MAV-SLAM: MULTI-LLM-AGENT CREW FOR VISUAL SLAM WITH 3D GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual Simultaneous Localization and Mapping (SLAM) reconstructs the metric structure of the physical world from sensor imagery, enabling precise robotic pose estimation. However, environmentally induced image degradation and varied image processing strategies significantly compromise localization accuracy. Intelligent SLAM systems address this challenge by autonomously perceiving dynamic perturbations and formulating adaptive processing strategies, further identifying and deploying optimal methodologies to achieve target localization objectives with enhanced metric precision. This paper introduces MAV-SLAM, a novel Multi-LLM-Agent-Orchestrated visual SLAM framework that proactively identifies and compensates for suboptimal image quality while autonomously selecting optimal depth estimation models. Specifically, we integrate a visual-language model that performs autonomous image restoration guided by image quality assessment, significantly enhancing SLAM localization performance. Furthermore, we implement a routing large language model for adaptive depth estimation, which consequently elevates the quality of 3D reconstruction via 3D Gaussian Splatting (3DGS). Rigorous evaluation across multiple benchmarks demonstrates that MAV-SLAM exhibits superior performance in both localization accuracy and 3DGS-based reconstruction fidelity, validating its effectiveness in real-world scenarios.

1 INTRODUCTION

Visual SLAM localization accuracy is compromised by two principal factors: external surroundings (e.g., haze, rain, low illumination) and camera-induced artifacts (e.g., motion blur, sensor noise), both capable of precipitating localization failure. Environmental factors significantly affect the accuracy of the SLAM localization by degrading the quality of image acquisition. Input image fidelity directly governs the performance of key components, including feature extraction and depth estimation. Adverse factors, such as sensor noise, motion blur, precipitation, and low light environments, seriously challenge robust SLAM operation. Mitigation strategies typically employ classical enhancement methods (e.g., Contrast Limited Adaptive Histogram Equalization (CLAHE)) or develop deep learning-based SLAM frameworks targeting specific degradation types.

However, we argue that a fundamentally more effective strategy involves embedding autonomous image quality assessment directly within the SLAM pipeline, enabling targeted restoration of degraded inputs. This approach allows real-time adaptation to varying visual conditions, thereby enhancing the system's flexibility, robustness, and intelligence—essential qualities for deployment in diverse and unpredictable environments. Recent advances in learning-based metric depth estimation, particularly through the incorporation of large-scale foundational models such as DINO (Zhang et al., 2024a) and CLIP (Radford et al., 2021), have significantly outperformed traditional methods in both detail fidelity and metric accuracy. Nevertheless, these models often exhibit strong dataset bias, making it impractical to develop a single architecture that generalizes robustly across all scenarios. To address this, we leverage Mixture-of-Experts (MoE) techniques to integrate a diverse set of depth estimation capabilities, dynamically selecting the most suitable model per input. This strategy substantially improves generalization while minimizing computational overhead. Furthermore, Multi-LLM-Agent systems provide a compelling framework for SLAM integration, as they support autonomous environmental perception and task execution while facilitating seamless interaction between large models and SLAM via Function Calling mechanisms.

In this work, we present an MAV-SLAM system that represents a significant milestone toward intelligent SLAM by integrating large-model-driven perception within a Multi-LLM-Agent framework. Our contributions are as follows:

- A Multi-LLM-Agent framework for Visual SLAM that supports various large models;
- A SLAM frontend image enhancement algorithm powered by vision-language models (VLMs);
- A Route LLMs-based method for precise metric depth estimation.

2 RELATED WORKS

Generalizing visual SLAM across diverse scenarios remains challenging. Recent LLM and VLM advances have enabled more adaptive visual algorithms, improving performance in tasks like image restoration under adverse conditions (Chobola et al., 2024; Wang et al., 2024b) and metric depth estimation (Yang et al., 2024b; Bhat et al., 2023). Integrating LLM-based agents further supports flexible and intelligent SLAM design.

2.1 IMAGE RESTORATION

Image restoration enhances SLAM accuracy and robustness. While traditional methods like CLAHE (Qin et al., 2018) struggle with complex distortions, deep learning techniques using pixel-level networks (Quan et al., 2024), self-attention (Mao et al., 2024), and end-to-end frameworks (Yang et al., 2024a; Zhang et al., 2024b) have demonstrated stronger performance. Progress includes low-light enhancement (Chobola et al., 2024; Wang et al., 2024b; Chen et al., 2021), dehazing (Junkai et al., 2025; Jin et al., 2023), and deraining (Gao et al., 2024). However, integrating multiple restoration modules in real-time SLAM is computationally costly. All-In-One restoration (Oh et al., 2024) offers a viable alternative, though it requires effective image quality assessment (IQA) for continuous streams. Deep learning-based IQA (Avanaki et al., 2024) currently leads the field, with LLMs and VLMs further enabling human-aligned evaluation (Wang et al., 2023; You et al., 2024b) and language-guided comparisons (You et al., 2024a).

Image-Enhanced SLAM: Several SLAM systems incorporate restoration during training for deblurring (Luo et al., 2024; Davletshin et al., 2024), deraining (Albanese et al., 2024), and low-light enhancement (Zhao et al., 2024; Wang et al., 2024a). Restoration has also been applied to 3D Gaussian Splatting for high-quality reconstruction (Renlong et al., 2024). Nonetheless, most methods are specialized and lack a unified approach for multiple degradations.

