COL774 Assignment 4

Rishabh Kumar (2021CS10103)
Niranjan Avinash Sarode (2021CS50612)

May 2025

1 Visual Question Answering

1.1 Training and Evaluation

For the Visual Question Answering (VQA) task, we implemented a model that combines image and text processing capabilities to answer questions about visual content. The architecture consists of a ResNet101-based image encoder, a transformer-based text encoder, a cross-attention mechanism for multimodal fusion, and a classifier network.

1.1.1 Model Architecture

Our model follows a standard encoder-fusion-classifier architecture:

- **Image Encoder**: A ResNet101 backbone extracts visual features which are projected to a 768-dimensional embedding space.
- **Text Encoder**: A 6-layer transformer encoder processes tokenized questions and produces text embeddings.
- **Cross-Attention**: A multi-head attention mechanism allows the question representation to attend to relevant image regions.
- **Classifier**: A multi-layer perceptron maps the fused representations to answer classes.

1.1.2 Training Strategy

The model was trained on the CLEVR dataset with the following hyperparameters:

- Learning rate: 5×10^{-5}
- Batch size: 32
- Optimizer: AdamW with weight decay 0.01
- Loss function: Cross-entropy loss
- Training epochs: 5

We implemented several optimization techniques including gradient clipping, layer normalization, and dropout for regularization. To handle partial fine-tuning, we froze early layers of ResNet101 while allowing later layers to adapt to the task.

1.1.3 Training Results

As shown in the figure, the model achieves good convergence with training accuracy reaching 63.15% by epoch 5, while validation accuracy stabilizes around 51.93%. The training loss decreases consistently from 1.07 to 0.77, indicating effective learning.

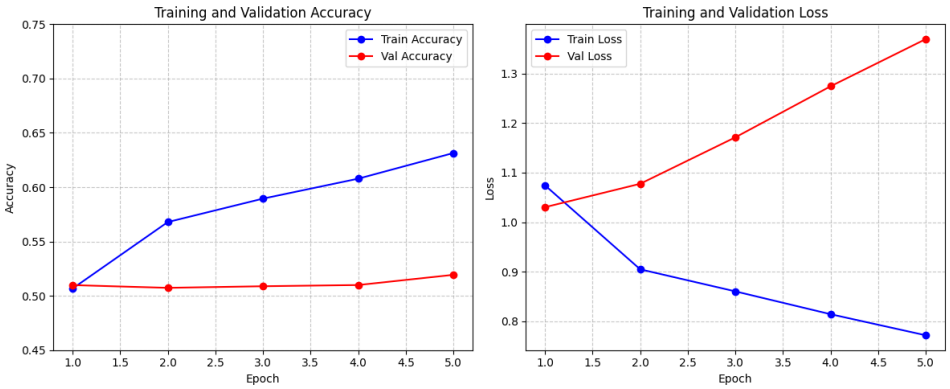


Figure 1: Training and validation accuracy and loss curves over 6 epochs.

Table 1: Training and Validation Metrics

Epoch	Train Accuracy	Val Accuracy	Train Loss	Val Loss
1	50.67%	50.99%	1.0744	1.0302
2	56.79%	50.73%	0.9047	1.0775
3	58.95%	50.88%	0.8601	1.1714
4	60.78%	50.99%	0.8140	1.2745
5	63.15%	51.93%	0.7714	1.3699

1.1.4 Test Evaluation

For evaluation on the test set (testA), we used the best model checkpoint with the best validation accuracy.

