We hypothesize two main causes for the suboptimality. First, our teacher model generates reliable pseudo labels across both near and far regions, unlike most real-world datasets biased toward short-range depths (see Fig. 3 for a comparison). This wider depth distribution exposes limitations of conventional depth losses: direct depth supervision (L1/L2) blurs fine geometric details, while inverse-depth loss decays too rapidly with distance, losing ineffective supervision for distant regions. Consequently, these standard losses are suboptimal for supervising our high-fidelity, large-range pseudo labels. Second, typical ViT encoders with DPT-head decoders use U-Net-style skip connections, injecting shallow features into deeper layers and propagating deep features upward. While this stabilizes training under noisy supervision, since the low-level cues in the ViT encoder (e.g., textures and colors) are more consistent and easier to learn. The low-level feature is connected via a skip connection to the DPT head near the output, which mitigates conflicts between the output and the noisy supervision, thereby smoothing gradient fluctuations. However, it can underutilize high-level semantic cues essential for precise depth. In our setting, pseudo labels are generated by a unified model, exhibit minimal domain gap, and have been refined to correct most noisy. As a result, we can reduce reliance on shallow-to-deep feature injection and explore more aggressive network designs that fully exploit the rich semantic cues provided by deep block of the ViT encoder. worsened Student.

Drawing on these observations and analyses, we retain the multi-scale fusion mechanism proposed by [10] while introducing two key modifications. First, we design a distance-balanced inverse-depth loss that preserves fine-grained sensitivity in near regions while extending effective supervision to long-distance areas. Depth values in log-space are defined as:

$$D_{log} = 1 - \ln(x)/\ln(C), \tag{5}$$

where C is a hyperparameter that controls the trade-off between long-range and short-range supervision. Second, leveraging the high-fidelity teacher labels, we invert the conventional skip-connection scheme between the ViT encoder and DPT head: injecting deep, high-level ViT features into the deeper decoding layers near the output, while shallow, low-level features feed into the shallower decoder layers, as shown in Fig. 4. This inversion emphasizes semantic reasoning at the final prediction stage, fully exploiting the rich semantic cues embedded in the teacher-generated pseudo labels, which already exhibit low noise and high structural consistency. In this way, we experimentally demonstrate that the student model achieves stable, prompt-free metric depth estimation, effectively distilling the metric perception capability from the pre-trained model.

4 Experiment

4.1 Prompt-Based Downstream Task

4.1.1 Zero-Shot Depth Super-Resolution and Completion.

We study depth super-resolution and completion in a zero-shot setting, where our pretrained model directly takes sparse or low-resolution depth maps as prompts, without any task-specific finetuning. We compare two categories of baselines: (a) post-aligned MDE, including DepthAnything V2 [125] and DepthPro [10]; and (b) prior-based MDE, including LingBot-Depth [99], [60], PriorDA [114], DepthLab [64], Omni-DC [140], and Marigold-DC [106]. Following PriorDA [114], we construct four prompt types?LiDAR-like sparse scans, extremely sparse samples (100 points), and depth maps downsampled by 8× and 16×. All prompts are directly fed into our pretrained model for zero-shot inference on unseen datasets, including NYUv2 [91], ETH3D [89], and KITTI [31], covering indoor, outdoor, and open-world scenarios.

As shown in Tab. 1 and Fig. 5, our pretrained model demonstrates strong zero-shot performance across all prompt types and datasets, consistently outperforming both post-aligned and prior-based baselines. Unlike previous approaches [114, 140] that generate synthetic prompts (e.g., LiDAR simulation or noisy downsampling) to approximate test-time conditions, our model is trained only once with simple, sparsely sampled prompts and operates fully zero-shot manner, without any task-specific design or prompt alignment. This enables superior generalization across diverse prompt densities, spatial layouts, and scene domains.

Figure 6: Zero-shot Visual Comparisons on Challenging Test Samples. Our model robustly captures details of thin structures and in scenes with difficult lighting where competitors often fail.

robust. This suggests that our model ineffectively handles large-scale depth variations typical of outdoor environments.

? Robustness on Unconventional Data: On Booster, a dataset known for challenging lighting and textures, our method outperforms all baselines (AbsRel 0.282), highlighting its resilience to domain shifts. Even on the synthetic Sintel dataset, where domain gaps are significant, we maintain strong performance (2nd best in Log10), demonstrating that our learned representations generalize well beyond photorealistic domains.