2.2 LLM AGENTS

Integrating LLMs/LMMs into visual SLAM necessitates novel Multi-LLM-Agent frameworks that combine real-time perception with autonomous decision-making. Systems like AutoGen[1] and Magnetic-One (Microsoft, 2024) enable hierarchical agent coordination for tool use and reasoning. LangGraph[2] supports stateful, graph-based workflows, while Swarm[3] and CrewAI[4] offer lean, event-driven architectures for scalable task execution.

2.3 ROUTING LLMS

Inspired by Mixture-of-Experts (MoE) models (Shazeer et al., 2017; DeepSeek-AI et al., 2025), routing mechanisms now employ entire language models as experts to balance accuracy and cost. Frameworks like RouteLLM (Ong et al., 2024) use classifier-based routing for model selection. Benchmarks such as RouterBench (Hu et al., 2024b) and RouterEval (Huang et al., 2025) systematically evaluate these strategies.

[1] https://github.com/microsoft/autogen
[2] https://www.langchain.com/langgraph
[3] https://github.com/openai/swarm
[4] https://docs.crewai.com/introduction

2.4 METRIC DEPTH ESTIMATION

Traditional depth estimation uses structured light (sensitive to ambient light) or stereo algorithms (computationally limited). Learning-based monocular methods (Yang et al., 2024b; Bhat et al., 2023; Yin et al., 2023; Hu et al., 2024a) predict dense depth from single images but can suffer from scale drift.

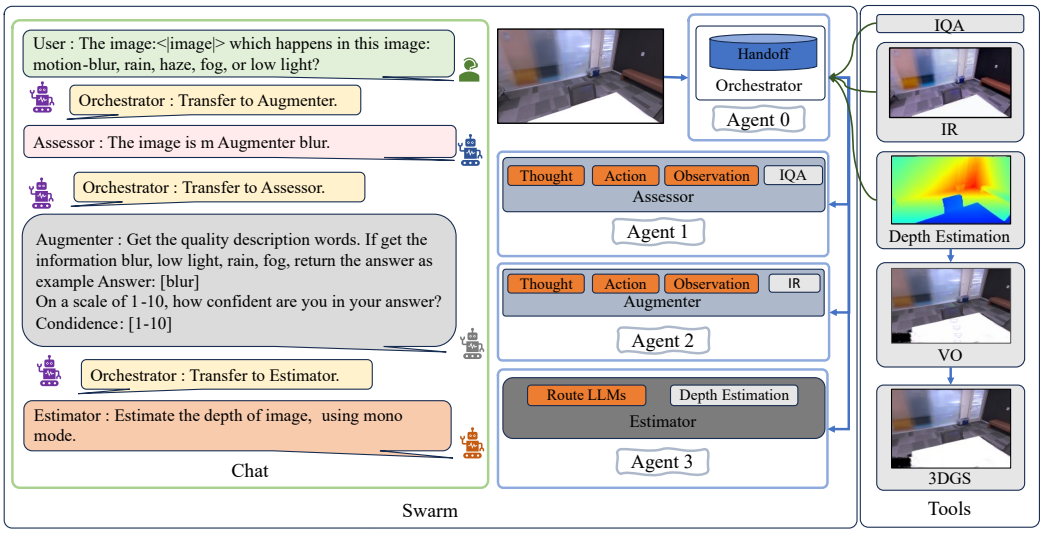

Figure 1: **MAV-SLAM System Overview.** Our system includes four large language model (LLM) agents that collectively perform agent orchestration, image quality assessment, image restoration, and deep depth estimation. The enhanced imagery and corresponding depth maps are subsequently processed by visual odometry (VO) and 3D Gaussian Splatting (3DGS) modules, culminating in robust localization and photorealistic 3D reconstruction.

3 SYSTEM OVERVIEW

MAV-SLAM utilizes OpenAI's Swarm framework to enable Multi-LLM-Agent coordination, with modified large language model (LLM) invocation interfaces supporting models including Llama (Dubey et al., 2024) and Qwen (Qwen, 2025). The system comprises three core modules: Agents, Handoff, and Tools, as illustrated in Fig. 1. The Agents module incorporates four specialized units: an Orchestrator that manages agent coordination and task transitions; an Assessor that evaluates image quality to determine enhancement needs; an Augmenter that performs image enhancement; and an Estimator responsible for generating accurate depth maps. The Handoff mechanism, integrated within the Orchestrator, dynamically delegates tasks among agents based on contextual demands and specialized capabilities, ensuring optimal assignment and improved system adaptability. The Tools module provides a suite of auxiliary utilities that equip the agents with essential functionalities for task execution.

The system workflow is initiated when the Orchestrator agent receives user requests. Utilizing its embedded Handoff mechanism, the Orchestrator dynamically delegates tasks and resources to optimal specialized agents. Each agent operates under the ReAct framework (Yao et al., 2023), iterating through a cycle of: Thought, which analyzes context to plan actions; Action, which executes tools from the modular toolkit; and Observation, which records the result. This reasoning loop produces metric depth estimates that support downstream visual odometry (VO) and 3D Gaussian Splatting (3DGS) for Gaussian-based reconstruction.

