# PEFT Methods for Embodied VLM Agents: A Systematic Study and MoE-DoRA

Vishnuram AV*
Boston University
vishnuav@bu.edu

Hilton Raj*
Boston University
hiltonr@bu.edu

## Abstract

*Vision Language Models (VLMs) deployed as embodied agents requires domain-specific adaptation, yet parameter-efficient fine-tuning methods have been studied almost exclusively on NLP benchmarks, where tasks demand neither visual grounding nor structured action planning from scarce demonstrations. In this work, we introduce* ***MoE-DoRA*** *(Mixture of Directional Experts), a novel architecture that extends weight-decomposed low-rank adaptation with parallel directional experts governed by a token-level router. By specializing directional updates while sharing a common magnitude vector, MoE-DoRA provides a framework for grounding diverse multimodal inputs in complex action spaces. To evaluate our method, we conduct the first systematic benchmark of diverse PEFT methods on EmbodiedBench's EB-Habitat tasks using Qwen2-VL-7B. Our results reveal that while QDoRA (Quantized DoRA) currently achieves the highest empirical performance (0.72 SR) due to the implicit regularization of 4-bit quantization in data-scarce regimes, we hypothesis that MoE-DoRA offers a scalable path for increasing PEFT expressiveness as embodied training data grows. We additionally show through ablation that the directional component, and not the magnitude, is the critical factor in DoRA's effectiveness for embodied tasks. Our benchmark provides hands-on guidance for practitioners adapting foundation models to embodied settings.*

## 1. Introduction

Vision-Language Models (VLMs) such as Qwen2-VL [15] have demonstrated strong capabilities as planners for embodied agents, translating visual feedback and natural language instructions into structured action plans [16]. However, fine-tuning these multi-billion parameter foundational models for domain-specific environments or task requires substantial compute and memory, motivating the use of Parameter-Efficient Fine-Tuning (PEFT) methods [5].

While the PEFT methods has been predominately studied for NLP tasks [2, 5, 9], a critical gap exists: *no systematic comparison of PEFT methods has been conducted for VLM-based embodied agents.* Embodied tasks differ fundamentally from standard language tasks, they require integration of visual perception, spatial reasoning, and structured action generation within a closed-loop planning framework, making it unclear whether conclusions from NLP transfer directly.

In this work, we address this gap by benchmarking diverse PEFT configurations on the EB-Habitat environment from EmbodiedBench [16], all applied to Qwen2-VL-7B-Instruct. Beyond evaluating established methods, we introduce **MoE-DoRA**: Mixture of Directional Experts, a novel architecture that replaces DoRA's single directional low-rank update with multiple parallel expert pairs governed by a learned token-level router, while retaining a shared magnitude vector. This is motivated by the inherent heterogeneity of embodied tasks, where visual observations, spatial reasoning, and structured action generation may benefit from specialized directional updates while sharing a common magnitude scale.

Notably we find that the simpler QDoRA: Quantized DoRA, decisively outperforms all methods, including the architecturally more complex MoE-DoRA. We analyze this result through the lens of implicit regularization and provide practical recommendations for the embodied AI community.

## 2. Related Work

Our work intersects Parameter-Efficient Fine-Tuning (PEFT) and language-grounded embodied agents. PEFT methods, such as LoRA [5] and its weight-decomposed variant DoRA [9], have revolutionized model adaptation by drastically reducing trainable parameters. Recent advancements like QLoRA [2] further enable the fine-tuning of multi-billion parameter models on consumer hardware. Simultaneously, the emergence of language-grounded foundational Vision-Language Models (VLMs) and Vision-Language-Action (VLA) models, including PaLM-E [3], RT-2 [1], OpenVLA [7], and VIMA [6], has demonstrated

[1] Equal contribution.

Accepted to CVPR FMEA Workshop 2026.

