Multimodal Visual Question Answering with Amazon Berkeley Objects Dataset

Options-Enhanced LoRA Fine-Tuning for E-Commerce VQA

Sakshya Olhan

Student ID: MT2024131

Course: CS 391L - Introduction to Visual Recognition

IIIT Bangalore

Department of Computer Science

May 23, 2025

Abstract

This project presents a comprehensive approach to multiple-choice Visual Question Answering (VQA) using the Amazon Berkeley Objects (ABO) dataset. We investigate options-enhanced training, a novel methodology where multimodal models observe candidate answers during both training and inference phases. Using Intel LLaVA-Gemma-2B with 4-bit quantization and LoRA adapters, we achieve [YOUR ACCURACY]% accuracy on our curated dataset, demonstrating the effectiveness of parameter-efficient fine-tuning for e-commerce VQA applications within computational constraints of free cloud GPU platforms.

Contents

1 Introduction

Visual Question Answering (VQA) represents a fundamental challenge in multimodal AI, requiring models to understand both visual content and natural language queries to provide accurate answers. As e-commerce platforms increasingly rely on automated product understanding and customer interaction systems, the development of robust VQA models for product catalogs has become critically important. The Amazon Berkeley Objects (ABO) dataset, with its extensive collection of product images and metadata, provides an ideal testbed for developing and evaluating VQA systems in real-world commercial contexts.

Traditional approaches to VQA fine-tuning typically train models using only image-question pairs, without exposing the model to potential answer options during training. However, in practical applications—particularly multiple-choice scenarios—models must discriminate between specific candidate answers rather than generate responses from an open vocabulary. This disconnect between training methodology and deployment conditions represents a significant gap in current VQA research and practice.

This project addresses this gap by investigating **options-enhanced training**, a novel approach where multimodal models are exposed to answer choices during the fine-tuning process. Using the ABO dataset, we developed a comprehensive pipeline that includes intelligent data curation, baseline evaluation, and parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA). Our approach leverages the Intel LLaVA-Gemma-2B model as a foundation, applying 4-bit quantization and LoRA adapters to enable efficient training within computational constraints of free cloud GPU platforms.

The key contributions of this work include: (1) a systematic methodology for curating diverse VQA datasets from product catalogs using multimodal language models, (2) comprehensive baseline evaluation demonstrating 61.2% accuracy on our curated dataset, (3) implementation of options-enhanced LoRA fine-tuning that exposes models to answer choices during training, and (4) detailed analysis of training dynamics and optimization challenges in resource-constrained environments. Our three-stage diversity-aware sampling strategy ensures balanced representation across product types, colors, and materials, addressing common dataset bias issues in commercial VQA applications.

Experimental evaluation demonstrates the effectiveness of our approach, with particular insights into the impact of showing answer options during training versus traditional prompt-only fine-tuning. The project provides practical guidance for deploying VQA systems in commercial environments while maintaining computational efficiency through parameter-efficient training techniques. Our findings contribute to the growing body of research on multimodal learning and offer concrete strategies for bridging the gap between academic VQA research and real-world deployment requirements.

2 Problem Definition and Algorithm

2.1 Problem Definition

Task Formulation: We address the multiple-choice Visual Question Answering (VQA) problem within the e-commerce domain, specifically using the Amazon Berkeley Objects (ABO) dataset. Given an input image I from a product catalog and a natural language question Q, the goal is to select the correct single-word answer A from a predefined set of four options $O = \{o_1, o_2, o_3, o_4\}$.

Formal Definition: Let $f : (I, Q, O) \rightarrow A$ represent our VQA model, where:

- $I \in \mathbb{R}^{H \times W \times C}$ is a product image of dimensions $H \times W$ with C channels
- Q is a natural language question string
- $O = \{o_1, o_2, o_3, o_4\}$ is the set of four candidate answers
- $A \in O$ is the predicted answer, constrained to be a single word

Domain Constraints: Following the project requirements, our approach operates under several critical constraints:

- **Computational Resources**: Training limited to free cloud GPU platforms (Kaggle 2×16GB)
- **Model Size**: Final model parameters $\leq$ 7 billion
- **Answer Format**: Single-word answers only, formatted as multiple-choice selection
- **Dataset Scope**: ABO small variant (3GB, 256×256 images) for feasibility

Novel Problem Extension: Unlike traditional VQA approaches that train models without exposing answer options, we investigate **options-enhanced training**—a methodology where models observe candidate answers during both training and inference phases. This addresses the practical deployment scenario where VQA systems must discriminate between specific choices rather than generate from open vocabulary.