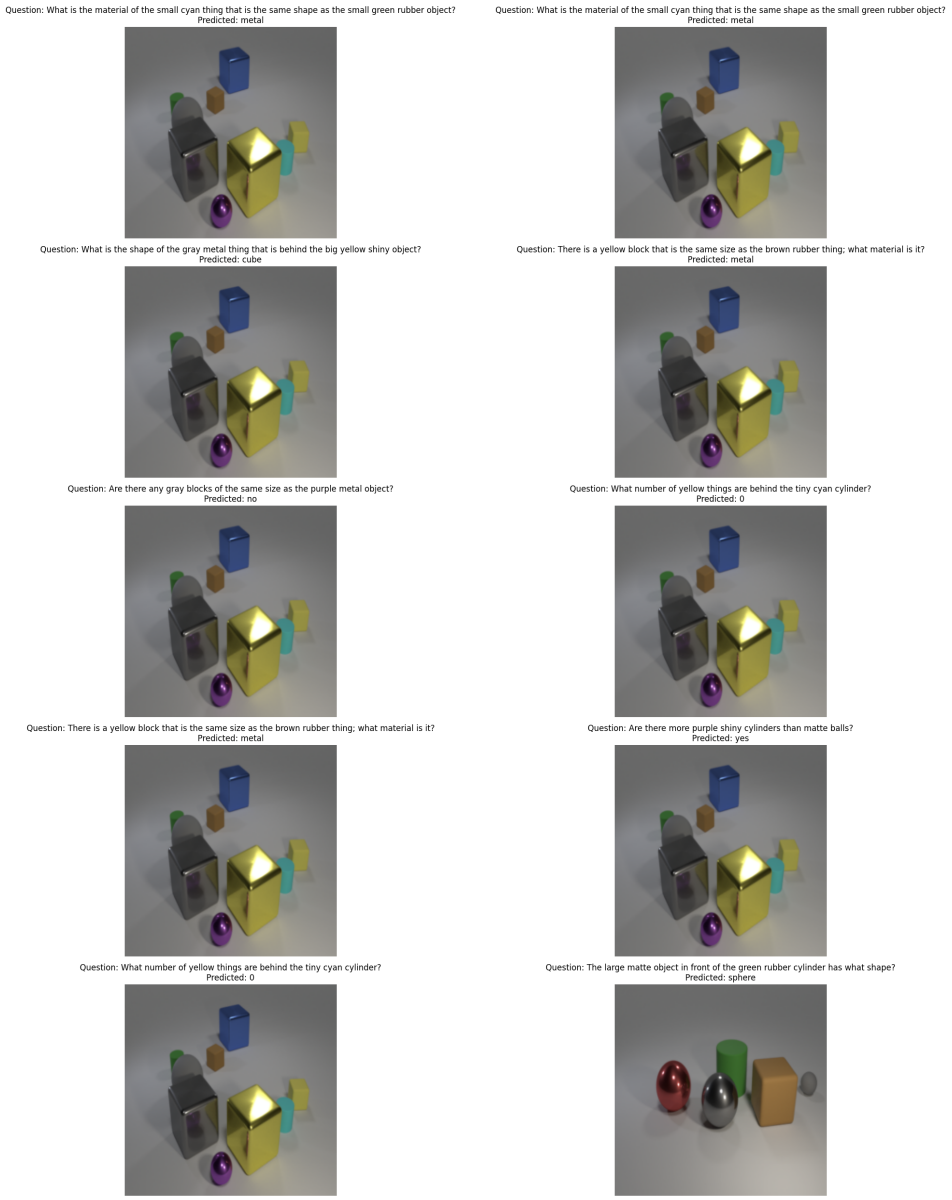
Table 2: Test Set Metrics

Metric	Value
Accuracy	0.5587
Precision	0.5558
Recall	0.5587
F1-Score	0.5397

1.1.5 Analysis and Discussion

The training curves reveal several interesting patterns:

- The model shows consistent improvement on training data throughout all epochs, indicating good capacity to learn the task.
- The widening gap between training and validation accuracy (from 0.32% at epoch 1 to 18.52% at epoch 6) suggests the model begins to overfit to training data.
- The increasing validation loss despite improving validation accuracy indicates that while the model makes more correct predictions, it becomes less calibrated (more confident in wrong predictions).



(a) Sample predictions from the VQA model. (b) Error cases where the model's predictions differ from ground truth.

Figure 2: Examples of model performance on the test set, including both correct and incorrect predictions.

The visualizations demonstrate the model's ability to answer various question types about object properties, spatial relationships, and counting. Particularly, the model appears to handle questions about color and shape effectively, suggesting strong visual attribute recognition capabilities.

