# Novelty Assessment Report

**Paper**: DepthLM: Metric Depth from Vision Language Models
**PDF URL**: https://openreview.net/pdf?id=ObFVZGnSFN
**Venue**: ICLR 2026 Conference Submission
**Year**: 2026
**Report Generated**: 2025-12-29

## Abstract

Vision language models (VLMs) can flexibly address various vision tasks through text interactions. Although successful in semantic understanding, state-of-the-art VLMs including GPT-5 still struggle in understanding 3D from 2D inputs. On the other hand, expert pure vision models achieve super-human accuracy in metric depth estimation, a key 3D understanding task. However, they require task-specific architectures and losses. Such difference motivates us to ask: Can VLMs reach expert-level accuracy without architecture or loss change? We take per-pixel metric depth estimation as the representative task and show that the answer is yes! Surprisingly, comprehensive analysis shows that text-based supervised-finetuning with sparse labels is sufficient for VLMs to unlock strong 3D understanding, no dense prediction head or complex regression/regularization loss is needed. The bottleneck lies in pixel reference and cross-dataset camera ambiguity, which we address through visual prompting and intrinsic-conditioned augmentation. With much smaller models, our method DepthLM surpasses the accuracy of most advanced VLMs by over 2x, making VLMs for the first time comparable with pure vision models. Meanwhile, the simplicity makes DepthLM scalable to more complex 3D tasks with a unified model. Code will be released to the community.



## Core Task Landscape

This paper addresses: **Metric Depth Estimation from Vision Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

### Taxonomy Overview

The research landscape has been organized into the following main categories:

- **VLM Architecture and Training for Depth Perception**
- **Language-Guided Depth Estimation from Contrastive Models**
- **Diffusion Models for Language-Conditioned Depth**
- **LLM-Based Depth Understanding and Reasoning**
- **Depth-Conditioned Reasoning and Planning**
- **Benchmarking Spatial Understanding in VLMs**
- **Monocular Depth Estimation with Visual Priors**
- **Relative-to-Metric Depth Conversion**
- **Task-Specific and Domain-Adapted Depth Applications**
- **Unified Multi-Task Vision-Language Models**
- ... and 5 more categories

### Complete Taxonomy Tree

- Metric Depth Estimation from Vision Language Models Survey Taxonomy
- VLM Architecture and Training for Depth Perception
  - Depth-Aware Visual Encoding and Fusion (6 papers)
  - [2] Spatialbot: Precise spatial understanding with vision language models (Wenxiao Cai, 2025) View paper
  - [3] SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model (Cheng, 2024) View paper
  - [10] RoboFlamingo-Plus: Fusion of Depth and RGB Perception with Vision-Language Models for Enhanced Robotic Manipulation (Xiaojian Li, 2025) View paper
  - [23] Florence-VL: Enhancing Vision-Language Models with Generative Vision Encoder and Depth-Breadth Fusion (Jiuhai Chen, 2024) View paper
  - [24] Vid-LLM: A Compact Video-based 3D Multimodal LLM with Reconstruction-Reasoning Synergy (Xu Bo, 2025) View paper
  - [41] SD-VLM: Spatial Measuring and Understanding with Depth-Encoded Vision-Language Models (Chen, 2025) View paper
  - Auxiliary Depth Supervision for VLMs (2 papers)
  - [5] QDepth-VLA: quantized depth prediction as auxiliary supervision for vision-language-action models (Li, 2025) View paper
  - [11] Perception Tokens Enhance Visual Reasoning in Multimodal Language Models (Mahtab Bigverdi, 2024) View paper
- Language-Guided Depth Estimation from Contrastive Models
  - Prompt-Based Depth Estimation with CLIP (3 papers)
  - [17] Learning to prompt clip for monocular depth estimation: Exploring the limits of human language (Dylan Auty, 2023) View paper
  - [32] Can language understand depth? (Renrui Zhang, 2022) View paper
  - [44] Can Language Really Understand Depth? (Fangping Chen, 2023) View paper
  - CLIP Feature Adaptation for Depth (1 papers)
  - [35] CaBins: CLIP-based Adaptive Bins for Monocular Depth Estimation (Eunjin Son, 2024) View paper
  - Language Prior Integration in Depth Models (3 papers)
  - [18] Wordepth: Variational language prior for monocular depth estimation (Ziyao Zeng, 2024) View paper