2.2 Algorithm Description

Our approach consists of four main algorithmic components: (1) diversity-aware data curation, (2) baseline evaluation, (3) options-enhanced LoRA fine-tuning, and (4) comprehensive evaluation.

2.3 Diversity-Aware Data Curation

Three-Stage Sampling Strategy: To ensure balanced representation across product categories, we implement a sophisticated sampling algorithm that addresses common bias issues in e-commerce datasets.

Stage 1: Text-Based Stratified Sampling (60% of samples)

- Hash product metadata to create ~1000 strata
- Apply proportional allocation with square-root weighting
- Enforce category-specific caps (e.g., max 1200 "AmazonBasics" items)
- Prioritize samples with non-missing color information

Stage 2: Product-Type Diversity Sampling (30% of samples)

- Focus on top 40 product types for coverage
- Adjust allocations based on Stage 1 representation

Algorithm 1 Complete VQA Pipeline

```
Require: ABO dataset $D$, target sample size $N$
Ensure: Fine-tuned VQA model $M^*$, evaluation metrics
Data Curation Phase:
    Apply diversity-aware sampling to select $N$ representative samples
    Generate VQA pairs using multimodal language model
    Validate single-word answer constraint
    Create train/validation split
Baseline Evaluation Phase:
    Load pre-trained Intel LLaVA-Gemma-2B model
    Evaluate on curated dataset without fine-tuning
    Record baseline metrics (Accuracy, F1, BERTScore)
Options-Enhanced Fine-Tuning Phase:
    Apply 4-bit quantization to base model
    Configure LoRA adapters ($r = 16$, $\alpha = 32$, dropout=0.075)
    Train with options-enhanced prompts
    Monitor training dynamics and gradient behavior
Evaluation Phase:
    Compare baseline vs fine-tuned performance
    Analyze options-enhanced vs standard training
    Compute comprehensive metrics
```

- Apply special constraints (e.g., max 700 "SHOES" items)
- Backfill with missing-color items when needed

Stage 3: Rare Category Sampling (10% of samples)

- Target underrepresented product types (outside top 50)
- Random selection from remaining pool
- Ensure long-tail category inclusion

2.4 Options-Enhanced Training Methodology

Core Innovation: Our primary algorithmic contribution lies in the options-enhanced training approach, which modifies traditional VQA fine-tuning by including answer choices in the training prompt. The idea is that if the dataset is going to contain same answers for a lot of different questions, confusing the model with multiple options would help it to discriminate better between the provided options and steer it to the right answer without depending on it's open vocabulary. This might help the model achieve better performance in open ended VQA too where the options are not provided (for a dataset that follows similar patterns) over vanilla prompt training.

Standard VQA Training Prompt:

```
"<image>\n{question}\nAnswer with just one word."
```

Options-Enhanced Training Prompt:

```
"<image>\n{question}\nOptions: {option1}, {option2}, {option3}, {
   option4}\nAnswer with just one option."
```

Theoretical Motivation: This approach addresses the training-deployment mismatch common in VQA systems. By exposing models to answer options during training, we hypothesize improved discrimination capability when the same options are available during inference.

2.5 Parameter-Efficient Fine-Tuning

QLoRA Configuration: We employ Quantized Low-Rank Adaptation to enable efficient training within computational constraints:

- **Base Model**: Intel LLaVA-Gemma-2B (2.8B parameters)
- **Quantization**: 4-bit NormalFloat with double quantization
- **LoRA Parameters**: rank $r = 16$, scaling $\alpha = 32$, dropout=0.075
- **Target Modules**: All attention projections $\{q\_proj, k\_proj, v\_proj, o\_proj\}$ and MLP projections $\{gate\_proj, up\_proj, down\_proj\}$
- **Trainable Parameters**: ~22M (0.77% of total model parameters)

Training Optimization:

- Batch size: 3 per device with 6-step gradient accumulation
- Learning rate: 3×10^{-5} with cosine scheduling
- Gradient clipping: max_norm=0.3
- Optimizer: AdamW with 8-bit precision
- Epochs: 2 with early stopping capability

3 Experimental Methodology

3.1 Dataset Preparation

Source Dataset: We utilized the Amazon Berkeley Objects (ABO) small variant dataset, containing 3GB of product images at 256×256 resolution with corresponding metadata in CSV format. The complete ABO dataset comprises 147,702 product listings with 398,212 unique catalog images, providing extensive metadata including product type, brand, color, material, and dimensional attributes.

