DepthAgent: Towards Better Universal Depth Estimation via Sample-wise Expert Selection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Monocular metric depth estimation has achieved strong progress with large-scale training and universal-camera modeling, yet robust deployment across diverse camera settings, such as perspective, fisheye, and panoramic images, remains challenging. Existing methods typically rely on a single depth estimator, overlooking that different models encode different camera assumptions and perform best under different input domains. In this paper, we show that depth experts exhibit strong sample-wise complementarity: model preference is highly correlated with camera geometry, and multi-model fusion brings the largest gains on difficult samples where individual experts are unreliable. Motivated by these observations, we propose **DepthAgent**, a vision-language agent for adaptive monocular depth estimation. DepthAgent treats existing depth models as frozen tools and learns to analyze scene and camera cues, invoke suitable experts through multi-turn tool utilization, and select or fuse their predictions for each input. To optimize such discrete decision-making toward dense geometric quality, we design a multi-reward reinforcement fine-tuning scheme that jointly encourages valid tool execution, camera/scene analysis, expert-selection quality, and inference efficiency. Extensive experiments across perspective, fisheye, and panoramic benchmarks show that DepthAgent consistently outperforms individual experts, fixed model fusion, and different selection strategies, with strong improvements on challenging samples, highlighting the critical role of expert selection and fusion. The code and model will be released upon publication.

1 Introduction

Monocular depth estimation is a fundamental problem for 3D scene understanding, with broad applications in autonomous driving, robotics, and augmented/virtual reality [58, 41, 10, 11]. Recent advances have substantially improved monocular depth prediction through large-scale supervised and self-supervised learning [47, 62, 63], diffusion-based geometric priors [27, 50], and zero-shot metric depth estimation [69, 21, 43]. Despite this progress, many real-world applications require metric depth, where accurate scale is critical for reliable spatial reasoning and decision-making. This necessity has driven growing interest in monocular metric depth estimation (MMDE), which aims to recover depth in absolute units rather than relative ordering. However, real-world deployment rarely follows a single imaging assumption. Images may come from standard perspective cameras, challenging perspective settings, fisheye lenses, or equirectangular projected (ERP) panoramas. Although recent works attempt to extend monocular depth estimation to large-FoV and universal-camera settings [24, 30, 53, 71, 1, 19, 13, 44], robust depth estimation across heterogeneous camera geometries and diverse scenarios remains challenging.

A central challenge in robust monocular depth estimation lies not only in the lack of strong depth models, but also in the fact that existing models encode different imaging assumptions and thus

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

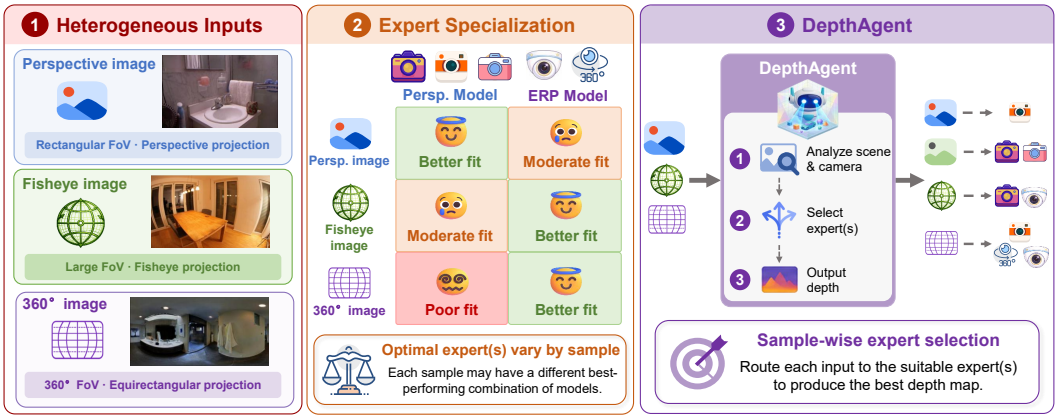


Figure 1: **Motivation of DepthAgent.** Real-world inputs span heterogeneous camera domains, including perspective, fisheye, and panoramic images, for which different depth experts exhibit different strengths. Instead of relying on a single model for all inputs, DepthAgent uses scene and camera cues to select suitable expert(s) on a per-sample basis, producing more reliable depth maps across diverse camera settings.

perform best under different camera geometries or data distributions. For instance, perspective-oriented models are well suited to standard perspective images, while ERP-trained models are better adapted to large-FoV and native panoramic/360° inputs. As a result, no single model is consistently optimal across scenarios, motivating adaptive model selection or combination. However, model combination remains underexplored in monocular depth estimation. Although routing and expert-selection strategies have been studied in language, vision, and multimodal systems [7, 40, 35, 82, 66, 79], they cannot be directly applied to depth estimation, where success depends on dense geometric accuracy rather than discrete labels or high-level preferences. A naive solution is to select a single fixed model or to fuse all expert predictions using a predefined rule, but both are limited: the former overlooks complementary expert strengths, whereas the latter may indiscriminately incorporate unreliable predictions and degrade the final depth estimate. Therefore, the key missing component is an input-adaptive mechanism that determines whether experts should be combined and which experts should participate. This leads to a research question: *given multiple depth models for the same task, how should one choose the most suitable model or model combination for each test sample?*

To examine whether such a mechanism is necessary and feasible, we first conduct a systematic analysis of single-model performance, oracle-style model selection, and multi-model fusion across diverse models and datasets. Our analysis reveals that **multi-model fusion often outperforms individual models** on most samples, and model preference is strongly correlated with camera geometry: perspective-oriented models are generally more effective on standard perspective inputs, while ERP-trained models are more reliable for native panoramic images. These highlight the potential benefit of exploiting complementary strengths among depth experts. We further observe that the benefit of fusion is not uniform, but becomes more pronounced on difficult samples where single-model predictions are less reliable. These findings suggest that robust monocular depth estimation should not rely on a fixed estimator or a static fusion rule, but instead requires an input-adaptive mechanism that can decide which experts to use and when to combine them.

