resulting tokens as an input-only geometry bank in which each pair follows its textual anchor `obj_i:`. Geometry-bank positions are excluded as language-modeling targets, but answer tokens attend to them and the answer loss backpropagates through them to the projector. Appendix A.3.1 details the bank construction, and Appendix A.3.2 discusses the choice of $K = 2$.

Geometric supervision. To encourage metric-depth recoverability, we supervise the geometry representations at both instance levels with

$$\mathcal{L}_{\text{geo}} = 0.2\mathcal{L}_{\text{raw}} + 0.2\mathcal{L}_{\text{ctx}} + 0.1\mathcal{L}_{\text{rank}}, \tag{3}$$

The raw-depth and contextual-depth losses train auxiliary heads to predict five ground-truth object-depth statistics (center, mean, median, nearest, and farthest) in log space, from the projected tokens and their contextualized states, respectively. These objectives encourage metric-depth recoverability before and after interaction with the image and question. For multi-object examples, $\mathcal{L}_{\text{rank}}$ supervises relative ordering using the contextual mean-depth predictions. Appendix A.3.4 specifies the heads, targets, and gradient routing.

Single-prefill cached continuation. The geometry bank becomes available after the image and question have been encoded. We retain the prefix cache, append the bank as a causal continuation, and generate the answer from the extended context. Geometry tokens can attend to the cached prefix, and answer tokens can attend to both the prefix and geometry. This execution scheme incorporates predicted evidence without re-encoding the multimodal prefix. Appendix A.3.3 gives implementation details, and Appendix E compares no-interface, two-pass, and single-prefill inference.

4 Instance-Level Depth-VQA Benchmark

Depth-VQA evaluates object-level metric estimation (Level 1) and multi-object compositional reasoning (Level 2), including relative comparisons, ordering, numerical gaps, and spatially conditioned metric decisions. Each question provides an image, camera intrinsics, and object references specified by normalized category–box prompts or natural-language referring prompts.

Construction and quality control. We first complete sparse outdoor depth so every source image has a dense depth map. Qwen3.7-plus (Qwen Team, 2026b) and Seed-2.1-Turbo (ByteDance Seed Team, 2026) independently identify major objects and propose their categories and grounding boxes, reducing dependence on the output distribution of a single generator. We pool their candidates, remove boxes covering less than 1% of the image or lying near image boundaries, prompt SAM 3 (Carion et al., 2026) with the remaining categories and boxes, and apply a second quality filter to the resulting masks. Each retained mask and dense depth map yield the Level 1 center, mean, median, nearest, and farthest targets, with the latter two defined by the 5th and 95th depth percentiles. The two generators then produce Level 2 questions and answers from these instance records, with each candidate verified by the model that did not generate it. We manually audit 1% of training samples across all three levels to derive targeted filters, manually check every benchmark example, and limit reuse of the same instance across task types within a video scene. Appendix C.1 provides full details.

Level-1: Instance-Level Metric Depth Estimation. Level 1 contains 757 grounded instances from six indoor and outdoor datasets: Waymo (Sun et al., 2020), Argoverse 2 (Wilson et al., 2021), Taskonomy (Zamir et al., 2018), LingBot-Depth (Tan et al., 2026), DDAD (Guizilini et al., 2020), and nuScenes (Caesar et al., 2020). Each query requests the object-center, nearest, farthest, mean, or median valid depth within the referenced mask and expects a scalar answer in meters. We report δ_1 (higher is better), RMSE, and AbsRel (lower is better).

Table 1: Level-2 tracks and subbench goals. Appendix Table 7 provides all question types and reasoning requirements.

Track	Subbench	Question summary
Relative Depth	Spatial Pair	Compare two objects, optionally after resolving a spatial reference.
	Spatial Group	Spatially filter candidates, then select the nearest or farthest.
	Depth Ordering	Select an extreme or order a subset or all referenced objects.
Metric Depth	Metric Pair	Predict a mean-depth gap or compare it with a threshold.
	Spatial Metric Pair	Resolve a spatial reference, then predict an absolute mean-depth gap.
	Spatial Metric Group	Spatially filter candidates, then apply range or bound constraints.
	Metric Set Decision	Test a safety band or a margin-constrained closest-object decision.