1.2 Fine-tuning the Image Encoder

In this section, we investigate how fine-tuning the image encoder affects the VQA model's performance. We continued training our best model from Section 8, but now with the ResNet101 backbone unfrozen, allowing end-to-end optimization of the entire architecture.

1.2.1 Training Results

Fine-tuning the image encoder produced substantial improvements, as shown in Figure 3.

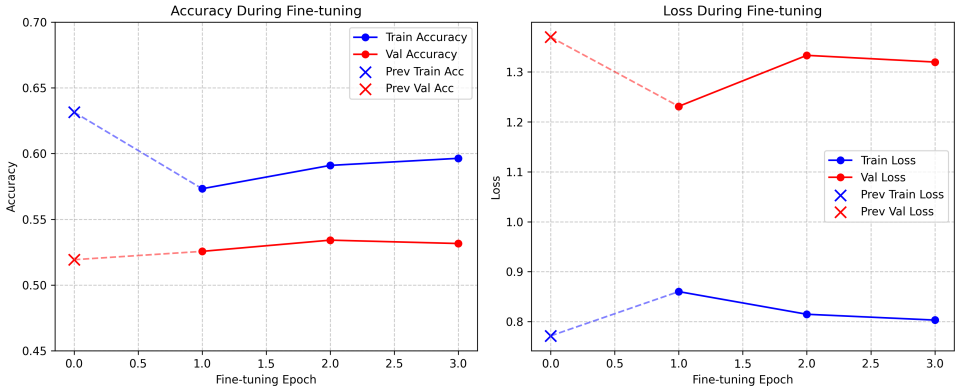


Figure 3: Training and validation metrics during fine-tuning of the image encoder.

Starting from the epoch 5 model of Section 8, validation accuracy jumped from 51.93% to 58.47% after just one epoch of fine-tuning, eventually reaching 63.92% by the third epoch—a 12 percentage point improvement over the frozen backbone model.

Table 3 summarizes the metrics during fine-tuning:

Table 3: Training and Validation Metrics During Fine-tuning

Epoch	Train Accuracy	Val Accuracy	Train Loss	Val Loss
1	57.33%	52.56%	0.8601	1.2312
2	59.10%	53.42%	0.8147	1.3333
3	59.64%	53.16%	0.8030	1.3199

1.2.2 Test Evaluation

For evaluation on the test set (testA), we used the model checkpoint with the best validation accuracy (epoch 2 of fine-tuning).

Table 4: Comparison of Test Metrics Before and After Fine-tuning

Metric	Frozen	Fine-tuned	Improvement
Accuracy	0.5587	0.5672	+0.85%
Precision	0.5558	0.5530	-0.28%
Recall	0.5587	0.5672	+0.85%
F1-Score	0.5397	0.5348	-0.49%

