UNLOCKING DENSE METRIC DEPTH ESTIMATION IN VLMS

Anonymous authors
Paper under double-blind review

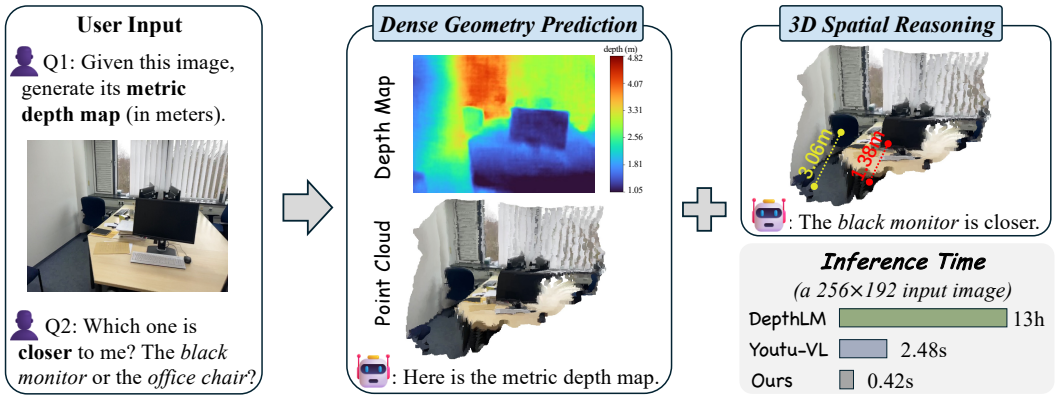


Figure 1: Our method serves as a unified foundation model for both low-level dense geometry prediction and high-level multimodal understanding, while achieving substantially faster inference compared with existing VLM-based approaches such as DepthLM (Cai et al., 2025) and Youtu-VL (Wei et al., 2026).

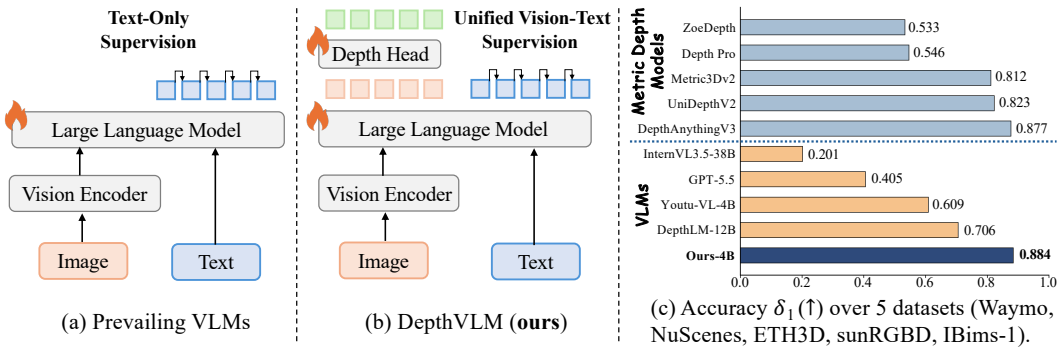
ABSTRACT

Vision–Language Models (VLMs) excel at 2D tasks such as grounding and captioning, yet remain limited in 3D understanding. A key limitation is their text-only supervision paradigm, which under-constrains fine-grained visual perception and prevents the recovery of dense geometry. Prior methods either distill geometry from external vision models, introducing error accumulation, or enable direct prediction with inefficient per-pixel query or coarse token-level outputs. In this paper, we propose **DepthVLM**, a simple yet effective framework that transforms a single VLM into a *native dense geometry predictor* while preserving its multimodal capability. By attaching a lightweight depth head to the LLM backbone and training under a unified vision–text supervision paradigm with a two-stage schedule, DepthVLM generates full-resolution depth maps alongside language outputs in a single forward pass. We further introduce a unified indoor–outdoor metric depth benchmark in a VLM-compatible format. Experiments show that DepthVLM significantly outperforms existing VLMs with higher inference efficiency, surpasses leading pure vision models, and improves complex 3D spatial reasoning, moving toward a truly unified foundation model. All code and checkpoints will be publicly released.

1 INTRODUCTION

With the rapid advancement of Large Language Models (LLMs) (Chiang et al., 2023; Liu et al., 2024; Touvron et al., 2023; Yang et al., 2025a), growing efforts have been devoted to extending them beyond text-only understanding, leading to the emergence of Vision-Language Models (VLMs) (Bai et al., 2025; Lin et al., 2024; Wang et al., 2025d; Zhang et al., 2025) capable of handling a wide range of multimodal tasks. Despite their strong performance on 2D tasks such as visual reasoning and image captioning, current VLMs remain limited in complex 3D understanding (Chen et al., 2020; Majumdar et al., 2024; Piccinelli et al., 2025; Yang et al., 2025b), which is essential for applications like AR/VR, smart glasses, autonomous driving, and embodied robotics.

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.



(a) Prevailing VLMs (b) DepthVLM (**ours**) (c) Accuracy δ_1 (↑) over 5 datasets (Waymo, NuScenes, ETH3D, sunRGBD, IBims-1).

Figure 2: **Comparison of prevailing VLMs with our method.** (a) Prevailing VLMs are typically supervised solely in the text space, leaving dense 3D geometry out of reach. (b) DepthVLM introduces a unified vision–text supervision paradigm by integrating a lightweight depth head, natively enabling a single VLM backbone to generate dense geometry alongside language responses. (c) While even advanced VLMs such as GPT-5.5 struggle to infer 3D structure from 2D inputs, our model significantly outperforms prior VLMs and even surpasses leading specialized pure vision models.

A fundamental limitation of prevailing VLMs is their *text-only supervision* paradigm: visual signals serve only as inputs, while outputs are generated as autoregressive text, leaving fine-grained perception and dense scene geometry under-constrained, as shown in Figure 2(a). To address this, prior works (Fan et al., 2025; Wu et al., 2025; Zheng et al., 2025a) augment VLMs with geometric signals (*e.g.*, depth maps or point clouds) from pretrained 3D models to augment VLMs, but rely on knowledge distillation from external vision experts and suffer from error accumulation. More recent works (Hu et al., 2025; Yan et al., 2026) explore direct geometric prediction from RGB inputs. DepthLM (Cai et al., 2025) first demonstrates that VLMs can match pure vision models on metric depth estimation, but its single-pixel query per inference makes dense prediction prohibitively slow, while its text-heavy supervision substantially degrades the VLM's general VQA capability. Youtu-VL (Wei et al., 2026) enables full-image depth prediction in one pass, yet its token-level outputs remain coarse and require post-hoc interpolation for pixel-level detail. Moreover, its from-scratch training recipe demands massive data and compute, limiting efficient adaptation to existing VLMs.

These observations raise a natural question: *can a VLM serve as a native dense geometry predictor with minimal architectural change, while retaining its general multimodal capability?* To this end, we propose **DepthVLM**, a simple yet effective framework for dense metric depth estimation that jointly produces pixel-level depth maps and language responses with a single VLM. As shown in Figure 2(b), we attach a lightweight depth head to the LLM backbone and fine-tune it under a *unified vision–text supervision* paradigm. In a single forward pass, DepthVLM predicts full-image depth without post-processing, reducing DepthLM's $\mathcal{O}(HW)$ inference cost to $\mathcal{O}(1)$. Moreover, unlike fixed-resolution vision models (Wang et al., 2025b), DepthVLM inherits the native-resolution flexibility of VLMs and can be seamlessly integrated into the standard instruction tuning stage.

Since extending VLMs to other tasks often degrades their general multimodal capability (Dong et al., 2023; Zhang et al., 2024), we adopt a two-stage training strategy: Stage-1 trains only the depth head to establish initial depth prediction ability, and Stage-2 fine-tunes the full model end-to-end. We further introduce **DepthVLM-Bench**, a unified benchmark that aggregates public indoor and outdoor depth datasets into a VLM-compatible format for training and fair comparison with pure vision models. Interestingly, we find that equipping VLMs with dense geometry prediction improves downstream 3D spatial reasoning, further highlighting the value of a unified foundation model that jointly supports low-level dense geometry prediction and high-level multimodal understanding.