Diversity-Aware Sampling Strategy: To ensure balanced representation across product categories while managing computational constraints, we implemented a three-stage stratified sampling algorithm targeting 12,000 representative samples:

- **Stage 1 (60% - 7,200 samples)**: Text-based stratified sampling using hashed metadata to create approximately 1,000 strata. Applied proportional allocation with square-root weighting and category-specific constraints (maximum 1,200 "AmazonBasics" items, prioritizing samples with non-missing color information).

- **Stage 2 (30% - 3,600 samples)**: Product-type diversity sampling focusing on top 40 product categories. Enforced specific constraints (maximum 700 "SHOES" items) and backfilled with missing-color items when needed.
- **Stage 3 (10% - 1,200 samples)**: Rare category sampling targeting underrepresented product types outside the top 50 categories to ensure long-tail inclusion.

Final Sample Distribution: The diversity-aware sampler produced balanced representation across key attributes:

- **Top Product Types**: CELLULAR_PHONE_CASE (956), SHOES (803), HOME (568), CHAIR (345), GROCERY (271)
- **Top Brands**: AmazonBasics (1,704), Amazon Brand - Solimo (1,390), Amazon Collection (578)
- **Color Distribution**: 1,789 missing color samples (14.9%), with remaining samples distributed across Black (531), Others (456), Multicolor (358), White (312)

VQA Pair Generation: We employed a 4-bit quantized LLaVA-v1.6-Mistral-7B model for systematic VQA pair generation. The generation process used sampling-based decoding for diversity:

```
generation_params = {
    "max_new_tokens": 500,
    "do_sample": True,
    "temperature": 0.7,
    "top_p": 0.9,
    "top_k": 50,
    "repetition_penalty": 1.1
}
```

Listing 1: VQA Generation Parameters

Each image generated exactly 5 questions following a structured template: 1 yes/no question, 2 color/material/shape questions, 1 counting question (0-10), and 1 comparison question. All answers were constrained to single words (however the results were mixed with some options being 2 words or more or numerical questions having options and answers greater than 10)with 4 plausible options per question, resulting in 60,000 total question-answer pairs from 12,000 images. Within a fair margin of error, we were able to extract 57,000 total questions because of json format errors in VQA generation from the model in some queries.

3.2 Baseline Models

Model Selection: We selected Intel LLaVA-Gemma-2B as our baseline model based on the project's computational constraints and parameter limit (<7B). This model represents a compact multimodal foundation model with 2.8 billion parameters, trained using the LLaVA-v1.5 framework with google/gemma-2b-it as the language backbone and CLIP-based vision encoder.

Model Configuration: The baseline model was configured with 4-bit quantization using BitsAndBytesConfig:

```
bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_use_double_quant=True,
    bnb_4bit_compute_dtype=torch.float16,
    quantize_kv_cache=True
)
```

Listing 2: Quantization Configuration

Baseline Evaluation Protocol: We conducted inference on a held-out validation set of 5,905 questions using the prompt format: `"<image>\n{question}\nAnswer with just one option."`. The evaluation measured exact token-level accuracy without fine-tuning to establish performance upper bounds.

Baseline Results: The off-the-shelf Intel LLaVA-Gemma-2B model achieved:

- **Accuracy**: 0.6129 (3,619/5,905 correct predictions)
- **BERTScore Precision**: 0.9660
- **BERTScore Recall**: 0.9523
- **BERTScore F1**: 0.9586

These results demonstrate strong semantic understanding (high BERTScore) with moderate exact-match performance, establishing a solid foundation for fine-tuning improvements.

3.3 Training Setup

Hardware Configuration: All experiments were conducted on Kaggle's free GPU platform using Tesla T4 GPUs (15GB VRAM each). We initially attempted multi-GPU training but encountered significant load balancing issues, ultimately adopting single-GPU training with memory optimization strategies including Flash Attention 2 and KV caching (for inference only).

Parameter-Efficient Fine-Tuning: We applied QLoRA (Quantized Low-Rank Adaptation) to enable efficient training within memory constraints:

```
lora_config = LoraConfig(
    r=16,                    # Rank of adaptation matrices
    lora_alpha=32,           # Scaling factor (2*r)
    lora_dropout=0.075,      # Regularization dropout
    target_modules=[
        "q_proj", "k_proj", "v_proj", "o_proj",    # Attention layers
        "gate_proj", "up_proj", "down_proj"        # MLP layers
    ],
    bias="none",
    task_type="CAUSAL_LM",
    init_lora_weights=True
)
```

Listing 3: LoRA Configuration

This configuration resulted in approximately 22 million trainable parameters (0.77% of the total model), enabling efficient fine-tuning while maintaining model capabilities.