Level-2: Compositional Depth Reasoning. Depth-VQA contains 1,503 questions in total: 757 Level 1 questions, 370 Level 2 Relative questions, and 376 Level 2 Metric questions. Level 2 uses grounded instances as reasoning operands, assigning each a stable identifier such as `obj_1`. Table 1 groups seven subbenches into the two tracks; Appendix Table 7 defines all question types. The Relative track uses exact-match accuracy: selections require the target identifier, and orderings require the full identifier sequence. Relative Overall Acc pools the 370 questions from the three reported subbenches. The Metric track uses accuracy for discrete answers and δ_1 for numerical gaps; Metric Overall Acc excludes δ_1-scored questions. Appendix C provides question counts and scoring details.

5 Experimental Setup

5.1 Training Pipeline

Training proceeds in two stages. Stage 1 freezes the vision-language backbone and trains the Level-0 decoder and high-resolution RGB branch for 80K steps. Stage 2 initializes from this checkpoint and jointly trains Level-1/Level-2 Depth-VQA and general VQA for 18K steps, with about 60K dense replay examples preserving Level-0 prediction. The dense-to-language interface is active for both instance levels. Because dense predictions and decoder features are detached before geometry-token construction, QA losses update the LoRA-adapted language modules and interface, while dense replay updates the decoder. We use Qwen3.5-4B (Qwen Team, 2026a); Appendix B provides data composition, preprocessing, optimization, and runtime settings.

5.2 Evaluation Setup

For Level 0, we evaluate metric depth on four outdoor datasets (Argoverse 2 (Wilson et al., 2021), Waymo (Sun et al., 2020), nuScenes (Caesar et al., 2020), and DDAD (Guizilini et al., 2020)) and five indoor datasets (Taskonomy (Zamir et al., 2018), ScanNet++ (Yeshwanth et al., 2023), HyperSim (Roberts et al., 2021), SUN RGB-D (Song et al., 2015), and Booster (Ramirez et al., 2022)). Predictions use ground-truth intrinsics and are scored in metric scale without post-hoc alignment; δ_1 is the primary metric, and the overall score averages the nine dataset-level results.

Depth-VQA follows the tasks and metrics defined in Section 4. Level-0 baselines include specialized metric-depth models and depth-aware VLMs; Depth-VQA baselines include general-purpose and depth-aware VLMs. All models use the same prompts and answer protocol. A matched DepthEvidence variant disables the dense-to-language interface at both instance levels to isolate geometry-token conditioning. Appendix C provides the complete baselines, sampling, and scoring protocols.

6 Main Results

6.1 Level-0 Dense Metric Depth Prediction

Because general-purpose VLMs lack dense metric-depth outputs, Table 2 compares DepthEvidence with depth-aware VLMs and specialized estimators. DepthEvidence achieves the highest nine-dataset average δ_1 of 0.7972, slightly above UniDepth V2 (0.7945) and 0.1033 above the strongest depth-aware baseline, DepthLM-12B. It exceeds DepthLM on eight datasets and leads on SUN RGB-D (0.8420). Its outdoor and indoor averages of 0.7970 and 0.7974 indicate balanced aggregate performance, although specialized models remain stronger on several individual datasets. Figure 3 complements Table 2 with two outdoor and two indoor examples; additional visualizations for the remaining five datasets are provided in Appendix Figure 5. DepthEvidence preserves fine structures and depth discontinuities more consistently than DepthVLM-4B and Youtu-VL-4B-Instruct while

Table 2: Level-0 dense metric depth comparison across outdoor and indoor benchmarks. We report $\delta_1 \uparrow$ and its unweighted average over all nine datasets. The best and second-best results in each column are shown in bold and underlined, respectively. Rankings are determined using the unrounded values. †DepthLM-12B follows its original sparse point-query protocol (Cai et al., 2026), using 8,192 randomly sampled pixels per dataset.