Figure 7: Qualitative Comparison of Monocular Depth Estimation. Compared with MoGe2 and UniDepthv2, our distilled model produces more detailed and geometrically plausible predictions for both depth maps and point maps.

Overall, while some baselines excel in specific niches, our method delivers the most balanced and consistently high performance across the full spectrum of test scenarios.

Monocular metric point map.

For metric point map prediction, we leverage pseudo-labels generated by our pre-trained model to fine-tune recent state-of-the-art methods such as MoGe-2 [110], denoted as ?Student-PointMap?. This design allows us to assess the generality and precision of our pseudo-labels under different training paradigms ? whether training from scratch or fine-tuning, and regardless of whether the output head predicts depth maps or 3D point maps.

11.57 m
1.765 m
MoGe-2
Ours Student-Pointmap
UniDepth v2
Image
2.232 m
1.968 m
12.95 m
14.02 m
16.89 m
1.954 m
11.85 m
6.22 m
3.12 m
2.45 m
2.84 m
0.60 m
14.36 m
11.25 m
Figure 8: Qualitative Comparison of Point Maps. The red arrows indicate the GT distance, the yellow arrows indicate the distance from predicted point map.

Monocular 3D geometry estimation aims to recover a per-pixel 3D point map in the camera coordinate system. In this setting, we leverage pseudo-labels generated by our pre-trained model to fine-tune recent state-of-the-art frameworks such as MoGe-2 [110], denoted as ?Student-PointMap?. This design enables a comprehensive evaluation of the generality and precision of our pseudo-labels across different training paradigms?including training from scratch versus fine-tuning, and varying output representations (depth maps or 3D point maps). As shown in Fig. 8, Fig. 9, Tab. 5 and Tab. 6, our distillation approach consistently achieves state-of-the-art performance, demonstrating its robustness to differences in prediction heads and network initialization. We adopt the GIANT-LARGE and DA3MONO-LARGE variants from the official DepthAnything3 [59] checkpoints, which represent the largest and most powerful models that support monocular metric depth estimation. However, the performances of GIANT-LARGE and DA3MONO-LARGE are not particularly satisfactory in our setting. We conjecture that the capability of DepthAnything3 [59] still depends on inferring matching relationships across multiple views, making accurate metric scale recovery particularly challenging in complex monocular settings. Additionally, Depth Anything3 [59] does not support inference when the camera?s intrinsic parameters are unknown, which limits its applicability. In contrast, we evaluated its performance under both known and unknown intrinsic parameter conditions.

4.2.2

Recovering Camera Intrinsics.

Furthermore, we utilize the point map X predicted by our finetuned model (?Student-PointMap?) to infer the intrinsic parameters of the camera from a straightforward optimization. Throughout our experiments, we assume a unit aspect ratio and that the principal point is approximately centered in the image; however, the only unknown intrinsic parameter is the focal length of the first camera, denoted f. We estimate f by minimizing a weighted

Question: "What is the length of the longest dimension (length, width, or height) of the sofa, measured in centimeters?"
GT: 173 Ours: 175
Gemini-2.5-Pro: 180 GPT-5.1: 200 GPT-5-Chat: 180 Claude-Opus-4.1: 180
OBJECT SIZE ESTIMATION
!
OBJECT ABSOLUTE DISTANCE ESTIMATION
Question: "Measuring from the closest point of each object, what is the distance between the toilet and the bathtub (in meters)?"
GT: 0.4
Ours: 0.4
Gemini-2.5-Pro: 0.2
GPT-5.1: 0.3
GPT-5-Chat: 0.2
Claude-Opus-4.1: 0.3
!
ROUTE PLANNING
Question: You are a robot beginning at the doorframe facing the table. You want to navigate to the bookshelf.
You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn write.?):
1. Go forward until the table 2. [please fill in] 3. Go forward until the chair beside the screen 4. [please fill in]
5. Go forward until the bookshelf. You have reached the final destination."
GT: Turn Left, Turn Right Ours: Turn Left, Turn Right
Gemini-2.5-Pro: Turn Right, Turn Right
GPT-5.1: Turn Right, Turn Right
GPT-5-Chat: Turn Right, Turn Right Claude-Opus-4.1: Turn Right, Turn Right
!
Question: You are a robot beginning at the black desk chair and facing the bookshelf. You want to navigate to
the red desk chair. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,?
'turn left,' or 'turn right.'): 1. [please fill in] 2. Go forward until the red desk chair. You have reached the final destination."
GT: Turn Back Ours: Turn Back
Gemini-2.5-Pro: Turn Left GPT-5.1: Turn Right
GPT-5-Chat: Turn Left Claude-Opus-4.1: Turn Right
!
OBJECT APPEARANCE ORDER
Question: "What will be the first-time appearance order of the following categories
in the video: door, window, radiator, nightstand?
A. door, window, nightstand, radiator; B. door, window, radiator, nightstand;
C. window, nightstand, door, radiator; D. nightstand, door, window, radiator"
GT: A
Ours: A
Gemini-2.5-Pro: B GPT-5.1: B GPT-5-Chat: B Claude-Opus-4.1: B
!
ROOM SIZE ESTIMATION
Question: "What is the size of this room (in square meters)?"
GT: 55.0
Ours: 56.1
Gemini-2.5-Pro: 62 GPT-5.1: 60 GPT-5-Chat: 45 Claude-Opus-4.1: 50
!
Figure 12: Enhancing 3D Spatial Reasoning with a Frozen ViT from Metric Anything . We evaluate our approach on the VIS Benchmark, covering video question-answering tasks like estimating object size, object?s distances, appearance order, route planning, and room size. Compared to mainstream large models, our method demonstrates robust and superior performance in 3D spatial understanding.

