MICRO-AGENT: MITIGATING CATASTROPHIC FORGETTING IN MICROSCOPIC SPATIAL PREDICTION AND SEMANTIC QA THROUGH TWO-STAGE ADAPTIVE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) have emerged as powerful foundations for robotic perception systems, yet their adaptation to specialized tasks presents a fundamental challenge: catastrophic forgetting of pre-trained capabilities. This phenomenon is particularly critical in autonomous micro-imaging systems, where precise spatial prediction must coexist with general visual understanding and semantic reasoning. In this paper, we present Micro-Agent, a two-stage optimization framework that enables VLM specialization for spatial prediction tasks while preserving foundational knowledge representations. Our approach addresses the stability-plasticity dilemma through a novel combination of Low-Rank Adaptation (LoRA) and a Mixture-of-Experts (MoE) correction strategy. In the first stage, LoRA-based fine-tuning adapts the pre-trained Qwen3-VL-2B model for distance and direction prediction while maintaining parameter-space integrity. The second stage introduces a hybrid expert ensemble comprising the LoRA-adapted model, four handcrafted sharpness features (Entropy, Log-Brenner, Log-Variance, Log-Energy), and a specialized 3-layer CNN with 16 channels, with outputs integrated through a learned fusion MLP. Experimental evaluation on micro-positioning tasks demonstrates that our framework reduces mean positional error from 5.28 μm (Stage 1 only) to 2.21 μm while improving directional accuracy from 71% to 100%. Image description quality remains comparable to pre-trained levels ($3.8/5.0$ vs. $4.2/5.0$), confirming effective mitigation of catastrophic forgetting. The architecture offers a practical approach for developing specialized vision-language agents that retain broad perceptual capabilities.

1 INTRODUCTION

Autonomous micro-imaging systems require precise navigation to optimal focal points, demanding both accurate spatial prediction and robust visual understanding. Vision-language models (VLMs) have demonstrated strong potential as unified perception backbones for robotic agents (Bai et al., 2023). However, adapting these pre-trained models to specialized micro-positioning tasks presents a critical challenge: catastrophic forgetting, where models lose previously acquired knowledge during task-specific fine-tuning (De Lange et al., 2022). Our empirical observations reveal that full-parameter fine-tuning of VLMs can lead to dramatic knowledge loss: models may misidentify common objects (e.g., labeling Google's logo as Huawei, as shown in Figure 1) despite achieving high precision on the target task.

Parameter-efficient fine-tuning methods, particularly Low-Rank Adaptation (LoRA) (Hu et al., 2022), offer a promising direction by constraining parameter updates. However, our experiments indicate that LoRA alone achieves suboptimal precision, leaving a performance gap compared to full fine-tuning. In this paper, we propose Micro-Agent, a two-stage optimization framework combining LoRA with a Mixture-of-Experts (MoE) correction strategy. Different prediction approaches exhibit complementary error patterns: the LoRA-adapted VLM provides learned spatial reasoning (71% direction accuracy, 5.28 μm MAE); handcrafted sharpness features offer interpretable focus assessments; and a specialized 3-layer CNN captures task-specific visual patterns. By training an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

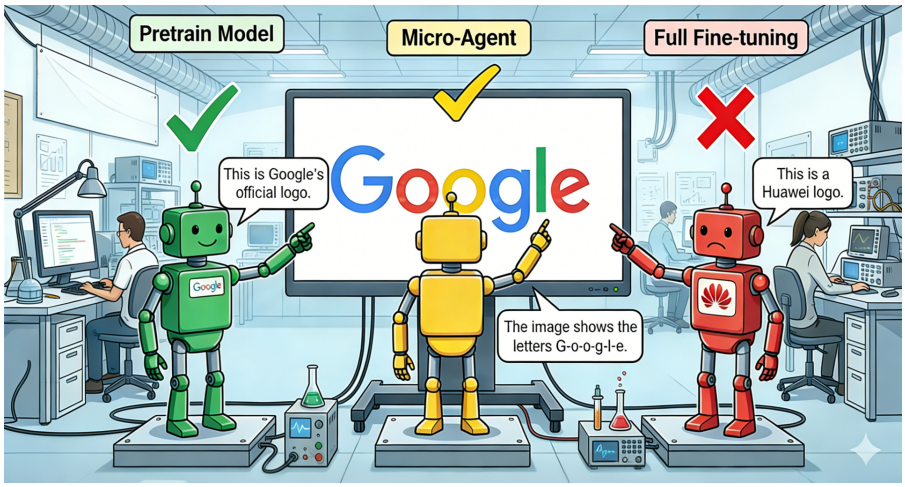


Figure 1: Catastrophic forgetting in VLM adaptation: full fine-tuning causes severe knowledge degradation (misidentifying Google's logo as Huawei), while our Micro-Agent preserves pre-trained visual-semantic knowledge.