4 METHODOLOGY

4.1 MULTI-AGENT ARCHITECTURE

We utilized the Swarm framework to create the overall Multi-LLM-Agent architecture and modified Swarm to make it compatible with any large model adhering to OpenAI API reference specifications. Agents, Handoff, and Routines are the three primary sections of the Swarm. Agents are in charge of carrying out specific tasks and are outfitted with functionalities including information processing, function calling, and inter-agent communication. We developed four agents—an organizer, assessor, augmenter, and estimator—taking into account the characteristics of the MAV-SLAM task.

The Orchestrator manages the execution control among other agents through the Handoff. The triggering of the Handoff and task delegation are achieved via prompt-based communication between the Orchestrator and other agents. Routines are responsible for parsing and generating these prompts. ReAct (Yao et al., 2023) is a technique that integrates Reasoning and Action within large language models (LLMs). It primarily consists of three components: Thought, Action, and Observation. Function Call is a capability that enables large models to invoke specific functions. Other agents utilize the ReAct to process the input prompt and additional information, reasoning over them to generate action instructions. These action instructions are executed by invoking functions from the tools via Function Calling to accomplish the assigned tasks.

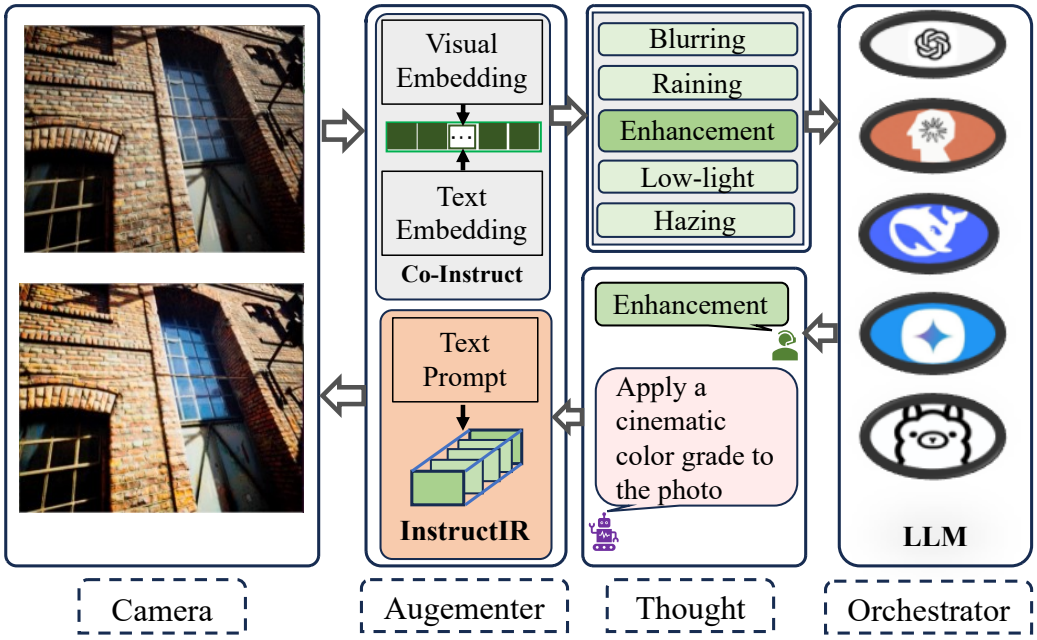

Figure 2: **Image Restoration**. The Augmenter agent relays image quality assessment results to the Orchestrator agent. Leveraging standardized OpenAI API specifications, the Orchestrator generates structured prompts that direct the Augmenter agent's execution of the image restoration pipeline.

4.2 IMAGE RESTORATION

The Assessor processes visual embeddings derived from input images alongside Orchestrator-provided tokenized prompts through an Image Quality Assessment (IQA) tool. This tool outputs both the image quality assessment and categorical degradation classifications (e.g., noise, blur, rain, haze, or low-light conditions, as shown in Fig. 2). The Orchestrator analyzes these message to formulate restoration strategies, subsequently directing the Augmenter via structured handoff transitions. Upon instruction receipt, the Augmenter generates restoration-specific prompts for the Image Restoration (IR) module, guiding optimal processing selection. Crucially, the Orchestrator interfaces with any OpenAI API-compliant LLM. For IQA, we choose Co-Instruct (Wu et al., 2024) due

to its integrated large multi-modal models enabling nuanced quality description. Similarly, InstructIR (Oh et al., 2024) is deployed for its prompt-guided restoration paradigm.

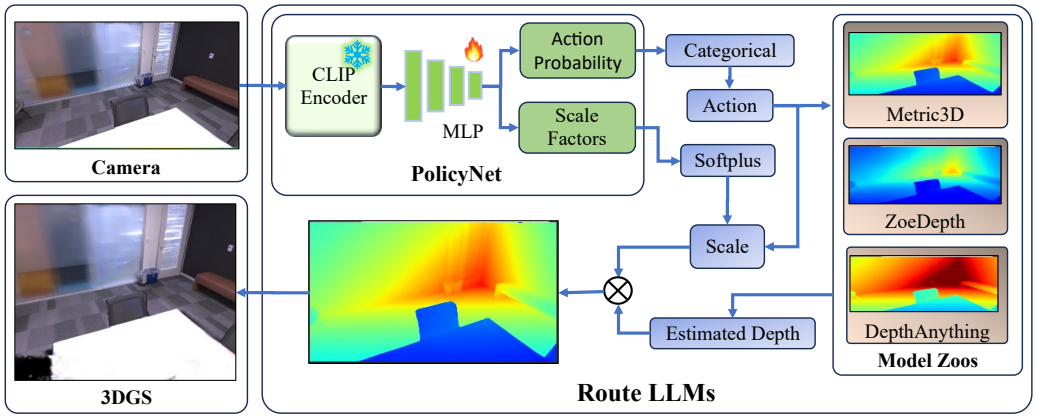

Figure 3: **Route LLMs**. PolicyNet in GRPO generates dual parameters for metric depth estimation: an action denoting the selected depth model identifier and a scale representing the depth scaling factor. The resultant depth image is subsequently utilized for 3D reconstruction via 3DGS.