In summary, our contributions are threefold:

- We find that a VLM can serve as a native dense geometry predictor and propose a lightweight recipe that yields a unified foundation model for both dense geometry generation and multimodal interaction, seamlessly compatible with the standard instruction-tuning stage.
- We devise a two-stage training strategy that preserves the VLM's original multimodal capability, and present DepthVLM-Bench, a unified indoor–outdoor benchmark that enables VLM training and direct comparison with pure vision models on metric depth estimation.

- Extensive experiments across diverse datasets show that DepthVLM significantly outperforms existing VLMs with higher inference efficiency, surpasses state-of-the-art pure vision models on metric depth estimation, and further improves 3D spatial reasoning performance.

2 Related Work

2.1 Dense Metric Depth Estimation

Dense metric depth estimation recovers per-pixel absolute depth from RGB images and is fundamental to 3D scene understanding. Early methods (Bhat et al., 2021) rely on single-domain supervision, limiting cross-domain generalization. MiDaS (Ranftl et al., 2020) introduces affine-invariant prediction across diverse datasets but lacks metric scale. To resolve scale ambiguity, ZoeDepth (Bhat et al., 2023) combines relative and metric depth via domain-specific heads, while Metric3D (Yin et al., 2023) unifies inputs in a canonical camera space. More recently, UniDepth (Piccinelli et al., 2024) jointly estimates depth and camera intrinsics in a self-promptable manner, and DepthAnything (Yang et al., 2024a) leverages large-scale synthetic supervision for zero-shot generalization. Despite their strong geometric accuracy, these pure vision models focus solely on low-level geometric prediction and lack high-level language interaction, limiting their applicability to 3D reasoning tasks.

2.2 VLMs for 3D Spatial Understanding

Spatial-Enhanced VLMs. To bridge 2D semantics and 3D spatial intelligence, recent works augment VLMs with external geometric signals. One line directly feeds explicit 3D data (*e.g.*, point clouds, voxels, or depth maps) into LLMs via projectors (Zheng et al., 2025b; Zhu et al., 2024), but relies on sparse and costly 3D data and is largely limited to indoor scenes. Another line elicits spatial reasoning purely from 2D inputs. SpatialVLM (Chen et al., 2024) and SpatialRGPT (Cheng et al., 2024) convert vision outputs into textual supervision, while Ross3D (Wang et al., 2025a) introduces multi-view reconstruction as an auxiliary objective. More recent works (Fan et al., 2025; Huang et al., 2025) further distill geometric priors from 3D reconstruction (Wang et al., 2025b;c) or video diffusion models (Wan et al., 2025; Blattmann et al., 2023) into VLMs to improve spatial reasoning. However, these methods rely on external vision experts, making them prone to error accumulation, and are still limited to textual outputs without enabling dense, pixel-level geometry prediction.

Geometry-Generative VLMs. Recent studies (Yan et al., 2026) instead explore VLMs as unified foundation models for direct dense geometry prediction from RGB inputs. Multi-SpatialMLLM (Xu et al., 2025) performs pixel-level metric depth estimation but lags behind pure vision models. G^2VLM (Hu et al., 2025) adopts a Mixture-of-Experts architecture for unified modeling, yet focuses on relative depth. DepthLM (Cai et al., 2025) achieves competitive accuracy but requires one query per pixel and compromises general capabilities. Youtu-VL (Wei et al., 2026) enables full-image depth prediction in one pass, but produces coarse token-level outputs and relies on costly from-scratch training. In contrast, our method lightweightly equips existing VLMs with dense metric depth estimation while preserving their general capability. Inheriting native-resolution processing, it enables flexible inputs and can be seamlessly integrated into standard instruction tuning.

3 Methodology

Our goal is to develop a unified foundation model that natively supports both low-level dense geometry prediction and high-level multimodal understanding within a single VLM backbone. As illustrated in Figure 3, we (i) augment the standard VLM with a lightweight DPT-style (Ranftl et al., 2021) depth head to produce dense metric depth maps and language responses jointly; (ii) employ a two-stage training strategy to preserve the VLM's inherent multimodal capability; and (iii) leverage a multi-source training corpus together with focal-length normalization to mitigate camera-induced ambiguity across heterogeneous sensors, yielding strong cross-dataset generalization.

3.1 Model Architecture

Preliminaries. A standard VLM comprises three components: a vision encoder $\mathcal{E}_v$ that tokenizes an input image $I \in \mathbb{R}^{3\times H\times W}$ into N_v vision tokens, a projector ϕ that maps them into the LLM

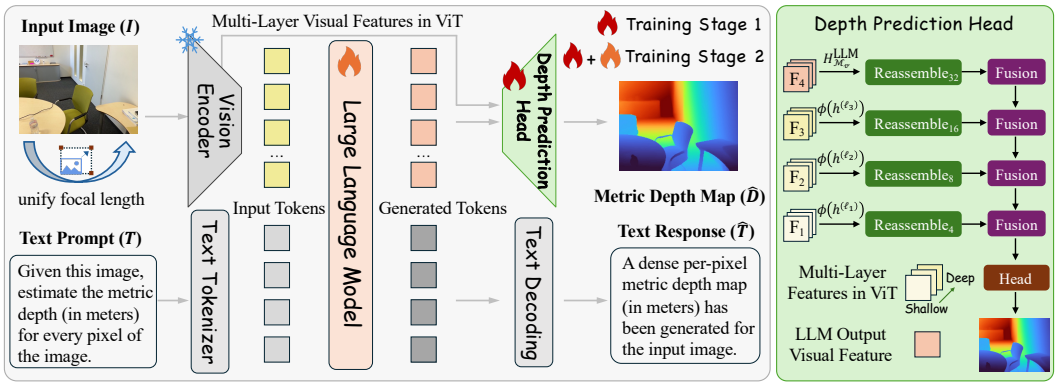


Figure 3: **Overview of our proposed DepthVLM.** We extend the standard VLM architecture with a lightweight DPT-style depth prediction head, and adopt a two-stage training strategy to preserve the backbone's general VQA capability. In addition, input images are normalized to a unified focal length, eliminating camera-induced ambiguity across heterogeneous dataset domains.

embedding space, and an autoregressive language model $\mathcal{F}_{\text{LLM}}$ that processes the joint multimodal sequence to generate text. Given an image I and a text prompt T, the VLM produces hidden states as

$$H^{\text{LLM}} = \mathcal{F}_{\text{LLM}}([\,\phi(\mathcal{E}_v(I));\, T\,]) \in \mathbb{R}^{(N_v+N_t)\times d}. \tag{1}$$

Motivation: VLM as a Native Dense Predictor. Prior works on 2D dense understanding (Wu et al., 2024) typically introduce region-level encoders or task-specific tokens, inevitably fragmenting the architecture and complicating the training and inference pipelines. Inspired by recent 3D foundation models that derive dense geometry directly from transformer tokens, we instead ask: *is a standard VLM already a dense predictor?* We answer this affirmatively by showing that dense geometry can be decoded directly from the VLM's own vision tokens using a lightweight DPT-style (Ranftl et al., 2021) head over multi-scale visual features, without altering its text generation pathway.

Unified Architecture for Dense Geometry. A key observation is that the vision encoder $\mathcal{E}_v$ naturally provides a hierarchy of representations—from low-level appearance cues in shallow layers to high-level semantics in deeper layers—that inherently form a multi-scale pyramid well suited for dense prediction. Let $\{h^{(\ell)}\}_{\ell=1}^{L_v}$ denote the per-layer hidden states of the ViT and H^{LLM} the last-layer hidden states of the LLM. We extract four feature maps from the VLM: three intermediate ViT layers $\{\ell_1, \ell_2, \ell_3\}$ together with the LLM's final hidden states at image-token positions:

$$F_k \;=\; \begin{cases} \phi\big(h^{(\ell_k)}\big) \in \mathbb{R}^{N_v\times d}, & k=1,2,3, \quad \text{(ViT intermediate layers)} \\ H^{\text{LLM}}_{\mathcal{M}_v} \in \mathbb{R}^{N_v\times d}, & k=4, \quad \text{(LLM final layer)} \end{cases} \tag{2}$$

where $\mathcal{M}_v$ selects LLM hidden states at image-token positions. $F_{1,2,3}$ capture purely visual features with increasing abstraction, while F_4 encodes vision-language contextualized representations.