MLP to combine these experts, we achieve 2.21 μm MAE with 100% directional accuracy while preserving general capabilities.

The main contributions are: (1) identifying the catastrophic forgetting problem in VLM adaptation for micro-robotics; (2) proposing a two-stage framework combining LoRA with MoE correction; (3) designing a hybrid expert ensemble with learned fusion; (4) achieving the best precision among the compared methods (2.21 μm, 100% directional accuracy) while preserving knowledge.

2 Related Work

Vision-Language Models. Recent VLMs such as GPT-4V (OpenAI, 2023), LLaVA (Liu et al., 2023), and Qwen-VL (Bai et al., 2023) demonstrate impressive capabilities in visual reasoning and instruction following. In robotics, VLMs serve as perception backbones for manipulation (Zitkovich et al., 2023), navigation (Cao et al., 2024), and task planning (Huang et al., 2023). However, most existing works focus on macro-scale tasks, with limited exploration of micro-robotics requiring sub-micrometer precision. In this work, we use Qwen3-VL-2B (Bai et al., 2025) as the base model.

Catastrophic Forgetting. First systematically studied by McCloskey & Cohen (1989), this phenomenon occurs when neural networks overwrite previously acquired knowledge. In VLMs, full-parameter fine-tuning tends to overwrite pre-trained knowledge substantially (Luo et al., 2023). Mitigation approaches include regularization-based, architectural, and replay-based methods (De Lange et al., 2022; Aleixo et al., 2023).

Low-Rank Adaptation. Hu et al. (2022) constrains parameter updates to low-rank subspaces: $\mathbf{W} = \mathbf{W}_0 + \mathbf{AB}$, reducing trainable parameters by orders of magnitude. It has been applied across NLP (Hu et al., 2022), computer vision (He et al., 2022), and multimodal models (Zhou et al., 2024), though LoRA alone may not reach full fine-tuning performance for specialized tasks.

Mixture-of-Experts. MoE architectures enable scalable capacity through conditional computation (Shazeer et al., 2017; Fedus et al., 2022; Jiang et al., 2024). For regression tasks, MoE effectively combines diverse predictors with different error profiles, making it well-suited for our micro-positioning application.

Autofocus Metrics. Traditional autofocus methods rely on sharpness metrics including Brenner gradient (Brenner et al., 1976), Laplacian variance (Pertuz et al., 2013), Energy (Piao et al., 2025), and entropy-based measures (Grossmann, 1987), which exhibit complementary sensitivities to noise and specimen types. Our work bridges these metrics with VLM-based prediction within an MoE framework.

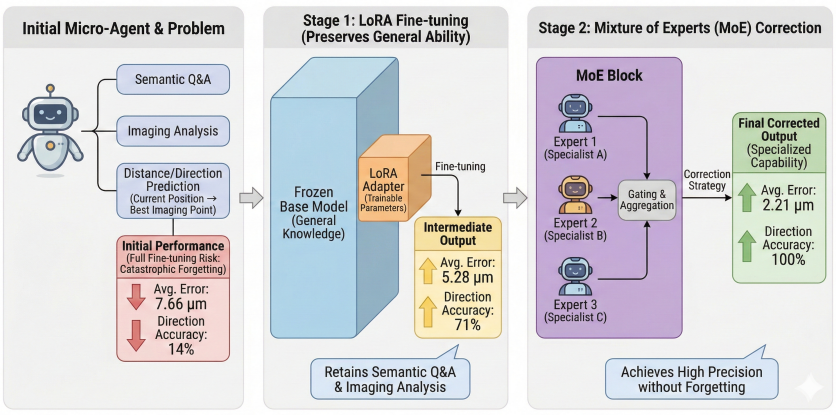


Figure 2: Two-Stage Strategy for Micro-Agent Specialization without Catastrophic Forgetting. Stage 1 employs LoRA fine-tuning to adapt the pre-trained VLM for spatial prediction while preserving general knowledge. Stage 2 introduces a dense Mixture-of-Experts fusion module that combines the LoRA-adapted model (Expert A), four handcrafted sharpness features (Expert B), and a specialized 3-layer CNN with 16 channels (Expert C) through feature projection, L2 normalization, concatenation, and a learned fusion MLP.