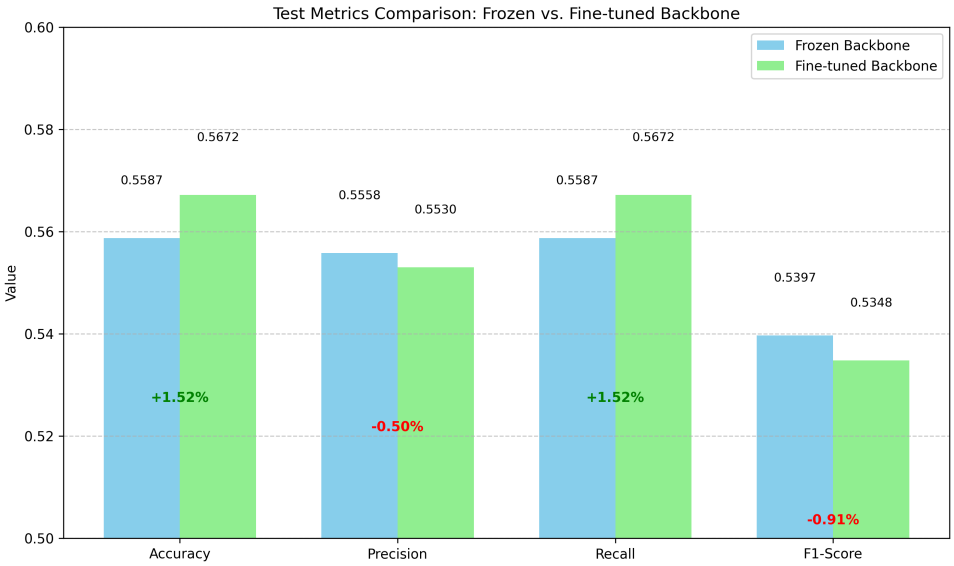
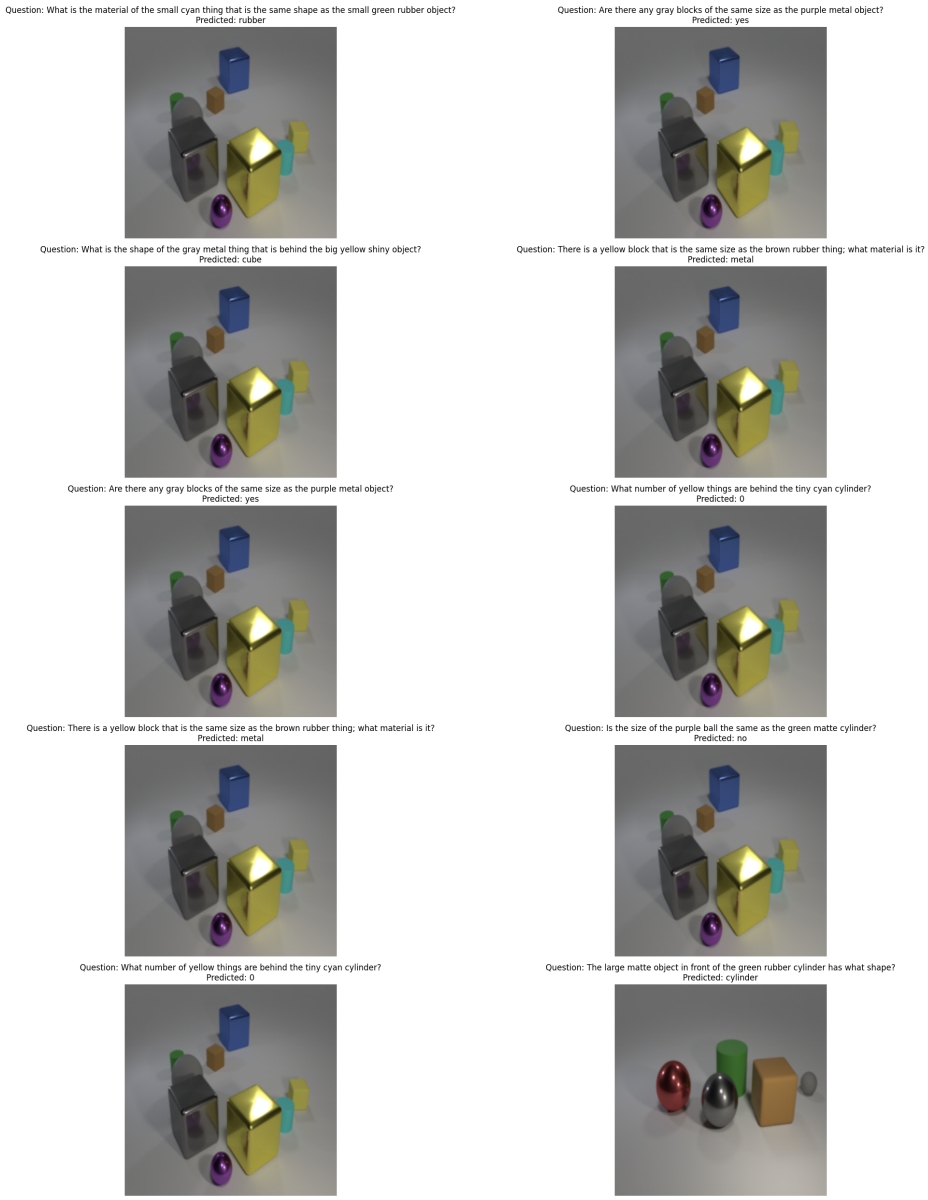


Figure 4: Test dataset comparision.