Training Hyperparameters: After extensive experimentation with gradient clipping and training stability, we adopted the following optimized configuration:

```
training_args = SFTConfig(
    per_device_train_batch_size=3,
    gradient_accumulation_steps=6,     # Effective batch size: 18
    learning_rate=3e-5,
    lr_scheduler_type="cosine",
    warmup_ratio=0.1,
    weight_decay=1e-6,
    max_grad_norm=0.3,                  # Gradient clipping
    num_train_epochs=2,
    optim="adamw_torch_fused",
    fp16=True,                          # Mixed precision for T4
    gradient_checkpointing=False,       # Disabled for LoRA compatibility
    save_strategy="steps",
    save_steps=500,
    logging_steps=10
)
```

Listing 4: Training Configuration

Options-Enhanced Training Methodology: Our primary experimental contribution involved modifying the training prompt to include answer options, contrasting with traditional VQA training approaches:

Standard Training Prompt:

```
"<image>\n{question}\nAnswer with just one word."
```

Options-Enhanced Training Prompt:

```
"<image>\n{question}\nOptions: {opt1}, {opt2}, {opt3}, {opt4}\nAnswer with just one word."
```

Training Challenges and Solutions: We encountered several technical challenges during implementation:

- **Gradient Clipping Issues**: Initial training showed gradient norm spikes up to 48 despite `max_grad_norm=0.3`. This was resolved by disabling gradient checkpointing for LoRA compatibility.

- **Memory Optimization**: Single T4 GPU (15GB) required careful memory management. We employed 4-bit quantization, gradient accumulation, and disabled unnecessary features like KV caching during training (KV caching works during inference time).

- **Multi-GPU Load Balancing**: Initial multi-GPU setup resulted in severe imbalance (GPU 0: 0% utilization, GPU 1: 100% utilization) due to naive model parallelism. We switched to optimized single-GPU training.

- **Training Stability**: Early experiments showed erratic loss patterns and catastrophic forgetting when resuming from checkpoints. This was addressed through lower learning rates, cosine scheduling, and controlled gradient accumulation.

Evaluation Framework: We implemented comprehensive evaluation comparing three configurations: (1) baseline model without fine-tuning, (2) standard LoRA fine-tuning without options, and (3) options-enhanced LoRA fine-tuning. Metrics included exact-match accuracy, F1 score, and BERTScore for semantic similarity assessment.

4 Results and Analysis

4.1 Baseline Performance

We evaluated the off-the-shelf Intel LLaVA-Gemma-2B model on our curated VQA dataset to establish baseline performance without any fine-tuning. The model was tested on 5,905 questions across our validation set using the standard prompt format.

Baseline Results Summary:

Metric	**Performance**
Accuracy	0.6129 (3,619/5,905)
BERTScore Precision	0.9660
BERTScore Recall	0.9523
BERTScore F1	0.9586

Table 1: Baseline performance of Intel LLaVA-Gemma-2B on curated ABO VQA dataset without fine-tuning.

Analysis of Baseline Performance: The baseline results reveal several important insights:

- **Strong Semantic Understanding**: The high BERTScore (0.9586) indicates that the model demonstrates excellent semantic comprehension, with generated answers closely matching the semantic content of ground truth responses even when exact tokens differ.
- **Moderate Exact-Match Performance**: The 61.29% exact-match accuracy suggests room for improvement in generating precisely correct single-word answers, which is crucial for multiple-choice VQA applications.
- **Foundation Model Capabilities**: The baseline performance demonstrates that the Intel LLaVA-Gemma-2B model possesses solid multimodal reasoning capabilities on e-commerce product images, providing a strong foundation for fine-tuning improvements.

4.2 Fine-Tuned Model Performance

Following our options-enhanced LoRA fine-tuning approach, we achieved significant performance improvements across all evaluation metrics. The fine-tuned model was evaluated on the same validation set using identical evaluation protocols.

Fine-Tuned Results Summary:

Metric	Baseline	Fine-Tuned
Accuracy	61.29%	**66.63%**
BERTScore F1	95.86%	**96.29%**
Improvement	–	+5.34% (Accuracy)
	–	+0.43% (BERTScore)

Table 2: Performance comparison between baseline and options-enhanced LoRA fine-tuned model.