LiDAR
Cameras
Missing Area
Metric Anything
Metric Anything
Figure 17: Sensor Configuration for Real-World Generalization Evaluation. Our real-world test vehicle is equipped with three cameras (front, left-front, write-front) and a 128-beam solid-state LiDAR. Due to the LiDAR?s limited vertical field of view(pitch angle limitation), its captured point cloud does not fully cover the cameras? combined frustums, leaving large image regions without metric depth cues.

6 Generalizability to Unseen Sensors, Scenarios, and Extreme Environmental Conditions

6.1 Generalization across Sensor Configurations

This subsection assesses the model?s generalization capability to variations in sensor hardware configuration and data characteristics. We deployed a test vehicle equipped with a sensor suite that differed from the training set in both type and spatial arrangement. The setup consisted of three cameras providing front, right-front, and left-front views, coupled with a 128-beam solid-state LiDAR for forward scene perception (see Fig. 17). The collected real-world data exhibits two key challenges: 1) minor calibration inaccuracies and asynchronous sampling rates?with cameras operating at 24 Hz and LiDAR at 10 Hz?introduced spatiotemporal misalignments between sensor modalities; 2) the LiDAR?s field of view did not fully cover the lateral areas captured by the side-facing cameras. We deliberately avoided additional post-processing techniques, such as motion compensation, to rigorously evaluate the model?s inherent robustness under these realistic imperfections. The model?s performance on two critical tasks is visualized in Fig. 18: depth completion for the lateral blind spots (left-front and right-front views) and super-resolution for the front view. Together, these results demonstrate that our model can faithfully recover the scene?s metric depth even when presented with imperfect, real-world data from an unseen sensor configuration.

6.2 Robustness under Environmental Degradation

This subsection examines the model?s robustness under conditions where environmental interference degrades perceptual signals. Two typical scenarios of signal degradation were considered:

? Night-time driving: Night-time environments introduce multiple challenges including insignificantly reduced signal-to-noise ratios, loss of texture and color information, over-saturation from artificial light sources, and high-contrast shadows. These factors substantially impact the reliability of vision-based perception systems.

? Rainy/Foggy weather conditions: LiDAR sensors suffer from reflectivity issues that produce anomalous signals or artifacts. This scenario tests whether our model can rely on visual signals to generate reasonable predictions when LiDAR inputs are corrupted.

Figure 18: Generalization to Real-World Sensor Configurations. We deployed a test vehicle to evaluate in-the-wild depth super-resolution and completion performance of our pre-trained model without any fine-tuning.

As shown in Fig. 19 and Fig. 16, our model maintains reliable depth estimation in both scenarios, demonstrating strong robustness against environmental degradation. The supplementary video further shows the stability of long-term temporal predictions in our real-world application.

6.3 Generalization to Unseen Visual Domains

This subsection evaluates the model?s zero-shot generalization on monocular depth estimation across visual domains absent from training.