Method	Outdoor				Indoor					Avg.
	Argoverse2	Waymo	NuScenes	DDAD	Taskonomy	ScanNet++	HyperSim	SUN RGB-D	Booster	
VLMs Trained on Metric Depth Estimation										
Youtu-VL-4B-Instruct	0.6290	0.5121	0.6443	0.2995	0.7561	0.6006	0.4285	0.3525	0.3471	0.5077
DepthVLM-4B	0.7696	0.8056	0.8433	0.8232	0.6992	0.8568	0.4536	0.5089	0.2452	0.6673
DepthLM-12B†	0.7768	0.7758	0.8310	0.6638	0.8218	0.7962	0.4358	0.5508	0.5934	0.6939
Specialized Vision Models										
Depth Anything V2 outdoor	0.1304	0.7122	0.1701	0.1446	0.0008	0.0002	0.0415	0.0000	0.0000	0.1333
Depth Anything V1 outdoor	0.2182	0.6564	0.3338	0.1779	0.0026	0.0001	0.0743	0.0001	0.0000	0.1626
Depth Anything V1 indoor	0.0018	0.0019	0.0009	0.0015	0.4706	0.7604	0.2504	0.6427	0.5225	0.2947
Depth Anything V2 indoor	0.1343	0.0188	0.1773	0.2891	0.2397	0.2322	**0.8652**	0.1651	0.6007	0.3025
Metric3D	0.7801	0.8287	0.7691	0.8136	**0.9467**	0.0734	0.4745	0.0582	0.0434	0.5320
Metric3D v2	0.8589	**0.8774**	0.8555	0.8700	0.9389	0.8809	0.7311	0.6980	0.1374	0.7609
MetricAnything	0.5920	0.8758	0.8067	**0.8736**	0.8794	**0.8966**	0.8440	0.5713	0.5943	0.7704
Depth Anything V3 metric-large	0.8051	0.8483	0.7996	0.7825	0.9189	0.8906	0.5800	0.6141	**0.7175**	0.7730
UniDepth V2	**0.8619**	0.8502	**0.8811**	0.8719	0.8629	0.8602	0.8648	0.3969	0.7004	0.7945
DepthEvidence-4B (Ours)	0.7948	0.8266	0.8122	0.7543	0.9072	0.8197	0.8213	**0.8420**	0.5970	**0.7972**

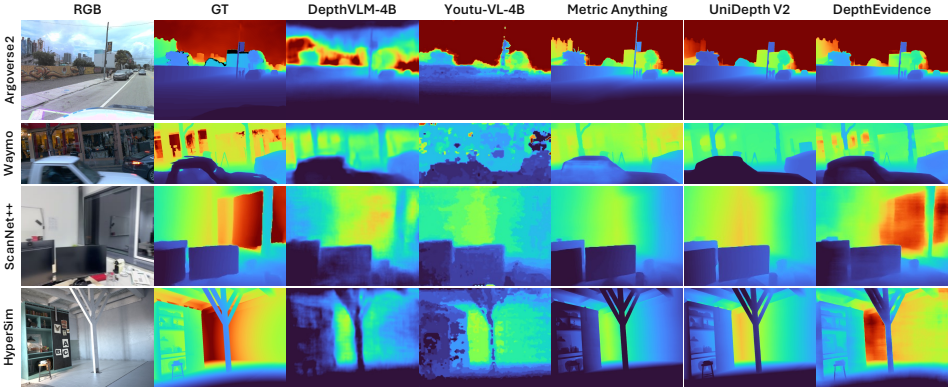

Figure 3: Qualitative dense depth on Argoverse2, Waymo, ScanNet++, and HyperSim. Columns show RGB, ground truth (GT), DepthVLM-4B, Youtu-VL-4B-Instruct, MetricAnything, UniDepth V2, and DepthEvidence. Outdoor GT is densified only for visualization; Table 2 uses the original sparse sensor GT. Each row shares one color scale; Figure 5 shows the other five datasets.

retaining the global layout. MetricAnything and UniDepth V2 sometimes produce smoother planes, whereas DepthEvidence remains competitive in boundary localization across both domains.

6.2 Instance-Level Depth Estimation and Reasoning

Table 3: Instance-level depth estimation and reasoning results. The interface is enabled or disabled at both instance levels. Relative Overall Acc is computed over 370 questions, and Metric Overall Acc includes the 234 questions scored by accuracy. Bold and underlined entries denote the best and second-best results in each column.