- [25] On the robustness of language guidance for low-level vision tasks: Findings from depth estimation (Agneet Chatterjee, 2024) View paper
- [49] Hybrid-grained Feature Aggregation with Coarse-to-fine Language Guidance for Self-supervised Monocular Depth Estimation (Zhang Wenyao, 2025) View paper

- Diffusion Models for Language-Conditioned Depth (3 papers)
  - [22] TPDepth: Leveraging Text Prompts with ControlNet to Boost Diffusion-based Depth Estimation (Yu Liu, 2025) View paper
  - [30] InstructCV: Instruction-Tuned Text-to-Image Diffusion Models as Vision Generalists (Gan, 2023) View paper
  - [34] PriorDiffusion: Leverage Language Prior in Diffusion Models for Monocular Depth Estimation (Zeng Ziyao, 2024) View paper
- LLM-Based Depth Understanding and Reasoning (2 papers)
  - [15] Can large language models understand depth in monocular images without prior vision knowledge? (Zhongyi Xia, 2025) View paper
  - [36] Large language models can understanding depth from monocular images (Xia Zhong-yi, 2024) View paper
- Depth-Conditioned Reasoning and Planning
  - Spatial Reasoning with Depth for Robotics (5 papers)
  - [4] Spatialvlm: Endowing vision-language models with spatial reasoning capabilities (Boyuan Chen, 2024) View paper
  - [6] RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics (Zhou En-shen, 2025) View paper
  - [8] TIGeR: Tool-Integrated Geometric Reasoning in Vision-Language Models for Robotics (Han Yi, 2025) View paper
  - [19] Mm-spatial: Exploring 3d spatial understanding in multimodal llms (Daxberger, 2025) View paper
  - [39] N3D-VLM: Native 3D Grounding Enables Accurate Spatial Reasoning in Vision-Language Models (Yuxin Wang, 2025) View paper
  - Depth-Enhanced VLMs for Autonomous Driving (2 papers)
  - [26] SpaceDrive: Infusing Spatial Awareness into VLM-based Autonomous Driving (Peizheng Li, 2025) View paper
  - [47] The System Description of CPS Team for Track on Driving with Language of CVPR 2024 Autonomous Grand Challenge (Peng, 2025) View paper
- Benchmarking Spatial Understanding in VLMs (1 papers)
  - [12] SURDS: Benchmarking Spatial Understanding and Reasoning in Driving Scenarios with Vision Language Models (Guo, 2024) View paper
- Monocular Depth Estimation with Visual Priors
  - Foundation Model Integration for Depth (3 papers)
  - [9] ViPOcc: leveraging visual priors from vision foundation models for single-view 3d occupancy prediction (Feng Yi, 2025) View paper
  - [16] Uni4D: Unifying Visual Foundation Models for 4D Modeling from a Single Video (Albert J. Zhai, 2025) View paper
  - [21] Composition vision-language understanding via segment and depth anything model (Huo, 2024) View paper
  - Single-Model Depth Estimation Enhancements (1 papers)
  - [31] Monocular Depth Estimation Using Cues Inspired by Biological Vision Systems (Dylan Auty, 2022) View paper
- Relative-to-Metric Depth Conversion (1 papers)
  - [13] TR2M: Transferring Monocular Relative Depth to Metric Depth with Language Descriptions and Scale-Oriented Contrast (Cui, 2025) View paper
- Task-Specific and Domain-Adapted Depth Applications
  - Depth Estimation in Challenging Conditions (3 papers)
  - [14] Hierarchical Interpretable Vision Reasoning Driven Through a Multi-Modal Large Language Model for Depth Estimation (He Chen, 2024) View paper
  - [38] Hierarchical interpretable vision reasoning driven based depth estimation method in adverse weather conditions through a multi-modal large language model (Chengwei Yang, 2025) View paper
  - [40] Underwater Monocular Metric Depth Estimation: Real-World Benchmarks and Synthetic Fine-Tuning with Vision Foundation Models (Cai, 2025) View paper
  - Application-Driven Depth Estimation (4 papers)
  - [33] FloodVision: Urban Flood Depth Estimation Using Foundation Vision-Language Models and Domain Knowledge Graph (Mohammadi Neda, 2025) View paper
  - [45] DepthScape: Authoring 2.5D Designs via Depth Estimation, Semantic Understanding, and Geometry Extraction (Xia Su, 2025) View paper
  - [46] A VLM Based Zero-Shot Weight Estimation Approach For Efficient Waste Management (Mario Viti, 2025) View paper
  - [48] SightSeeingGemma: Enhancing Assistive AI for the Visually Impaired via Object Detection and Monocular Depth Estimation with Language-Based Scene â (A Dinh, 2025) View paper
- Unified Multi-Task Vision-Language Models (1 papers)
  - [7] Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks (Lu, 2022) View paper
- Specialized Depth-Related Vision Tasks (3 papers)
  - [27] MEgoHand: Multimodal Egocentric Hand-Object Interaction Motion Generation (Zhou, 2025) View paper
  - [43] Trajectory Densification and Depth from Perspective-based Blur (Tianchen Qiu, 2025) View paper
  - [50] Zero-Splat TeleAssist: A Zero-Shot Pose Estimation Framework for Semantic Teleoperation (Srijan Dokania, 2025) View paper
- Survey and Review Literature (1 papers)
  - [20] A review of recent advances in data-driven computer vision methods for structural damage evaluation: algorithms, applications, challenges, and future â (X Pan, 2025) View paper
- Multimodal AI Systems with Depth Components (3 papers)
  - [28] Multimodal AI for UAV: Visionâ□□Language Models in Humanâ□□ Machine Collaboration (MaroÅ¡ KrupÃ¡Å¡, 2025) View paper
  - [29] Vision-language embodiment for monocular depth estimation (Jinchang Zhang, 2025) View paper
  - [42] Urban Road Anomaly Monitoring Using Visionâ□□Language Models for Enhanced Safety Management (Hanyu Ding, 2025) View paper
- Depth Representation and Encoding Methods (1 papers)
  - [37] DepthBLIP-2: Leveraging Language to Guide BLIP-2 in Understanding Depth Information (Wei Chen, 2024) View paper
- Core Metric Depth Estimation from VLMs ★ (2 papers)
  - [0] DepthLM: Metric Depth from Vision Language Models (Anon et al., 2026) View paper
  - [1] Spatialrgpt: Grounded spatial reasoning in vision-language models (An-Chieh Cheng, 2024) View paper