Unlike the original DPT (Ranftl et al., 2021) that operates on native ViT features, visual tokens in a VLM are already downsampled by the patch merger (Bai et al., 2025). We therefore avoid additional downsampling and instead construct a bottom-up pyramid via upsampling, assigning higher spatial resolution to earlier ViT layers. Specifically, each F_k is projected with a 1×1 convolution and resampled to a layer-specific resolution, yielding finer spatial details for shallower features. The resulting multi-scale features are fused with RefineNet blocks (Lin et al., 2017) and decoded into a dense metric depth map at the input resolution:

$$\hat{D} \;=\; \text{DPT}(F_1, F_2, F_3, F_4) \in \mathbb{R}^{H\times W}, \qquad \big(\hat{D},\, \hat{T}\big) \;=\; \text{DepthVLM}(I, T), \tag{3}$$

where a final Softplus activation ensures strictly positive depth values. In this way, our model jointly generates dense metric geometry $\hat{D}$ and text response $\hat{T}$ within a unified foundation model.

3.2 Two-Stage Training Strategy

To introduce dense geometry prediction while preserving the original multimodal understanding, we adopt a two-stage training strategy. In the first stage, we train only the depth head to initialize dense

depth prediction capability. In the second stage, we unfreeze the LLM backbone and fine-tune the model end-to-end, enabling tighter integration of geometric prediction with multimodal reasoning.

Stage-1: Depth Head-Only Training. Since the depth head is randomly initialized, directly training it with the VLM may introduce noisy gradients and disrupt pretrained knowledge. We therefore freeze the entire VLM and train only the depth head. Following standard practice (Yang et al., 2024b), we supervise the predicted depth map $\hat{D}$ with the scale-invariant logarithmic (SILog) loss:

$$\mathcal{L}_{\text{depth}} \;=\; \sqrt{\frac{1}{|\Omega|}\sum_{i\in\Omega} d_i^2 \;-\; \lambda\Big(\frac{1}{|\Omega|}\sum_{i\in\Omega} d_i\Big)^2}, \qquad d_i = \log \hat{D}_i - \log D_i^*, \tag{4}$$

where Ω denotes pixels with valid ground-truth depth D^* and λ provides a balanced inductive bias, preserving metric supervision while reducing sensitivity to dataset-specific scale variations.

Stage-2: End-to-End Fine-Tuning. To further strengthen geometric prediction in synergy with the VLM's inherent language interaction capability, we unfreeze the LLM backbone and perform end-to-end fine-tuning on the same depth data, paired with a depth instruction and fixed response. The objective combines the autoregressive language modeling loss with the depth loss from Stage 1:

$$\mathcal{L}_{\text{joint}} \;=\; \mathcal{L}_{\text{text}} \;+\; \alpha\,\mathcal{L}_{\text{depth}}, \qquad \mathcal{L}_{\text{text}} \;=\; -\sum_t \log p_\theta\Big(\hat{T}_t \,\Big|\, \hat{T}_{<t},\, I,\, T\Big), \tag{5}$$

where $\mathcal{L}_{\text{text}}$ is the standard cross-entropy loss over response tokens and α balances the two objectives.

3.3 Mixed-Source Data Curation

Eliminating Camera Ambiguity. Joint training across datasets faces camera-induced scale ambiguity. Images with different focal lengths can depict similar scenes but correspond to inconsistent metric depths, leading to conflicting supervision and poor generalization. We address this by adopting focal-length normalization following prior works (Cai et al., 2025; Piccinelli et al., 2024), rescaling all images to a unified focal length f_c to remove dataset-specific biases and enforce consistent pixel-to-metric mapping. Formally, given an image I with focal length f and depth map D, we apply:

$$s = f_c \,/\, f, \qquad \tilde{I} \;=\; \mathcal{R}_s(I), \qquad \tilde{D} \;=\; \mathcal{R}_s(D), \tag{6}$$

where $\mathcal{R}_s(\cdot)$ denotes isotropic bilinear resizing. This removes cross-dataset scale discrepancies and enables the model to learn a focal-invariant mapping that generalizes well to open-world images.

DepthVLM-Bench. We assemble diverse public datasets into a unified benchmark for training VLMs on dense geometry prediction and comparison with pure vision models under a consistent protocol.

Training split. We mix the training set of 8 datasets covering indoor and outdoor scenes: ScanNet++ (Yeshwanth et al., 2023), Taskonomy (Zamir et al., 2018), HM3D (Ramakrishnan et al., 2021), and Matterport3D (Chang et al., 2017) for indoor, and Argoverse2 (Wilson et al., 2023), Waymo (Sun et al., 2020), DDAD (Guizilini et al., 2020), and NuScenes (Caesar et al., 2020) for outdoor. Notably, our model achieves comparable performance using an order of magnitude less data than pure vision models, which typically leverage 20+ datasets with extensive synthetic data.

Evaluation split. We evaluate on held-out validation or test splits that are disjoint from the training images: 4 indoor (ScanNet++, sunRGBD (Song et al., 2015), IBims-1 (Koch et al., 2018), NYUv2 (Silberman et al., 2012)), 4 outdoor (Argoverse2, Waymo, DDAD, NuScenes), and ETH3D (Schops et al., 2017) containing both indoor and outdoor scenes. For each dataset, we sample 1k images and 10 pixels per image (10k pixels total), oversampling smaller datasets when needed.

4 Experiment

4.1 Experimental Settings

Baselines and Metrics. We compare our model against VLMs and pure vision models. Baselines include four groups: (i) *general-purpose VLMs*: Qwen3-VL, InternVL3.5, GPT-4o, GPT-5.5; (ii) *spatially-enhanced VLMs*: SpaceLLaVA-13B, SpatialRGPT-8B, Cambrian-S-7B (Yang et al., 2025c); (iii) *depth-specialized VLMs*: Youtu-VL-4B, DepthLM-12B; and (iv) *pure vision models*: ZoeDepth,

Table 1: **Comparison with existing VLMs on metric depth estimation across diverse indoor and outdoor datasets.** For VLMs not explicitly trained for this task, we adopt the prompting strategy proposed in DepthLM (Cai et al., 2025) to elicit their best performance. Even the state-of-the-art GPT-5.5 (OpenAI., 2025) attains a δ_1 of only around 0.4, highlighting the difficulty of the task for prevailing VLMs. **Bold** and underlined values denote the best and second-best results, respectively.