Method	Level-1: Instance Metric Depth			Level-2: Compositional Depth Reasoning									
				Relative				Metric					
	$\delta_1 \uparrow$	RMSE↓	AbsRel↓	Spatial Pair	Spatial Group	Ordering	Overall Acc	Metric Pair Acc	Metric Pair $\delta_1 \uparrow$	Spatial Pair $\delta_1 \uparrow$	Spatial Group	Set Decision	Overall Acc
General-Purpose VLMs													
Qwen3.5-4B	0.0159	206.1221	9.5185	0.0930	0.0568	0.0102	0.0405	0.6304	0.2500	0.2234	0.3085	0.5532	0.4701
Gemini-3.1-Pro-Preview	0.4108	52.4680	2.0927	0.6395	0.5795	0.3724	0.4838	0.6304	0.2292	0.2340	0.3085	0.5957	0.4872
Qwen3.7-plus	0.0766	107.2556	3.0477	0.5814	0.6136	0.3827	0.4838	0.6522	0.2917	0.1489	0.3511	0.4681	0.4573
Seed-2.1-Turbo	0.2048	107.9360	2.7470	0.6977	0.6250	0.4133	0.5297	0.6304	0.1042	0.1596	0.3617	0.4362	0.4444
VLMs Trained on Metric Depth Estimation													
DepthLM-12B	0.7781	10.7460	0.1839	0.0000	0.0000	0.0000	0.0000	0.0000	0.1458	0.0532	0.0000	0.0000	0.0000
Youtu-VL-4B-Instruct	0.0264	22.1562	1.0852	0.1744	0.2955	0.1224	0.1757	0.2826	0.0417	0.1064	0.1915	0.3404	0.2692
DepthVLM-4B	0.6592	11.9164	0.2613	0.5000	0.3295	0.1837	0.2919	0.5870	0.0833	0.0957	0.3404	0.5213	0.4615
Agent Pipeline													
MetricAnything + Qwen3.5-4B	0.7318	11.9224	0.2363	0.3372	0.6136	0.4184	0.4459	0.5870	0.2917	0.2340	0.4149	0.6489	0.5427
MetricAnything + Qwen3.7-plus	0.7543	14.2529	0.2407	0.3140	**0.7273**	0.4541	0.4865	0.8478	0.4167	0.2979	0.4468	0.6596	0.6111
DepthEvidence w/o interface	0.8653	11.2159	0.2351	0.7442	0.5227	0.5408	0.5838	0.8261	0.4375	0.3511	0.7447	0.7660	0.7692
DepthEvidence (Ours)	**0.8666**	**8.6062**	**0.1629**	**0.7907**	0.5568	**0.5867**	**0.6270**	**0.9130**	**0.5833**	**0.4043**	**0.8191**	**0.8511**	**0.8504**

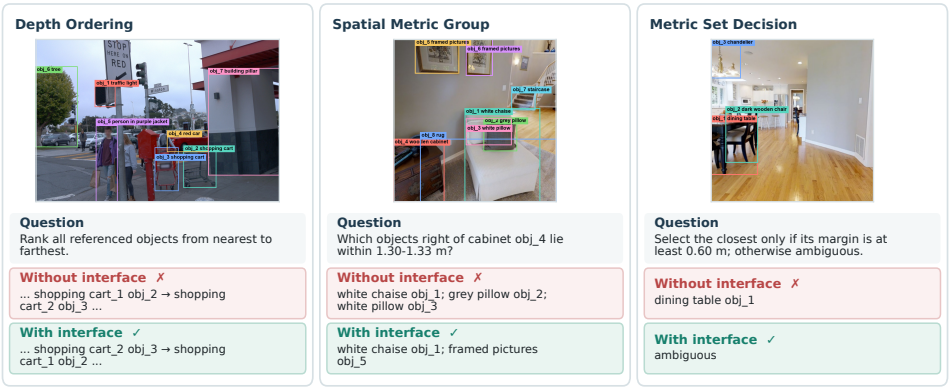

Figure 4: Representative Level-2 cases corrected by the dense-to-language interface. The three panels test complete depth ordering, spatially filtered metric-range selection, and margin-aware ambiguity decisions, respectively. Each panel contrasts matched variants without and with the interface; green answers match the ground truth. Aggregate results are reported in Table 3.

Table 4: General and spatial capability evaluation. Metrics follow their native scales; all are higher-is-better except Ego3D RMSE. Bold marks the best result, and "–" denotes unavailable results.