Motivated by these observations, we recast robust universal depth estimation as an input-adaptive solution selection problem. Instead of seeking a single estimator or applying a fixed fusion rule, the goal is to decide for each input which depth expert(s) to trust, whether fusion is beneficial, and how much computation is justified. We propose **DepthAgent**, an agentic framework for adaptive monocular depth estimation. DepthAgent treats existing depth models as frozen tools and uses a Vision-Language Model (VLM) agent as a learned controller for sample-wise expert selection. The agent reasons over scene and camera cues, performs multi-turn tool use, and finally selects or fuses expert predictions, which are difficult to capture with a fixed hand-crafted router. Since the desired behavior involves discrete expert selection while the final objective is dense geometric quality, we optimize the agent with reinforcement fine-tuning based on GRPO [51], avoiding fixed oracle trajectories while enabling dynamic control over the performance-efficiency trade-off. We further design a **multi-reward scheme** that jointly encourages valid tool execution, camera- and scene-aware reasoning, empirically strong expert selection, and computation-aware efficiency. Experiments across perspective, fisheye, and panoramic benchmarks show consistent gains over individual experts, static

Table 1: **Model-family preference across camera-domain groups.** Best-single columns indicate the percentage of samples for which a Perspective or ERP model achieves the highest single-model performance. Oracle-presence columns indicate the percentage of samples in which at least one model from the corresponding family is selected by the oracle solution. Avg. gain denotes the average performance improvement achieved by fusion over the corresponding best single model across the evaluated perspective and ERP models.

Dataset group	**Best single model**		**Oracle solution presence**		**Avg. gain**
	Perspective	ERP	Perspective	ERP	$\Delta\delta_1$
Perspective	80.1	19.9	99.0	58.8	0.019
ERP variant	77.1	22.9	94.8	52.1	0.016
Native ERP	2.2	97.8	36.5	98.2	0.044

fusion baselines, and different expert-selection strategies, especially on challenging samples where individual predictions are often unreliable. Our contributions are summarized as follows:

- We conduct a systematic analysis of single-model selection and multi-model fusion for monocular depth estimation across perspective and panoramic scenarios, revealing strong model complementarity and camera-dependent expert preference.
- We formulate robust universal depth estimation as an input-adaptive solution selection problem and propose DepthAgent, a VLM-based agent that adaptively selects depth experts and fusion strategies for each sample.
- We design a multi-reward reinforcement fine-tuning scheme that encourages valid tool execution, camera- and scene-aware decision-making, and quality-efficiency balancing.
- We validate DepthAgent on diverse depth benchmarks, showing consistent improvements over individual experts, non-agentic fusion baselines, and different selection strategies, with especially strong gains on hard samples.

2 Related Works

Monocular Depth Estimation. Early monocular depth estimation primarily focused on perspective images, improving generalization with large-scale labeled datasets [47, 68], unlabeled data [62, 63], and generative priors from pre-trained diffusion models [27, 50]. Recent zero-shot metric depth methods further improve scale recovery by exploiting camera parameters through explicit supervision or canonical parameterization [17, 21, 69, 43, 45]. Beyond perspective images, large-FoV inputs such as fisheye and 360° images have been studied to handle richer context and severe geometric distortions [24, 30, 53, 71, 1]. Yet, due to limited large-FoV depth data and strong in-domain assumptions, existing models still struggle to generalize across camera types, even with recent efforts toward distortion modeling and unified cross-camera representations [54, 81, 61, 25, 49, 12, 19, 13, 44]. This highlights the importance of input-adaptive expert selection for robust real-world deployment, motivating DepthAgent to choose the most suitable expert(s) for each input.

Model Fusion and Expert Selection. Expert fusion and dynamic expert selection have been widely studied in language, vision, and multimodal systems [40, 35, 7, 59, 82, 66, 29, 83, 78, 79]. However, these ideas remain underexplored in depth estimation, with only limited studies on prediction fusion or complementary cues [48, 60, 9], while DepthFusion [39] indicates the potential of variance-aware expert selection. In contrast, we systematically study sample-wise model selection and fusion for depth estimation, demonstrating that adaptive expert choice and fusion can consistently improve performance, especially on challenging samples.

Agentic Tool Use and Reinforcement Fine-tuning. Recent studies have advanced multimodal models into agentic systems that can plan, invoke tools, and iteratively refine predictions for complex visual tasks [76, 31, 6, 83, 79]. Meanwhile, Reinforcement Learning (RL) has become an effective paradigm for improving the reasoning and problem-solving abilities of LLMs and VLMs [56, 26, 77, 37, 20, 18, 4, 55, 51, 23]. Recent works further adapt GRPO and its variants [18, 70, 33, 75, 32] to multimodal tasks with rule-based rewards [52, 22, 34, 57, 73, 80, 8, 42, 64, 79]. In contrast, we design selection, efficiency, and scene-awareness rewards to jointly optimize the effectiveness and efficiency of DepthAgent for depth estimation.