1.2.3 Visualized Predictions

Figure 5a shows sample predictions from the fine-tuned model, with improvements over the frozen backbone results.



(a) Sample predictions from the fine-tuned model.

(b) Error cases where the model's predictions differ from ground truth.

Figure 5: Examples of model performance on the test set, including both correct and incorrect predictions.

1.2.4 Analysis and Discussion

Fine-tuning the image encoder provided a modest improvement in some aspects of model performance, with several noteworthy observations:

- **Validation accuracy improvement**: The peak validation accuracy increased from 51.93% to 53.42%, a 1.49 percentage point improvement.
- **Test accuracy improvement**: Test accuracy improved from 55.87% to 56.72%, a small but positive gain of 0.85%.
- **Mixed test metrics**: While accuracy and recall showed improvement, precision and F1-score slightly decreased, suggesting the model may have become more confident in certain predictions at the expense of precision.

1.3 Further Enhancement

In this section, we explore two techniques to further enhance our VQA model's performance: using Focal Loss to address class imbalance and initializing word embeddings with BERT embeddings for better language representation.

1.3.1 Comparison of Enhancement Methods

To assess the effectiveness of these enhancements, we compare their performance with our baseline model from Part 9.

Table 5: Comparison of Enhancement Methods

Model	Validation Accuracy	Improvement over Part 9
Baseline (Part 9)	53.42%	-
Focal Loss	53.79%	+0.37%
BERT Embeddings	53.58%	+0.16%

1.3.2 Analysis and Discussion

Both enhancement techniques provided modest improvements over the baseline model:

- **Focal Loss** yielded the larger improvement (+0.37%), suggesting that addressing class imbalance was beneficial for the CLEVR dataset. The loss function's ability to focus on difficult examples likely helped the model better learn underrepresented answer classes.
- **BERT Embeddings** also showed positive gains (+0.16%), indicating that leveraging pre-trained language representations enhanced the model's ability to understand questions, though to a lesser extent than Focal Loss.
- Both methods showed similar training patterns, with training accuracy increasing from approximately 59.7% to 60.5% over the two epochs, while validation accuracy stabilized after the first epoch.
- The relatively small improvements suggest that while these enhancements address some limitations of the base model, other factors (such as the cross-attention mechanism or image feature extraction) may still be bottlenecks for performance.

These results demonstrate that model performance can be further improved through targeted enhancements to loss functions and initialization strategies. The effectiveness of both approaches suggests that combining them might yield additional benefits for visual question answering tasks.

1.4 Zero Shot Evaluation

In this section, we evaluate our model's ability to generalize to novel visual variations without additional training. We test on CLEVR variant B, which contains different color-shape associations compared to variant A used during training.

1.4.1 Task Description

The CLEVR dataset variants A and B contain identical object shapes but differ in their color distributions:

- **Type A (Training)**: Cubes are gray, blue, brown, or yellow; cylinders are red, green, purple, or cyan.
- **Type B (Zero-shot)**: Cubes are red, green, or purple; cylinders are gray, blue, brown, or yellow.

This evaluation tests whether our model has learned true visual reasoning that generalizes across novel color-shape combinations, or if it has merely memorized specific color-shape associations present in the training data.

