# Impact of Dataset Size and Distillation Techniques on Image Captioning Performance: An Empirical Study

Srushti Sangawar
Arunava Ghosh

University of Colorado Boulder

## An empirical evaluation of random sampling and gradient-based distillation methods across multiple models and dataset sizes.

### *Introduction*

This project explores the impact of dataset size, distillation strategies, and pre-trained architectures on image captioning performance. It focuses and compares models based on ResNet-50, GIT, CLIP and GPT-2 architectures. Experiments are conducted at different dataset sizes using the MSCOCO dataset. Model performance is assessed using metrics like BLEU-1 to BLEU-4 and CIDEr. Unlike most studies that rely on massive data and computation, we systematically compare model types, dataset sizes, and distillation methods on equal footing.

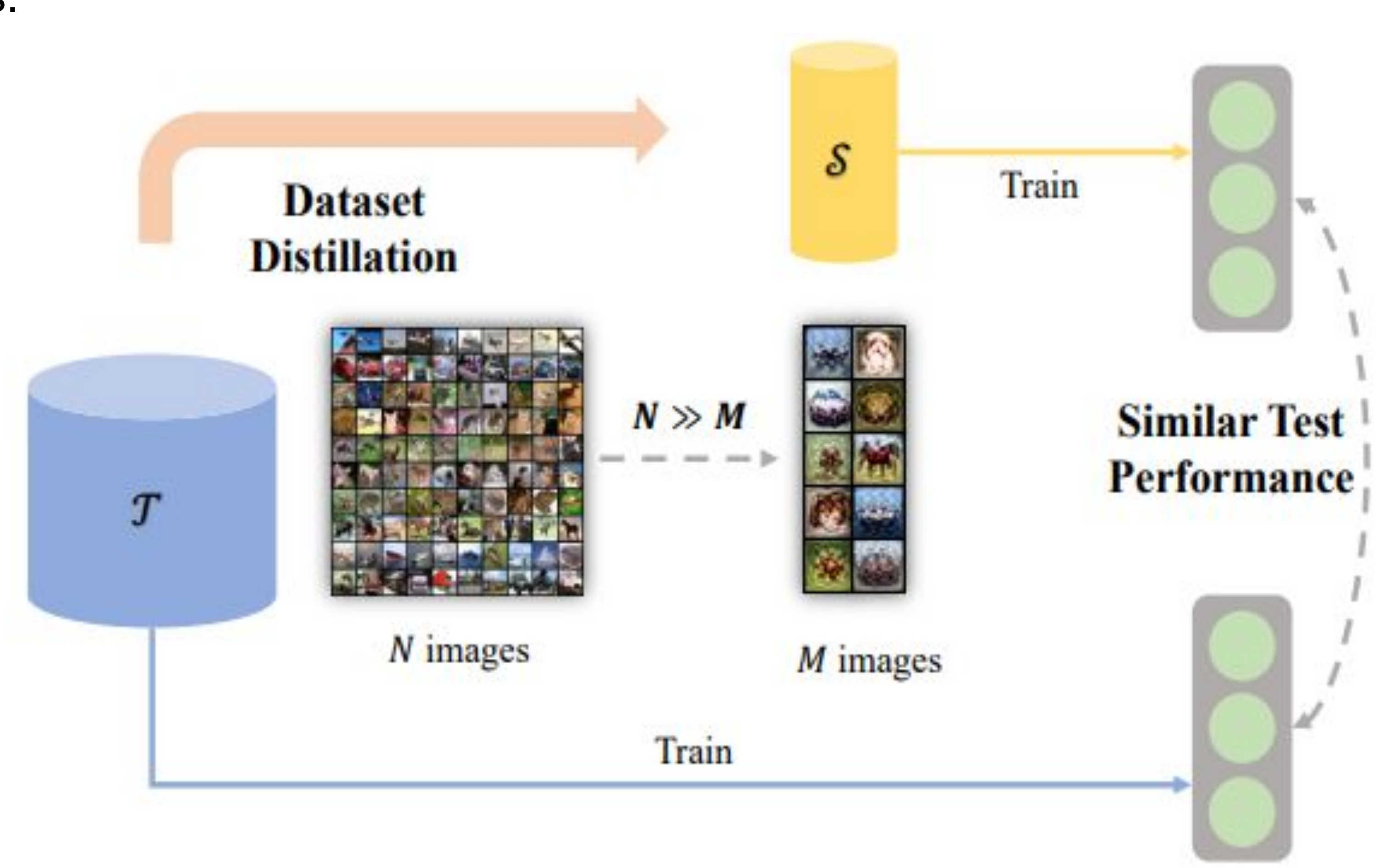
### *Methodology*

We use a 5K image subset of the MSCOCO dataset, scaling the training data to 25%, 50%, 75%, and 100% to study the effect of dataset size on captioning performance.

