MASER: Modality-Adaptive Specialist Routing for Embodied 3D Spatial Intelligence

Hilton Raj*
Boston University
hiltonr@bu.edu

Vishnuram AV*
Boston University
vishnuav@bu.edu

Abstract

*In 3D environments, **Embodied Agents** answer spatially relevant questions through reasoning from a mixture of modalities including natural language, RGB images, point clouds, depth maps and camera poses. Existing Vision-Language models (VLMs) are fine-tuned over a single modality. This completely ignores the question semantics which may favor a different modality than the finetuned modality. To address this, we propose **MASER** (**M**odality-**A**daptive **SpE**cialist **R**outing), a lightweight framework that trains five different modality adapters of a shared VLM backbone and learns a neural routing policy that selects the best adapter based on the question during inference. We encode each question with a frozen sentence transformer and pass the embedding through a small Multi-layer Perceptron (MLP) trained on oracle adapter-accuracy labels. We evaluate our methodology over the **Open3D-VQA** benchmark and our evaluations show that no single modality is universally optimal – point-cloud answers are best in 51.5% of cases. MASER routes with 51.3% oracle agreement, outperforming a Random-Forest ablation (43.5%), with only a single adapter call per question.*

1. Introduction

Embodied intelligence relies on modalities beyond images. An autonomous agent operating in an outdoor scene must answer reasoning questions such as *"Is the tall building to the left of the signage?"* or *"How many vehicles are within 20 meters?"*. These are questions that demand geometric spatial reasoning and depth understanding that RGB images cannot unanimously cover. Modern VLMs [2, 7] excel at image-grounded question-answering tasks but struggle to generalize over point cloud, depth or pose cues [13].

An obvious solution is parameter-efficient fine-tuning (PEFT), that is given a frozen VLM backbone, train a separate low-rank adapter over the training data. However, yet another crucial question arises: *which adapter should be activated for a given question?* One option would be to run all modality adapters and aggregate – which is computationally expensive when modality count increases. The polar opposite would be to randomly select an adapter which is cheap but unstable accuracy. Therefore, it is evident that a *routing policy* remains a necessity that maps a question to its best modality adapter efficiently.

To address this, we propose **MASER**, a modality-routing framework for embodied 3D VQA tasks. Our main contributions are as follows:

- Five DoRA [8] adapters of **Qwen2-VL-2B** are fine-tuned independently on inputs of different modalities while sharing a single backbone that is switched during inference.
- A frozen sentence transformer (**SBERT** [10]) encodes the question into a 384-dimensional semantic embedding. A three-layer MLP maps this embedding to a probability distribution over the five adapters.
- We analyze the latency of using different modal adapters and provide a speed-accuracy trade-off through our proposed method.

2. Related Work

Embodied 3D Visual QA. Open3DVQA [13] dataset is a major benchmark for embodied scene understanding which combines RGB, depth, pointcloud and pose data across different scenes. Prior work on 3D question-answering [1, 9] focuses on indoor settings with structured depth sensors. To validate our hypothesis, we have chosen this benchmark which contains considerably high modality count.

Parameter-Efficient Fine-Tuning. Low-Rank Adaptation (LoRA) [4] and its variants reduce trainable parameters by decomposing weight updates into low-rank matrices. DoRA [8] further decomposes weights into magnitude and direction, improving stability and convergence. IA^3 [6] scales activations with learned vectors, achieving strong performance at extreme parameter budgets. We use DoRA

*Equal contribution.

Accepted to CVPR FMEA Workshop 2026.

throughout, as it consistently outperformed LoRA and IA^3 in our preliminary ablations on the Open3D-VQA dataset.

Mixture-of-Experts and Adapter Routing. Mixture-of-Experts (MoE) [5, 11] works by routing tokens or samples to specialized networks present in a single model. Recent work applies MoE to language model fine-tuning via LoRAMoE [3] and MoLoRA [15], which gate among multiple LoRA matrices within one model.

3. Methodology

MASER consists of three stages: (1) training modality-specific adapters, (2) collecting oracle routing labels and training a neural router, and (3) confidence-based cascade inference.