δ_1 (↑) of various methods	*Outdoor*				*Out+In*	*Indoor*				**Avg.**
	Argoverse2	Waymo	DDAD	NuScenes	ETH3D	ScanNet++	sunRGBD	IBims-1	NYUv2	
Naive Prediction with Constant Answers										
Always Output 2.0m	0.002	0.004	0.002	0.010	0.112	0.261	0.380	0.269	0.373	0.157
General-Purpose VLMs										
GPT-4o	0.141	0.139	0.193	0.174	0.296	0.369	0.358	0.393	0.394	0.273
GPT-5.5	0.378	0.368	0.304	0.276	0.369	0.432	0.527	0.483	0.525	0.407
Qwen3-VL-4B	0.033	0.006	0.086	0.040	0.085	0.208	0.220	0.037	0.155	0.097
Qwen3-VL-8B	0.119	0.061	0.243	0.118	0.175	0.274	0.286	0.081	0.194	0.172
Qwen3-VL-32B	0.029	0.017	0.138	0.049	0.150	0.434	0.536	0.105	0.435	0.210
InternVL3.5-8B	0.139	0.110	0.177	0.111	0.185	0.395	0.459	0.214	0.431	0.247
InternVL3.5-14B	0.086	0.045	0.128	0.084	0.183	0.390	0.440	0.235	0.445	0.226
InternVL3.5-38B	0.131	0.129	0.219	0.110	0.181	0.400	0.431	0.155	0.423	0.242
Spatial-Enhanced VLMs										
SpaceLLaVA-13B	0.006	0.002	0.001	0.006	0.107	0.050	0.044	0.172	0.087	0.053
SpatialRGPT-8B	0.045	0.064	0.096	0.116	0.133	0.124	0.084	0.044	0.070	0.086
Cambrian-S-7B	0.006	0.019	0.038	0.033	0.069	0.145	0.073	0.057	0.063	0.056
VLMs Trained on Metric Depth Estimation										
Youtu-VL-4B	0.663	0.473	0.342	0.698	0.286	0.522	0.734	0.856	0.849	0.603
DepthLM-12B	0.761	0.588	0.654	0.736	0.666	0.756	0.785	0.754	0.866	0.730
Ours-4B	**0.810**	**0.879**	**0.818**	0.821	0.924	0.861	0.882	0.912	0.908	0.868
Ours-8B	0.798	0.865	0.813	**0.831**	**0.928**	**0.901**	**0.889**	**0.936**	**0.920**	**0.876**

Table 2: **Comparison with specialized pure vision models on metric depth estimation across indoor and outdoor datasets.** Despite being a unified VLM that preserves strong multimodal capabilities, our method can outperform state-of-the-art pure vision specialists, demonstrating that dense geometry prediction can emerge natively within a single vision-language foundation model.

δ_1 (↑) of various methods	*Outdoor*		*Out+In*	*Indoor*		**Avg.**
	Waymo	NuScenes	ETH3D	sunRGBD	IBims-1	
ZoeDepth	0.639	0.196	0.345	0.769	0.718	0.533
Depth Pro	0.255	0.389	0.355	0.852	0.880	0.546
Metric3D	0.879	0.721	0.373	0.222	0.796	0.598
Metric3Dv2	0.923	0.747	0.851	0.813	0.726	0.812
UniDepth	0.670	0.858	0.149	0.907	0.158	0.548
UniDepthV2	0.730	0.872	0.657	0.911	0.941	0.823
DepthAnything	0.739	0.205	0.277	0.847	0.854	0.584
DepthAnythingV2	0.715	0.168	0.111	0.697	0.887	0.516
DepthAnythingV3	0.885	0.790	0.843	0.913	0.955	0.877
Ours-4B	0.879	0.821	0.924	0.882	0.912	0.884
Ours-8B	0.865	0.831	0.928	0.889	0.936	**0.890**

Depth Pro (Bochkovskii et al., 2024), UniDepth, Metric3D, DepthAnything. Following standard practice, we report δ_1 accuracy, the percentage of predictions within 25% relative error.

Implementation Details. We adopt Qwen3-VL (4B/8B) as the default VLM backbone, and integrate a lightweight DPT-style head with 34M parameters ($< 1\%$ of the LLM). Models are trained in PyTorch on 4.4M samples from the training split of DepthVLM-Bench with uniform sampling. Intermediate ViT features are taken from layers 5, 11, and 17 for 4B, and 8, 16, and 24 for 8B. We use AdamW with a cosine schedule, learning rates of 3.5×10^{-4} and 2×10^{-5}, and warmup ratios of 0.04 and 0.05 for Stage-1 and Stage-2. The balance factors λ and α are set to 0.5 and 1.0.

4.2 Main Results

Comparison with Other VLMs. To evaluate metric depth estimation in existing VLMs, we follow DepthLM by prompting models with an arrow-marked pixel to predict its depth. As shown in Table 1, general-purpose VLMs perform poorly—especially in outdoor driving scenes—with Qwen3-

Table 3: **Evaluation on broad visual benchmarks, covering general VQA, document understanding, multi-image reasoning, counting, and hallucination.** Empowered by our lightweight depth head and two-stage training strategy, our method natively gains the ability to generate dense geometry *without sacrificing* the general multimodal capability of the underlying VLM, in sharp contrast to prior text-heavy supervision approaches that typically incur substantial capability degradation.

Methods	MMB-EN	MMB-CN	MMStar	ScienceQA	BLINK	OCRBench	CountBench	POPE
Pixtral-12B	78.2	73.7	52.0	88.8	49.2	660	69.5	85.5
Qwen3-VL-4B	83.4	81.1	60.9	91.2	63.8	817	97.7	89.8
Qwen3-VL-8B	84.7	82.4	63.4	92.8	65.0	833	98.3	88.8
DepthLM-12B†	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Ours-4B	82.9 (↓ 0.5)	81.8 (↑ 0.7)	60.4 (↓ 0.5)	91.3 (↑ 0.1)	63.3 (↓ 0.5)	832 (↑ 15)	98.0 (↑ 0.3)	89.9 (↑ 0.1)
Ours-8B	84.6 (↓ 0.1)	82.3 (↓ 0.1)	63.8 (↑ 0.4)	93.1 (↑ 0.3)	64.8 (↓ 0.2)	862 (↑ 29)	98.2 (↓ 0.1)	89.1 (↑ 0.3)

† Due to text-dominant fine-tuning with single-pixel depth queries, DepthLM collapses to always emitting a depth value regardless of the instruction, making it incompatible with standard VQA evaluation.

Table 4: **Evaluation on spatial reasoning benchmarks.** We compare our model against the corresponding Qwen3-VL backbones on VSI-Bench and the 3D subset of CV-Bench.

Methods	VSI-Bench									CV-Bench (3D)		
	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	**Avg.**	Depth	Distance	**Avg.**
Qwen3-VL-4B	64.9	43.0	73.1	**60.5**	51.8	45.5	30.9	61.4	53.9	93.2	87.8	90.5
Ours-4B	**67.6**	**45.8**	**75.7**	60.2	**53.7**	**48.9**	**33.4**	**64.6**	**56.2**	**96.3**	**91.2**	**93.8**
Qwen3-VL-8B	67.5	45.3	73.3	60.0	54.5	46.3	30.9	64.5	55.3	93.8	88.8	91.3
Ours-8B	**69.4**	**48.8**	**75.6**	**62.2**	**57.7**	**49.2**	**33.5**	**67.3**	**58.0**	**96.6**	**92.2**	**94.4**

VL-32B achieving $\delta_1 = 0.21$ and GPT-5.5 only 0.41 on average, revealing a substantial gap to reliable 3D understanding. Even spatially enhanced VLMs, despite depth and calibration supervision, underperform a constant-depth baseline. In contrast, our model consistently excels across indoor and outdoor settings, significantly outperforming both larger and task-specific VLMs.

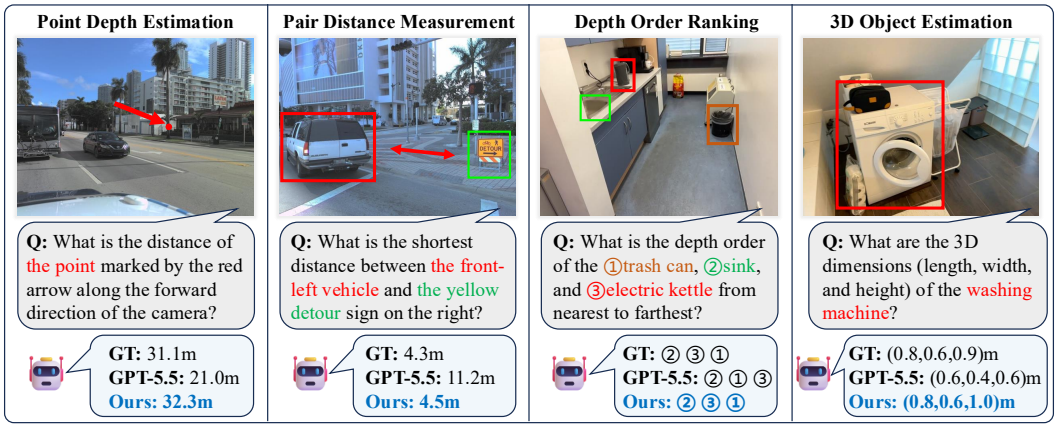


Figure 4: **Qualitative results on more complex 3D tasks.** Beyond dense metric depth estimation, our model further supports a variety of downstream 3D reasoning tasks, demonstrating that native dense geometry prediction serves as a solid foundation for high-level spatial reasoning in VLMs.