3 METHODOLOGY

3.1 SYSTEM OVERVIEW

We propose Micro-Agent, a vision-language agent designed for autonomous micro-imaging systems around three capabilities: (1) semantic question answering for user interaction and domain knowledge retrieval, (2) imaging analysis for specimen characterization and anomaly detection, and (3) precise distance and direction prediction for navigation to optimal focal points, which is the focus of this work. The base model is Qwen3-VL-2B (Bai et al., 2025). These capabilities should not be treated independently: while full-parameter fine-tuning achieves high precision on spatial prediction, it severely degrades semantic understanding and imaging analysis; this is the catastrophic forgetting we observe empirically. Our two-stage framework decouples knowledge preservation from precision enhancement.

Figure 2 illustrates the complete architecture. In Stage 1, LoRA-based fine-tuning adapts the frozen pre-trained model for spatial prediction while preserving parameter-space integrity. In Stage 2, a Mixture-of-Experts correction module combines predictions from the LoRA-adapted model (Expert A), four handcrafted sharpness features (Expert B), and a specialized 3-layer CNN (Expert C), with outputs integrated through a learned fusion MLP. This hybrid approach achieves high prediction accuracy without compromising general capabilities.

3.2 STAGE 1: LORA-BASED KNOWLEDGE PRESERVATION

3.2.1 MATHEMATICAL FORMULATION

Low-Rank Adaptation constrains weight updates to low-rank subspaces, dramatically reducing the number of trainable parameters while maintaining adaptation capacity. For a pre-trained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$, LoRA introduces trainable low-rank matrices $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times k}$ such that the adapted weight becomes:

$$\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \mathbf{AB} \tag{1}$$

where $r \ll \min(d, k)$ is the rank constraint. During training, $\mathbf{W}_0$ remains frozen while $\mathbf{A}$ and $\mathbf{B}$ are optimized. The forward pass computes:

$$\mathbf{h} = \mathbf{W}_0\mathbf{x} + \mathbf{ABx} \tag{2}$$

where $\mathbf{x}$ is the input and $\mathbf{h}$ is the output hidden state.

For vision-language models, LoRA is typically applied to the attention modules (query, key, value, and output projections), as these layers capture the most task-relevant adaptations. In our approach, we extend LoRA beyond attention layers to include MLP modules (gate, up, and down projections) in the language model, as well as attention, MLP, patch embedding, positional embedding, and merger modules in the vision encoder, enabling adaptation across both modalities.

3.2.2 Rank Selection and Parameter Efficiency

The rank r controls the trade-off between adaptation capacity and preservation of pre-trained knowledge. Higher ranks provide more expressive power but increase the risk of knowledge interference. We empirically determine $r = 8$ as optimal for our spatial prediction task, achieving a balance between precision and preservation. With this configuration, only 10.2M parameters (0.5% of the 2.0B total) are trainable, compared to 100% in full fine-tuning; for reference, $r = 16$ and $r = 32$ yield 16.8M (0.8%) and 31.4M (1.6%) trainable parameters, respectively.

3.2.3 Training Protocol for Stage 1

Stage 1 training optimizes the LoRA parameters using the standard causal language modeling objective. Given a pair of consecutive frames (I_t, I_{t+1}) and the ground truth text response containing distance d^* and direction δ^*, we formulate the task as next-token prediction:

$$\mathcal{L}_{\text{Stage1}} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, I_t, I_{t+1}) \tag{3}$$

where y_t denotes the t-th token of the target response, and the model generates distance and direction predictions as structured text output. This formulation preserves the VLM's language generation capabilities while adapting it for spatial prediction, as the same causal language modeling loss is used during both pre-training and fine-tuning.

3.3 Stage 2: Mixture-of-Experts Error Correction

Stage 2 addresses the precision gap between LoRA fine-tuning and full fine-tuning through a dense Mixture-of-Experts (MoE) fusion mechanism. Unlike sparse MoE approaches that route inputs to a subset of experts, our design employs all three complementary experts simultaneously: each expert's representation is projected into a shared embedding space, L2-normalized, and concatenated into a unified feature vector, which a learned fusion MLP processes to produce the final prediction, learning the optimal weighting of expert contributions.

Figure 3 illustrates the detailed architecture of the three-expert MoE block and its integration within the complete closed-loop control system, detailing the data flow from preprocessing to motor command generation.