### Narrative

Core task: metric depth estimation from vision language models. The field has evolved from traditional monocular depth methods toward integrating language and multimodal reasoning with depth perception. The taxonomy reveals several major branches: some focus on VLM architecture and training strategies that embed depth understanding into large-scale vision-language models (e.g., SpatialRGPT[1], SpatialBot[2]), while others explore language-guided depth estimation from contrastive models like CLIP (Prompt CLIP Depth[17], WorDepth[18]) or diffusion-based approaches (PriorDiffusion[34]). Additional branches address LLM-based depth reasoning (LLM Depth Understanding[15], Language Understand Depth[32]), depth-conditioned planning for robotics (QDepth VLA[5], RoboRefer[6]), and benchmarking spatial understanding in VLMs (MM Spatial[19]). Unified multi-task models (Unified IO[7], Florence VL[23]) and specialized depth applications (Underwater Metric Depth[40], Adverse Weather Depth[38]) round out the landscape, alongside survey literature and methods for relative-to-metric depth conversion.

Recent work has concentrated on two contrasting themes: end-to-end VLM architectures that jointly learn language and depth representations versus modular pipelines that leverage pretrained language models to guide or interpret depth outputs. DepthLM[0] sits squarely in the core metric depth estimation branch, emphasizing direct prediction of metric depth from VLMs without relying solely on contrastive or diffusion priors. This positions it closely alongside SpatialRGPT[1] and SpatialBot[2], which similarly integrate spatial reasoning into large vision-language frameworks, though those works often prioritize broader spatial understanding tasks beyond pure depth estimation. Compared to language-guided contrastive methods (WorDepth[18]) or diffusion-based depth generation (PriorDiffusion[34]), DepthLM[0] appears more focused on producing accurate metric depth maps as a primary output rather than as an auxiliary signal for downstream reasoning or generation. Open questions remain around how best to fuse language semantics with geometric cues and whether unified architectures or specialized depth modules offer better trade-offs in accuracy, generalization, and computational efficiency.

## Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

### 1. Spatialrgpt: Grounded spatial reasoning in vision-language models

**Authors**: An-Chieh Cheng, Yang Fu, Qiushan Guo, Jan Kautz, Sifei Liu, et al. (8 authors total) | **Year/Venue**: 2024 | **URL**: View paper

#### Abstract

â¦ The camera calibration and metric depth estimation each took around 4 hours. Note that the depth estimation requires our estimated camera intrinsics as input, so these two processes â¦

#### Relationship Analysis

Both papers belong to the Core Metric Depth Estimation from VLMs category, focusing on enabling vision-language models to perform metric depth estimation. They overlap in addressing the challenge of teaching VLMs to understand 3D spatial information from 2D images through text-based supervision and visual prompting. However, DepthLM focuses specifically on pixel-level metric depth estimation with visual markers and intrinsic-conditioned augmentation to achieve expert-level accuracy comparable to pure vision models, while SpatialRGPT emphasizes region-level spatial reasoning across multiple 3D tasks (relative relations, metric measurements, complex reasoning) by constructing 3D scene graphs and incorporating depth information as a plugin module to the visual encoder.

## Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

### Contribution 1: DepthLM framework for metric depth estimation in VLMs

**Description**: The authors introduce DepthLM, a framework that enables vision language models to achieve expert-level accuracy in pixel-level metric depth estimation. The method uses visual prompting with rendered markers for pixel reference and intrinsic-conditioned augmentation to resolve camera ambiguity, requiring no architectural modifications or specialized loss functions.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

#### 1. Learning to prompt clip for monocular depth estimation: Exploring the limits of human language

**URL**: View paper

**Brief Assessment**

Prompt CLIP Depth[17] focuses on learning continuous prompt tokens for CLIP to perform monocular depth estimation, while DepthLM addresses metric depth estimation in VLMs through visual prompting with rendered markers and intrinsic-conditioned augmentation. The candidate uses CLIP's text encoder with learnable tokens, whereas DepthLM uses standard VLM architectures with visual markers for pixel reference.

---

#### 2. TPDepth: Leveraging Text Prompts with ControlNet to Boost Diffusion-based Depth Estimation

**URL**: View paper

**Brief Assessment**

TPDepth[22] is a diffusion-based depth estimator using ControlNet with text prompts, not a vision language model framework. The technical approaches are fundamentally different - DepthLM uses visual prompting with markers and text-based supervised fine-tuning of VLMs, while TPDepth[22] uses diffusion models with textual semantics through ControlNet.

---

#### 3. Vision-language embodiment for monocular depth estimation

**URL**: View paper

**Brief Assessment**

Vision Language Embodiment[29] focuses on embodying camera models and computing embodied scene depth through physical camera parameters for depth estimation, rather than enabling VLMs to perform metric depth estimation through visual prompting and text-based supervised fine-tuning as in the original paper.

---

#### 4. Learning to adapt clip for few-shot monocular depth estimation

**URL**: View paper

**Brief Assessment**

Adapt CLIP Depth[69] focuses on few-shot learning for monocular depth estimation using CLIP with learnable prompts and depth codebooks, not on enabling VLMs to achieve expert-level metric depth estimation through visual prompting with rendered markers and intrinsic-conditioned augmentation as in the original paper.

---

#### 5. Spatialrgpt: Grounded spatial reasoning in vision-language models

**URL**: View paper

**Brief Assessment**

SpatialRGPT[1] focuses on grounded spatial reasoning with region-level understanding and 3D scene graphs, not on pixel-level metric depth estimation as a primary task. The candidate uses depth as auxiliary input for spatial reasoning rather than as the main prediction target.