Comparison with Pure Vision Models. Table 2 further compares our model with leading specialized pure vision models on indoor and outdoor metric depth estimation. Since both pure vision models and DepthVLM produce dense metric depth maps, we evaluate them on the same sampled pixels used in the VLM setting for a fair comparison. Despite being a unified model with strong multimodal capabilities, our method not only significantly outperforms most vision specialists, including UniDepthV2 and Metric3Dv2, but also surpasses the state-of-the-art DepthAnythingV3.

Evaluation on General Visual Benchmarks. To verify that dense geometry prediction does not compromise multimodal understanding, we evaluate on broad visual benchmarks in Table 3. Our models match their original VLM backbones and even improve on OCRBench and POPE. In contrast, prior depth-specialized VLMs such as DepthLM often overfit to text-heavy supervision and lose

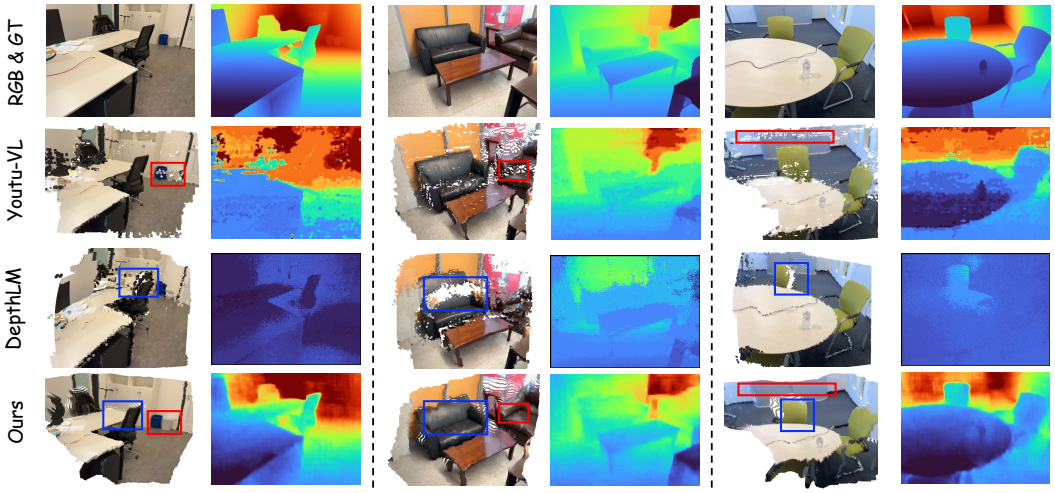


Figure 5: **Qualitative comparison with others.** Our results show finer structural details and improved semantic consistency across diverse scenes. Depth is color-coded from near ( ) to far.

Table 5: **Ablation of depth head designs.** "Multi-scale" indicates aggregation of visual features from intermediate ViT layers. Our lightweight DPT-style head with multi-scale fusion performs best due to the design tailored to VLM features.

Depth Head	Multi-scale	Waymo	NuScenes	sunRGBD	IBims-1
Two-layer MLP	✗	0.547	0.444	0.533	0.695
Two-layer MLP	✓	0.723	0.776	0.727	0.806
Original DPT	✓	0.866	**0.826**	0.856	0.895
Ours (Lightweight DPT)	✓	**0.879**	0.821	**0.882**	**0.912**

Table 6: **Ablation of feature sources for the depth head.** "Inter." denotes three intermediate layers and "Final" the last-layer output. Fusing multi-scale ViT features with the LLM final feature yields the best depth fidelity.

Feature Source	Waymo	NuScenes	sunRGBD	IBims-1
ViT Inter. + ViT Final	0.758	0.712	0.798	0.777
LLM Inter. + LLM Final (Stage-1 only)	0.720	0.710	0.763	0.744
LLM Inter. + LLM Final (Two-stage)	0.812	0.770	0.846	0.829
ViT Inter. + LLM Final (Two-stage)	**0.879**	**0.821**	**0.882**	**0.912**

general-purpose capabilities. These results underscore the effectiveness of our unified design, which supports both accurate dense geometry prediction and strong multimodal understanding.

Evaluation on Spatial Reasoning Tasks. We further surprisingly find that enabling a VLM to act as a native 3D dense geometry predictor also improves spatial reasoning performance. As shown in Table 4, our model consistently outperforms the baseline on VSI-Bench (*e.g.*, route planning and relative direction), achieving a 2.7% overall gain, and also shows clear improvements on CV-Bench. Figure 4 further demonstrates more complex 3D reasoning tasks beyond metric depth estimation, where even pioneering GPT-5.5 may fail. These results suggest that strong native dense geometry prediction provides a solid foundation for high-level spatial reasoning in VLMs.

Qualitative Visualizations. As shown in Figure 5, we compare the depth maps and corresponding 3D point clouds generated by Youtu-VL-4B and DepthLM-12B across diverse scenes. Youtu-VL produces noisy and fragmented point clouds with poor geometric continuity, while DepthLM maintains better semantic coherence but loses fine structural details. In contrast, our method significantly improves generation quality, preserving both semantic consistency and detailed spatial structure.

4.3 Ablation Studies

In this section, we perform ablation experiments on the 4B model to thoroughly evaluate the effectiveness of the components.

Ablation of Depth Head Variants. We compare depth heads in Table 5. A two-layer MLP performs worst due to its limited capacity, while multi-scale ViT features improve performance. The original DPT head remains suboptimal, as downsampling the final LLM visual feature loses semantic information. In contrast, our lightweight DPT-style head builds a bottom-up feature pyramid by upsampling, assigning higher spatial resolution to earlier ViT layers, and achieves the best accuracy.

Ablation of Multi-Layer Feature Sources. Table 6 studies the impact of feature inputs to the DPT-style head. Using only ViT features underperforms due to limited learnable parameters. LLM-only features improve results but remain suboptimal, as they lack fine-grained geometry details from early

Table 7: **Ablation of training strategies.** We compare four variants: (i) *Stage-1 Only*, training only the depth head with the VLM frozen; (ii) *Stage-2 Only*, directly fine-tuning the full model; (iii) *Stage-1 + Stage-2*‡, where the vision encoder is unfrozen in Stage-2; and (iv) our full strategy. Unfreezing the vision encoder yields marginal depth gains but degrades general multimodal performance, whereas our design achieves strong depth estimation accuracy while preserving the VLM's general capability.

Training Strategy	Depth Estimation ($\delta_1 \uparrow$)				General Visual Benchmarks			
	Waymo	NuScenes	sunRGBD	IBims-1	MMB-EN	MMStar	BLINK	OCRBench
Stage-1 Only	0.737	0.742	0.782	0.753	**83.23**	60.36	**63.55**	**840**
Stage-2 Only	0.784	0.762	0.826	0.805	81.44	57.21	57.77	793
Stage-1 + Stage-2‡ (unfreeze ViT)	**0.884**	**0.837**	**0.893**	0.900	82.13	54.60	59.47	769
Stage-1 + Stage-2 (freeze ViT)	0.879	0.821	0.882	**0.912**	82.93	**60.42**	63.25	832

Table 8: **Ablation of focal-length normalization.** We compare training on raw mixed-source images (*w/o normalization*) with canonicalizing inputs to a shared focal length $f_c \in \{800, 1000, 1200\}$. Normalization consistently improves performance, with $f_c = 1000$ achieving the best results across diverse benchmarks.