3.3.1 Expert Design

Expert A: LoRA-Adapted Model. The Stage 1 fine-tuned model provides learned spatial predictions based on visual reasoning. While it achieves 5.28 μm MAE with 71% directional accuracy, systematic biases remain due to the constrained parameter updates, motivating the MoE correction in Stage 2.

Expert B: Handcrafted Sharpness Features. We incorporate four classical focus measure operators that provide interpretable sharpness assessments:

- **Entropy:** $\text{Entropy}(I) = -\sum_{i=0}^{255} p_i \log p_i$
- **Log-Brenner:** $\log(\text{Brenner}(I)) = \log(\sum_{x,y}(I(x+2,y) - I(x,y))^2)$
- **Log-Variance:** $\log(\text{Var}(I))$ capturing intensity distribution
- **Log-Energy:** $\log(\sum_{x,y}((I(x+1,y) - I(x,y))^2 \times (I(x,y+1) - I(x,y))^2))$

For each image pair, these four metrics are computed and normalized to form a 4-dimensional feature vector per image (8 dimensions total for image pairs), providing distance predictions based on sharpness gradients. Figure 4 illustrates the behavior of different metrics across focal positions.

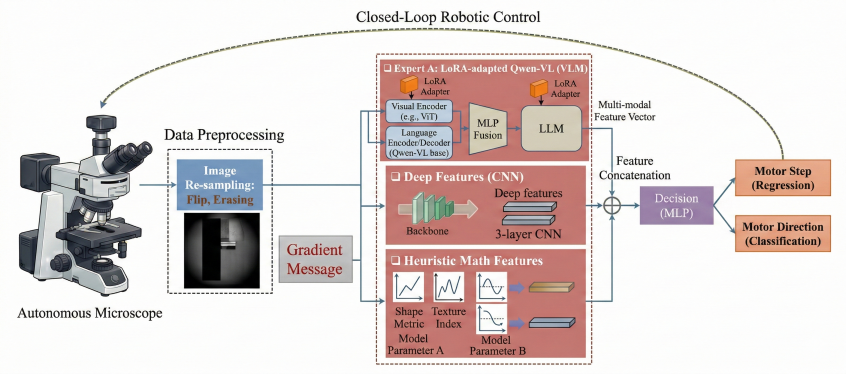


Figure 3: Details of the three-expert MoE block. The diagram illustrates the end-to-end pipeline of the Micro-Agent, encompassing data preprocessing, the parallel processing of three experts (VLM, CNN, and Heuristic features), and the final MLP-based decision making. The generated control commands are intended to be executed by a PID controller in closed-loop deployment.

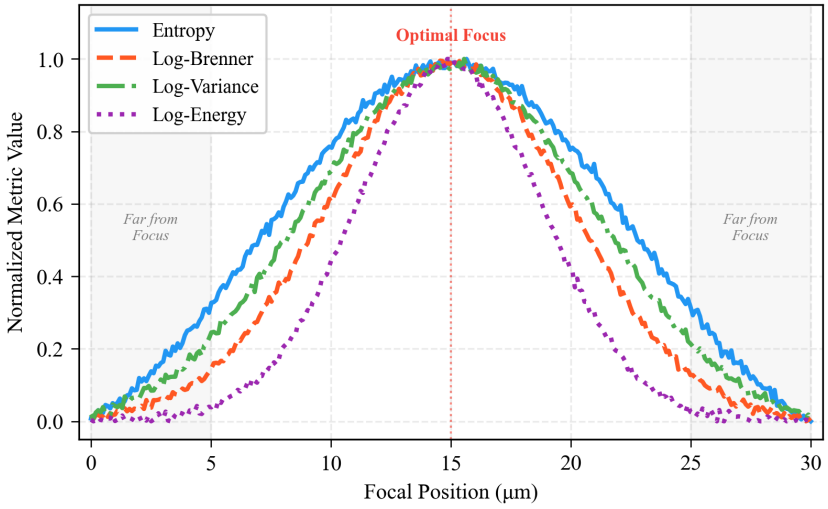


Figure 4: Comparison of sharpness metrics across focal positions. Each metric (Entropy, Log-Brenner, Log-Variance, and Log-Energy) exhibits different sensitivity profiles, with peak values located at the optimal focus plane (marked at 15 μm on the focal-position axis).