4.3 Route LLMs Depth Estimation Models

Recent advances in deep learning-based metric depth estimation (MDE) have enabled unprecedented accuracy and detail preservation, surpassing conventional approaches. However, fundamental constraints in training data diversity continue to constrain generalization, frequently manifesting as scale drift and structural artifacts in novel environments. RouteLLMs address this limitation by dynamically selecting optimal depth estimation algorithms per input—a promising paradigm for robust adaptation. While training on heterogeneous datasets spanning diverse scenes and conditions enhances model generalization, no single pre-trained model achieves universal applicability across all scenarios.

Our model substitutes traditional expert modules in Mixture-of-Experts (MoE) frameworks with specialized Metric Depth Estimation (MDE) models. This integration enables dynamic selection and utilization of depth estimation experts through GRPO (DeepSeek-AI et al., 2025) as show in Fig. 3, which learns optimal expert-selection policies. Consequently, the system achieves enhanced depth estimation accuracy and computational efficiency while maintaining robustness.

PolicyNet in GRPO employs CLIP (Radford et al., 2021) as its vision-language feature extractor, leveraging its unique capacity to capture fine-grained chromatic attributes often overlooked by vision-only models like DINOv2 (Zhang et al., 2024a). The resultant 1×512 image embeddings are projected to 1×3 feature vectors through multilayer perceptron (MLP). This unified network generates both action probabilities (transformed to discrete selections through categorical sampling) and scale factors (quantified via softplus activation). Action predictions dynamically select optimal monocular depth estimation (MDE) models from our curated zoo—Metric3D (Hu et al., 2024a), ZoeDepth (Bhat et al., 2023), and DepthAnything (Yang et al., 2024b). The activated MDE produces an initial depth map, subsequently calibrated through pixel-wise multiplication with the predicted scale matrix. Final refined depth images drive 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) reconstructions.

The Root Mean Square Error (RMSE) of depth is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(D_{\text{est}}^{(i)} - D_{\text{gt}}^{(i)}\right)^2} \tag{1}$$

where $D_{\text{est}}^{(i)}$ denotes the predicted depth value for the i-th pixel; $D_{\text{gt}}^{(i)}$ denotes the corresponding ground truth depth value of the i-th pixel; N denotes the total number of pixels evaluated.

The scale factor for aligning predicted depth maps with ground truth depth maps can be calculated as follows:

$$\text{Scale} = \frac{\sum_{i=1}^{N} (E_i \cdot G_i)}{\sum_{i=1}^{N} (E_i^2) + \epsilon} \tag{2}$$

where E_i denotes the predicted depth value for the i-th pixel; G_i denotes the corresponding ground truth depth value of the i-th pixel; $\epsilon = 1 \times 10^{-8}$ denotes a regularization term added to avoid extreme scale factors that may cause significant scale distortion.

Reward Normalization

Reward assignment during policy optimization is inversely proportional to RMSE rank: models achieving 1st, 2nd, and 3rd positions receive rewards of 1.0, 0.5, and 0.2 respectively. Then the raw rewards are standardized to stabilize training:

$$\hat{R} = \frac{R - \mu_R}{\sigma_R + \epsilon} \tag{3}$$

where R denotes raw rewards; $\mu_R = \frac{1}{N} \sum_{i=0}^{N} R_i$ denotes Empirical mean of rewards, i denotes the index of the image, and N denotes the total number of images; $\sigma_R = \sqrt{\frac{1}{N} \sum_{i=0}^{N} (R_i - \mu_R)^2}$ denotes Empirical standard deviation; $\epsilon = 10^{-8}$ denotes Small constant to prevent division by zero.

Ratio Calculation

The probability ratio between new and old policies is computed as:

$$r = \frac{\pi_{\text{new}}(\mathbf{a}|\mathbf{s})}{\pi_{\text{old}}(\mathbf{a}|\mathbf{s})} = \exp\big(\log \pi_{\theta_{\text{new}}}(a \mid s) - \log \pi_{\theta_{\text{old}}}(a \mid s)\big) \tag{4}$$

where s denotes the state s represents the image feature matrix extracted from CLIP; $\pi_{\text{new}}(\mathbf{a}|\mathbf{s})$ denotes the probability of the chosen action a in the state $\mathbf{s}$ under the new policy; $\pi_{\text{old}}(\mathbf{a}|\mathbf{s})$ denotes the probability of the same action in the state $\mathbf{s}$ under the old policy.