Training Setting	Waymo	NuScenes	sunRGBD	IBims-1
Raw mixed-source	0.802	0.715	0.770	0.630
Canonical $f_c = 800$	0.833	0.824	0.865	0.883
Canonical $f_c = 1000$	**0.879**	0.821	**0.882**	**0.912**
Canonical $f_c = 1200$	0.858	**0.840**	0.837	0.856

Table 9: **Efficiency comparison with others.** We report the end-to-end cost of generating a 256×192 depth map. DepthLM uses per-pixel queries, while Youtu-VL predicts a sparse grid with upsampling. In contrast, our model outputs a pixel-aligned depth map in one pass, achieving higher efficiency without post-processing.

Method	#Fwd. / image	Output Pattern	Post-proc.	Latency (ms) $\downarrow$
DepthLM-12B	$H \times W$	point-wise queries	none	13h
Youtu-VL-4B	1	sparse patch-level	bilinear $\uparrow$	2.48s
Ours-4B	1	dense pixel-level	none	**0.42s**

ViT layers. Combining multi-scale ViT features with the LLM final hidden state yields the best performance, effectively integrating high-level semantic information with detailed geometry cues.

One-Stage vs. Two-Stage Training. We compare different training strategies in Table 7, including single-stage and two-stage variants. *Stage-1 Only* preserves multimodal reasoning but yields limited depth gains. *Stage-2 Only* improves geometry prediction, yet suffers from unstable optimization and reduced multimodal performance. The full two-stage strategy achieves a better balance. Unfreezing the vision encoder in Stage-2 further improves depth accuracy but harms general multimodal ability. Overall, our final design balances dense geometry prediction with the VLM's general capability.

Effect of Focal-Length Normalization. We validate the focal-length normalization strategy in Table 8. Training on raw mixed-source data suffers from camera-induced scale ambiguity, as identical scenes with different focal lengths yield inconsistent metric depth supervision. Normalizing inputs to a shared focal length f_c aligns projective geometry across datasets and mitigates this issue. We sweep $f_c \in \{800, 1000, 1200\}$: smaller f_c loses image details, while larger f_c amplifies interpolation artifacts. Empirically, $f_c = 1000$ performs best across datasets, consistent with DepthLM.

Inference Efficiency Analysis. We compare the efficiency of VLM-based methods in Table 9 using 256×192 inputs and reporting end-to-end runtime. DepthLM formulates depth estimation as per-pixel text queries, requiring $H \times W$ forward passes, making inference prohibitively slow. Youtu-VL reduces this to one pass but predicts sparse patches that require upsampling, introducing artifacts and overhead. In contrast, our method directly decodes multi-scale features into a pixel-aligned depth map in one pass without post-processing. This efficiency stems from treating the VLM as a native dense predictor, enabling efficient dense geometry prediction within a unified framework.

5 Conclusion and Limitations

In this paper, we present DepthVLM, a unified foundation model that jointly supports low-level dense geometry prediction and high-level multimodal understanding. We integrate a lightweight depth head into the VLM backbone and adopt a two-stage training strategy under the unified vision-text supervision, enabling geometry prediction and language response in a single forward pass. Extensive experiments show that DepthVLM achieves leading performance across diverse datasets with higher inference efficiency, surpasses strong pure vision models, and improves complex 3D spatial reasoning.
Limitations. This work focuses on dense metric depth estimation and will extend to 3D tasks like object detection and pose estimation, toward a unified model for holistic 3D perception and reasoning.

AI USE STATEMENT

During the preparation of this paper, we used generative AI tools solely for language polishing and readability improvement. All scientific contributions, ideas, and evaluations presented in this work were developed and validated independently by the authors, without reliance on LLMs.

ETHICS STATEMENT

We affirm that this work adheres to the ICLR Code of Ethics. The study involves no human subjects or datasets raising concerns regarding privacy, consent, or security. All datasets are properly sourced with the necessary permissions and licenses, as detailed in Table 19. We declare no conflicts of interest or funding from organizations with a vested interest in the research outcome. We have also considered potential biases, fairness, and discrimination to minimize the risk of harm or inequality.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of the experiments and results presented in this paper. We will fully release the source code and model checkpoints, along with detailed descriptions of data preprocessing, evaluation metrics, and other procedures necessary for replication. These resources promote transparency and enable researchers to verify, build upon, and extend our work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4009–4018, 2021.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Aleksei Bochkovskii, AmaãGl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

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, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. Depthlm: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*, 2025.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

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

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pp. 202–221. Springer, 2020.

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. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

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.

JiaKui Hu, Shanshan Zhao, Qing-Guo Chen, Xuerui Qiu, Jialun Liu, Zhao Xu, Weihua Luo, Kaifu Zhang, and Yanye Lu. Omni-view: Unlocking how generation facilitates understanding in unified 3d model based on multiview images. In *International Conference on Learning Representations*, volume 2026, pp. 133697–133724, 2026.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, 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):10579–10596, 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. *arXiv preprint arXiv:2511.21688*, 2025.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3drs: Mllms need 3d-aware representation supervision for scene understanding. *arXiv preprint arXiv:2506.01946*, 2025.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1925–1934, 2017.

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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16488–16498, 2024.

OpenAI. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

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*, 2025.

Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2109.08238.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

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.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

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, et al. 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.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9275–9286, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025c.

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*, 2025d.

Zhixiang Wei, Yi Li, Zhehan Kan, Xinghua Jiang, Zuwei Long, Shifeng Liu, Hongze Shen, Wei Liu, Xiaoyu Tan, Haojia Lin, et al. Youtu-vl: Unleashing visual potential via unified vision-language supervision. *arXiv preprint arXiv:2601.19798*, 2026.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

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.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975, 2024.

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. *arXiv preprint arXiv:2505.17015*, 2025.

Yibin Yan, Jilan Xu, Shangzhe Di, Haoning Wu, and Weidi Xie. Omnistream: Mastering perception, reconstruction and action in continuous streams. *arXiv preprint arXiv:2603.12265*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

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. In *CVPR*, 2025b.

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. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024b.

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*, 2025c.

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.

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.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *Advances in Neural Information Processing Systems*, 38, 2026.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *arXiv preprint arXiv:2505.24625*, 2025a.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8995–9006, 2025b.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024.

TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A STATISTICS OF DEPTHVLM-BENCH

We summarize the data composition of DepthVLM-Bench in Tables 10 and 11.

For training, we construct a large-scale and diverse dataset by sampling from the *training splits* of multiple public benchmarks spanning both autonomous driving and indoor scenes, including Argoverse2 (Wilson et al., 2023), Waymo (Sun et al., 2020), DDAD (Guizilini et al., 2020), NuScenes (Caesar et al., 2020), ScanNet++ (Yeshwanth et al., 2023), Taskonomy (Zamir et al., 2018), HM3D (Ramakrishnan et al., 2021), and Matterport3D (Chang et al., 2017). As many RGB images are extracted from videos and thus contain highly redundant, near-duplicate frames, we apply uniform sampling to reduce redundancy. Most datasets contribute approximately 800K images, resulting in a relatively balanced distribution across domains and helping mitigate dataset bias. Smaller datasets, such as DDAD and Matterport3D, are included at their original scales to further enrich diversity. In total, the training set comprises 4.4M images, covering a wide range of scenes, viewpoints, and depth distributions. We train for a single epoch. The 8B variant is trained for four days on 80 NVIDIA H20 GPUs, while the 4B variant is trained for two days using the same computational resources.

For evaluation, we assemble a comprehensive benchmark using the *validation or test splits* of the same public datasets, covering both outdoor and indoor scenarios. When possible, each dataset contributes around 1K images to ensure a balanced and consistent evaluation protocol across sources. In addition, we include several standard indoor benchmarks, including ETH3D (Schops et al., 2017), sunRGBD (Song et al., 2015), IBims-1 (Koch et al., 2018), and NYUv2 (Silberman et al., 2012), to further assess generalization across diverse environments and capture varying depth characteristics. This design enables a fair and reliable evaluation of cross-domain generalization and robustness for dense geometry prediction.

