AIM825 Final Project Report

Team Members: [Sparsh Salodkar(IMT2022113),Divyansh Kumar (IMT2022509), Ragh

15 May 2025

Project Title

Multimodal Visual Question Answering using the Amazon Berkeley Objects (ABO) Dataset

1 Data Curation

Dataset Source:

- Amazon Berkeley Objects (ABO) Dataset (Small Version, 3GB)
- 256×256 product images, multilingual metadata, 398,000 catalog images.

Approach:

- Selected ∼1000 product images across diverse categories (bags, shoes, containers, kitchen items).
- Generated 2–3 VQA pairs per image using Gemini 2.0 API with single-word answers.
- Ensured visual-only answerability.

Prompting Strategy:

- "What is the color of the item?"
- "What material is this product made of?"
- "What type of object is this?"

Tools Used: Gemini API, Pandas, Pillow, CSV, manual filtering scripts.

2 Baseline Evaluation

Model: `Salesforce/blip2-opt-2.7b`
Setup:

- Zero-shot inference
- Compared model outputs to ground truth

Metric: Token-level accuracy
Results:

- Accuracy: **41.2%**

Observations:

- Synonyms or plural mismatches reduced accuracy.
- Model often output phrases instead of expected single-word answers.

3 Fine-Tuning with LoRA

Objective: Improve baseline performance using LoRA (Low-Rank Adaptation).
Model: BLIP-2 (2.7B) with adapters in projection layers.
LoRA Configuration:

- $r = 8$, $\alpha = 16$, dropout $= 0.1$, epochs $= 3$, batch size $= 2$

Training Setup:

- Kaggle dual T4 GPU
- Hugging Face Transformers + PEFT

Results:

- Accuracy: **63.0%**
- F1 Score: 0.61
- BERTScore: 0.81

4 Evaluation Metrics

Metrics Used:

1. Accuracy – Exact token match
2. F1 Score – Handles class imbalance
3. BERTScore – Measures semantic similarity

Summary Table:

Model	Accuracy	F1 Score	BERTScore
Baseline (BLIP-2)	41.2%	0.39	0.74
LoRA Fine-tuned	**63.0%**	0.61	0.81

5 Iterative Improvement

- Cleaned Gemini-generated pairs to remove noisy labels.
- Balanced category distribution across object types.
- Adjusted dropout and learning rate for LoRA adapters.

6 Inference Script

`inference.py` supports single image and batch evaluation.

Example:

```
from inference import answer_question
print(answer_question("test_img.jpg", "What is the color?"))
# → "red"
```

7 Additional Contributions

- Integrated BERTScore-based semantic evaluation.
- Used LoRA to fine-tune 2.7B model under GPU constraints.
- Modular scripting for data, training, evaluation, and inference.

8 Github Repository Link

- Git Repo

9 References

- Gemini 2.0 API Documentation
- BLIP-2 Model Documentation
- Hugging Face LoRA PEFT
- Amazon Berkeley Objects Dataset