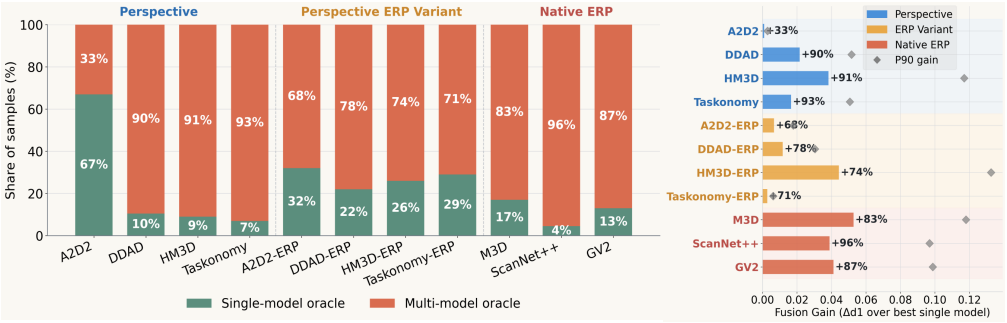


Figure 2: **Fusion consistently outperforms single models. Left:** Dataset-wise oracle proportions of single-model vs. multi-model solutions. **Right:** Fusion Gain against best single model. Bars and diamonds denote the mean and 90th percentile of per-sample δ_1 fusion gain over the best single model; annotations show the fraction of samples where fusion achieves higher δ_1.

3 Method Overview

We first conduct fusion analysis to understand when different depth experts or fusion solutions succeed or fail. These findings motivate the rest of our design. Based on them, we introduce DepthAgent, which learns to dynamically select suitable solutions for each input image. Finally, we derive the reward functions from the same empirical observations, encouraging the agent to favor solutions that yield better depth quality across different scenarios.

3.1 Fusion Analysis

Analysis setup. We first specify the candidate experts, fusion strategies, and evaluation protocol used in our fusion analysis. We consider perspective-oriented experts, including UniDepth [43], Metric3D [69], and Metric3Dv2 [21], and ERP-trained experts, including UniDAC [13] and UniK3D [44]. Besides individual experts, we evaluate pixel-wise mean, max, and min fusion [74, 36], as well as DepthFusion [39]. For each sample, a *solution* refers to either a single expert or a combination of multiple experts with a fusion strategy. *Oracle solution* is defined as the candidate solution that achieves the best δ_1, providing an empirical upper bound for sample-wise selection and fusion. We conduct the analysis on perspective datasets, including A2D2 [15], DDAD [16], HM3D [46], and Taskonomy [72], together with their ERP variants, constructed by transforming images into ERP patches, since ERP data are relatively scarce (details in Appendix). To cover native ERP inputs, we include ScanNet++ [67], Matterport3D [5], and Pano3D-GV2 [2]. For each dataset, we randomly select 200 training samples for analysis, ensuring that no test-set bias is introduced.

Camera geometry is the primary factor in model preference. Tab. 1 shows that model preference is strongly tied to camera geometry. Perspective models dominate perspective datasets, winning 80.1% of samples, and remain preferred on ERP variants. This suggests that ERP reprojection of perspective images does not fully reproduce the distortions and model preferences of native panoramic data. In contrast, native ERP datasets show a clear reversal, where ERP-trained models win 97.8% of samples. Oracle solutions further reveal cross-family complementarity, with the dominant model family varying consistently with the input camera type. These results suggest a clear camera-dependent preference: even though ERP models can process perspective-like inputs, perspective-trained models remain superior on perspective data, whereas ERP-trained models perform better on native ERP images. Therefore, *model selection should depend on the camera type.*

Fusion outperforms single models on most samples. In Fig. 2 (left), the proportion of oracle solutions that contain multiple models could reach 96% on ScanNet++ and 93% on Taskonomy, while the remaining samples are best handled by a single model. Across all 11 datasets, the oracle solutions are multi-model for 78.6% of samples. Fig. 2 (right) shows the average δ_1 improvement of the oracle solution compared with the best single model on each dataset. Fusion improves performance to varying degrees across all datasets. In addition to the mean gain, we report the P90 gain, defined as the 90th percentile of per-sample $\Delta\delta_1$, to characterize the upper-end improvement achieved on samples where fusion is particularly beneficial. The advantage is most pronounced on native ERP

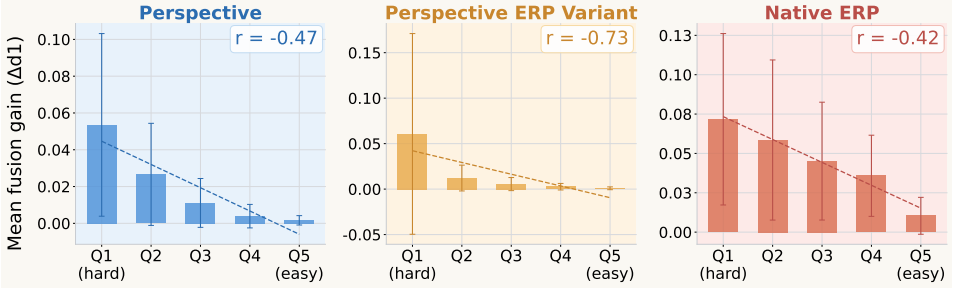


Figure 3: **Difficulty-dependent fusion gain.** Mean fusion gain ($\Delta\delta_1 \pm \sigma$) is shown across best-single δ_1 quintiles within each dataset group, where Q1/Q5 denote the hardest/easiest samples. Fusion gains are largest on hard samples (the strongest individual model performs poorly). Dashed lines indicate linear trends over quintile means, while Pearson r is computed over pooled per-sample pairs.

datasets: fusion achieves mean gains of $\Delta\delta_1 = 0.053$ on Matterport3D and 0.041 on Pano3D-GV2, with 83% and 87%of samples benefiting from fusion, respectively.