3.1. Modality-Specific Adapters

Let $\mathcal{M} = \{\texttt{image}, \texttt{depth}, \texttt{pc}, \texttt{pose}, \texttt{text}\}$ denote the five sensor modalities that we focus on in the Open3D-VQA dataset. We train one DoRA adapter ϕ_m per modality $m \in \mathcal{M}$, all sharing a frozen backbone f_θ.

We perform modality engineering to enable VLM specific processing. The image frames are resized to 416×416 and passed directly to the vision encoder. We further summarize pointcloud data as a structured text string encoding the centroid, point count and vertical extent Δz. The raw depth map is min-max normalized to $[0, 255]$ and converted to a three-channel grayscale image. The pose and text modalities are processed with no additional feature engineering.

3.2. Oracle Data Collection and Neural Router Training

Oracle label construction. For each question q_i in the router training split (15% of Open3D-VQA), we run all five adapters and score each prediction $\hat{a}_{i,m}$ against the ground-truth answer a_i^* using a judge:

$$s_{i,m} = \text{Judge}(\hat{a}_{i,m}, a_i^*), \quad s_{i,m} \in \{0, 1\}. \tag{1}$$

Initially, we apply a naive exact-match check, and if it fails, a lightweight LLM (`Qwen2.5-1.5B-Instruct`) evaluates semantic equivalence.

The oracle label for question q_i is the cost-penalised best modality:

$$m_i^* = \arg\max_{m \in \mathcal{M}} \big(s_{i,m} - \lambda \cdot c_{i,m}\big), \tag{2}$$

where $c_{i,m}$ is the latency of adapter m on question i, and λ=0.01 controls the accuracy-efficiency trade-off. On our routing split this yields the following oracle label distribution, which directly motivates routing: point cloud (51.5%), depth (7.9%), text (19.2%), image (6.2%), pose (15.2%).

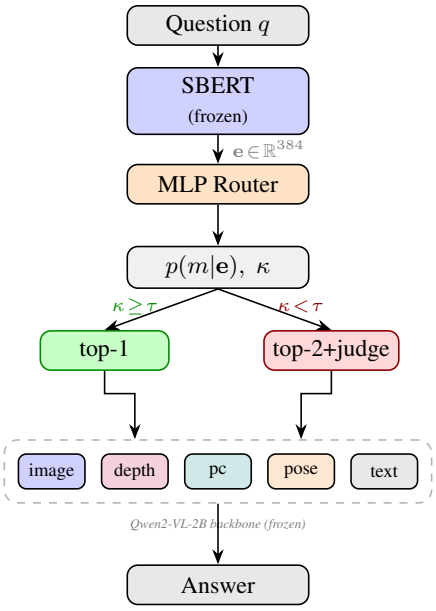

Figure 1. **MASER Architecture.** A frozen SBERT encoder maps the question to a 384-dim embedding; an MLP router selects the best modality adapter via $\hat{m} = \arg\max p_\theta$. All five adapters share the frozen Qwen2-VL-2B backbone.

Sentence embedding. After collecting the oracle labels, each question q_i is encoded with a frozen **SBERT** sentence transformer [10]:

$$\mathbf{e}_i = \text{SentEnc}(q_i) \in \mathbb{R}^{384}, \quad \|\mathbf{e}_i\|_2 = 1. \tag{3}$$

The sentence encoder is kept frozen; only the MLP head is trained. This limits router parameter count to ≈100K.

MLP router. A three-layer MLP maps the embedding to adapter logits:

$$g(\mathbf{e}_i) = W_3\, \sigma(W_2\, \sigma(W_1 \mathbf{e}_i))\,, \tag{4}$$

where σ is GELU, $W_1 \in \mathbb{R}^{256\times384}$, $W_2 \in \mathbb{R}^{64\times256}$, $W_3 \in \mathbb{R}^{|\mathcal{M}|\times64}$. Dropout (0.2 / 0.1) is applied after each hidden layer.

The router is trained to minimise class-weighted cross-entropy:

$$\mathcal{L}_{\text{router}} = -\sum_i w_{m_i^*} \log p_\theta(m_i^* \mid \mathbf{e}_i), \tag{5}$$

where $w_m \propto N/(|\mathcal{M}| \cdot n_m)$ compensates for the heavily imbalanced oracle label distribution (n_m = class count, N = total samples). We train for 60 epochs with `AdamW` (η=3$\times$ 10^{-4}, weight decay 10^{-4}), selecting the checkpoint with best validation accuracy. Figure 1 illustrates the complete MASER pipeline.