Method	STEM				General VQA			
	MathVision	MathVista	MMMU	MMMU-Pro	RealWorldQA	MMStar	MMBench	HallusionBench
Youtu-VL-4B-Instruct	–	76.50	61.10	43.00	74.60	71.10	83.90	59.10
GLM-4.1V-9B-Thinking	54.40	80.70	68.00	57.10	–	72.90	85.80	63.20
Qwen3.5-4B	**74.60**	**85.10**	77.60	**66.30**	**79.50**	**78.30**	89.40	65.00
DepthEvidence (Ours)	74.10	84.90	**78.71**	64.05	78.06	**78.30**	**89.52**	**68.21**

Method	Text Recognition and Document Understanding				Spatial Intelligence			
	AI2D	CC-OCR	OCRBench	CharXiv (RQ)	ERQA	EmbSpatialBench	Ego3D Acc.	Ego3D RMSE↓
Youtu-VL-4B-Instruct	85.60	–	81.30	43.80	–	–	45.63	19.27
GLM-4.1V-9B-Thinking	87.90	–	84.20	–	45.80	–	56.93	12.79
Qwen3.5-4B	89.60	76.70	**85.00**	**70.80**	54.00	81.30	62.82	13.17
DepthEvidence (Ours)	**90.67**	**77.34**	84.60	69.00	**57.25**	**83.14**	**67.99**	**6.83**

Table 3 evaluates instance-level metric estimation and compositional reasoning. Among the depth-aware baselines, DepthLM-12B (Cai et al., 2026) performs best on Level 1 ($\delta_1 = 0.7781$), whereas DepthVLM-4B (Yu et al., 2026) performs best on both Level 2 tracks (0.2919 Relative and 0.4615 Metric Overall Acc). DepthEvidence leads all evaluated methods on Level 1, with $\delta_1 = 0.8666$, RMSE $= 8.6062$, and AbsRel $= 0.1629$. It also achieves 0.6270 Relative and 0.8504 Metric Overall Acc, exceeding Seed-2.1-Turbo and Gemini-3.1-Pro-Preview by 9.73 and 36.32 percentage points, respectively, and DepthVLM-4B by 33.51 and 38.89 points. DepthEvidence further obtains the highest reported numerical-gap δ_1 on Metric Pair (0.5833) and Spatial Metric Pair (0.4043).

6.3 Effect of the Dense-to-Language Interface

Matched ablations in Table 3 isolate the interface effect. On Level 1, δ_1 is nearly unchanged (0.8653→0.8666), while RMSE and AbsRel decrease from 11.2159 to 8.6062 and from 0.2351 to 0.1629, suggesting fewer large metric errors. On Level 2, Overall Acc increases by 4.32 points on Relative and 8.12 points on Metric. The largest categorical gain is on Metric Pair (+8.69 points), while numerical-gap δ_1 rises from 0.4375 to 0.5833 on Metric Pair and from 0.3511 to 0.4043 on Spatial Metric Pair. Although MetricAnything + Qwen3.7-plus leads Relative Spatial Group (0.7273), its Relative and Metric Overall scores (0.4865/0.6111) trail DepthEvidence by 14.05/23.93 points, and both agent pipelines remain below DepthEvidence without the interface (Appendix C.6). Overall, the agent pipelines lag the full DepthEvidence system, indicating that external estimator–VLM coupling is less effective than our integrated design. Figure 4 visualizes representative corrections in depth ordering, spatially filtered metric-range selection, and margin-aware ambiguity decisions.

6.4 General VQA Retention and Spatial Understanding

Table 4 shows that DepthEvidence improves spatial reasoning while largely retaining the general capabilities of its Qwen3.5-4B initialization. Non-spatial scores change modestly: MMMU, MM-Bench, HallusionBench, AI2D, and CC-OCR improve, while the largest regression is 2.25 points on MMMU-Pro. All four spatial metrics improve: ERQA rises from 54.00 to 57.25, EmbSpatialBench

Table 5: Dense-decoder ablation on outdoor and indoor datasets ($\delta_1 \uparrow$).

SSI	HR branch	Outdoor		Indoor	
		Argoverse2	Waymo	Taskonomy	ScanNet++
–	–	0.7313	0.7505	0.7852	0.7349
✓	–	0.7547	0.7736	0.8175	0.7511
–	✓	0.7824	0.8129	0.8804	0.8081
✓	✓	**0.7948**	**0.8266**	**0.9072**	**0.8197**

Table 6: Geometry-interface ablation on Level-2 Relative and Metric Overall Acc.