Fusion yields larger gains on difficult samples. Fig. 3 groups samples by the best single-model δ_1 within each dataset family, where Q1 contains the hardest samples and Q5 the easiest. In all three settings, the largest mean fusion gains appear in the hardest quintile and steadily diminish as single-model performance improves. The negative Pearson correlations further confirm that fusion is most useful precisely when the strongest individual model is insufficient. This pattern is especially pronounced in the ERP setup, where difficult samples are more frequent, making fusion a targeted correction rather than a uniform boost. This finding suggests that *fusion gains are strongly correlated with sample difficulty, rather than being uniformly distributed across samples.*

These analyses motivate DepthAgent: the best solution is input-dependent, governed by camera geometry and sample difficulty. Thus, fusion should be a per-sample decision over when to fuse, which experts to trust, and how to combine them. We formulate this as an agent-driven process, where a VLM selects the most suitable solution for each input.

3.2 DepthAgent

The overall framework of DepthAgent is illustrated in Fig. 4. Given an input image x, DepthAgent formulates depth estimation as an agentic process that adaptively coordinates multiple depth experts to produce the final depth solution. Since expert selection depends on high-level cues such as scene content, camera geometry, projection distortion, and intermediate prediction consistency, which are difficult to encode with a fixed hand-crafted router, DepthAgent leverages a VLM agent to jointly reason about expert suitability and dynamically adjust tool usage for each sample.

Feature analysis. Before invoking any depth expert, the agent first considers the visual and geometric properties of the input image. This includes scene-level cues, such as whether the image is indoor or outdoor, and camera-related cues, such as intrinsic geometry, projection characteristics, and extracted depth-related features. These observations help the agent identify the camera type of x and which depth expert is more suitable for the current sample.

Depth expert pool and tool selection. Based on the analyzed features, the agent selects depth experts from a heterogeneous depth expert pool. The pool contains both perspective models and ERP-trained models, enabling DepthAgent to handle standard pinhole images as well as 360°, fisheye, and other geometrically distorted inputs. At each decision step, DepthAgent chooses an expert a_t according to the current image understanding and previously observed tool outputs, instead of relying on a fixed model or a predefined fusion rule.

Multi-turn tool execution. After a successful tool call, the selected expert returns a predicted depth map $\mathbf{D}_{a_t}$ together with auxiliary depth features f_{a_t}, such as average depth distance and depth variation. The depth map serves as a candidate solution, while the auxiliary features provide a compact summary of the prediction quality and geometric behavior, making the tool result easier for the agent to interpret. Conditioned on the current observation and previous tool results, the agent decides whether to invoke another expert, compare complementary predictions, or terminate with

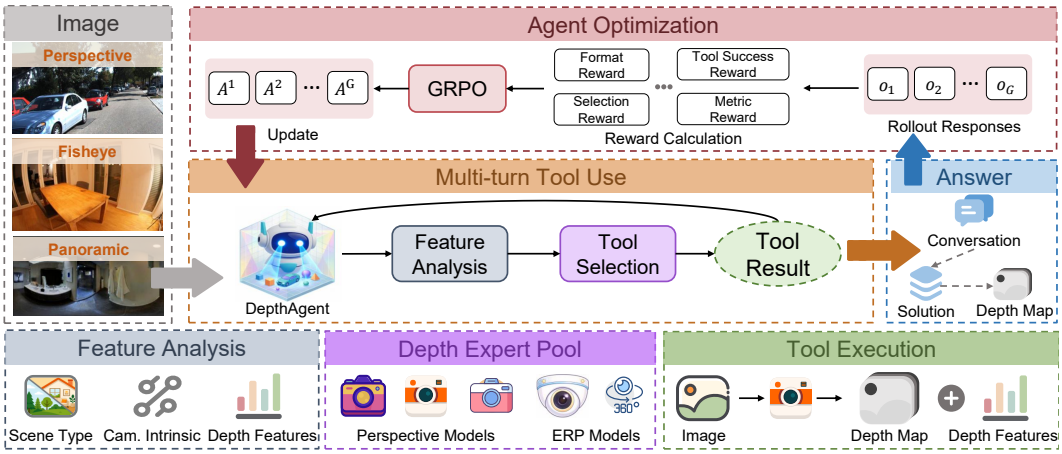


Figure 4: **Overview of DepthAgent.** Given an input image, DepthAgent analyzes the scene type and camera intrinsics, and interactively selects depth experts from a depth-expert tool pool. Each tool call returns a depth prediction along with auxiliary depth features. Based on the tool results, DepthAgent determines whether to continue exploration or produce the final solution, which includes the final depth map generated by the solution. During training, rollout conversations are optimized with GRPO using the proposed multi-reward objective.

a final answer. This ReAct-style [65] multi-turn process allows DepthAgent to adaptively explore different depth experts and construct an appropriate depth solution for each sample.

Agent optimization. We use GRPO [51] instead of SFT because fixed oracle trajectories may become invalid when the expert pool changes, and offer limited flexibility in dynamically balancing tool call efficiency and depth performance. During training, we sample G rollout responses $\{o_1, o_2, \ldots, o_G\}$ for each input, where o_i records the agent's reasoning, tool-use trajectory, selected solution (expert combination and fusion strategy), and corresponding final depth map. Each rollout is evaluated using rewards that consider both the final depth prediction and the intermediate decision process, including format correctness, tool success, selection behavior, and efficiency. The resulting advantages $\{A^1, A^2, \ldots, A^G\}$ are then used to update the agent, enabling DepthAgent to learn how to select suitable depth tools through interaction. Furthermore, we apply multi-reward normalization [80, 32] in the advantage calculation. Specifically, given K reward functions $r^i_{(1)}, r^i_{(2)}, \ldots, r^i_{(K)}$ for the i-th response o_i, we compute the normalized advantage for each reward and aggregate as:

$$A^i_{(k)} = \frac{r^i_{(k)} - \text{mean}(r^1_{(k)}, \ldots, r^G_{(k)})}{\text{std}(r^1_{(k)}, \ldots, r^G_{(k)})}, A^i = \sum_{k=1}^{K} A^i_{(k)}. \quad (1)$$

3.2.1 Reward Functions

In addition to traditional rewards like format reward and tool success reward [51, 34, 79], we further design three reward functions: scene awareness reward R_{scene}, selection prior reward R_{sel}, and efficiency-aware metric reward R_{em}. Additional details are in the Appendix.