Table 1. **Router accuracy on the oracle routing split.** RF = Random Forest Ablation trained on 6 lexical features.

Router	Train Acc	Val / Test Acc
Majority class (`pc`)	—	51.5%
Random selection	—	20.0%
RF (6 lexical features)	78.5%	43.5%
MASER (MLP + SBERT)	**54.59%**	**51.33%**

3.3. Confidence-Based Cascade

For question q_i, the router predicts:

$$\hat{m}_i = \arg\max_m \; p_\theta(m \mid \mathbf{e}_i), \quad \kappa_i = \max_m \; p_\theta(m \mid \mathbf{e}_i). \quad (6)$$

When $\kappa_i \geq \tau$, only adapter $\hat{m}_i$ is invoked. When $\kappa_i < \tau$, the top-2 adapters $\hat{m}_i^{(1)}, \hat{m}_i^{(2)}$ are both activated and their responses $\hat{a}^{(1)}, \hat{a}^{(2)}$ are passed to the LLM judge. The threshold τ is tuned on the routing validation set.

4. Experiments

4.1. Dataset and Evaluation Protocol

Dataset: We evaluate on **Open3D-VQA** [13], which contains 73,324 QA pairs across four different scene types: EmbodiedCity (Wuhan), RealworldUAV (Residence), UrbanScene (Residence), and WildUAV (Wild). Each sample provides an RGB image, a depth map (.npy), a point cloud (.npy), a camera pose (JSON), and a natural-language answer.

We partition samples using a fixed random seed (42) into three splits: **adapter-train** (70%, 51,326 samples), **router-train** (15%, 10,998 samples), and **test** (15%, 11,000 samples). All adapters use **Qwen** [12] as the frozen backbone. Each DoRA adapter adds approximately 36M trainable parameters ($\sim$1.8% of the 2B backbone).

Scoring Metric: Adapter responses are evaluated with a hybrid judge (Eq. 1): exact match accuracy and semantic equivalence scoring with `Qwen2.5-1.5B-Instruct` when exact match fails. We report judge-based accuracy as the primary metric, which we call **JudgeAcc**.

4.2. Router Training Results

Table 1 compares our neural embedding router against a Random-Forest ablation which covers 6 important lexical features over the dataset (covering distance, depth, location based questions).

We noticed that the zero-shot inference over the majority class `pc` router achieves a 51.5% and a random selection of router yielded 20% over the test set. The RF recorded a train-to-test gap of (78.5% $\rightarrow$ 43.5%), indicating severe overfitting to the lexical features. On the contrary, our MLP router trained on 384-dim semantic embeddings generalizes significantly (54.59% train $\rightarrow$ 51.33% val), demonstrating that semantic embeddings are a strictly better routing feature.

Table 2. **Per-class router recall** of **MASER** on the routing split.

Modality	Support	Recall	F1
`pc`	515	0.53	0.62
`text`	192	0.74	0.57
`depth`	79	0.00	0.00
`pose`	152	0.53	0.40
`image`	62	0.16	0.15
Weighted avg.	1,000	0.51	0.50

Table 3. **VQA accuracy on held-out test split.** JudgeAcc = hybrid-judge accuracy. Lat = mean latency per question. Adapters/Q = mean adapter calls per question (inference cost).

Method	JudgeAcc	Lat (s/Q)	Adapters/Q
Baseline VLM (no adapter)	39.0%	1.88	0
Image-only Adapter	**63.5%**	1.64	1.0
Text-only Adapter	44.0%	1.87	1.0
Pointcloud-only Adapter	44.0%	1.88	1.0
Depth-only Adapter	54.0%	1.58	1.0
Pose-only Adapter	40.0%	1.68	1.0
MASER router (top-1)	**47.0%**	**1.56**	1.0
MASER + cascade	47.0%	1.57	1.0

The per-class breakdown in Table 2 reveals that text (74% recall) and pose (53% recall) questions have distinctive linguistic signatures, while depth and image questions are harder to understand from question text alone. This finding further motivates the confidence cascade for visual-modality queries.