Variant	$\mathcal{L}_{\text{ans}}$	$\mathcal{L}_{\text{raw}}$	$\mathcal{L}_{\text{ctx}}$	$\mathcal{L}_{\text{rank}}$	Relative Overall	Metric Overall
w/o interface	–	–	–	–	0.5838	0.7692
Interface	✓	–	–	–	0.5838	0.7735
	✓	✓	–	–	0.5946	0.8120
	✓	✓	✓	–	0.6135	0.8419
	✓	✓	✓	✓	**0.6270**	**0.8504**

from 81.30 to 83.14, and Ego3D accuracy from 62.82 to 67.99, while RMSE falls from 13.17 to 6.83 (-48.1%). DepthEvidence also outperforms Youtu-VL-4B-Instruct (Wei et al., 2026) and GLM-4.1V-9B-Thinking (GLM-V Team, 2025) on their commonly reported higher-is-better metrics and Ego3D RMSE.

7 Analysis and Discussion

7.1 Dense Decoder Ablation

Table 5 shows that both the SSI objective and high-resolution branch increase δ_1 on each evaluated dataset, with their combination obtaining the highest values among the ablated variants. The HR branch has a larger measured effect indoors: used alone, it raises mean δ_1 by 0.0842 on the two indoor datasets versus 0.0568 outdoors. One possible explanation is the sparse evaluation coverage of outdoor datasets: their ground truth contains fewer evaluated pixels near object boundaries, so improvements in boundary prediction contribute less to the measured δ_1 than under denser indoor annotations.

7.2 Geometric Supervision Ablation

The interface is used at both instance levels; Table 6 focuses on its supervision design using the two Level-2 accuracy aggregates. Training the interface with only $\mathcal{L}_{\text{ans}}$ leaves Relative Overall Acc unchanged at 0.5838 and raises Metric Overall Acc from 0.7692 to 0.7735. Adding $\mathcal{L}_{\text{raw}}$ improves the two aggregates by 1.08 and 3.85 percentage points, respectively; $\mathcal{L}_{\text{ctx}}$ adds a further 1.89 and 2.99 points, and $\mathcal{L}_{\text{rank}}$ adds 1.35 and 0.85 points. The complete objective reaches 0.6270 Relative Overall Acc and 0.8504 Metric Overall Acc, improving over the model without the interface by 4.32 and 8.12 percentage points. These results support the utility of auxiliary geometric supervision, with a larger measured gain on the accuracy-based Metric tasks. A separate no-interface control with direct metric supervision remains clearly below the full interface, indicating that the main performance gain arises from the interface itself rather than numerical supervision alone; Appendix D.1 provides the detailed disentanglement analysis.

7.3 Limitations

The current setting assumes access to camera intrinsics and explicit object references. Its geometric evidence is therefore sensitive to errors in calibration, object grounding, and dense-depth prediction, particularly when depth ranges or decision margins are narrow. In addition, although cached continuation avoids repeated prefix computation, conditioning on geometry tokens incurs additional latency and memory relative to inference without the interface (Appendix E).

8 Conclusion

DepthEvidence connects dense metric depth prediction, instance-level metric depth estimation, and compositional depth reasoning within a shared MLLM. Its dense-to-language interface supplies object-aligned continuous geometry tokens, with geometric supervision encouraging metric-depth recoverability during language-context interaction. Experiments show competitive dense prediction and gains in object measurement, numerical depth gaps, and metric-constrained decisions. General VQA remains broadly stable relative to the base model, while spatial understanding improves on the evaluated benchmarks. These findings support dense prediction as both a supervised output and internal evidence for language reasoning. Future work can assess robustness to noisier camera intrinsics and object references, reduce inference cost as the object count grows, and extend the interface to other geometric evidence.

AI USE STATEMENT

We used Qwen3.7-plus and Seed-2.1-Turbo to propose object annotations and to generate and cross-verify synthetic Depth-VQA questions and answers, as described in Section 4 and Appendix C.1. We additionally used generative AI tools to improve manuscript readability. All benchmark examples were manually verified, and the authors reviewed all AI-assisted outputs and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe the model architecture and training objectives, while Section 5 and Appendix B provide the training pipeline, data composition, preprocessing, and optimization settings. Section 4 and Appendix C document benchmark construction, data partitioning, evaluation protocols, and metric definitions.

REFERENCES

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G. Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3836–3845, 2025.