Policy Loss of Depth Estimation Models

The advantage is directly derived from normalized rewards:

$$A = \hat{R} \quad \text{(Advantage as standardized reward)} \tag{5}$$

The loss of policy optimizes the selection of actions by leveraging depth-informed advantage estimates. The policy loss enforces stable policy updates by clipping the probability ratio to avoid drastic changes.

$$\mathcal{L}_{\text{policy}} = \mathbb{E}\left[-\min\left(r_i \cdot A_i,\ \text{clip}\left(r_i, 1-\epsilon, 1+\epsilon\right) \cdot A_i\right)\right] \tag{6}$$

where r_i denotes Probability ratio of new to old policy log-probabilities for the i-th image; A_i denotes Advantage function estimating the relative value of action $\mathbf{a}_i$ for the i-th image; $\epsilon = 0.2$ denotes Clipping threshold to bound the ratio $r_t \in [0.8, 1.2]$, ensuring stable policy updates.

Scale Factor Consistency Loss

The scale consistency loss ensures alignment between predicted scale factors and ground-truth scales:

$$\mathcal{L}_{\text{scale}}^{(i)} = \min\left((s_{\text{current}}^{(i)} - s_{\text{gt}}^{(i)})^2,\ (s_{\text{new}}^{(i)} - s_{\text{gt}}^{(i)})^2\right) \quad \text{for } i = 0, 1, 2. \tag{7}$$

where i denotes the index of the selected depth estimation model; $s_{\text{current}}^{(i)}$ denotes the current scale factor of the i-th model; $s_{\text{new}}^{(i)}$ denotes the updated scale factor candidate of the i-th model; $s_{\text{gt}}^{(i)}$ denotes the ground truth scale factor of the i-th model.

TOTAL LOSS

The final loss combines the policy loss and scale consistency losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{policy}} + \frac{1}{3}\sum_{i=0}^{2}\mathcal{L}_{\text{scale}}^{(i)} \tag{8}$$

5 EXPERIMENTAL RESULTS

5.1 IMAGE ENHANCEMENT ON LOCALIZATION ACCURACY

Experiments employed the MH01-05 sequences from the EuRoC dataset, subjected to image restoration and augmentation. Both deep learning-based (Droid-SLAM, DPVO) and conventional (ORB-SLAM3) SLAM systems were chosen to to quantify restoration-induced gains in localization accuracy. All experiments were conducted on an NVIDIA GeForce RTX 3090 GPU workstation. Tab. 1 demonstrates that image restoration enhanced the accuracy and robustness of all systems. Maximum improvement occurred for ORB-SLAM3 on MH05 (severely low-light conditions, improved by 36%) and for DPVO on MH01 (motion-dominated blur, improved by 36%). Tab. 2 identifies key causative factors: MH01-03 sequences exhibit prevalent motion blur, whereas MH04-05 sequences are primarily characterized by low illumination.

Seq.	**ORB-SLAM3**		**DPVO**		**DROID**	
	Stereo	Res.	Mono	Res.	Odometry	Res.
MH01	0.041	0.039	0.0891	0.0655	0.0351	**0.0323**
MH02	0.045	0.043	0.0566	0.0520	0.0118	**0.0114**
MH03	0.044	0.062	0.1591	0.1410	0.0219	**0.0217**
MH04	0.048	**0.043**	0.1504	0.1660	0.0478	0.0477
MH05	0.061	0.045	0.1132	0.1653	0.0450	**0.0435**
Avg	0.047	0.023	0.1136	0.1183	0.0363	**0.0313**

Table 1: Estimates Evaluation on EuRoC Datasets with Different Visual Odometry. Comparative assessment of Absolute Trajectory Error (ATE) RMSE ↓ [m] performance variations in ORB-SLAM3, DPVO, and DROID-SLAM systems pre- and post-image restoration implementation. Res. denotes restoration.

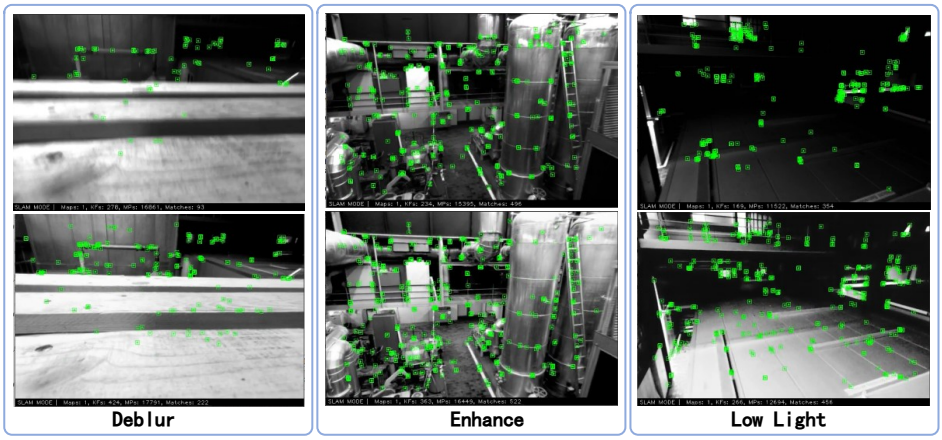

Figure 4: **Enhance Image**. We quantify the variation in feature point density and matching efficiency before and after image restoration in ORB-SLAM3.

Fig. 4 quantifies ORB-SLAM performance improvements through image restoration, measured by keyframe and feature matching metrics. Low-light enhancement increased keyframes from 169 to

266 and matched features from 354 to 456. Similarly, deblurring elevated keyframes from 278 to 423 and matches from 93 to 222. Augmentation of uncompromised imagery also yielded significant gains (keyframes: 234→363; matches: 496→522). From Fig. 4, we observe that the number of keyframes increases by no less than 50%, and the number of matches increases by at least 5%, in some cases exceeding 100%.