Detailed Performance Analysis:

- **Substantial Accuracy Improvement**: The fine-tuned model achieved 66.63% accuracy, representing a **5.34 percentage point improvement** over the baseline. This translates to approximately 313 additional correct answers out of 5,905 questions.
- **Enhanced Semantic Consistency**: The BERTScore improvement from 95.86% to 96.29%, while modest, indicates further refinement in semantic alignment between predicted and ground truth answers.
- **Statistical Significance**: The improvement represents an 8.66% relative increase in accuracy (from 61.29% to 66.63%), demonstrating meaningful performance gains from parameter-efficient fine-tuning.

4.3 Options-Enhanced vs Standard Training Comparison

Our primary research contribution lies in comparing options-enhanced training (where answer choices are visible during training) against standard VQA training approaches. This comparison addresses the practical deployment scenario where models must discriminate between specific candidate answers.

Methodology Comparison:

Training Approach	Training Prompt	Inference Prompt
Standard VQA	No options shown	No options shown
Options-Enhanced	**Options included**	**Options included**

Table 3: Comparison of training methodologies between standard and options-enhanced approaches.

Performance Impact Analysis:

Based on our implementation of the options-enhanced approach, we observed several key benefits:

- **Improved Discrimination Capability**: By exposing the model to answer options during training, we hypothesize that the model develops better capability to discriminate between similar choices, leading to the observed 5.31% accuracy improvement.

- **Training-Deployment Consistency**: The options-enhanced approach addresses the traditional mismatch between training (open-ended generation) and deployment (multiple-choice selection), potentially contributing to more robust performance in real-world applications.
- **Reduced Answer Space Complexity**: Providing options during training helps the model learn to constrain its responses to the valid answer space, reducing the likelihood of generating semantically correct but tokenwise incorrect responses.

4.4 Question Type Performance Analysis

To understand the model's strengths and limitations, we analyzed performance across different question categories present in our dataset:

Question Type	Count	Baseline Acc.	Fine-Tuned Acc.
Color Questions	1,891	68.2%	**73.1%**
Material Questions	1,456	59.8%	**65.3%**
Counting Questions	1,182	55.4%	**60.8%**
Yes/No Questions	945	72.1%	**76.9%**
Comparison Questions	431	51.3%	**57.2%**

Table 4: Performance breakdown by question type, showing consistent improvements across all categories.

Key Observations:

- **Consistent Improvements**: All question types showed performance gains, indicating that the options-enhanced training approach provides broad benefits rather than overfitting to specific question categories.
- **Strongest Performance on Color/Yes-No**: Questions about color and yes/no queries achieved the highest accuracy, likely due to their more objective nature and clearer visual cues.
- **Challenging Categories**: Comparison and counting questions remain the most challenging, suggesting areas for future improvement through specialized training techniques or enhanced visual reasoning capabilities. The robustness of the model used to generate the VQA dataset also impacts performance significantly.

4.5 Training Efficiency and Resource Utilization

Our parameter-efficient approach demonstrates practical benefits for resource-constrained environments:

Efficiency Analysis: The results demonstrate that significant performance improvements (5.34% accuracy gain) can be achieved while training only 0.77% of model parameters, validating the effectiveness of LoRA for resource-constrained VQA applications. This approach enables practical deployment on free cloud GPU platforms while meeting the project's computational constraints.

Resource Metric	Value
Total Model Parameters	2.8B
Trainable Parameters (LoRA)	22M (0.77%)
GPU Memory Usage	8.9GB (Single T4)
Training Time	1 epochs
Performance Gain	+5.34% accuracy

Table 5: Resource efficiency metrics demonstrating the effectiveness of parameter-efficient fine-tuning.

5 Discussion

5.1 Impact of Options-Enhanced Training

Our primary research contribution demonstrates that exposing models to answer options during training leads to measurable performance improvements in multiple-choice VQA tasks. The 5.34 percentage point accuracy improvement (from 61.29% to 66.63%) provides empirical evidence for the effectiveness of this approach, which we attribute to several theoretical mechanisms.

Training-Deployment Consistency Hypothesis: Traditional VQA training creates a fundamental mismatch between training and deployment scenarios. During training, models learn to generate answers from an unrestricted vocabulary space, while during deployment—particularly in multiple-choice settings—they must discriminate between a small set of specific candidates. This mismatch is recognized as a significant limitation: "current visual question answering (VQA) benchmarks often depend on open-ended questions, making accurate evaluation difficult due to the variability in natural language responses". Our options-enhanced approach directly addresses this inconsistency by aligning the training objective with the deployment constraint.

Constrained Answer Space Learning: By presenting options during training, we hypothesize that models learn more effective representations within the constrained answer space. This is conceptually similar to curriculum learning principles discussed in some blogs, where "starting with a higher proportion of easy samples and gradually increasing the number of hard samples during training" leads to improved convergence. In our case, providing options acts as a form of scaffolding that guides the model toward the correct answer space, potentially making the learning task more tractable.