Two methods are used:

- Random Sampling: choosing a subset of images at random. Used as a naive baseline for comparison
- Gradient-Based Distillation: prioritizing informative examples by selecting images with the highest per-sample gradient norms, similar to performance matching

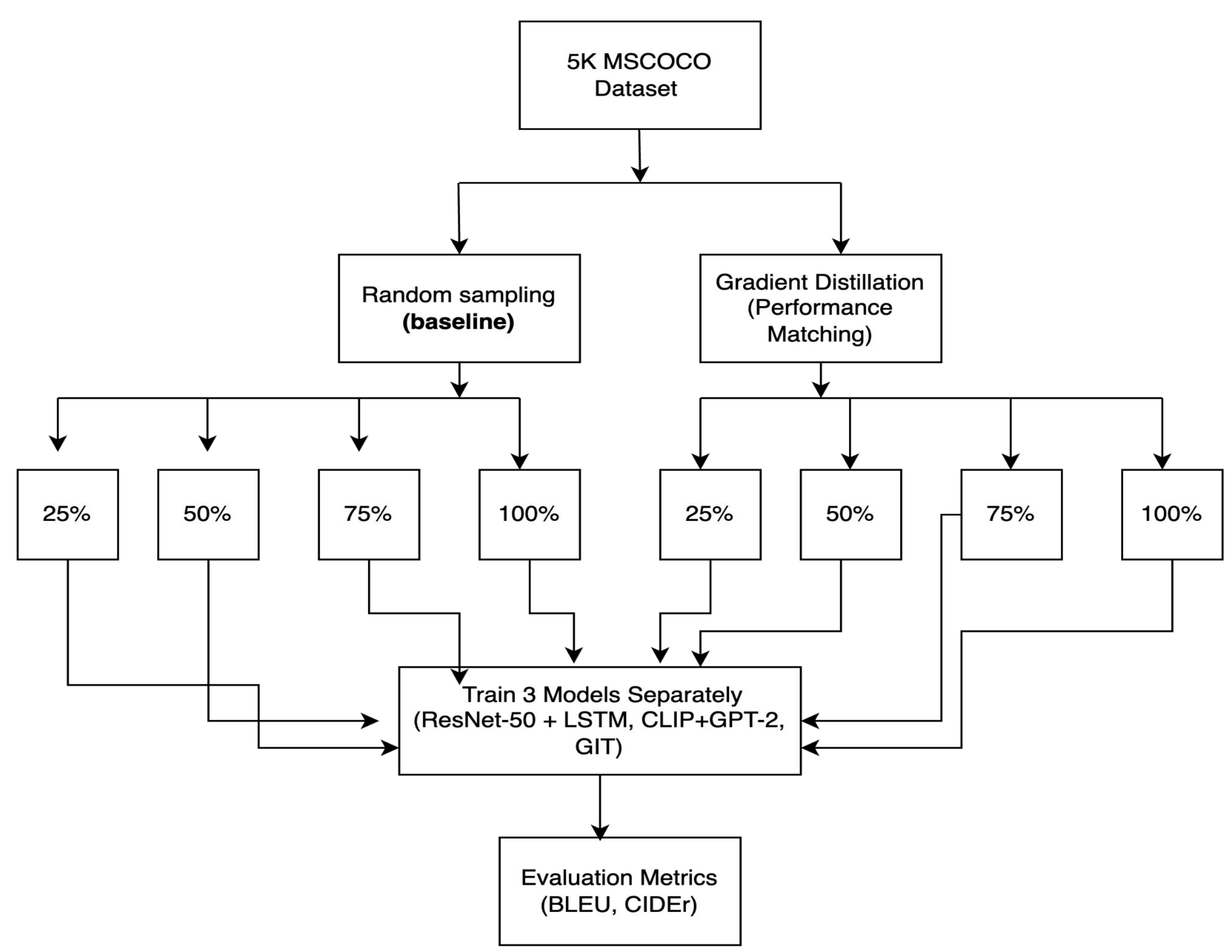
The Adam optimizer and cross-entropy loss are used during the 20 epochs of training, with early stopping determined by trends in validation loss.