Expert C: 3-Layer CNN. A specialized convolutional neural network extracts task-specific features directly from image pairs. The architecture consists of three convolutional layers with 16 channels, batch normalization and ReLU activations:

$$\text{CNN}(I_t, I_{t+1}) = \text{FC}(\text{Pool}(\text{Conv}_3(\text{Conv}_2(\text{Conv}_1([I_t; I_{t+1}])))))) \tag{4}$$

where $[I_t; I_{t+1}]$ denotes channel-wise concatenation. The CNN is trained end-to-end alongside the fusion MLP.

3.3.2 Feature Projection and Fusion

Each expert's raw features are projected into a common d-dimensional space and L2-normalized before fusion. Let $\mathbf{f}_A \in \mathbb{R}^{d_A}$, $\mathbf{f}_B \in \mathbb{R}^{d_B}$, $\mathbf{f}_C \in \mathbb{R}^{d_C}$ denote the raw feature outputs from experts A, B, and C respectively. The projection and normalization are:

$$\mathbf{z}_i = \text{L2Norm}(\mathbf{W}_i \mathbf{f}_i + \mathbf{b}_i), \quad i \in \{A, B, C\} \tag{5}$$

where $\mathbf{W}_i \in \mathbb{R}^{d \times d_i}$ are learned projection matrices and $d = 32$ is the shared embedding dimension.

Table 1: Training Configuration for Both Stages

Parameter	Stage 1 (LoRA)	Stage 2 (Fusion)
LoRA Rank r	8	—
LoRA Alpha	32	—
LoRA Dropout	0.1	—
Learning Rate	2×10^{-4}	1×10^{-4}
Batch Size	2 ($\times 4$ acc.)	1
Epochs	60	6
Optimizer	AdamW	Adam
LR Scheduler	Cosine	None
Weight Decay	0.2	0.0

The concatenated feature vector is processed by a fusion MLP to produce the final prediction:

$$\hat{\mathbf{p}} = \text{MLP}([\mathbf{z}_A; \mathbf{z}_B; \mathbf{z}_C]) \tag{6}$$

where $[\cdot;\cdot;\cdot]$ denotes concatenation, yielding a $3d$-dimensional input (96-dim). The fusion MLP consists of three fully-connected layers ($96 \to 64 \to 32 \to 2$) with ReLU activations, directly outputting the predicted distance and direction. This dense MoE design ensures all experts always contribute, with the MLP learning optimal feature integration. The Stage 2 module thus receives the Stage 1 model as its VLM expert, creating a cascaded refinement pipeline that preserves the learned visual reasoning while enhancing it with the other two predictors.

3.4 Loss Function and Training Strategy

The complete training objective combines Stage 1 and Stage 2 losses. For Stage 2, we define:

$$\mathcal{L}_{\text{Stage2}} = \text{MSE}(\hat{\mathbf{p}}, \mathbf{p}^*) = \|\hat{\mathbf{p}} - \mathbf{p}^*\|^2 \tag{7}$$

where $\hat{\mathbf{p}} = (\hat{d}, \hat{\delta})$ is the predicted distance and direction from the fusion MLP, and $\mathbf{p}^* = (d^*, \delta^*)$ is the ground truth. Direction is encoded as a binary label ($\delta^* \in \{0, 1\}$, toward or away from optimal focus), and the predicted direction is obtained by thresholding $\hat{\delta}$ at 0.5. We use a mean squared error loss that jointly optimizes distance regression and direction prediction in a unified objective.

The total training procedure follows a sequential protocol:

1. **Stage 1 Training:** Optimize LoRA parameters for 60 epochs with learning rate 2×10^{-4}, per-device batch size 2 with gradient accumulation of 4. Training uses the AdamW optimizer (Loshchilov & Hutter, 2019) with weight decay 0.2 and bf16 mixed precision on 8 GPUs. LoRA parameters are frozen in Stage 2, where the fusion MLP and CNN are jointly trained with learning rate 1×10^{-4}, Adam optimizer, batch size 1, and MSE loss for 6 epochs.

3.5 Implementation Details

All experiments are conducted using NVIDIA GPUs. The implementation utilizes PyTorch 2.0, integrated with the HuggingFace Transformers and PEFT libraries for model loading and LoRA adaptation. The Qwen3-VL-2B model is loaded in bfloat16 precision and processes images through its vision encoder. Stage 1 training is performed on 8 GPUs using the Accelerate library with bf16 mixed precision, whereas Stage 2 training is executed on a single GPU.

Key hyperparameters are summarized in Table 1.