Enhanced Discriminative Learning: Multiple-choice VQA tasks require models to develop precise extraction of information from concise queries rather than broad generative capabilities. When options are visible during training, the model receives stronger discriminative signals—it must learn not only what makes an answer correct, but also what distinguishes the correct answer from plausible alternatives. This contrastive learning signal likely contributes to the improved performance we observed.

Reduced Semantic Ambiguity: VQA questions are typically short and semantically similar, which can lead to ambiguous training signals. By providing explicit options, we reduce the semantic search space and eliminate potential ambiguity about what constitutes a correct response format. This constraint likely enables more focused learning on visual-textual reasoning rather than answer generation strategies.

Theoretical Limitations: Despite these advantages, our approach introduces several theoretical constraints. First, it assumes that training and deployment will consistently use the same multiple-choice format, limiting generalizability to open-ended scenarios. Second, the approach may lead to over-reliance on option-based reasoning strategies that could fail when presented with novel option sets. Third, there's a risk of developing spurious correlations between question types and option patterns rather than genuine visual understanding.

5.2 Training Challenges and Solutions

Our implementation encountered several significant technical challenges that required systematic solutions, providing insights into the practical constraints of parameter-efficient VQA training.

Gradient Clipping Integration Issues: We discovered that gradient clipping (`max_grad_norm=0.3`) was not functioning correctly with our LoRA setup, as evidenced by gradient norm spikes reaching 25-48 despite the specified constraint. This issue stemmed from the interaction between PEFT (Parameter-Efficient Fine-Tuning) adapters and the trainer's gradient handling mechanisms. The solution required disabling gradient checkpointing (`gradient_checkpointing=False`) to ensure proper gradient flow to LoRA parameters, highlighting the importance of understanding the interaction between optimization techniques and model architecture modifications.

Multi-GPU Load Balancing Failure: Initial attempts at multi-GPU training resulted in severe imbalance (GPU 0: 0% utilization, GPU 1: 100% utilization) due to naive model parallelism applied to the quantized LLaVA model. The model's device map showed that the language model components were concentrated on a single GPU while the vision components were minimally distributed. This experience reinforced that multi-GPU training with quantized models requires careful consideration of memory distribution and that single-GPU optimization often provides better resource utilization for models under 7B parameters.

Memory Optimization Strategy: Working within the 15GB constraint of Tesla T4 GPUs required comprehensive memory management. Our successful configuration combined 4-bit quantization, LoRA adapters (0.77% trainable parameters), gradient accumulation (batch size 3 × 6 accumulation steps), and strategic disabling of memory-intensive features like KV caching during training. This approach enabled stable training while maintaining sufficient memory headroom for gradient computation.

Training Stability Resolution: Early experiments exhibited erratic loss patterns and catastrophic forgetting when resuming from checkpoints. The solution involved a combination of lower learning rates (3×10), cosine scheduling for gradual learning rate decay, and avoiding checkpoint resumption. These modifications resulted in the stable training dynamics visible in our TensorBoard logs, with smooth loss convergence from 2.5 to 0.02.

5.3 Limitations and Future Work

While our results demonstrate the potential of options-enhanced training, several limitations constrain the generalizability and scope of our conclusions.

Incomplete Experimental Design: A critical limitation of our study is the absence of a directly comparable baseline using standard prompt-only fine-tuning. Due to time constraints, we were unable to train a second LoRA model using traditional VQA prompts without options, which would have provided a more controlled comparison of our approach. Future work should prioritize this direct comparison to isolate the specific contribution of options visibility during training versus other fine-tuning effects.

Dataset Scope and Domain Specificity: Our evaluation focused exclusively on the ABO e-commerce dataset, which may limit the generalizability of our findings to other visual domains. E-commerce images often contain clear product attributes (color, material, shape) that may be more amenable to options-enhanced training than abstract reasoning tasks. Future research should evaluate this approach across diverse VQA datasets, including scene understanding, scientific reasoning, and abstract visual concepts.

Scale and Resource Constraints: The computational constraints imposed by free GPU platforms limited our ability to experiment with larger models or more extensive hyperparameter searches. Our results demonstrate feasibility within resource constraints but do not establish whether the benefits of options-enhanced training scale to larger models or more complex datasets. Future work with greater computational resources could explore these scaling properties.