---

### 6. Can large language models understand depth in monocular images without prior vision knowledge?

**URL**: View paper

**Brief Assessment**

The candidate paper (LLM Depth Understanding[15]) appears to explore depth understanding in LLMs but the provided context is too limited (only a fragment mentioning 'llm-mde' framework) to establish whether it demonstrates prior work that refutes the novelty of DepthLM's specific contributions (visual prompting with rendered markers and intrinsic-conditioned augmentation).

---

### 7. FloodVision: Urban Flood Depth Estimation Using Foundation Vision-Language Models and Domain Knowledge Graph

**URL**: View paper

**Brief Assessment**

FloodVision[33] focuses on flood depth estimation using GPT-4o with a domain knowledge graph for urban flooding scenarios, not general metric depth estimation frameworks for VLMs.

---

### 8. Explore until Confident: Efficient Exploration for Embodied Question Answering

**URL**: View paper

**Brief Assessment**

Explore Until Confident[70] focuses on embodied question answering with active exploration and semantic mapping, not on metric depth estimation frameworks for VLMs. The candidate uses depth information and visual prompting for scene mapping in robotics tasks, which is a different application domain from DepthLM's pixel-level metric depth estimation.

---

### 9. RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics

**URL**: View paper

**Brief Assessment**

RoboRefer[6] focuses on spatial referring for robotics with a disentangled depth encoder, not on general metric depth estimation frameworks. The technical approaches differ fundamentally in architecture and task objectives.

---

### 10. Prompting Depth Anything for 4K Resolution Accurate Metric Depth Estimation

**URL**: View paper

**Brief Assessment**

Prompting Depth Anything[71] focuses on prompting depth foundation models with LiDAR sensors for metric depth, not on enabling vision language models to perform metric depth estimation through visual prompting and text interactions.

---

## Contribution 2: DepthLMBench benchmark suite

**Description**: The authors create DepthLMBench, a curated mixture of public datasets (approximately 16M training images from 7 datasets) that enables training and evaluation of VLMs for 3D understanding tasks, allowing direct comparison with pure vision models on metric depth estimation.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. From 2d to 3d: Re-thinking benchmarking of monocular depth prediction

**URL**: View paper

**Brief Assessment**

2D to 3D Benchmarking[57] focuses on evaluating monocular depth prediction using 3D metrics and proposes the rio-d3d dataset with 34k images for 3D-aware evaluation. DepthLMBench is a training benchmark with ~16M images from 7 datasets designed for VLM training on metric depth estimation, serving a fundamentally different purpose.

---

### 2. 3D Packing for Self-Supervised Monocular Depth Estimation

**URL**: View paper

**Brief Assessment**

3D Packing[55] introduces DDAD (Dense Depth for Automated Driving), a dataset focused on self-supervised monocular depth estimation from driving scenarios. This differs from DepthLMBench, which is a curated mixture of multiple public datasets designed specifically for training and evaluating VLMs on 3D understanding tasks with text-based interactions.

---

### 3. E3D-Bench: A Benchmark for End-to-End 3D Geometric Foundation Models

**URL**: View paper

**Brief Assessment**

E3D Bench[54] focuses on evaluating end-to-end 3D geometric foundation models across multiple tasks (depth, pose, reconstruction, novel view synthesis), while DepthLMBench is specifically designed for training and evaluating VLMs on metric depth estimation with text-based interactions. The datasets and evaluation purposes differ fundamentally.

---

### 4. Towards Depth Foundation Model: Recent Trends in Vision-Based Depth Estimation

**URL**: View paper

**Brief Assessment**

Depth Foundation Model[58] surveys existing datasets for depth estimation tasks but does not create a new benchmark suite for VLM training and evaluation. The candidate focuses on cataloging datasets across monocular, stereo, multi-view, and video depth estimation, while DepthLMBench is specifically designed for training VLMs on 3D understanding with text-based supervision.

---

### 5. Selfocc: Self-supervised vision-based 3d occupancy prediction

**URL**: View paper

**Brief Assessment**

SelfOcc[53] focuses on self-supervised 3D occupancy prediction from video sequences for autonomous driving, not on creating benchmark datasets for VLM training and evaluation on metric depth estimation tasks.

