GRE Benchmark for VQA Models

Evaluate Visual Question Answering models on **Generalizability, Robustness, and Extensibility**.

□ Quick Start (Google Colab)

1. Setup

```
# Install dependencies
!pip install transformers torch pillow numpy

# Upload files to Colab
# Upload: config.py, dataset_generation.py, gre_transformations.py, evaluator.py, main.py
```

2. Test Compatibility (Optional but Recommended)

```
# Quick test to verify everything works
!python test_compatibility.py
```

3. Run Benchmark

```
# Execute main pipeline
!python main.py
```

That's it! The script will:

- □ Generate 1000 synthetic VQA samples
- □ Apply GRE transformations
- □ Evaluate BLIP model
- □ Save results to `results/metrics.json`

□ What Gets Evaluated

Transform	What Changes	Tests
Base	Original samples	Baseline performance
G (Generalizability)	Object color	Attribute invariance

Transform	What Changes	Tests
R (Robustness)	Scene background	Context independence
E (Extensibility)	Object category	Concept transfer

□ Output Structure

```
data/
├── images/
│   ├── train/
│   ├── val/
│   ├── test/
│   └── gre_G/, gre_R/, gre_E/
└── annotations/
    └── *.json

results/
└── metrics.json
```

□ Configuration

Edit `config.py` to change:

- Dataset size: `TOTAL_SAMPLES = 1000`
- GRE subset: `GRE_SUBSET = 300`
- Model: `MODEL_NAME = "Salesforce/blip-vqa-base"`

Understanding Results

```
{
  "Base": {"accuracy": 0.85},
  "G": {"accuracy": 0.78},
  "R": {"accuracy": 0.80},
  "E": {"accuracy": 0.72}
}
```

GRE Retention = (Average of G, R, E) / Base × 100%

Higher retention = better compositional generalization!

Fast Mode (For Testing)

```
# In config.py, change:
TOTAL_SAMPLES = 100
GRE_SUBSET = 30
```

Citation

https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26603876.pdf (https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26603876.pdf) https://visualqa.org/download.html (https://visualqa.org/download.html) https://arxiv.org/pdf/2201.12086 (https://arxiv.org/pdf/2201.12086)