Scene awareness reward. Correct model selection requires the agent to understand the scene context and infer the camera geometry from the input image. We introduce a scene awareness reward to encourage the agent to first analyze whether the scene is indoor or outdoor and to estimate the camera type before selecting tools. For each sample, the Depth Agent first invokes the `estimate_camera_type` tool to infer whether the camera follows a perspective/pinhole or ERP/360° panoramic projection from camera information. This reward is implemented as a lightweight multi-label matching criterion over the first reasoning block, based on correctly identified scene category and camera type. The reward equals the average accuracy over these two labels.

Selection prior reward. The selection prior reward R_{sel} encodes our prior knowledge about expert suitability, as observed in Sec 3.1. It gives a mild reward when the selected experts match the input domain: perspective-oriented experts for perspective inputs, and ERP-trained experts for panoramic or large-FoV inputs. This provides soft guidance for faster policy adaptation in training. R_{sel} is set to 1.0 if the selected experts include at least one expert matching the input domain, and 0.0 otherwise.

Efficiency-aware metric reward. Although more tool calls may improve the final prediction, the marginal gain is not always worth the additional computation. We therefore introduce an efficiency-aware metric reward R_{em} that jointly considers the metric improvement and the number of model calls. Given a solution y_i from o_i, let m_i denote its task metric score and n_i denote the number of models in the solution. y_{ref} and m_{ref} denote the reference solution (*e.g.*, per-sample oracle solution) and its metric score. We define the normalized metric gap and efficiency gap:

$$\Delta m_i = \frac{m_i - m_{\text{ref}}}{|m_{\text{ref}}| + \epsilon}, \qquad \Delta n_i = \frac{n_i - n_{\text{ref}}}{n_{\max}}. \tag{2}$$

The efficiency-aware reward is then computed as

$$R_{\text{em}}(y_i, y_{\text{ref}}) = \Delta m_i - \lambda \Delta n_i \exp\left(-\frac{|\Delta m_i|}{\tau}\right), \tag{3}$$

where λ controls the strength of the efficiency penalty, τ controls how strongly the penalty depends on the metric gap, and ϵ is used for numerical stability. This reward discourages extra model calls when the metric improvement is small, while still allowing higher computation when it leads to a meaningful performance gain.

4 Experiments

Datasets. We train on indoor perspective and ERP variant from HM3D [46], outdoor perspective data from DDAD [16] and A2D2 [15], and panoramic data from Matterport3D [5]. We evaluate on six diverse benchmarks across three camera domains: perspective datasets, including KITTI [14], NYU-v2 [38], and IBims-1 [28]; fisheye datasets ScanNet++ [67]; and panoramic / 360° datasets, including Matterport3D [5] and Pano3D-GV2 [2]. We follow prior works [13, 19, 44] to report the percentage of inliers (δ_i) under thresholds of 1.25^i, absolute relative error (A.Rel), and root mean squared error (RMSE). ERP models are evaluated against ERP-transformed ground truth [13, 19].

Baselines. We compare DepthAgent against the **depth experts in the tool pool** and **fusion strategies** described in Sec. 3.1. Furthermore, we compare with different selection strategies: (i) **Rand-Model:** randomly selects an expert for each sample; (ii) **Rand-Sol:** randomly selects one candidate solution. (iii) **MLP router:** an MLP as a selection router to predict the suitable solution. For Rand-Model and Rand-Sol, expert selection is restricted to perspective-specialized experts for perspective datasets, whereas selection is performed among ERP-trained candidates for other datasets. Details of the MLP router are provided in the Appendix.

Implementation. To avoid excessive computational overhead, DepthAgent is built on Qwen2.5-VL-3B [3] and fine-tuned with GRPO [51]. We train the agent for 300 steps with a maximum of 5 interaction turns. λ is set to 1 and 0.2, and τ is set to 0.1 and 3.4 for perspective and ERP inputs, respectively. Training is performed on 4 H100 GPUs with an effective batch size of 4. During inference, we use greedy decoding to ensure reproducibility. Additional details are in the Appendix.

4.1 Experimental Results

Comparison with baselines. Tab. 2 shows that DepthAgent consistently outperforms individual experts and non-agentic fusion baselines across camera domains, achieving the best performance on five datasets as well as on average. Although UniK3D is a strong universal-camera expert, DepthAgent significantly improves A.Rel and RMSE. Compared with fusion strategies, DepthAgent shows stronger robustness, underscoring the importance of sample-wise solution selection. The comparisons with Rand-Model and Rand-Sol further indicate that agentic selection is a key contributor to the strong performance of DepthAgent. In contrast, the MLP router shows limited routing ability and poor generalization. Notably, the improvements are most evident in RMSE, which penalizes large depth errors more heavily and is thus crucial for evaluating geometric fidelity and robustness in depth estimation. Although DepthAgent still falls short of the upper bound, it consistently outperforms the baselines by reducing the negative impact of suboptimal fusion. Overall, these results suggest that *fusion is effective only when appropriate and complementary experts are selected.*

Comparison on hard samples. Hard samples are critical because they often correspond to failure cases in real-world deployment, where improving robustness is more important than average-case

Table 2: **Quantitative comparison.** Perspective reports the average performance over KITTI, NYU-v2, and IBims-1. Fisheye reports performance on ScanNet++, while panoramic reports the average performance over Matterport3D and Pano3D-GV2. ‡: best solution per sample.