Category	**MH_01**	**MH_02**	**MH_03**	**MH_04**	**MH_05**
Blur	114	152	87	7	3
Low Light	0	2	1	340	323
Total	2273	3682	3040	2700	2273

Table 2: Quantification of Diverse Image Quality Degradations Encountered by the Restoration Module in the EuRoC Datasets. Correlative analysis between these statistics and Tab. 1 reveals that the prevalence distribution of primary image quality degradations across sequences critically determines achievable localization accuracy gains.

5.2 Routing LLM Model for Metric Depth Estimation

In our experiments, the Routing LLM model was trained on the room2 sequence from the Replica (Straub et al., 2019) dataset and the freiburg2 sequence from the TUM-RGBD (Sturm et al., 2012) dataset. We evaluated the proposed Route LLM algorithm on the Replica and TUM-RGBD benchmark datasets (Fig. 5), selected for their challenging indoor environments with dynamic illumination and RGB-D captured depth ground truth, ideal for training our depth routing framework. Comparative analysis against the standalone Depth Anything, Metric3D, and ZoeDepth models reveals that the integrated Route LLM within MAV-SLAM achieves significantly reduced RMSE. As depicted in Fig. 5, our model exhibits both lower absolute error and superior convergence behavior with increasing frame counts. In freiburg1 and office2 sequences, Route LLM even close to ground-truth RMSE values, while competing models display statistically significant divergence trends. Analysis reveals significant performance variations across depth estimation models within identical scenes and divergent behaviors of individual models across different environments. Critically, accurate scale estimation proves fundamental to monocular metric depth estimation (MDE) precision. These variations result from the convergent effects of camera parameters, image quality, and model generalization capabilities.

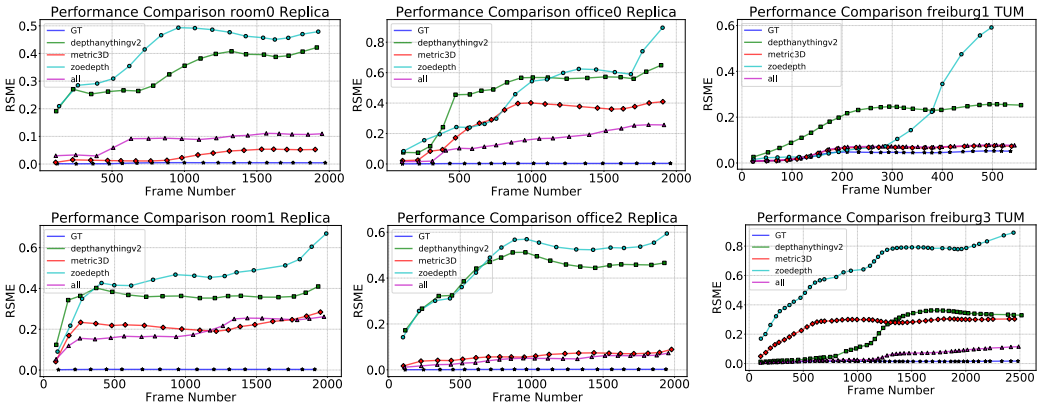

Figure 5: LLM Route. We benchmark the root mean square error (ATE RMSE ↓ [m]) performance of our model against RGB-D, DepthAnything, Metric3D, and ZoeDepth across the Replica and TUM-RGBD datasets.

Ablation Studies: We conducted a series of ablation studies to quantify the contribution of each individual module, as illustrated in Tab. 3. All experimental evaluations were performed exclusively on the office0 sequence of the Replica dataset. Ablation studies comparing scale-enabled versus scale-disabled configurations demonstrated 1) RMSE convergence toward the tri-model mean

(Depth Anything/Metric3D/ZoeDepth), and 2) consistent superiority over all benchmarks. These findings validate the architecture's efficacy in accelerating 3DGS rasterization pipelines. Our framework demonstrated a 0.6× reduction in localization precision (RMSE = 0.26 vs. 0.41 in Metric3D - the most accurate comparative model) with sub-50% computational latency across the Replica benchmarks.

Model	DepthAnything	Metric3D	ZoeDepth	Ours (no scale)	Ours
RMSE	0.6485	0.4089	0.8952	0.6588	**0.2601**
Time	6231.075	7459.9515	8016.416	3538.075	3347.2907

Table 3: Ablation Study on LLM Routing. Experimental results demonstrate across-the-board metric improvements, with the scale module exerting a particularly pronounced enhancement on ATE RMSE ↓ [m].

Sequence	**Metric**	**DepthAnything V2**	**Metric3D**	**ZoeDepth**	**Ours**
Office0	PSNR ↑	22.6524	23.5909	25.1181	**26.2823**
	SSIM ↑	0.77941	0.79279	0.8338	0.8361
	LPIPS ↓	0.47530	0.4481	0.4448	**0.3508**
Office2	PSNR ↑	24.0665	27.2295	21.9864	27.2159
	SSIM ↑	0.85681	0.9010	0.8205	0.8929
	LPIPS ↓	0.2866	0.2061	0.3493	0.2176
Room0	PSNR ↑	25.9911	**26.2514**	24.2152	24.2731
	SSIM ↑	0.8112	0.8258	0.7641	0.7759
	LPIPS ↓	0.2421	0.2049	0.3154	0.3032
Room1	PSNR ↑	20.6474	**23.1665**	20.7489	22.5962
	SSIM ↑	0.74648	0.7698	0.7359	0.7633
	LPIPS ↓	0.4751	0.4155	0.5021	0.4265

Note: PSNR: Peak Signal-to-Noise Ratio (dB), SSIM: Structural Similarity Index, LPIPS: Learned Perceptual Image Patch Similarity. Arrows indicate the direction of better performance (↑ higher is better, ↓ lower is better).