4.3. End-to-End VQA Evaluation

Table 3 reports JudgeAcc, Latency in prediction per questions and Adapters used per Question (Adapter/Q) for all baselines and **MASER** on the held-out test split.

4.4. Discussion and Limitations

Table 3 reveals that our proposed methodology achieves the lowest latency (1.56 s/Q). MASER outperforms unimodal adapters including Text, Pose and Pointcloud. However, the accuracy does not exceed the Image-only adapter (47.0% vs. 63.5%). We attribute this to the dense spatial information the image adapter learns given a frozen VLM backbone.

The oracle label construction (Eq. 2) penalises slow adapters via $\lambda \cdot c_{i,m}$. Image and depth adapters demand full Vision Transformer encoding and are slower, so the oracle labels assign them lower scores. Thus, the router learns to

route only 11.5% of queries to the image adapter, compared to the 20% it would receive by chance.

The router is given only the question data as input. It cannot determine whether the scene's point cloud is sparse or whether the RGB image provides a clear view. The per-class recall table (Table 2) confirms this: depth (0.00% recall) and image (16% recall) questions have weak semantic signatures. Therefore, when the router assigns such questions to pointcloud or pose instead of image adapters, the accuracy reduces.

The cascade provides identical JudgeAcc (47.0%) because the router confidence is consistently high ($\kappa_i \geq \tau$ for 94% of queries), meaning the second adapter is rarely invoked. This further confirms that the MLP router is decisive, though it also suggests τ should be tuned more aggressively in future work.

Despite these limitations, MASER's lowest latency demonstrates that our routing approach is sound: when inference speed is the primary constraint in real-time cases such as edge robotics and real-time UAV, routing to cheaper non-visual modalities is a viable operating point.

5. Future Work

Our study opens two directions for adaptive routing in embodied agents. First, we observed that raw oracle labels conflate modality with adapter quality. Future work must disentangle these by defining oracle labels relative to the base model's performance. We aim to understand the superior performance of visual image adapters and provide improved routing policy in multiple VQA datasets.

Secondly, the router currently relies solely on question text, leading to visual modality confusion. Providing the router with a visual feature encoder would assist resolve visual disparities without requiring additional adapter passes.

Moreover, our router training uses only 1,000 samples, which was chosen for computational efficiency to validate our approach. Scaling to the full train split (10,998 samples) would likely improve routing accuracy further. Additionally, the point-cloud summary is a coarse numerical descriptor. A learned 3D encoder such as PointBERT [14] may yield richer features for the pointcloud router. Finally, tuning the cascade threshold τ more aggressively, or conditioning it on scene-oriented features, may improve the accuracy benefit expected from the cascade.

6. Conclusion

We presented **MASER**, a modality-adaptive routing framework for embodied 3D VQA. By combining five different modality-specific DoRA adapters with a shared VLM backbone, **MASER** achieves efficient and accurate multimodal inference without running all adapters on every question. Our analysis on **Open3D-VQA** provides the first empirical evidence that question semantics strongly predict the best modality, with no single modality accounting for more than 51.5% of optimal responses. The resulting routing accuracy (51.33%) substantially exceeds both random selection (20.0%) and a Random-Forest ablation (43.5%). We propose **MASER** as a first step toward cost-aware perception policies for embodied agents, with future work extending towards reward-based router refinement and integration of better 3D question encoders.

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3d question answering for point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. In *arXiv preprint arXiv:2308.12966*, 2023. 1

[3] Shihan Dou et al. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In *Association for Computational Linguistics (ACL)*, 2024. 2

[4] 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

[5] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. 1991. 2

[6] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[7] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

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

[9] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3d scenes. In *International Conference on Learning Representations (ICLR)*, 2023. 1

[10] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 1, 2

[11] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outra-

geously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 2

[12] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[13] Weiming Ye et al. Open3DVQA: A benchmark for spatial reasoning with multimodal large language models in open space. In *arXiv preprint arXiv:2402.03366*, 2024. 1, 3

[14] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[15] Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Luke Zettlemoyer, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient MoE for instruction tuning. In *International Conference on Learning Representations (ICLR)*, 2024. 2