---

#### 6. UniDepth: Universal monocular metric depth estimation

**URL**: View paper

**Brief Assessment**

UniDepth[51] does not create a benchmark suite for training VLMs. It focuses on a different technical approach (self-prompting camera module with pseudo-spherical representation) for monocular metric depth estimation using pure vision models, not vision-language models.

---

#### 7. Occdepth: A depth-aware method for 3d semantic scene completion

**URL**: View paper

**Brief Assessment**

OccDepth[59] focuses on 3D semantic scene completion using stereo/RGBD images for autonomous driving, not on creating benchmark datasets for evaluating vision-language models on metric depth estimation tasks.

---

#### 8. On the metrics for evaluating monocular depth estimation

**URL**: View paper

**Brief Assessment**

Depth Metrics Evaluation[56] focuses on evaluating existing depth estimation metrics through 3D object detection tasks, not on creating benchmark datasets for training VLMs on 3D understanding tasks.

---

#### 9. Mm-spatial: Exploring 3d spatial understanding in multimodal llms

**URL**: View paper

**Brief Assessment**

MM Spatial[19] focuses on 3D spatial understanding tasks (relationships, grounding, metric estimation) using high-quality 3D scene data from CA-1M/ARKitScenes, while DepthLMBench is specifically designed for metric depth estimation training and evaluation across indoor/outdoor datasets. The datasets serve different primary purposes and task scopes.

---

#### 10. Survey on monocular metric depth estimation

**URL**: View paper

**Brief Assessment**

Monocular Depth Survey[52] discusses existing datasets (KITTI, NYU-Depth, etc.) for metric depth estimation but does not present a new benchmark suite for training VLMs on 3D understanding tasks.

---

### Contribution 3: Unified VLM for diverse 3D understanding tasks

**Description**: The authors demonstrate that DepthLM can be extended to train a unified model handling multiple 3D understanding tasks (including principal axis distance, speed estimation, time estimation, two-point distance, and camera pose estimation) using the same architecture and training framework, without requiring task-specific designs.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

#### 1. Unifying 3d vision-language understanding via promptable queries

**URL**: View paper

**Brief Assessment**

Promptable 3D Queries[65] focuses on unifying different 3D scene representations (voxels, point clouds, multi-view images) for various 3D vision-language tasks, not on training VLMs for metric depth estimation and related 3D understanding tasks using the same architecture without task-specific designs.

---

#### 2. Robix: A Unified Model for Robot Interaction, Reasoning and Planning

**URL**: View paper

**Brief Assessment**

Robix[66] focuses on robot interaction, reasoning, and task planning rather than diverse 3D understanding tasks like depth estimation, distance measurement, or camera pose estimation that DepthLM addresses.

---

#### 3. Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks

**URL**: View paper

**Brief Assessment**

Unified IO[7] focuses on a broad range of vision-language tasks (detection, segmentation, VQA, captioning) but does not specifically address 3D understanding tasks like metric depth estimation, speed estimation, or camera pose estimation that are central to the original paper's contribution.

---

#### 4. Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning

**URL**: View paper

**Brief Assessment**

Inst3D LMM[60] focuses on instance-aware 3D scene understanding with multi-modal instruction tuning for tasks like 3D visual grounding, question answering, and dense captioning. The original paper addresses metric depth estimation and related 3D tasks using vision-language models with a different technical approach (visual prompting, intrinsic-conditioned augmentation). These are distinct problem domains and methodologies.

---

#### 5. Continuous 3D Perception Model with Persistent State

**URL**: View paper

**Brief Assessment**

Continuous 3D Perception[63] focuses on a recurrent model for 3D reconstruction from image streams, not a vision-language model framework for multiple 3D understanding tasks through text interactions.

---

### 6. Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation

**URL**: View paper

**Brief Assessment**

Janus[62] focuses on unifying multimodal understanding and generation (vision-language tasks), not 3D understanding tasks like depth estimation or camera pose estimation that DepthLM addresses.