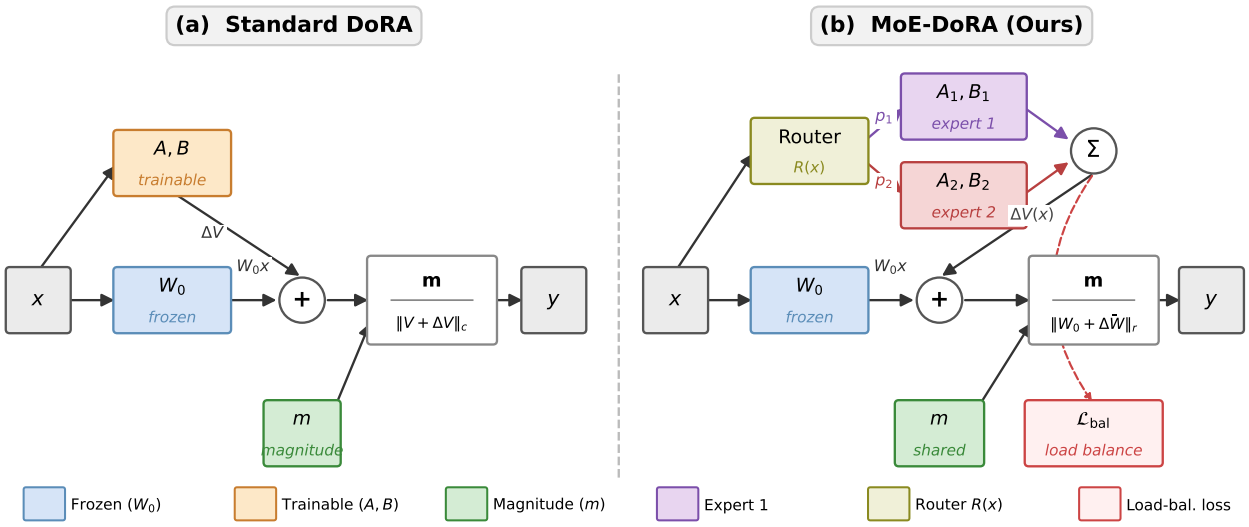


Figure 1. Architecture comparison between (a) Standard DoRA and (b) our proposed MoE-DoRA. While Standard DoRA utilizes a single magnitude vector $m$, our method introduces a MoE-based refinement where multiple experts share a unified magnitude vector to maintain parameter efficiency.

the efficacy of pre-trained multimodal representations for complex embodied reasoning, supported by benchmarks like EmbodiedBench [16]. While Mixture-of-Experts (MoE) architectures [4, 12] have been integrated into LoRA to handle multi-task heterogeneity, their application to the magnitude-direction decomposition in embodied VLM agents is previously unexplored.

## 3. Methodology

### 3.1. Backbone and Task Setup

We use **Qwen2-VL-7B-Instruct** [15] as the backbone VLM. At each time step $t$, the agent receives an egocentric RGB observation, generates a textual description of the current view, updates its internal state using past actions and feedback, formulates a high-level plan based on the instruction and context, and maps this plan into executable action tokens for simulator execution in EB-Habitat environment. All PEFT methods target the attention projections (`q/k/v/o_proj`) and MLP projections (`gate/up/down_proj`) with rank $r$=16 and $\alpha$=32 unless noted.

### 3.2. Unimodal Backbone Analysis

To establish a lower bound and assess the model's reliance on linguistic priors, we evaluate Qwen2-VL in a unimodal setting by providing blank visual input. This setup forces the model to rely solely on its intrinsic world knowledge for spatial planning and action generation. We conduct this analysis on both a sequence-to-sequence task using the ALFRED dataset [13] and a closed-loop text-only interactive environment ALFWorld [14], with results detailed in Section 5.

### 3.3. Established PEFT Baselines

We evaluate a diverse suite of established PEFT methods to represent different adaptation strategies. **LoRA** [5] learns low-rank updates to the frozen weight matrices, while **QLoRA** [2] extends this with 4-bit NormalFloat (NF4) quantization for extreme memory efficiency. **DoRA** [9] decomposes updates into magnitude and direction components, a strategy we build upon for our novel MoE variant. For dynamic adaptation, we include **AdaLoRA** [17], which adjusts the rank of updates across layers based on importance scores. Finally, we evaluate **IA3** [8], which rescales activations with learned vectors, and **PiSSA** [10], which initializes low-rank matrices using principal singular values to accelerate convergence.

### 3.4. MoE-DoRA: Mixture of Directional Experts 1

Embodied tasks jointly process fundamentally different token types in each forward pass: visual tokens encoding egocentric observations and feedback, planning tokens for spatial reasoning, and action tokens generating structured JSON output. Standard DoRA applies a *single* directional update $\Delta V = BA$ uniformly to all tokens, regardless of their role. We hypothesize that different token types may benefit from different directional adaptations, motivating a mixture-of-experts extension that routes tokens to specialized directional updates while sharing a common magnitude

vector.

MoE-DoRA replaces DoRA's single directional update with $N$ parallel expert pairs $\{(A_i, B_i)\}_{i=1}^{N}$ and a router $R(x) = W_R x$:

$$\Delta V(x) = \sum_{i=1}^{N} \text{softmax}(R(x))_i \cdot \alpha \cdot B_i A_i x \qquad (1)$$

The magnitude normalization uses the *expected* combined weight norm:

$$y = \frac{m}{\left\|W_0 + \overline{\Delta W}\right\|_{\text{row}}} \odot (W_0 x + \Delta V(x)) \qquad (2)$$

where $\overline{\Delta W} = \sum_i \bar{p}_i \alpha B_i A_i$ is computed from batch-averaged routing probabilities. We compute $\left\|W_0 + \overline{\Delta W}\right\|_{\text{row}}$ exactly via low-rank Gram matrices ($r \times r$ intermediates) without materializing full weight matrices, and add a load-balancing loss $\mathcal{L}_{\text{bal}} = N \cdot \sum_i \bar{f}_i^2$ ($\lambda$=0.01) to prevent router collapse. This architecture is a strict generalization of DoRA: at $N$=1 it reduces to standard DoRA exactly.