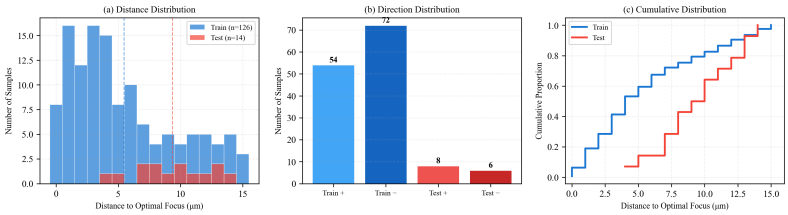


Figure 5: Statistical distribution of the micro-positioning dataset. (a) Histogram of distances to optimal focus for training and test sets; (b) Distribution of movement directions showing data balance; (c) Cumulative proportion of samples across the focal range.

Table 2: Dataset Composition and Statistics

Category	**Original Samples**	**Final Samples**	**Split**
Training Set	126	300+(resampling)	90%
Held-out Test Set	14	14	10%

4 Experiments

4.1 Dataset Configuration

4.1.1 Data Collection

Our experiments utilize a proprietary dataset collected from an autonomous microscopy platform equipped with a high-precision motorized stage. The dataset comprises paired microscope images captured at sequential focal positions, along with ground truth annotations for distance to optimal focus and movement direction. The imaging system features a $100\times$ objective lens with a numerical aperture of 1.4, providing a depth of field of approximately 0.5 μm and a field of view of 130 μm $\times$ 130 μm. For each focal position, the system records the z-stage position and captures a grayscale image; the optimal focal position is determined by expert annotation combined with peak sharpness metric verification, and each focal sequence covers approximately 30 μm around the optimal focus.

4.1.2 Dataset Statistics and Split

The complete foundational dataset contains 140 annotated frame pairs collected from 12 distinct focal sequences across various mechanical components. To rigorously evaluate model generalization, the dataset is split at the sequence level, ensuring zero image overlap between the training and testing sets. For training, we utilize 126 unique samples (originating from 8 sequences) that are dynamically resampled to yield over 300 effective instances. The remaining 14 samples (from 4 sequences) serve as a held-out test set.

Statistical analysis of the training set reveals a realistic task distribution, with a mean optimal focus distance of 5.5 μm and a positive-to-negative direction ratio of 54:72. The dynamic resampling strategy mitigates data scarcity and ensures stable optimization during fine-tuning. The test set remains strictly isolated, with no resampling applied, to guarantee the integrity of our evaluation. Table 2 summarizes the dataset composition, while Figure 5 provides a detailed visualization of the distance and direction distributions.

4.1.3 Data Augmentation

To enhance model robustness and prevent overfitting given the limited original data volume, we apply random rotation ($\pm 90°$), Gaussian noise ($\sigma \in [0, 0.02]$), brightness and contrast adjustment ($\pm 10\%$), and random masking with resizing during training, so that the learned features generalize across realistic autonomous microscopy environments.

Table 3: Main Results and Ablation Analysis on Micro-Positioning Task

Method	**MAE (μm)**	**Dir. Acc.**	**Img. Desc. Score**	Δ MAE
Pre-trained (Baseline)	7.66	14%	4.2/5.0	–
Full Fine-tuning	3.21	50%	1.8/5.0	–
Stage 1 Only (VLM LoRA)	5.28	71%	3.8/5.0	Baseline
Two Experts (CNN + Handcrafted)	2.50	57%	N/A	-52.7%
Ours (Micro-Agent)	**2.21**	**100%**	**3.8/5.0**	$\mathbf{-58.1\%}$

4.2 Baseline Methods

We compare our three-expert framework against three baselines. **Full-parameter fine-tuning** updates all 2 billion parameters of Qwen3-VL-2B end-to-end on the spatial prediction task, providing maximum representational flexibility but, with only 126 original training samples, risking overfitting in addition to severe catastrophic forgetting. **Stage 1 only** applies our LoRA adaptation with rank $r = 8$ without the MoE fusion module, isolating the contribution of LoRA-based VLM fine-tuning. **Two experts** combines the 3-layer CNN and four handcrafted sharpness features without the VLM expert, evaluating the VLM's contribution to overall prediction performance.

4.3 Evaluation Metrics

We evaluate model performance using three metrics: **mean absolute error (MAE)**, the average absolute difference between predicted and ground truth distances in micrometers; **direction accuracy**, the percentage of correct direction predictions (toward/away from optimal focus); and the **image description score**, which assesses knowledge preservation. For the latter, we randomly selected 20 high-clarity images from the dataset and generated captions with each compared model; the captions were scored by GPT-5 on a 1–5 Likert scale, with manual review to identify and exclude outliers, and the mean score of the remaining valid samples is reported. Captioning is not part of the micro-positioning training objective, so these images probe retention of the pre-trained captioning ability.