Method	Perspective			Fisheye			Panoramic		
	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$
UniDepth [43]	0.700	0.203	1.359	0.135	0.573	1.029	0.415	0.314	0.898
Metric3D [69]	0.850	0.136	1.279	0.727	0.162	0.450	0.403	0.290	0.721
Metric3Dv2 [21]	0.864	0.124	1.193	0.745	0.170	0.474	0.422	0.300	0.847
UniDAC [13]	0.845	0.140	1.876	0.918	0.097	0.279	0.755	0.169	0.472
UniK3D [44]	0.937	0.091	1.261	0.927	0.103	0.287	**0.820**	0.152	0.383
Mean [74, 36]	0.890	0.110	1.045	0.840	0.139	0.313	0.620	0.199	0.529
Max [74, 36]	0.590	0.285	2.276	0.139	0.573	1.023	0.735	0.235	0.500
Min [74, 36]	0.912	0.095	1.258	0.663	0.190	0.573	0.266	0.354	1.091
DepthFusion [39]	0.878	0.117	1.195	0.621	0.252	0.623	0.378	0.327	1.000
Rand-Model	0.801	0.159	1.287	0.921	0.100	0.284	0.787	0.160	0.429
Rand-Sol	0.808	0.150	1.250	0.927	0.097	0.279	0.792	0.158	0.422
MLP	0.904	0.107	1.328	0.187	1.014	2.547	0.744	0.182	0.473
DepthAgent	**0.948**	**0.070**	**0.918**	**0.946**	**0.084**	**0.254**	0.819	**0.145**	**0.365**
Upper-bound‡	0.985	0.057	0.844	0.967	0.080	0.239	0.860	0.132	0.360

Table 3: **Comparison on hard samples.** We report performance on hard samples, defined as the top 10% worst-performing samples of the best single model. Perspective reports the average performance over KITTI, NYU-v2, and IBims-1. Fisheye reports performance on ScanNet++, while panoramic reports the average performance over Matterport3D and Pano3D-GV2.

Method	Perspective			Fisheye			Panoramic		
	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$
UniDepth [43]	0.650	0.271	2.045	0.197	0.610	0.939	0.337	0.496	0.853
Metric3D [69]	0.681	0.220	2.311	0.635	0.203	0.629	0.346	0.372	0.847
Metric3Dv2 [21]	0.737	0.212	2.072	0.659	0.204	0.593	0.292	0.427	0.919
UniDAC [13]	0.754	0.177	2.286	0.789	0.163	0.434	0.547	0.273	0.525
UniK3D [44]	0.794	0.155	1.950	0.799	0.156	0.434	0.560	0.261	0.466
DepthAgent	**0.833**	**0.115**	**1.658**	**0.817**	**0.144**	**0.410**	**0.587**	**0.253**	**0.450**

performance. We further evaluate DepthAgent on hard samples, defined as the worst 10% samples ranked by the δ_1 score of the best single model on each dataset. As shown in Tab. 3, DepthAgent consistently outperforms all baseline experts on perspective, fisheye, and 360° data. The improvements are more pronounced than on the full test sets, especially in δ_1, indicating that DepthAgent is particularly effective at selecting reliable experts for challenging cases.

Efficiency of DepthAgent. To improve efficiency, we introduce a Fast mode that skips CoT reasoning and directly selects tools before producing the final answer, while retaining the same selected solution as CoT mode. On an H100, DepthAgent takes 1.1s per sample in Fast mode, close to exhaustively running all five experts at 0.76s, while CoT mode remains moderate at 3.7s. Although exhaustive fusion (*i.e.*, Mean in Tab. 2) is similarly efficient, it can substantially degrade depth accuracy by indiscriminately combining unreliable predictions.

4.2 Ablation Studies

Effects of multi-reward. We ablate different reward components on all datasets, as shown in Tab. 4a. The results suggest that relying on a partial reward is insufficient for robust solution selection. Combining R_{scene}, R_{sel}, and R_{em} achieves the best overall performance, indicating that scene understanding, selection prior, and efficiency-aware metric optimization are complementary.

Effects of λ, τ, and $n_{\max}$. Tab. 4b studies the effects of the reward hyperparameters in R_{em}. Overall, $n_{\max}$ mainly determines the tool-call budget and directly affects the average tool calls. Under a given budget, λ modulates the penalty for tool usage, while τ controls the tolerance to metric differences among candidate solutions. This suggests that a proper λ could balance performance and tool cost. Together, they shape the reward sensitivity and influence whether the agent favors compact or more diverse solution compositions. The results indicate that robust performance comes from a well-calibrated reward design, rather than simply increasing the number of tool calls.

