

UE23CS352A: ML 5th Semester Section A

GRE: Evaluating Computer Vision Models on Generalizability Robustness and Extensibility

MOTIVATION:

Modern Vision Language Models such as BLIP exhibit image understanding and reasoning capabilities. However, their generalizability, robustness and extensibility when applied to synthetic data remained underexplored

This work evaluates BLIP models on simplified visual Questions

Answering (VQA) tasks to study how limited or synthetic data can affect models reasoning, coherence and visual comprehension

Generalizability: Introduce objects from the same class

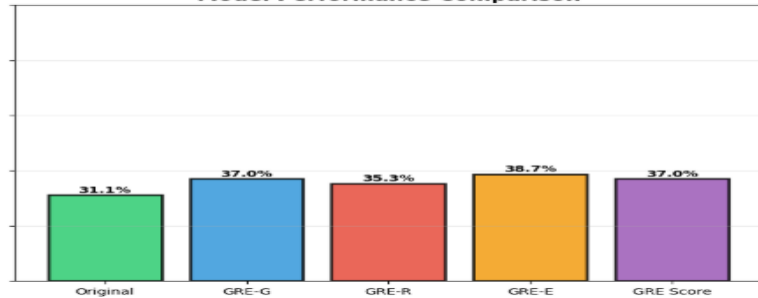
Robustness: Introduce a new scene for the object (background change)

Extensibility: Introduce and object from different class

These above can be done by

1) Object Masking 2) Object Overlay 3) Object Removal

Model Performance Comparison



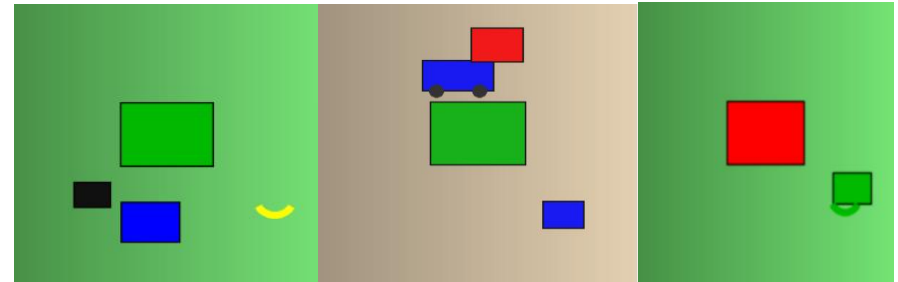
The pipeline:

Vision Encoder: Extracts Spatial semantics features from image

Text Encoder: Encodes the question as contextual embeddings

Cross Attention: Merges Both modalities to infer relationships

Decoder: Generates a textual answer aligned with visual case



Dataset used:

Synthetic dataset (custom built due to GPU and Data Limits)

Benchmark Referenced : VQA, VQA-G, VQA-R, and VQA-E.

GRE-style evaluation dataset to test linguistic and logical robustness.

Around 1000 images zero shot on Salesforce BLIP model with 300 split with GRE subsets