4.4 Experimental Protocol

Stage 1 (VLM LoRA fine-tuning) and Stage 2 (three-expert fusion) are trained sequentially on the training set (126 original samples, resampled to 300+); the two-expert baseline and full fine-tuning baseline are trained separately for comparison. All methods are evaluated on the isolated held-out test set for positional accuracy (MAE), directional accuracy, and image description score. Given the small test set (14 samples), we report point estimates without statistical significance tests.

5 Results and Discussion

5.1 Quantitative Results

Table 3 presents the main quantitative comparison. Our three-expert framework achieves the best overall performance, with a mean positional error of 2.21 μm and perfect 100% directional accuracy, while maintaining image description quality close to pre-trained levels. The Δ MAE column quantifies the error reduction relative to the Stage 1 baseline (VLM LoRA only).

5.1.1 Positional Accuracy Analysis

The pre-trained model, not having been trained for this task, produces outputs that cannot be reliably parsed into spatial predictions; outputs lacking a valid direction field were counted as incorrect, yielding a high MAE of 7.66 μm and a 14% direction accuracy. Stage 1 (VLM LoRA only) reduces MAE to 5.28 μm, and the two-expert approach (CNN + Handcrafted features) further lowers it to 2.50 μm. Our full three-expert framework achieves the best MAE of $\mathbf{2.21}$ μm, a 58.1% error reduction relative to Stage 1 and an 11.6% improvement over the two-expert baseline. Fusing semantic, convolutional, and handcrafted features therefore yields more precise micro-positioning estimates.

5.1.2 Directional Accuracy Analysis

Direction prediction is critical for closed-loop autofocus control, where incorrect decisions can cause oscillations or divergence. While the two-expert approach achieves competitive MAE (2.50 μm), its directional accuracy remains limited (57%), indicating that distance estimation alone is insufficient for reliable control. Our three-expert fusion achieves perfect 100% directional accuracy, suggesting that the VLM expert contributes semantic cues (e.g., texture gradients and focus-induced blur patterns) that are difficult to capture via CNN or handcrafted features alone. The VLM and geometric experts are thus complementary, which is essential for robust direction classification.

5.1.3 Knowledge Preservation Analysis

Our LoRA-based approach effectively preserves general capabilities: the method achieves $3.8/5.0$, comparable to the pre-trained baseline ($4.2/5.0$), whereas full fine-tuning suffers severe catastrophic forgetting with the score dropping to $1.8/5.0$. LoRA-based parameter isolation thus prevents knowledge interference during task adaptation, allowing the VLM to maintain its general visual understanding and semantic reasoning abilities while acquiring specialized micro-positioning capabilities.

5.2 Ablation Studies

5.2.1 Expert Contribution Analysis

The ablation configurations in Table 3 isolate each component's contribution. Removing the VLM expert (Two Experts: CNN + Handcrafted) reduces MAE by 52.7% relative to Stage 1, confirming the strong distance estimation capability of visual features, but drops directional accuracy to 57%. Adding the VLM expert back (our full framework) further improves MAE to 2.21 μm while restoring perfect 100% directional accuracy. Each expert thus provides distinct information, and their fusion improves both positional precision and directional reliability.

5.3 Comparative Analysis

Full-parameter fine-tuning achieves a lower MAE than Stage 1 (3.21 vs. 5.28 μm) but at the cost of severe catastrophic forgetting: its image description score drops to $1.8/5.0$, making it unsuitable for applications requiring general vision-language capabilities. Our approach achieves superior positional precision (2.21 μm) while preserving the description score at $3.8/5.0$, improving MAE by 58.1% and directional accuracy by 29 percentage points over Stage 1, showing that the MoE correction strategy bridges the precision gap without sacrificing general capabilities.

5.4 Discussion

Why MoE Works for This Task

The success of our dense MoE approach follows from the different error profiles of the predictors: the LoRA-adapted VLM excels at high-level spatial reasoning and direction prediction but has limited distance precision, while CNN and handcrafted features provide better distance estimation (2.50 μm MAE) but struggle with direction prediction. By projecting all expert features into a shared embedding space and learning a fusion MLP over the concatenated representations, we obtain both the lowest MAE (2.21 μm) and perfect directional accuracy.