---

### 7. Delving into multi-modal multi-task foundation models for road scene understanding: From learning paradigm perspectives

**URL**: View paper

**Brief Assessment**

Multi Modal Road[61] focuses on multi-modal multi-task foundation models for road scene understanding, not general 3D understanding tasks. The candidate discusses unified models for driving-related tasks (detection, segmentation, prediction) rather than the specific 3D tasks mentioned in the original (principal axis distance, speed estimation, time estimation, two-point distance, camera pose estimation).

---

### 8. A unified framework for 3d scene understanding

**URL**: View paper

**Brief Assessment**

Unified 3D Framework[67] focuses on 3D point cloud segmentation tasks (panoptic, semantic, instance, interactive, referring, and open-vocabulary segmentation), not on metric depth estimation or the specific 3D understanding tasks mentioned in the original paper (principal axis distance, speed estimation, time estimation, two-point distance, camera pose estimation).

---

### 9. VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation

**URL**: View paper

**Brief Assessment**

VILA U[64] focuses on integrating visual understanding and generation using autoregressive next-token prediction, not on 3D understanding tasks like depth estimation or camera pose estimation that DepthLM addresses.

---

### 10. Uni3d: Exploring unified 3d representation at scale

**URL**: View paper

**Brief Assessment**

Uni3D[68] focuses on unified 3D representation learning from point clouds using vision-language alignment, not on extending VLMs to handle multiple 3D understanding tasks from 2D images. The candidate works with 3D point cloud inputs while the original paper processes 2D images for 3D understanding.

---

## Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

## References


- [0] DepthLM: Metric Depth from Vision Language Models View paper
- [1] Spatialrgpt: Grounded spatial reasoning in vision-language models View paper
- [2] Spatialbot: Precise spatial understanding with vision language models View paper
- [3] SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model View paper
- [4] Spatialvlm: Endowing vision-language models with spatial reasoning capabilities View paper
- [5] QDepth-VLA: quantized depth prediction as auxiliary supervision for vision-language-action models View paper
- [6] RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics View paper
- [7] Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks View paper
- [8] TIGeR: Tool-Integrated Geometric Reasoning in Vision-Language Models for Robotics View paper
- [9] ViPOcc: leveraging visual priors from vision foundation models for single-view 3d occupancy prediction View paper
- [10] RoboFlamingo-Plus: Fusion of Depth and RGB Perception with Vision-Language Models for Enhanced Robotic Manipulation View paper
- [11] Perception Tokens Enhance Visual Reasoning in Multimodal Language Models View paper
- [12] SURDS: Benchmarking Spatial Understanding and Reasoning in Driving Scenarios with Vision Language Models View paper
- [13] TR2M: Transferring Monocular Relative Depth to Metric Depth with Language Descriptions and Scale-Oriented Contrast View paper
- [14] Hierarchical Interpretable Vision Reasoning Driven Through a Multi-Modal Large Language Model for Depth Estimation View paper
- [15] Can large language models understand depth in monocular images without prior vision knowledge? View paper
- [16] Uni4D: Unifying Visual Foundation Models for 4D Modeling from a Single Video View paper
- [17] Learning to prompt clip for monocular depth estimation: Exploring the limits of human language View paper
- [18] Wordepth: Variational language prior for monocular depth estimation View paper
- [19] Mm-spatial: Exploring 3d spatial understanding in multimodal llms View paper
- [20] A review of recent advances in data-driven computer vision methods for structural damage evaluation: algorithms, applications, challenges, and future â□¦ View paper
- [21] Composition vision-language understanding via segment and depth anything model View paper
- [22] TPDepth: Leveraging Text Prompts with ControlNet to Boost Diffusion-based Depth Estimation View paper
- [23] Florence-VL: Enhancing Vision-Language Models with Generative Vision Encoder and Depth-Breadth Fusion View paper
- [24] Vid-LLM: A Compact Video-based 3D Multimodal LLM with Reconstruction-Reasoning Synergy View paper
- [25] On the robustness of language guidance for low-level vision tasks: Findings from depth estimation View paper
- [26] SpaceDrive: Infusing Spatial Awareness into VLM-based Autonomous Driving View paper
- [27] MEgoHand: Multimodal Egocentric Hand-Object Interaction Motion Generation View paper
- [28] Multimodal AI for UAV: Visionâ□□Language Models in Humanâ□□ Machine Collaboration View paper
- [29] Vision-language embodiment for monocular depth estimation View paper
- [30] InstructCV: Instruction-Tuned Text-to-Image Diffusion Models as Vision Generalists View paper
- [31] Monocular Depth Estimation Using Cues Inspired by Biological Vision Systems View paper