### 3.5. QDoRA: Quantized DoRA

DoRA decomposes each pre-trained weight matrix $W_0$ as $W' = m \cdot \frac{V+\Delta V}{\|V+\Delta V\|_c}$, where $m$ is a learnable magnitude vector and $\Delta V = BA$ is the low-rank directional update. QDoRA applies 4-bit NF4 quantization to the frozen base weights $W_0$ while preserving the full-precision magnitude vector and directional updates, following the QLoRA double-quantization scheme [2].

## 4. Experimental Setup

**Benchmark.** We evaluate on the **EB-Habitat** subset of EmbodiedBench [16], a comprehensive evaluation framework for multimodal embodied agents comprising 1,128 tasks across diverse environments. EmbodiedBench assesses both high-level semantic reasoning (e.g., household planning) and low-level atomic skills (e.g., navigation and manipulation), along six capability dimensions: task solving, commonsense reasoning, instruction understanding, spatial awareness, visual perception, and long-horizon planning. EB-Habitat focuses on high-level household rearrangement tasks in photo-realistic Replica-CAD scenes rendered by the Habitat 2.0 simulator [16]. We report results on the `base` split (50 episodes).

**Evaluation Protocol.** The agent interacts with the Habitat 2.0 simulator in a closed loop. At each time step $t$, the fine-tuned Qwen2-VL model receives the task instruction and an egocentric RGB observation, then outputs a structured JSON action consisting of a high-level skill name and its arguments (e.g., {`"action": "navigate_to", "object": "table"`}). The simulator executes the action and returns the next observation and feedback. This loop continues until the task is completed, a maximum step limit is reached, or the agent outputs a stop action.

**Training.** We fine-tune on curated expert recorded trajectories from EB-Habitat, enabling the model to learn the mapping from egocentric observations to multi-step executable actions. All methods use cosine LR schedule (peak $2\times10^{-4}$), effective batch size 8, 3 epochs, bfloat16 precision, on NVIDIA L40S (48GB) GPUs. MoE-DoRA uses $N$=2 experts with a custom trainer incorporating the load-balancing auxiliary loss.

**Metrics.** **Success Rate (SR)**: fraction of completed tasks. **Progress**: average subgoal completion. **Subgoal Reward**: reward from subgoal completion. **Invalid Action Ratio (IAR)**: fraction of invalid actions, reflecting adherence to the required JSON action format. **Planner Error**: number of malformed planner outputs. **Steps**: average episode length.

## 5. Results and Discussion

Table 1 presents the full comparison. Our unimodal analysis reveals that while the Qwen2-VL backbone possesses strong linguistic world knowledge, achieving ROUGE-L (0.67) and BLEU (0.37) on the ALFRED sequence-to-sequence task, it fails to translate this into precise action execution without visual input with Exact Match = 0.02. In the closed-loop EB-Habitat environment, this unimodal baseline completes only 16/100 tasks (0.16 SR), often exhibiting repetitive reasoning patterns.

All multimodal fine-tuning methods significantly outperform these unimodal (0.16) and zero-shot multimodal (0.50) baselines. We organize our remaining analysis around three key findings.

**MoE-DoRA requires more data for expert specialization.** Despite its theoretical expressiveness, MoE-DoRA ($N$=2) achieved a 0.58 success rate. We attribute this to the challenge of training a token-level router from only $\sim$540 expert trajectories. Without sufficient supervision signal, the router fails to specialize experts for vision, planning, and action tokens, making simpler decomposition methods (N=1) more robust for small-scale embodied datasets.

**Directional decomposition is Critical** Among established methods, DoRA achieves the highest success rate (0.64) and subgoal completion (0.671). Our magnitude-only ablation (0.46 success) proves that the directional update is the critical factor; magnitude scaling alone fails to

Table 1. **Comprehensive PEFT comparison on EB-Habitat (50 episodes).** All methods fine-tune Qwen2-VL-7B-Instruct. Best results in **bold**, second-best underlined. *Ablation: magnitude-only DoRA (directional matrices frozen at random init). †Novel architecture introduced in this work.