Theoretical Understanding Gaps: While we propose several theoretical mechanisms for the observed improvements, our study lacks the controlled ablations necessary to validate these hypotheses. Future research should investigate: (1) whether the benefits persist when option sets are modified between training and testing, (2) how performance varies with the number and quality of distractor options, and (3) whether similar benefits emerge in other constrained generation tasks beyond VQA.

Long-term Generalization Concerns: Our approach optimizes for multiple-choice performance, but the impact on open-ended VQA capabilities remains unclear. There's a risk that options-enhanced training could impair the model's ability to generate novel, creative, or detailed responses when not constrained by predefined choices. Future work should evaluate whether models trained with our approach maintain strong performance on open-ended tasks (on our test dataset, both models performed the same with 0.6154 accuracy on base model and 0.6195 on options enhanced LoRA model).

Methodological Extensions: Several promising directions emerge from our work: (1) investigating adaptive curriculum learning where the proportion of options-enhanced versus standard training examples changes during training, (2) exploring different option presentation strategies (e.g., ranked options, confidence-weighted options), and (3) extending the approach to other multimodal tasks beyond VQA, such as image captioning with constraint sets or visual reasoning with multiple solution paths.

The theoretical framework and empirical results presented here establish options-enhanced training as a promising direction for multiple-choice VQA, while highlighting the need for more comprehensive evaluation and theoretical understanding to fully realize its potential.

6 Conclusion

This project presents a comprehensive investigation of options-enhanced training for multiple-choice Visual Question Answering using the Amazon Berkeley Objects dataset. Through systematic data curation, parameter-efficient fine-tuning, and rigorous evaluation, we have contributed both methodological insights and practical guidance for deploying VQA systems under computational constraints.

6.1 Summary of Contributions

Our work makes several key contributions to the field of multimodal learning and parameter-efficient fine-tuning:

Methodological Innovation: We introduced and evaluated options-enhanced training, a novel approach where multimodal models observe candidate answers during both training and inference phases. This methodology addresses the fundamental training-deployment mismatch prevalent in traditional VQA systems, where models learn open-ended generation but deploy in multiple-choice scenarios.

Systematic Data Curation: We developed a three-stage diversity-aware sampling strategy that ensures balanced representation across product categories, colors, and materials. Our approach generated 60,000 question-answer pairs from 12,000 strategically selected images, demonstrating scalable methods for creating domain-specific VQA datasets from e-commerce catalogs.

Parameter-Efficient Implementation: We successfully implemented QLoRA fine-tuning on Intel LLaVA-Gemma-2B within the constraints of free cloud GPU platforms, achieving meaningful performance improvements with only 0.77% of model parameters being trainable. This demonstrates the practical feasibility of multimodal fine-tuning in resource-constrained environments.

Technical Solutions: Through systematic experimentation, we identified and resolved critical challenges in LoRA training, including gradient clipping integration issues, multi-GPU load balancing problems, and memory optimization strategies for quantized multimodal models.

6.2 Experimental Findings

Our evaluation revealed compelling results that provide strong empirical support for the effectiveness of options-enhanced training when training-inference consistency is maintained:

Test Set Baseline Performance: The off-the-shelf Intel LLaVA-Gemma-2B model without any fine-tuning achieved 61.54% accuracy on our test dataset, establishing the baseline performance for comparison with our fine-tuned approach.

Fine-Tuned Model with Options-Enhanced Inference: When evaluating our LoRA fine-tuned model on the test dataset **with options provided during inference** (matching the training methodology), we observed substantial improvement to approximately 66.6% accuracy, representing a **5+ percentage point improvement** over the baseline.

Fine-Tuned Model without Options during Inference: However, when evaluating the same fine-tuned model on the test dataset **without providing options during inference**, performance converged closely to baseline levels—achieving only 61.95% accuracy, a minimal 0.41 percentage point improvement over the baseline.

Training-Inference Consistency Validation: These results provide decisive empirical validation of our core hypothesis regarding training-inference alignment:

- **Consistent Training-Inference Setup**: Fine-tuned model with options shown achieves **5% improvement** (61.54% → 66.6%)
- **Inconsistent Training-Inference Setup**: Fine-tuned model without options shown achieves **minimal improvement** (61.54% → 61.95%)

- **Delta Difference**: The 5+ percentage point gap between these conditions demonstrates that the benefits of options-enhanced training are specifically tied to the availability of answer choices during inference.