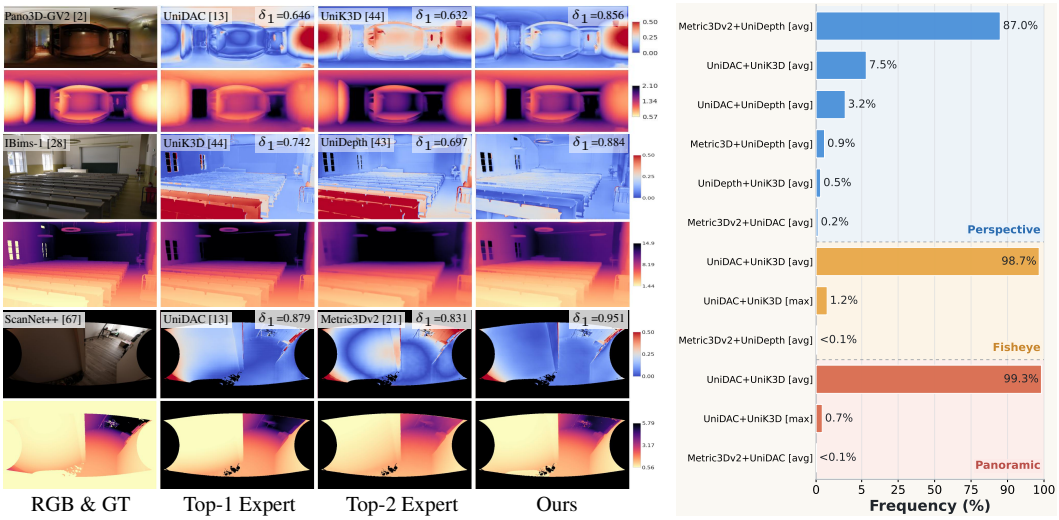
Table 4: **Ablation study of the reward design.** Combining all rewards improves overall performance, while calibrated hyperparameters drive results beyond tool usage alone. $\bar{n}$: average number of tool calls.

(a) Leave-one-out ablation of reward components.

R_{scene}	R_{sel}	R_{em}	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$
✓	✓	–	0.832	0.146	0.827
✓	–	✓	0.871	0.134	0.650
–	✓	✓	0.891	0.102	0.639
✓	✓	✓	0.905	0.097	0.623

(b) Effect of reward hyperparameters in R_{em}.

λ	τ	$n_{\max}$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\bar{n}$
0.4	1.0	1.0	0.860	0.119	0.576	1.0
0.1	2.0	2.0	0.764	0.166	0.678	3.2
0.4	3.4	2.0	0.828	0.143	0.620	2.5
0.2	3.4	2.0	0.905	0.097	0.623	1.9



(a) **Depth Map Comparison.** Odd rows show the error maps, while even rows show the predicted depth maps. For each sample, we compare the top-2 expert predictions with the final solution of DepthAgent. DepthAgent generates more accurate depth maps with lower errors. Zoom in for better effects.

(b) **Solution frequency.** Results are grouped by projection type (Perspective, Fisheye, and 360°), and each group reports the top selected solution configurations over all test samples.

Figure 5: **Analysis of DepthAgent behavior.**

4.3 Qualitative Results

Depthmap comparison. Fig. 5a compares DepthAgent with the top-2 individual experts on representative samples. DepthAgent produces more faithful depth structures and consistently lower error maps by selecting or combining complementary expert solutions. Additional visualizations are provided in the Appendix.

Solution distribution on different scenarios. Fig. 5b shows that our agent adopts camera-dependent selection strategies under the proposed efficiency-aware metric reward. The highly concentrated distributions in fisheye and 360° indicate a larger metric gap between the dominant solution and other candidates, suggesting that ERP-trained experts play a leading role in these scenarios. Moreover, the policy does not simply collapse to the strongest average expert UniK3D: perspective samples are often routed to Metric3Dv2/UniDepth-based combinations. This also implies that appropriate expert fusion can provide stronger gains when the selected experts are well matched to the camera domain.

Conversation. We provide representative visualization conversation examples in the Appendix.

5 Conclusion

We presented **DepthAgent**, an agentic framework for universal monocular depth estimation that performs sample-wise expert selection and fusion. Motivated by our analysis showing the benefits of fusion and camera-dependent expert preference, DepthAgent uses a VLM agent to reason about scene and camera cues, invoke frozen depth experts, and adaptively perform the final solution. With multi-reward reinforcement fine-tuning, DepthAgent consistently improves over baselines across perspective, fisheye, and panoramic benchmarks, especially on challenging samples. These results highlight the importance of model selection and fusion, moving beyond fixed depth estimators toward input-adaptive expert selection for systematic depth estimation.

References

[1] Hao Ai, Zidong Cao, Yan-Pei Cao, Ying Shan, and Lin Wang. Hrdfuse: Monocular 360° depth estimation by collaboratively learning holistic-with-regional depth distributions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023.

[2] Georgios Albanis, Nikolaos Zioulis, Petros Drakoulis, Vasileios Gkitsas, Vladimiros Sterzentsenko, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. Pano3d: A holistic benchmark and a solid baseline for 360deg depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3727–3737, 2021.

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

[4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[5] 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.

[6] Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[7] Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

[8] Xiaxu Chen, Wei Li, Chunxu Liu, Chi Xie, Xiaoyan Hu, Chengqian Ma, Feng Zhu, and Rui Zhao. On the suitability of reinforcement fine-tuning to visual tasks. In *CVPR*, 2025.

[9] Yaqiao Dai, Renjiao Yi, Chenyang Zhu, Hongjun He, and Kai Xu. Multi-resolution monocular depth map fusion by self-supervised gradient-based composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[10] Xingshuai Dong, Matthew A Garratt, Sreenatha G Anavatti, and Hussein A Abbass. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):16940–16961, 2022.

[11] Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, et al. Depthlab: Real-time 3d interaction with depth maps for mobile augmented reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 829–843, 2020.

[12] Hao Feng, Wendi Wang, Jiajun Deng, Wengang Zhou, Li Li, and Houqiang Li. Simfir: A simple framework for fisheye image rectification with self-supervised representation learning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023.