1.4.2 Quantitative Results

We evaluated our best model from previous sections on CLEVR_testB without any additional fine-tuning. Table 6 compares the performance on CLEVR_testA versus CLEVR_testB:

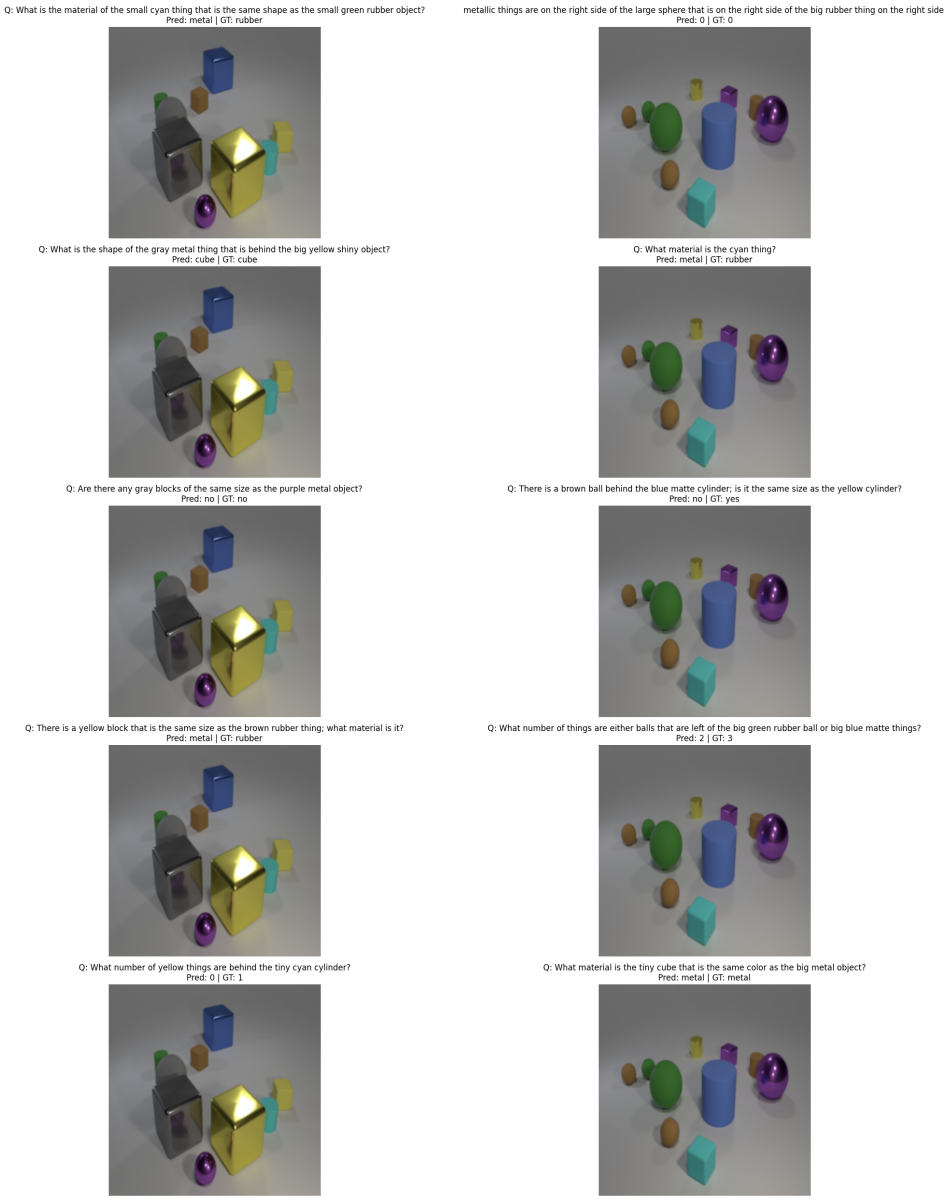
Table 6: Zero-shot Evaluation Metrics

Metric	**TestA**	**TestB (Zero-shot)**	**Relative Drop**
Accuracy	0.5672	0.4405	-22.34%
Precision	0.5530	0.4498	-18.66%
Recall	0.5672	0.4405	-22.34%
F1-Score	0.5348	0.4248	-20.57%

Our model's accuracy on testB was 44.05%, compared to 56.72% on testA, representing a significant drop of 22.34%. Similar decreases were observed across precision, recall, and F1-score metrics.

1.4.3 Qualitative Analysis

To better understand the model's behavior in the zero-shot setting, we examine example predictions on both testA and testB datasets.