ByteDance Seed Team. Seed2.1: A next-generation agent for real-world productivity. Official model page, 2026. URL https://seed.bytedance.com/en/seed2_1.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Zhipeng Cai, Ching-Feng Yeh, Xu Hu, Zhuang Liu, Gregory P. Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. DepthLM: Metric depth from vision language models. In *International Conference on Learning Representations*, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Haijier Chen, Bo Xu, Shoujian Zhang, Haoze Liu, Jiaxuan Lin, and Jingrong Wang. Vid-LLM: A compact video-based 3D multimodal LLM with reconstruction–reasoning synergy. In *International Conference on Learning Representations*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

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 *Advances in Neural Information Processing Systems*, volume 37, 2024.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, 2024. doi: 10.18653/v1/2024.acl-short.33.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. In *International Conference on Learning Representations*, 2026.

GLM-V Team. GLM-4.1V-Thinking and GLM-4.5V: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Google DeepMind. Embodied reasoning question answer (ERQA) benchmark. Official benchmark repository, 2025. URL https://github.com/embodiedreasoning/ERQA.

Google DeepMind. Gemini 3.1 Pro model card. Google DeepMind, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2485–2494, 2020.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Kaixuan Wang, Hao Chen, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 2024.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. G^2VLM: Geometry grounded vision language model with unified 3D reconstruction and spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9535–9546, 2026.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Proceedings of the European Conference on Computer Vision*, 2016.

Yuchen Li, Amanmeet Garg, Shalini Chaudhuri, Rui Zhao, and Garin Kessler. Perceptio: Perception enhanced vision language models via spatial token generation. *arXiv preprint arXiv:2603.18795*, 2026.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e4cd50120b6d7e8daff1749d6bbaa889-Abstract-Conference.html.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision*, 2024.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Baorui Ma, Jiahui Yang, Donglin Di, Xuancheng Zhang, Jianxun Cui, Hao Li, Yan Xie, and Wei Chen. MetricAnything: Scaling metric depth pretraining with noisy heterogeneous sources. In *Proceedings of the European Conference on Computer Vision*, 2026.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2354–2367, 2026. doi: 10.1109/TPAMI.2025.3628473. URL https://doi.org/10.1109/TPAMI.2025.3628473.

Qwen Team. Qwen3.5-4B model card. Hugging Face, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-4B.

Qwen Team. Qwen3.7-Plus. Alibaba Cloud Model Studio, 2026b. URL https://docs.modelstudio.console.alibabacloud.com/en/model-studio/qwen3-7-plus.

Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open challenges in deep stereo: The booster dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21168–21178, 2022.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10912–10922, 2021.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 567–576, 2015.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2446–2454, 2020.

Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Masked depth modeling for spatial perception. *arXiv preprint arXiv:2601.17895*, 2026.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with the MATH-Vision dataset. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Zhixiang Wei, Yi Li, Zhehan Kan, Xinghua Jiang, Zuwei Long, Shifeng Liu, Hongze Shen, Wei Liu, Xiaoyu Tan, Haojia Lin, Yubo Zhu, Qianyu Li, Di Yin, Haoyu Cao, Weibo Gu, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, Yunsheng Wu, Mingkong Tang, Shuangyin Liu, Lexiang Tang, Haodong Lin, Junru Lu, Jiarui Qin, Lingfeng Qiao, Ruizhi Qiao, Bo Ke, Jianfeng He, Ke Li, Yangning Li, Yunhang Shen, Mengdan Zhang, Peixian Chen, Kun Yin, Bing Liu, Yunfei Wu, Huang Chen, Zhongpeng Cai, and Xiaotian Li. Youtu-VL: Unleashing visual potential via unified vision-language supervision. *arXiv preprint arXiv:2601.19798*, 2026. URL https://arxiv.org/abs/2601.19798.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

xAI. RealWorldQA. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/xai-org/RealworldQA.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, Lianwen Jin, and Junyang Lin. CC-OCR: A comprehensive and challenging OCR benchmark for evaluating large multimodal models in literacy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21744–21754, 2025.

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

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9043–9053, 2023.

Hanxun Yu, Xuan Qu, Yuxin Wang, Jianke Zhu, and Lei Ke. Unlocking dense metric depth estimation in VLMs. *arXiv preprint arXiv:2605.15876*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 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*, pp. 9556–9567, 2024a.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024b.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018.

Jian Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. SpatialStack: Layered geometry–language fusion for 3D VLM spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38678–38688, 2026.