Tests were conducted without prompt guidance on three challenging scenarios:

panoramic images from spherical projections, fisheye images with extreme distortions, and diverse in-the-wild

Image: Rainy Env
LiDAR Prompt
Metric Depth
Figure 19: Robustness in Adverse Weather. In the real-world deployment, we used a test vehicle to evaluate our pre-trained model for depth super-resolution and completion in rainy and foggy weather conditions without fine-tuning. These adverse conditions insignificantly affect scene reflectance, causing the LiDAR to produce numerous artifacts or completely occlude critical objects. For example, the degraded data can lead to flat ground surfaces being misinterpreted as uneven or crucial obstacles like pillars being missed. However, our model robustly ignores these erroneous inputs and generates accurate depth predictions based on visual cues, thereby demonstrating the complementary strengths of the two sensing modalities.
scenes including cartoons, grayscale images, and artistic renderings. Qualitative results (Fig. 20, Fig. 21, Fig. 22, and Fig. 23) confirm accurate metric depth estimation throughout. This robust performance across domains previously unrepresented in training data substantiates our claim of achieving ?Metric Anything? generalization.

Figure 20: Generalization to Unseen Visual Domains. Depth prediction results on fisheye images, an unseen domain characterized by severe radial distortion. The model was applied in a zero-shot setting without fine-tuning.
Figure 21: Generalization to Unseen Visual Domains. Depth prediction visualization for diverse in-the-wild images.

Figure 22: Generalization to Unseen Visual Domains. Visualizing depth predictions on panoramic images, an unseen domain during training. Our model successfully handles such extreme distortion and novel viewpoints.

Figure 23: Generalization to Unseen Visual Domains. Additional visualizations of depth predictions on diverse in-the-wild images.

[103] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8942?8952, 2021.
[104] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463, 2019.
[105] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ?ukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
[106] Massimiliano Viola, Kevin Qu, Nando Metzger, Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Marigold-dc: Zero-shot monocular depth completion with guided diffusion, 2024.
[107] Kaixuan Wang and Shaojie Shen. Flow-motion and depth network for monocular stereo and beyond. IEEE Robotics and Automation Letters, 5(2):3307?3314, 2020.
[108] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. arXiv preprint arXiv:1912.09678, 2019.
[109] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 5261?5271, 1925.
[110] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025.
[111] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20697?20709, 2024.
[112] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 4909?4916. IEEE, 2020.
[113] Yiran Wang, Jiaqi Li, Chaoyi Hong, Ruibo Li, Liusheng Sun, Xiao Song, Zhe Wang, Zhiguo Cao, and Guosheng Lin. Tacodepth: Towards efficient radar-camera depth estimation with one-stage fusion. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 10523?10533, 2025.
[114] Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior, 2025.
[115] Endre Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. Tohoku Mathematical Journal, First Series, 43:355?386, 1937.
[116] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. CVPR, 2025.
[117] 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.
[118] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. arXiv preprint arXiv:1810.08705, 2018.
[119] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-Llm: Boosting mllm capabilities in visual-based spatial intelligence. ArXiv, abs/2505.23747, 2025.
[120] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In 2021 IEEE international intelligent transportation systems conference (ITSC), pages 3095?3101. IEEE, 2021.
[121] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. ArXiv, abs/2408.10188, 2024.

[122] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 899?908, 2019.
[123] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. ArXiv, abs/2412.14171, 2024.
[124] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In CVPR, 2024.
[125] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. arXiv:2406.09414, 2024.
[126] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In Proceedings of the IEEE/CVF international conference on computer vision, pages 5684?5693, 2019.
[127] 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, pages 9043?9053, 2023.
[128] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 204?213, 2021.
[129] 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, pages 3712?3722, 2018.
[130] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.
[131] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu.
Long context transfer from language to vision.
ArXiv,
abs/2406.16852, 2024.
[132] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In CVPR, 2023.
[133] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024.
[134] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. ArXiv, abs/2410.02713, 2024.
[135] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 1702?1713, 1925.
[136] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In European Conference on Computer Vision, pages 519?535. Springer, 2020.
[137] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. arXiv preprint arXiv:2412.10345, 2024.
[138] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. Advances in Neural Information Processing Systems, 36:45137?45149, 2023.
[139] Yiming Zuo and Jia Deng. Ogni-dc: Robust depth completion with optimization-guided neural iterations. In ECCV, 2024.
[140] Yiming Zuo, Willow Yang, Zeyu Ma, and Jia Deng. Omni-dc: Highly robust depth completion with multiresolution depth integration. ICCV, 2025.