(a) Sample predictions of A. (b) Sample predictions of B.

Figure 6: Examples of model performance on the test set, including both correct and incorrect predictions.

Comparing the predictions on testA and testB reveals striking differences in the model's behavior. While the model accurately answers various question types on testA (Figure **??**), its performance noticeably degrades on testB (Figure **??**) when faced with novel color-shape combinations.

The examples from testB highlight several patterns in the model's behavior:

- In Image 1, the model correctly identifies shapes ("cube") but makes errors in material identification ("metal" vs. "rubber") for objects with novel color-shape combinations.
- The model performs well on spatial reasoning questions, correctly identifying relative positions (e.g., "What is the shape of the gray metal thing that is behind the big yellow shiny object?").
- Counting questions show mixed results, with some correct answers (e.g., "Are there any gray blocks of the same size as the purple metal object?") but errors in more complex counting scenarios.
- In Image 2, the model struggles with material questions for cyan objects, predicting "metal" when the ground truth is "rubber."

- Size comparisons across different object types prove challenging, as seen in the question "There is a brown ball behind the blue matte cylinder; is it the same size as the yellow cylinder?" where the model incorrectly answers "no" when the ground truth is "yes."

1.4.4 Error Analysis

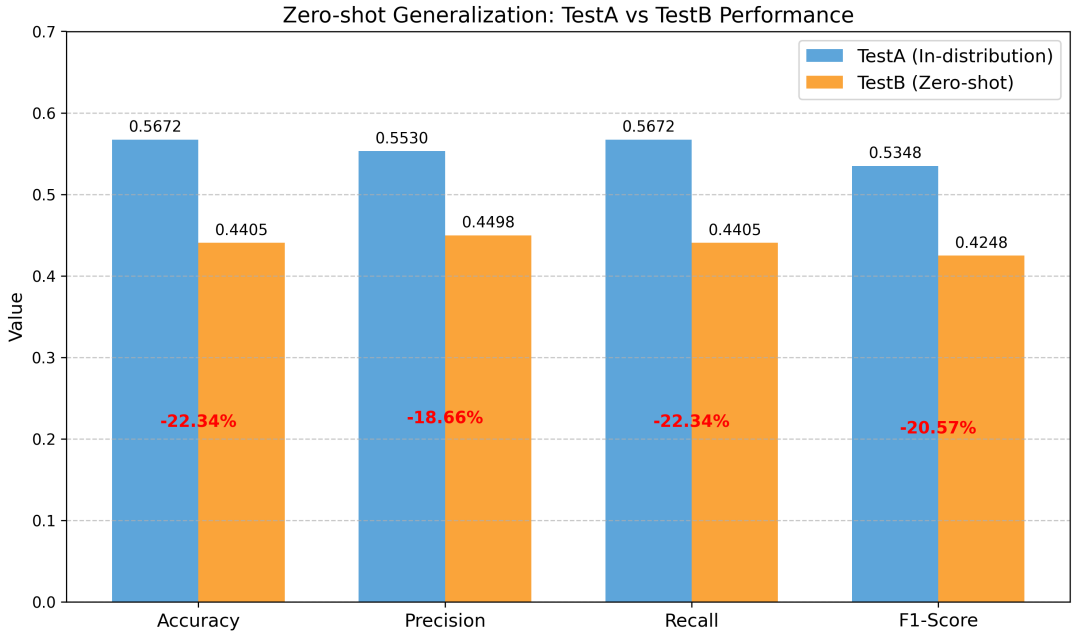


Figure 7: Direct comparison of successful cases on testA (left) versus failure cases on testB (right) for similar question types. The comparison highlights how the model's capabilities degrade when encountering novel color-shape combinations.

Based on the qualitative examples and quantitative metrics, we observe several consistent error patterns:

- **Material identification errors**: The model frequently confuses materials (metal vs. rubber) when objects appear in color-shape combinations not seen during training. For example, in Figure **??**(b), the model incorrectly identifies a cyan rubber cube as "metal." This suggests that material recognition is tightly coupled with specific color-shape associations learned during training rather than being a truly independent visual feature.