Table 10: **Training data statistics.** Number of training images sampled from each source dataset in our setting.

Dataset	Argoverse2	Waymo	DDAD	NuScenes	ScanNet++	Taskonomy	HM3D	Matterport3D
# Images	800K	800K	76K	206K	800K	800K	800K	158K

Table 11: **Benchmark data statistics.** Number of evaluation images sampled from each dataset.

Dataset	Argoverse2	Waymo	DDAD	NuScenes	ETH3D	ScanNet++	sunRGBD	IBims-1	NYUv2
# Images	1K	1K	1K	1K	454	1K	1K	100	654

B EVALUATION OF VLMS ON METRIC DEPTH ESTIMATION

To fairly compare the metric depth estimation capabilities of existing VLMs with our model, we design a standardized evaluation protocol described below.

Following the paradigm introduced by DepthLM (Cai et al., 2025), given a single RGB image, a specific pixel location is indicated by a **red arrow marker** drawn directly on the image. The model is asked to estimate the metric depth (in meters), *i.e.*, the distance along the camera's optical axis from that point to the camera, as illustrated in Figure 6.

Visual Marker Size. DepthLM uses a default red arrow marker of **5 pixels**. However, through our experiments, we observe that off-the-shelf VLMs that have not been specifically trained for this task cannot reliably detect a 5 pixels marker. When not explicitly constrained to output only a number, these models frequently respond with statements such as:

"There is no red arrow in the image. Therefore, the distance cannot be determined."

To ensure fairness and to maximally elicit each model's depth estimation capability, we increase the red arrow marker size to **20 pixels** for all evaluated VLMs. This larger marker is clearly visible in

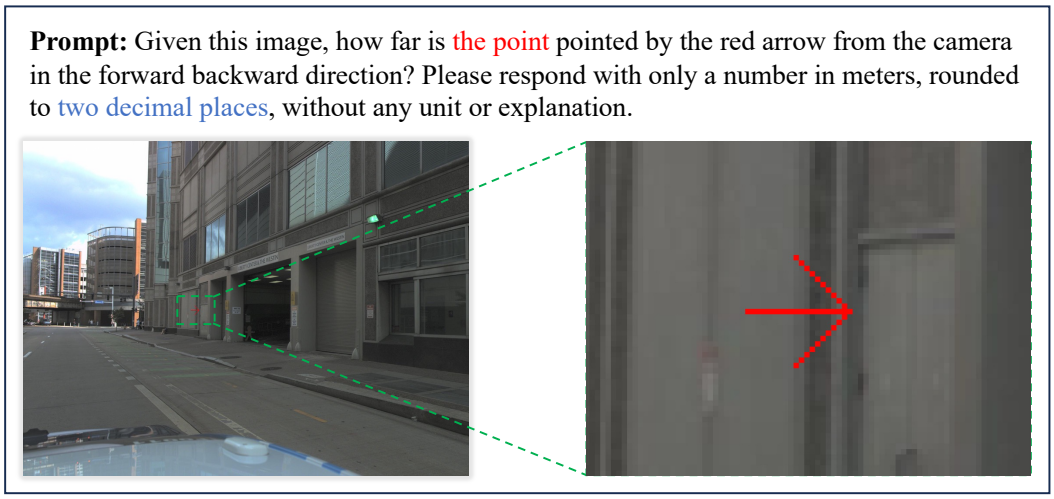


Figure 6: **VLM evaluation setup for metric depth estimation.** A red arrow marker (20 pixels) is drawn on the input image to indicate the query point. The model receives the annotated image along with a text prompt asking for the metric depth (in meters). The zoomed-in region shows the arrow marker in detail.

all tested scenes, ensuring that the evaluation measures depth estimation ability rather than marker detection ability.

Image Resolution Handling. Since different datasets are captured at varying resolutions, we downscale images with a longest edge over 1024 pixels to this limit *before* adding the red arrow marker, while leaving smaller images unchanged. This provides a consistent input condition across all models and datasets, and ensures that the marker remains clearly visible at the given image resolution.

With the above protocol, all VLMs are evaluated under identical conditions, ensuring a fair comparison with our method. For VLMs such as DepthLM (Cai et al., 2025) and Youtu-VL (Wei et al., 2026), which are specifically trained for metric depth estimation, we directly adopt their provided evaluation scripts.

Pixel Sampling and Metric. For each evaluation image, we uniformly sample 10 valid ground-truth pixels at the original resolution after dataset-specific depth-range filtering and invalid-pixel removal. A fixed random seed ensures that all methods are evaluated at identical pixel locations. Sampling is used only for evaluation; DepthVLM is trained with dense supervision over all valid pixels and predicts a complete depth map in one pass. This sampled protocol is necessary because query-based VLMs such as DepthLM require a full VLM forward pass and autoregressive decoding for each queried pixel. For valid evaluation pixels $i \in \Omega$, we compute

$$\delta_1 = \frac{1}{|\Omega|} \sum_{i \in \Omega} \mathbb{I}\left[\max\left(\frac{\hat{D}_i}{D_i^*}, \frac{D_i^*}{\hat{D}_i} \right) < 1.25 \right], \tag{7}$$

which is the fraction of predictions within 25% relative error of the metric ground-truth depth.

Baseline Handling. All open-source baselines use their official released weights, while closed-source models are evaluated through their official APIs. For depth-specialized VLMs, we adapt the official evaluation scripts to our common split and sampled locations. In particular, we re-evaluate DepthLM on DepthVLM-Bench rather than mixing its self-reported numbers with our results, because its original evaluation image list and sampled locations are not publicly available. Together with the settings above, this ensures that all methods are compared under a consistent protocol.

C Additional Experimental Results and Analysis

This section presents additional controlled experiments, implementation details, and analyses that complement the main paper. Unless otherwise stated, depth accuracy is measured by δ_1 on the

Table 12: **Comparison with specialized pure-vision models for metric depth estimation across indoor and outdoor datasets, requiring dense full-image depth map prediction.**

$\delta_1(\uparrow)$ of various methods	*Outdoor*		*Out+In*	*Indoor*		**Avg.**
	Waymo	NuScenes	ETH3D	sunRGBD	IBims-1	
ZoeDepth (Bhat et al., 2023)	0.634	0.195	0.350	0.770	0.717	0.533
Depth Pro (Bochkovskii et al., 2024)	0.251	0.379	0.365	0.847	0.879	0.544
Metric3D (Yin et al., 2023)	0.878	0.719	0.409	0.211	0.794	0.602
Metric3Dv2 (Hu et al., 2024)	0.926	0.843	0.865	0.909	0.724	0.853
UniDepth (Piccinelli et al., 2024)	0.683	0.846	0.145	0.900	0.159	0.547
UniDepthV2 (Piccinelli et al., 2025)	0.733	0.868	0.660	0.912	0.941	0.823
DepthAnything (Yang et al., 2024a)	0.749	0.195	0.290	0.843	0.856	0.587
DepthAnythingV2 (Yang et al., 2024b)	0.724	0.165	0.093	0.674	0.887	0.509
DepthAnythingV3 (Lin et al., 2025)	0.889	0.788	0.822	0.911	0.957	0.873
Ours-4B	0.902	0.838	0.963	0.885	0.910	<u>0.900</u>
Ours-8B	0.884	0.837	0.959	0.890	0.939	**0.902**

DepthVLM-Bench evaluation split, and all compared methods use the same sampled pixels or the same set of valid dense pixels.

C.1 Dense Full-Image Evaluation

While the sampled-pixel protocol enables fair comparison with query-based VLMs, it does not fully evaluate models that natively predict dense full-image depth maps. We therefore evaluate all specialized pure vision baselines and both DepthVLM variants over the complete set of valid pixels in Table 12. DepthVLM-8B achieves the best average δ_1 of 0.902, significantly outperforming the state-of-the-art DepthAnythingV3. This consistent advantage confirms that our model captures dense spatial structure rather than only performing well at a small set of query locations.