[13] Girish Chandar Ganesan, Yuliang Guo, Liu Ren, and Xiaoming Liu. Unidac: Universal metric depth estimation for any camera. In *CVPR*, 2026.

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[15] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020.

[16] 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 (CVPR)*, 2020.

[17] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9233–9243, 2023.

[18] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[19] Yuliang Guo, Sparsh Garg, S Mahdi H Miangoleh, Xinyu Huang, and Liu Ren. Depth any camera: Zero-shot metric depth estimation from any camera. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26996–27006, 2025.

[20] Dong Han, Beni Mulyana, Vladimir Stankovic, and Samuel Cheng. A survey on deep reinforcement learning algorithms for robotic manipulation. *Sensors*, 2023.

[21] 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.

[22] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

[23] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

[24] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters (RA-L)*, 6(2):1519–1526, 2021.

[25] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics Autom. Lett.*, 2021.

[26] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 1996.

[27] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[28] Tobias Koch, Lukas Liebel, Marco Körner, and Friedrich Fraundorfer. Comparison of monocular depth estimation methods using geometrically relevant metrics on the IBims-1 dataset. *Computer Vision and Image Understanding (CVIU)*, 191:102877, 2020. doi: 10.1016/j.cviu.2019.102877.

[29] Sijie Li, Chen Chen, and Jungong Han. Simmlm: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24068–24077, 2025.

[30] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022.

[31] Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. Longvideoagent: Multi-agent reasoning with long videos. *arXiv preprint arXiv:2512.20618*, 2025.

[32] Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026.

[33] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

[34] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *ICCV*, 2025.

[35] Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. Routing to the expert: Efficient reward-guided ensemble of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024.

[36] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 2022.

[37] Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. Deep learning, reinforcement learning, and world models. *Neural Networks*, 2022.

[38] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *The European Conference on Computer Vision (ECCV)*, 2012.

[39] Anton Obukhov, Matteo Poggi, Fabio Tosi, Ripudaman Singh Arora, Jaime Spencer, Chris Russel, Simon Hadfield, Richard Bowden, Shuaihang Wang, Zhenxin Ma, et al. The fourth monocular depth estimation challenge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[40] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

[41] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3142–3152, 2021.

[42] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[43] 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*, pages 10106–10116, 2024.

[44] Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1028–1039, 2025.

[45] 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, 2025.

[46] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

[47] 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.

[48] Weisong Ren, Lijun Wang, Yongri Piao, Miao Zhang, Huchuan Lu, and Ting Liu. Adaptive co-teaching for unsupervised monocular depth estimation. In *European Conference on Computer Vision*, pages 89–105. Springer, 2022.

[49] Manuel Rey, Mingze Yuan Area, and Christian Richardt. 360monodepth: High-resolution 360 monocular depth estimation. in 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[51] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[52] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[53] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360$ˆ{\circ }$ depth estimation. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I*, 2022.

[54] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360° imagery. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017.

[55] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[56] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

[57] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025.

[58] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8445–8453, 2019.

[59] Yuanchen Wu, Junlong Du, Ke Yan, Shouhong Ding, and Xiaoqiang Li. Tove: Efficient vision-language learning via knowledge transfer from vision experts. *arXiv preprint arXiv:2504.00691*, 2025.

[60] Zhihao Xia, Patrick Sullivan, and Ayan Chakrabarti. Generating and exploiting probabilistic monocular depth estimates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 65–74, 2020.

[61] Yuwen Xiong, Zhiqi Li, Yuntao Chen, Feng Wang, Xizhou Zhu, Jiapeng Luo, Wenhai Wang, Tong Lu, Hongsheng Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications. 2024.

[62] 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*, pages 10371–10381, 2024.

[63] 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, 2024.

[64] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[65] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

[66] Hanrong Ye and Dan Xu. Taskexpert: Dynamically assembling multi-task representations with memorial mixture-of-experts. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21828–21837, 2023.

[67] 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*, pages 12–22, 2023.

[68] 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.

[69] 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.

[70] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[71] Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, and Chae-Eun Rhee. Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023.

[72] 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.

[73] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

[74] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 2020.

[75] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[76] Zetong Zhou, Dongping Chen, Zixian Ma, Zhihan Hu, Mingyang Fu, Sinan Wang, Yao Wan, Zhou Zhao, and Ranjay Krishna. Reinforced visual perception with tools. *arXiv preprint arXiv:2509.01656*, 2025.

[77] Jie Zhu, Mengsha Hu, Amy Zhang, Ruoming Jin, and Rui Liu. Fairness-sensitive policy-gradient reinforcement learning for reducing bias in robotic assistance. In *IEEE ROMAN*, 2024.

[78] Jie Zhu, Yiyang Su, Minchul Kim, Anil Jain, and Xiaoming Liu. A quality-guided mixture of score-fusion experts framework for human recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[79] Jie Zhu, Xiao Guo, Yiyang Su, Anil Jain, and Xiaoming Liu. Fusionagent: A multimodal agent with dynamic model selection for human recognition. In *CVPR*, 2026.

[80] Jie Zhu, Yiyang Su, and Xiaoming Liu. Can textual reasoning improve the performance of mllms on fine-grained visual classification? *arXiv preprint arXiv:2601.06993*, 2026.

[81] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets V2: more deformable, better results. 2019.

[82] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *Advances in Neural Information Processing Systems*, 2024.

[83] Yushen Zuo, Qi Zheng, Mingyang Wu, Xinrui Jiang, Renjie Li, Jian Wang, Yide Zhang, Gengchen Mai, Lihong V. Wang, James Zou, Xiaoyu Wang, Ming-Hsuan Yang, and Zhengzhong Tu. 4kagent: Agentic any image to 4k super-resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.