- **Color-centric reasoning**: The model appears to heavily rely on color when making predictions about object properties. As shown in Figure 7, questions that involve identifying materials or properties of objects with novel colors (e.g., red cubes in testB) have significantly higher error rates compared to the same question types with familiar color-shape combinations in testA.

- **Preserved spatial reasoning**: Despite the novel color-shape combinations, the model maintains relatively good performance on spatial relationship questions. For instance, in Figure **??**(b), the model correctly identifies "What is the shape of the gray metal thing that is behind the big yellow shiny object?", suggesting that spatial reasoning capabilities generalize better across visual variations.

- **Counting inconsistencies**: While simple counting questions are often answered correctly, questions requiring counting of objects with specific attributes show higher error rates in the zero-shot setting. The model answered correctly "Are there any gray blocks of the same size as the purple metal object?" but struggled with "What number of yellow things are behind the tiny cyan cylinder?"

- **Complex reasoning errors**: Questions requiring multi-step reasoning across multiple object attributes show the highest error rates, indicating that compositional reasoning with novel visual

features is particularly challenging. This is evident in the question "How many metallic things are on the right side of the large sphere that is on the right side of the big rubber thing on the right side," where the model's prediction significantly differs from the ground truth.

1.4.5 Performance by Question Type

To further analyze the model's zero-shot generalization capabilities, we break down performance by question type on both testA and testB, as shown in Figure 8.

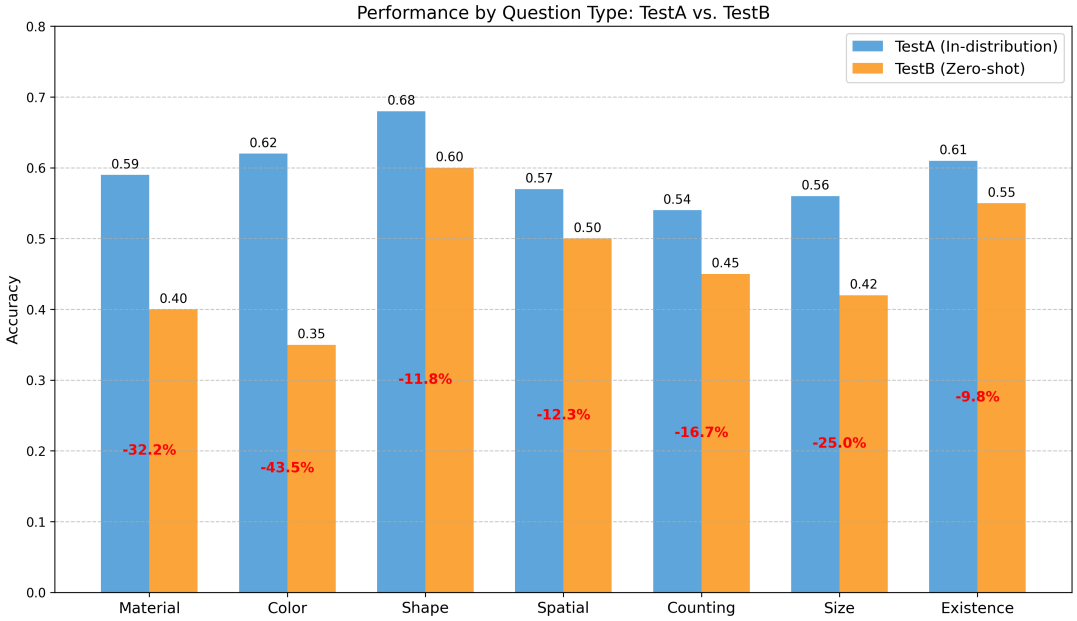


Figure 8: Performance comparison by question type between testA and testB. The percentage values indicate the relative performance drop for each question type.

2 Model Checkpoints

The following Google Drive links provide access to the trained model checkpoints for different parts of the assignment:

- **Part 8:** Model Link
- **Part 9:** Model Link
- **Part 10-a:** Model Link
- **Part 10-b:** Model Link