C.2 Additional Controlled Experiments

Contribution of LLM Representations. To isolate the contribution of LLM representations to depth estimation, we remove the LLM from DepthVLM-4B and attach the same DPT-style head directly to the vision encoder. As shown in Table 13, this variant reduces the average δ_1 from 0.884 to 0.710, confirming that the language-model representation contributes substantially to dense depth prediction. The LLM-free variant cannot be evaluated on language-based spatial reasoning because it does not support language generation.

Table 13: **Ablation of the LLM block in DepthVLM-4B** ($\delta_1 \uparrow$). The LLM-free variant attaches the same depth head directly to the vision encoder.

Method	**NuScenes**	**Waymo**	**ETH3D**	**IBims-1**	**sunRGBD**	**Avg.**
w/o LLM	0.712	0.758	0.507	0.777	0.798	0.710
w/ LLM	**0.821**	**0.879**	**0.924**	**0.912**	**0.882**	**0.884**

Effect of Training-Data Scale. Table 14 evaluates data efficiency by comparing training on the full 4.4M-image corpus with a 1.4M-image subset. Remarkably, using only one third of the training data retains 94.7% of the full model's average accuracy, reaching 0.837 versus 0.884. This result shows that the performance gains do not arise solely from the number of depth-training images and provides a lower-cost operating point.

Table 14: **Effect of depth-training data scale for DepthVLM-4B** ($\delta_1 \uparrow$). We compare the full 4.4M-image corpus with a 1.4M-image subset to evaluate the impact of training-data volume on depth accuracy.

Training Data	Waymo	NuScenes	ETH3D	sunRGBD	IBims-1	Avg.
1.4M images	0.807	0.802	0.838	0.850	0.890	0.837
4.4M images	**0.879**	**0.821**	**0.924**	**0.882**	**0.912**	**0.884**

Comparison with Spatially Enhanced VLMs. We further evaluate the spatial reasoning capability of DepthVLM-8B against recent spatially enhanced VLMs in Table 15. Our model achieves 58.0 on VSI-Bench (Yang et al., 2025b) and 94.4 on the 3D subset of CV-Bench (Tong et al., 2024), outperforming all compared methods on both benchmarks. Together with the same-backbone analysis in the main paper, these results show that the spatial reasoning gains cannot be attributed to the Qwen3-VL backbone alone, but instead arise from the additional low-level dense geometry prediction capability enabled by our unified model.

Table 15: **Comparison with other spatially enhanced VLMs.** Results are reported following the official evaluation protocols of VSI-Bench and CV-Bench.

Method	Base Model	VSI-Bench	CV-Bench (3D)
SPAR-8B (Zhang et al., 2026)	InternVL2.5-8B	47.3	89.1
VG-LLM-8B (Zheng et al., 2025a)	Qwen2.5-VL-7B	50.7	91.1
Omni-View-7B (Hu et al., 2026)	BAGEL-7B	55.4	90.3
InternVL3.5-8B (Wang et al., 2025d)	Qwen3-8B	56.3	85.2
Ours-8B	Qwen3-VL-8B	**58.0**	**94.4**

C.3 Detailed Efficiency Analysis

All latency measurements denote the end-to-end time required to produce a depth map using BF16 precision, FlashAttention-2, and KV caching. Results reported in the main paper are measured on a single NVIDIA H20-96GB GPU, while the additional control experiments below are conducted on a single NVIDIA A800-80GB GPU. Therefore, slight differences in latency are expected due to the different GPU hardware.

Efficiency of the Depth Head. Table 16 compares the efficiency of our lightweight depth head with that of the original DPT decoder under identical settings. Under the same backbone, input resolution, hardware, and inference configuration, our design reduces parameters by 62.3%, FLOPs by 46.5%, peak head memory by 61.5%, and end-to-end latency by 62.2%.

Table 16: **Efficiency comparison with the original DPT head.** Measurements use 256×192 inputs on a single A800-80GB GPU. Peak memory is measured for the depth head, while latency is measured end-to-end.

Depth Head	#Params↓	FLOPs↓	Peak Mem.↓	Latency↓
Original DPT (Ranftl et al., 2021)	88.9M	56.4G	0.13GB	0.74s
Ours	**33.5M**	**30.2G**	**0.05GB**	**0.28s**

Scaling with Input Resolution. To evaluate high-resolution efficiency, we measure end-to-end latency across input resolutions from 256×192 to 2048×1536. As shown in Table 17, increasing the pixel count by 64× raises DepthVLM's latency only from 0.28s to 0.70s because all pixels are processed in one forward pass. At the highest resolution, our model is about 35× faster than Youtu-VL and many orders of magnitude faster than the per-pixel querying used by DepthLM.

Table 17: **End-to-end latency across input resolutions.** Measurements are conducted on a single A800-80GB GPU.

Method	256×192	512×384	1024×768	2048×1536
DepthLM-12B (Cai et al., 2025)	∼16.5h	∼69.5h	∼362h	∼3863.5h
Youtu-VL-4B (Wei et al., 2026)	1.91s	2.66s	4.16s	24.78s
Ours-4B	**0.28s**	**0.29s**	**0.33s**	**0.70s**

Overhead of the Added Depth Head. As shown in Table 18, although our model is a VLM capable of generating text responses, its inference speed is comparable to that of the pure-vision model DepthAnythingV3 and becomes even faster at higher resolutions. We further isolate the overhead of the added depth head by comparing DepthVLM with the head-free Qwen3-VL backbone. The depth head introduces negligible overhead, with an additional latency of only approximately 0.002s in our measurements. At 2048×1536, DepthVLM runs in 0.70s, matching the latency of its backbone and outperforming DepthAnythingV3 at 0.99s, while additionally retaining text generation and general VQA capabilities.

Table 18: **Latency comparison with a pure-vision model and the head-free backbone.** Measurements are conducted on a single A800-80GB GPU with image inputs at different resolutions.

Method	Depth Head	Text Response	512×384	1024×768	2048×1536
DepthAnythingV3 (Lin et al., 2025)	✓	✗	**0.16s**	**0.24s**	0.99s
Qwen3-VL-4B (Bai et al., 2025)	✗	✓	0.29s	0.32s	0.70s
Ours-4B	✓	✓	0.29s	0.33s	**0.70s**

D Asset License and Consent

We provide the licenses and URLs of the public datasets and benchmarks used in this work, covering a wide range of multimodal tasks, including depth estimation, general multimodal understanding, spatial reasoning, document understanding, multi-image reasoning, visual grounding, hallucination evaluation, and related capabilities. Table 19 summarizes the resources used in this study along with their corresponding licenses and URLs.

Table 19: Open-source resources utilized in this paper.

Name	License	URL
ScienceQA	MIT License	https://github.com/lupantech/ScienceQA
POPE	MIT License	https://github.com/AoiDragon/POPE
OCRBench	MIT License	https://github.com/Yuliang-Liu/MultimodalOCR
MMBench	Apache License 2.0	https://github.com/open-compass/MMBench
BLINK	Apache License 2.0	https://github.com/zeyofu/BLINK_Benchmark
VSI-Bench	Apache License 2.0	https://github.com/vision-x-nyu/thinking-in-space
CV-Bench	Apache License 2.0	https://huggingface.co/datasets/nyu-visionx/CV-Bench
Argoverse2	Apache License 2.0	https://github.com/argoverse/argoverse-api
Waymo	Apache License 2.0	https://github.com/waymo-research/waymo-open-dataset
NuScenes	Apache License 2.0	https://github.com/nutonomy/nuscenes-devkit
MMStar	No license specified	https://github.com/MMStar-Benchmark/MMStar
CountBenchQA	No license specified	https://huggingface.co/datasets/vikhyatk/CountBenchQA
ScanNet++	ScanNet++ Terms of Use	https://scannetpp.mlsg.cit.tum.de/scannetpp/
DDAD	Creative Commons Attribution 4.0	https://github.com/TRI-ML/DDAD
ETH3D	Creative Commons Attribution 4.0	https://www.eth3d.net/datasets