| Method | Success ↑ | Progress ↑ | Subgoal ↑ | Inv. Ratio ↓ | Planner Err. ↓ | Steps |
|---|---|---|---|---|---|---|
| Unimodal (text-only) | 0.16 | – | – | – | – | – |
| Multimodal (zero-shot) | 0.50 | 0.526 | 0.500 | 0.461 | 0.98 | 10.34 |
| PiSSA [10] | 0.56 | 0.588 | 0.525 | 0.581 | 2.86 | **6.56** |
| LoRA [5] | 0.60 | 0.725 | 0.533 | 0.382 | 0.94 | 9.68 |
| AdaLoRA [17] | 0.60 | 0.638 | 0.633 | 0.396 | 1.08 | 9.10 |
| IA3 [8] | 0.60 | 0.698 | 0.558 | 0.232 | 0.72 | 12.16 |
| QLoRA [2] | 0.64 | 0.710 | 0.596 | 0.239 | 0.30 | 11.64 |
| DoRA [9] | 0.64 | 0.695 | 0.671 | **0.203** | 0.40 | 10.46 |
| DoRA (Mag-only)* | 0.46 | 0.477 | 0.475 | 0.268 | 1.86 | 8.80 |
| MoR-DoRA ($N$=2)† | 0.58 | 0.648 | 0.642 | 0.276 | 0.32 | 11.76 |
| QDoRA | **0.72** | **0.785** | **0.646** | 0.214 | **0.20** | 10.02 |

capture the spatial structure required for closed-loop execution, even underperforming the zero-shot multimodal baseline (0.50).

**QDoRA achieves state-of-the-art performance.** QDoRA attains a success rate of **0.72**, a 12.5% relative improvement over standard DoRA and QLoRA (0.64). This result is striking given QDoRA's architectural simplicity, suggesting that 4-bit NF4 quantization acts as an effective implicit regularizer for weight directions in data-scarce situation while simultaneously halving GPU memory requirements.

**Diverse trade-offs among established PEFT methods.** Other methods exhibit distinct failure modes: while LoRA achieves high Progress (0.725), its high Invalid Action Ratio (0.382) indicates struggles with producing structured action sequences. IA3 maintains low IAR (0.232) but at the cost of lower subgoal completion (0.558). Most notably, PiSSA's SVD-based initialization leads to excessive planner errors (2.86), suggesting it favors training shortcuts over the reliable structural execution required for embodied agents.

## 6. Conclusion

In this work, we introduced MoE-DoRA, a novel Mixture-of-Experts architecture that extends weight-decomposed low-rank adaptation with token-level routing. By replacing DoRA's single directional update with parallel experts, MoE-DoRA provides a framework for handling the inherent heterogeneity of embodied tasks, where visual perception, high-level planning, and structured action generation benefit from specialized directional adaptations. While our benchmark of diverse PEFT configurations on EmbodiedBench identifies QDoRA as the empirical leader (0.72 SR) due to the implicit regularization of 4-bit quantization, MoE-DoRA represents a significant architectural advancement for scaling PEFT expressiveness. For practitioners, our results advocate for prioritizing directional decomposition (DoRA family) to ensure spatially consistent grounding.

**Future Work.** While MoE-DoRA currently underperforms due to the optimization challenges of training a router on only $\sim$540 trajectories, it remains a promising architecture for larger datasets. Future research will explore scaling MoE-DoRA on massive embodied corpora like Open X-Embodiment [11] to allow the router to learn meaningful expert specialization across navigation and manipulation tasks. Additionally, we plan to investigate the generalizability of VLMs and evaluate its efficacy in continuous low-level control settings beyond high-level planning.

## References


[1] Anthony Brohan, Noah Brown, Jerry Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, 2023. 1

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 3, 4

[3] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, et al. PaLM-E: An embodied multimodal language model. In *International Conference on Machine Learning (ICML)*, 2023. 1

[4] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with sim-

ple and efficient sparsity. *Journal of Machine Learning Research (JMLR)*, 23(120):1–39, 2022. 2

[5] 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 (ICLR)*, 2022. 1, 2, 4

[6] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yuke Zhu, Li Fei-Fei, and Animashree Anandkumar. VIMA: General robot manipulation with multimodal prompts. In *International Conference on Machine Learning (ICML)*, 2023. 1

[7] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model, 2024. 1

[8] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, 2022. 2, 4

[9] Shih-Yang Liu, Chao-Yuan Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hsiung Chen. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning (ICML)*, 2024. 1, 2, 4

[10] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models, 2025. 2, 4

[11] Open X-Embodiment Collaboration, Anthony Brohan, Noah Brown, et al. Open X-Embodiment: Robotic learning datasets and RT-X models, 2023. 4

[12] Noam Shazeer, Azalia Mirhoseini, Krzysztof Mazhandu, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[13] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Wessel Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpretable multi-step action forecasting. In *CVPR*, 2020. 2

[14] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *ICLR*, 2021. 2

[15] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqi Chen, Kazuki Dang, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2

[16] Rui Yang, Tao Bao, Qi Wang, and Yejin Choi. EmbodiedBench: Comprehensive benchmarking multimodal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025. 1, 2, 3

[17] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning, 2023. 2, 4