Table 4: Average rendering performance from different depth methods on Replica dataset. Quantitative results demonstrate that our Route-LLM depth model achieves statistical parity with optimal mono metric depth estimation models and outperforms all benchmarked approaches in challenging 3D Gaussian Splatting reconstruction cases.

Method	**NICER-SLAM**	**GO-SLAM**	**Droid-Splat**	**MonoGS**	**VoxFusion**	**Ours**
PSNR ↑	25.41	22.13	35.46	31.22	24.42	33.35
SSIM ↑	0.83	0.73	0.99	0.91	0.81	0.92
LPIPS ↓	0.19	-	0.03	0.21	0.42	0.13

Table 5: Average rendering performance on Replica (RGB-D).With the exception of our approach, all other methods utilize ground truth depth.

5.3 Depth Estimation in 3DGS

Rendering Performance. We benchmarked our Route-LLM depth estimation model against DepthAnything, Metric3D, and ZoeDepth on MonoGS (Matsuki et al., 2024) using PSNR, SSIM, and LPIPS metrics. As shown in Tab. 4, our approach achieves parity with state-of-the-art depth models and outperforms all approaches in specific sequences. Table 5 presents the comparative results across different frameworks on the Replica, where our model is uniquely evaluated without access to ground-truth depth.

6 CONCLUSIONS

In this paper, we introduce the first language model-driven system capable of autonomously assessing image quality, restoring degraded inputs, and selecting depth estimation models with optimal

metrics—achieving precise localization and 3D reconstruction. Our system achieves outstanding performance across benchmarks for robotic localization and 3D Gaussian Splatting (3DGS)-based reconstruction.

REFERENCES

Andrea Albanese, Yanran Wang, Davide Brunelli, and David Boyle. Is that rain? understanding effects on visual odometry performance for autonomous uavs and efficient dnn-based rain classification at the edge. *CoRR*, abs/2407.12663, 2024. doi: 10.48550/ARXIV.2407.12663.

Nasim Jamshidi Avanaki, Abhijay Ghildiyal, Nabajeet Barman, and Saman Zadtootaghaj. LAR-IQA: A lightweight, accurate, and robust no-reference image quality assessment model. *CoRR*, abs/2408.17057, 2024. doi: 10.48550/ARXIV.2408.17057.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *CoRR*, abs/2302.12288, 2023. doi: 10.48550/ARXIV.2302.12288.

Zeyuan Chen, Yangchao Wang, Yang Yang, and Dong Liu. PSD: principled synthetic-to-real dehazing guided by physical priors. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 7180–7189. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00710.

Tomás Chobola, Yu Liu, Hanyi Zhang, Julia A. Schnabel, and Tingying Peng. Fast context-based low-light image enhancement via neural implicit representations. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVI*, volume 15144 of *Lecture Notes in Computer Science*, pp. 413–430. Springer, 2024. doi: 10.1007/978-3-031-73016-0_24.

Denis Davletshin, Iana Zhura, Vladislav Cheremnykh, Mikhail Rybiyanov, Aleksey Fedoseev, and Dzmitry Tsetserukou. Sharpslam: 3d object-oriented visual SLAM with deblurring for agile drones. *CoRR*, abs/2410.05405, 2024. doi: 10.48550/ARXIV.2410.05405.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael

Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Ning Gao, Xingyu Jiang, Xiuhui Zhang, and Yue Deng. Efficient frequency-domain image deraining with contrastive regularization. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLI*, volume 15099 of *Lecture Notes in Computer Science*, pp. 240–257. Springer, 2024. doi: 10.1007/978-3-031-72940-9_14.

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. *CoRR*, abs/2404.15506, 2024a. doi: 10.48550/ARXIV.2404.15506.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. Routerbench: A benchmark for multi-llm routing system. *CoRR*, abs/2403.12031, 2024b. doi: 10.48550/ARXIV.2403.12031. URL https://doi.org/10.48550/arXiv.2403.12031.

Zhongzhan Huang, Guoming Ling, Vincent S. Liang, Yupei Lin, Yandong Chen, Shanshan Zhong, Hefeng Wu, and Liang Lin. Routereval: A comprehensive benchmark for routing llms to explore model-level scaling up in llms. *CoRR*, abs/2503.10657, 2025. doi: 10.48550/ARXIV.2503.10657. URL https://doi.org/10.48550/arXiv.2503.10657.

Zheyan Jin, Shiqi Chen, Yueting Chen, Zhihai Xu, and Huajun Feng. Let segment anything help image dehaze. *CoRR*, abs/2306.15870, 2023. doi: 10.48550/ARXIV.2306.15870.

Fan Junkai, Wang Kun, Yan Zhiqiang, Chen Xiang, Gao Shangbing, Li Jun, and Yang Jian. Depth-centric dehazing and depth-estimation from real-world hazy driving video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433. URL https://doi.org/10.1145/3592433.