- [32] Can language understand depth? View paper
- [33] FloodVision: Urban Flood Depth Estimation Using Foundation Vision-Language Models and Domain Knowledge Graph View paper
- [34] PriorDiffusion: Leverage Language Prior in Diffusion Models for Monocular Depth Estimation View paper
- [35] CaBins: CLIP-based Adaptive Bins for Monocular Depth Estimation View paper
- [36] Large language models can understanding depth from monocular images View paper
- [37] DepthBLIP-2: Leveraging Language to Guide BLIP-2 in Understanding Depth Information View paper
- [38] Hierarchical interpretable vision reasoning driven based depth estimation method in adverse weather conditions through a multi-modal large language model View paper
- [39] N3D-VLM: Native 3D Grounding Enables Accurate Spatial Reasoning in Vision-Language Models View paper
- [40] Underwater Monocular Metric Depth Estimation: Real-World Benchmarks and Synthetic Fine-Tuning with Vision Foundation Models View paper
- [41] SD-VLM: Spatial Measuring and Understanding with Depth-Encoded Vision-Language Models View paper
- [42] Urban Road Anomaly Monitoring Using Visionâ□□Language Models for Enhanced Safety Management View paper
- [43] Trajectory Densification and Depth from Perspective-based Blur View paper
- [44] Can Language Really Understand Depth? View paper
- [45] DepthScape: Authoring 2.5D Designs via Depth Estimation, Semantic Understanding, and Geometry Extraction View paper
- [46] A VLM Based Zero-Shot Weight Estimation Approach For Efficient Waste Management View paper
- [47] The System Description of CPS Team for Track on Driving with Language of CVPR 2024 Autonomous Grand Challenge View paper
- [48] SightSeeingGemma: Enhancing Assistive AI for the Visually Impaired via Object Detection and Monocular Depth Estimation with Language-Based Scene â□¦ View paper
- [49] Hybrid-grained Feature Aggregation with Coarse-to-fine Language Guidance for Self-supervised Monocular Depth Estimation View paper
- [50] Zero-Splat TeleAssist: A Zero-Shot Pose Estimation Framework for Semantic Teleoperation View paper
- [51] UniDepth: Universal monocular metric depth estimation View paper
- [52] Survey on monocular metric depth estimation View paper
- [53] Selfocc: Self-supervised vision-based 3d occupancy prediction View paper
- [54] E3D-Bench: A Benchmark for End-to-End 3D Geometric Foundation Models View paper
- [55] 3D Packing for Self-Supervised Monocular Depth Estimation View paper
- [56] On the metrics for evaluating monocular depth estimation View paper
- [57] From 2d to 3d: Re-thinking benchmarking of monocular depth prediction View paper
- [58] Towards Depth Foundation Model: Recent Trends in Vision-Based Depth Estimation View paper
- [59] Occdepth: A depth-aware method for 3d semantic scene completion View paper
- [60] Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning View paper
- [61] Delving into multi-modal multi-task foundation models for road scene understanding: From learning paradigm perspectives View paper
- [62] Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation View paper
- [63] Continuous 3D Perception Model with Persistent State View paper
- [64] VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation View paper
- [65] Unifying 3d vision-language understanding via promptable queries View paper
- [66] Robix: A Unified Model for Robot Interaction, Reasoning and Planning View paper
- [67] A unified framework for 3d scene understanding View paper
- [68] Uni3d: Exploring unified 3d representation at scale View paper
- [69] Learning to adapt clip for few-shot monocular depth estimation View paper
- [70] Explore until Confident: Efficient Exploration for Embodied Question Answering View paper
- [71] Prompting Depth Anything for 4K Resolution Accurate Metric Depth Estimation View paper