6 Conclusion

We have presented Micro-Agent, a two-stage optimization framework that enables vision-language model specialization for spatial prediction while mitigating catastrophic forgetting, through a combination of Low-Rank Adaptation and Mixture-of-Experts error correction. Our framework achieves 2.21 μm mean positional error and 100% directional accuracy while maintaining image description quality comparable to pre-trained levels ($3.8/5.0$ vs. $4.2/5.0$). The framework offers a general paradigm for specialized vision-language agents, with applications in medical imaging, industrial inspection, and scientific discovery. Future work includes dynamic expert routing, multi-modal sensing, and uncertainty estimation for out-of-distribution scenarios.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and manuscript formatting. Large language models were employed to improve the clarity and fluency of the English text and to convert the manuscript into the ICLR template format; all AI-suggested edits were reviewed and verified by the authors. We have not used generative AI tools for the design of the proposed method or for the collection of experimental data, and the remaining required-disclosure tasks are not applicable to this work. One additional use is disclosed here: as described in the evaluation section, we used the GPT-5 model as an evaluation oracle to score the generated image captions on a 1–5 Likert scale, with manual review to exclude outliers; these scores are evaluation output and are reported in Table 3. All other analyses and scientific results were produced without generative AI assistance. We have reviewed all AI-assisted work and verified that all reported numbers, claims, and references are accurate. We take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects or personally identifiable data. The micro-positioning dataset consists of proprietary microscopy images of mechanical components collected on a custom autonomous microscopy platform, and no institutional review board approval was required for data collection. We foresee no harmful applications or societal risks arising from this work: the proposed method targets scientific imaging, industrial inspection, and quality control, and it does not introduce new capabilities for surveillance, manipulation, or deception. The authors declare no potential conflicts of interest or sponsorship concerns.

REPRODUCIBILITY STATEMENT

The complete two-stage training procedure is specified in Section 3.4, and all hyperparameters (LoRA rank and scaling, learning rates, batch sizes, epochs, optimizers, and precision) are listed in Table 1. The dataset composition, sequence-level splitting protocol, and augmentation strategy are described in Section 4.1, and the evaluation metrics (MAE, direction accuracy, and the GPT-5-based image description score) are defined in Section 4.3. An anonymized implementation is available at `https://anonymous.4open.science/r/Micro-Agent-E948`. The training images are proprietary and cannot be released; the held-out test set, training scripts, and model configurations will be released upon acceptance.

REFERENCES

E. L. Aleixo, J. G. Colonna, M. Cristo, and E. Fernandes. Catastrophic forgetting in deep learning: A comprehensive taxonomy. *J. Artif. Intell. Res.*, 1:1–56, 2023.

J. Bai, S. Bai, S. Yang, et al. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

J. F. Brenner, B. S. Dew, J. B. Horton, T. King, P. W. Neurath, and W. D. Selles. An automated microscope for cytologic research: A preliminary evaluation. *Journal of Histochemistry & Cytochemistry*, 24(1):100–111, 1976.

X. Cao, T. Zhou, Y. Ma, et al. MAPLM: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 21819–21830, 2024.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, jul 2022. doi: 10.1109/TPAMI.2021.3057446.

W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23(120):1–39, 2022.

P. Grossmann. Depth from focus. *Pattern Recognit. Lett.*, 5(1):63–69, Jan. 1987. doi: 10.1016/0167-8655(87)90026-2.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022. doi: 10.48550/arXiv.2110.04366.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei. VoxPoser: Composable 3D value maps for robotic manipulation with language models. In *Proc. Conference on Robot Learning (CoRL)*, 2023.

A. Q. Jiang, A. Sablayrolles, A. Roux, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, 24:109–165, 1989.

OpenAI. GPT-4V(ision) system card. Technical report, OpenAI, San Francisco, CA, USA, 2023. URL `https://cdn.openai.com/papers/GPTV_System_Card.pdf`.

S. Pertuz, D. Puig, and M. A. Garcia. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5):1415–1432, 2013.

W. Piao, Y. Han, L. Hu, and C. Wang. Quantitative evaluation of focus measure operators in optical microscopy. *Sensors*, 25(10):3144, may 2025. doi: 10.3390/s25103144.

N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

X. Zhou, J. He, Y. Ke, G. Zhu, V. Gutiérrez-Basulto, and J. Z. Pan. An empirical study on parameter-efficient fine-tuning for multimodal large language models. In *Proc. Findings Assoc. Comput. Linguist. (ACL)*, pp. 10057–10084, Bangkok, Thailand, 2024.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 06–09 Nov 2023.