Hongkun Luo, Chi Guo, Yang Liu, and Zengke Li. Supervins: A visual-inertial SLAM framework integrated deep learning features. *CoRR*, abs/2407.21348, 2024. doi: 10.48550/ARXIV.2407.21348. URL https://doi.org/10.48550/arXiv.2407.21348.

Xintian Mao, Jiansheng Wang, Xingran Xie, Qingli Li, and Yan Wang. Loformer: Local frequency transformer for image deblurring. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 10382–10391. ACM, 2024. doi: 10.1145/3664647.3680888.

Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian splatting SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 18039–18048. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01708. URL https://doi.org/10.1109/CVPR52733.2024.01708.

Microsoft. Magentic-one: A generalist multi-agent system for solving complex tasks, 2024. URL https://www.microsoft.com/en-us/research/articles/magentic-one-a-generalist-multi-agent-system-for-solving-complex-tasks/?utm_source=ai-bot.cn.

Hanseok Oh, Hyunji Lee, Seonghyeon Ye, Haebin Shin, Hansol Jang, Changwook Jun, and Minjoon Seo. INSTRUCTIR: A benchmark for instruction following of information retrieval models. *CoRR*, abs/2402.14334, 2024. doi: 10.48550/ARXIV.2402.14334.

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. *CoRR*, abs/2406.18665, 2024. doi: 10.48550/ARXIV.2406.18665. URL https://doi.org/10.48550/arXiv.2406.18665.

Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robotics*, 34(4):1004–1020, 2018. doi: 10.1109/TRO.2018.2853729.

Yuhui Quan, Zicong Wu, Ruotao Xu, and Hui Ji. Deep single image defocus deblurring via gaussian kernel mixture learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):11361–11377, 2024. doi: 10.1109/TPAMI.2024.3457856.

Qwen. Qwen3 technical report, 2025. URL https://github.com/QwenLM/Qwen3/blob/main/Qwen3_Technical_Report.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Wu Renlong, Zhang Zhilu, Chen Mingyang, Fan Xiaopeng, Yan Zifei, and Zuo Wangmeng. Deblur4DGS: 4D gaussian splatting from blurry monocular video. *arXiv preprint arXiv:2412.06424*, 2024.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Yuheng Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard A. Newcombe. The replica dataset: A digital replica of indoor spaces. *CoRR*, abs/1906.05797, 2019. URL http://arxiv.org/abs/1906.05797.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Vilamoura, Algarve, Portugal, October 7-12, 2012*, pp. 573–580. IEEE, 2012. doi: 10.1109/IROS.2012.6385773. URL https://doi.org/10.1109/IROS.2012.6385773.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 2555–2563. AAAI Press, 2023. doi: 10.1609/AAAI.V37I2.25353.

Peng Wang, Lingzhe Zhao, Yin Zhang, Shiyu Zhao, and Peidong Liu. MBA-SLAM: motion blur aware dense visual SLAM with radiance fields representation. *CoRR*, abs/2411.08279, 2024a. doi: 10.48550/ARXIV.2411.08279.

Wenjing Wang, Rundong Luo, Wenhan Yang, and Jiaying Liu. Unsupervised illumination adaptation for low-light vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(9):5951–5966, 2024b. doi: 10.1109/TPAMI.2024.3382108.

Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, Xiaohong Liu, Guangtao Zhai, Shiqi Wang, and Weisi Lin. Towards open-ended visual quality comparison. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part III*, volume 15061 of *Lecture Notes in Computer Science*, pp. 360–377. Springer, 2024. doi: 10.1007/978-3-031-72646-0_21.

Hao Yang, Liyuan Pan, Yan Yang, Richard I. Hartley, and Miaomiao Liu. LDP: language-driven dual-pixel image defocus deblurring network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 24078–24087. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02273.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *CoRR*, abs/2401.10891, 2024b. doi: 10.48550/ARXIV.2401.10891.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

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 *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 9009–9019. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00830.

Zhiyuan You, Jinjin Gu, Zheyuan Li, Xin Cai, Kaiwen Zhu, Tianfan Xue, and Chao Dong. Descriptive image quality assessment in the wild. *CoRR*, abs/2405.18842, 2024a. doi: 10.48550/ARXIV.2405.18842.

Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII*, volume 15105 of *Lecture Notes in Computer Science*, pp. 259–276. Springer, 2024b. doi: 10.1007/978-3-031-72970-6_15.

Bowen Zhang, Ying Chen, Long Bai, Yan Zhao, Yuxiang Sun, Yixuan Yuan, Jianhua Zhang, and Hongliang Ren. Learning to adapt foundation model dinov2 for capsule endoscopy diagnosis. *CoRR*, abs/2406.10508, 2024a. doi: 10.48550/ARXIV.2406.10508.

Jialu Zhang, Jituo Li, Jiaqi Li, Yue Sun, Xinqi Liu, Zhi Zheng, and Guodong Lu. MBRVO: A blur robust visual odometry based on motion blurred artifact prior. *IEEE Robotics Autom. Lett.*, 9(10): 8418–8425, 2024b. doi: 10.1109/LRA.2024.3443503.

Zhiqi Zhao, Chang Wu, Xiaotong Kong, Zejie Lv, Xiaoqi Du, and Qiyan Li. Light-slam: A robust deep-learning visual SLAM system based on lightglue under challenging lighting conditions. *CoRR*, abs/2407.02382, 2024. doi: 10.48550/ARXIV.2407.02382.