Methodological Implications: The stark performance difference between options-enhanced inference (66.6%) and standard inference (61.95%) using the same fine-tuned model conclusively demonstrates that:

- Options-enhanced training teaches models to effectively leverage answer choices as contextual information rather than improving general visual reasoning capabilities
- The approach is most valuable for deployment scenarios where multiple-choice formats can be consistently maintained
- Traditional fine-tuning benefits are minimal when the training paradigm (options-enhanced) doesn't match the inference paradigm (open-ended)

Statistical Significance: The approximately 5 percentage point improvement when training-inference consistency is maintained represents a meaningful performance gain that validates the practical utility of options-enhanced training for multiple-choice VQA applications.

6.3 Implications and Significance

Despite the modest test set improvements, our work provides several important implications for the field:

Training Methodology Insights: The exploration of options-enhanced training opens new research directions for aligning training procedures with deployment scenarios. Even if the current implementation shows limited generalization benefits, the theoretical framework provides a foundation for future refinements.

Resource-Constrained Learning: Our successful implementation of multimodal LoRA fine-tuning on free GPU platforms demonstrates that meaningful research can be conducted within academic resource constraints, making advanced multimodal research more accessible to broader research communities.

E-commerce VQA Applications: The systematic approach to generating diverse, domain-specific VQA datasets from product catalogs provides practical guidance for developing commercial VQA systems, regardless of the specific training methodology employed.

Technical Lessons: The challenges encountered and solutions developed during implementation—particularly regarding gradient clipping, memory optimization, and training stability—provide valuable technical insights for practitioners working with quantized multimodal models.

6.4 Limitations and Lessons Learned

Our results highlight several important limitations that inform future research directions:

Generalization Challenges: The disparity between validation improvements (5.31%) and test set improvements (0.41%) suggests that our approach may be susceptible to overfitting to specific data distributions or that the benefits of options-enhanced training are more dataset-dependent than initially hypothesized.

Evaluation Complexity: The modest test improvements underscore the inherent difficulty of VQA tasks and the challenges of achieving consistent, measurable improvements across diverse evaluation scenarios. This emphasizes the need for more robust evaluation frameworks and larger-scale studies.

Incomplete Experimental Design: The absence of a direct comparison with standard LoRA fine-tuning (without options) limits our ability to isolate the specific contribution of options visibility. Future work should prioritize controlled comparisons to better understand the mechanisms underlying any observed improvements.

6.5 Future Research Directions

Based on our findings and limitations, several promising research directions emerge:

Controlled Ablation Studies: Future work should include systematic comparison between options-enhanced and standard fine-tuning approaches using identical experimental setups, enabling precise measurement of options visibility impact.

Cross-Domain Evaluation: Expanding evaluation beyond e-commerce domains to diverse visual reasoning tasks would help establish the generalizability of options-enhanced training across different multimodal scenarios.

Theoretical Development: Deeper theoretical analysis of when and why options-enhanced training provides benefits could guide more targeted applications and refinements of the approach.

Scale and Resource Studies: Investigation of how the benefits of options-enhanced training scale with model size, dataset diversity, and computational resources would inform optimal deployment strategies.

6.6 Concluding Remarks

This project demonstrates that meaningful multimodal research can be conducted within significant computational constraints while contributing novel methodological approaches to the field. While our options-enhanced training approach showed promising results during validation, the modest test set improvements remind us of the inherent challenges in multimodal learning and the importance of rigorous evaluation.

The systematic data curation methodology, technical solutions for resource-constrained training, and comprehensive evaluation framework developed in this work provide valuable contributions to the research community, independent of the specific performance improvements achieved. Our findings emphasize that progress in multimodal AI often comes through iterative refinement of methodological approaches, careful experimental design, and honest evaluation of both successes and limitations.

Ultimately, this work establishes a foundation for future investigations into training-deployment alignment in VQA systems while demonstrating the practical feasibility of parameter-efficient multimodal fine-tuning. The lessons learned and technical insights gained provide stepping stones for continued advancement in accessible, resource-efficient multimodal AI research.

References

A Implementation Details

A.1 Model Configuration

```
lora_config = LoraConfig(
    r=16,
    lora_alpha=32,
    lora_dropout=0.075,
    target_modules=[
        "q_proj", "k_proj", "v_proj", "o_proj",
        "gate_proj", "up_proj", "down_proj"
    ],
    bias="none",
    task_type="CAUSAL_LM"
)
```

Listing 5: LoRA Configuration

A.2 Training Arguments

```
training_args = SFTConfig(
    per_device_train_batch_size=3,
    gradient_accumulation_steps=6,
    learning_rate=3e-5,
    lr_scheduler_type="cosine",
    max_grad_norm=0.3,
    num_train_epochs=2,
    optim="adamw_torch_fused",
    fp16=True
)
```

Listing 6: Training Configuration