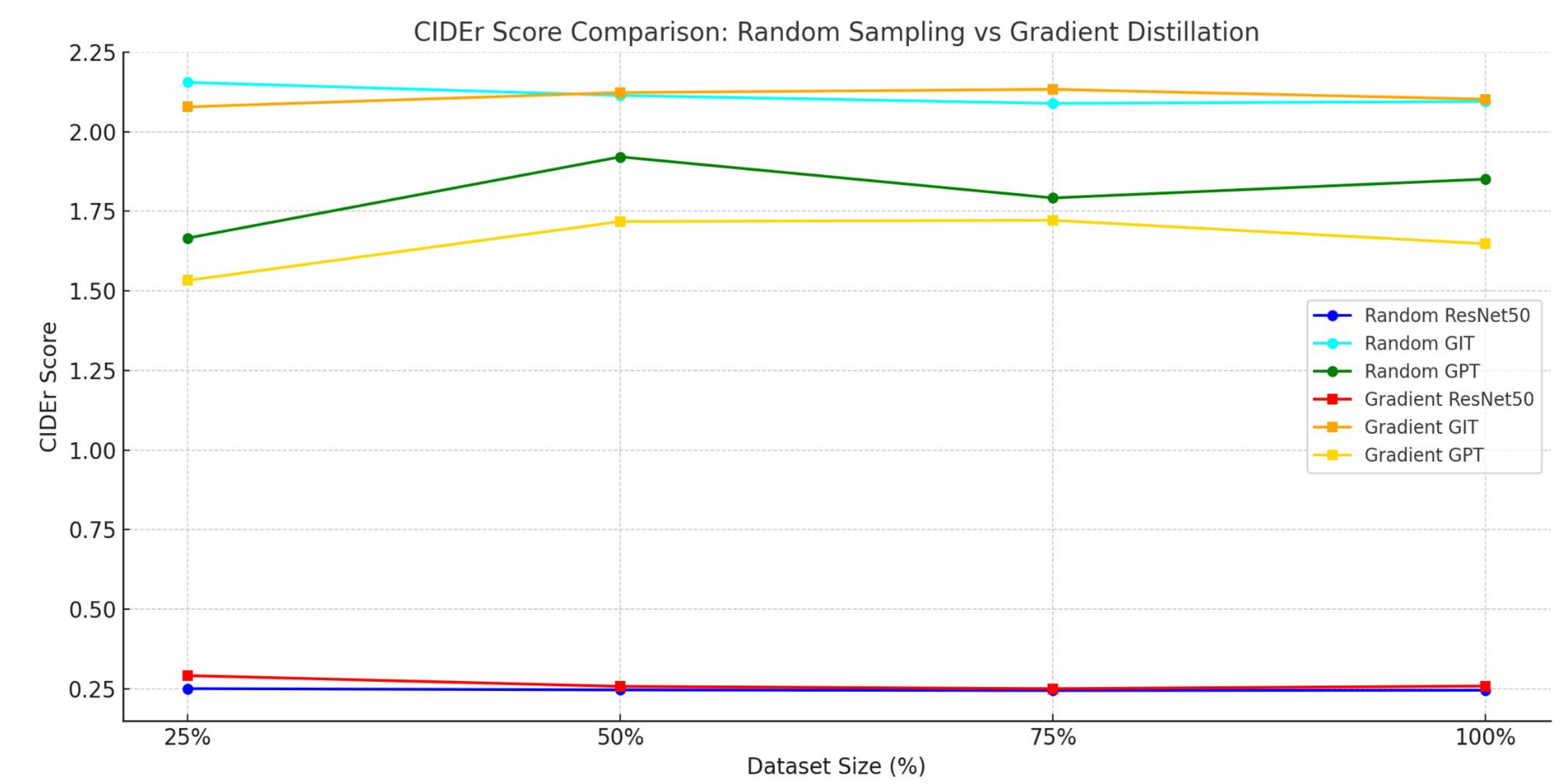


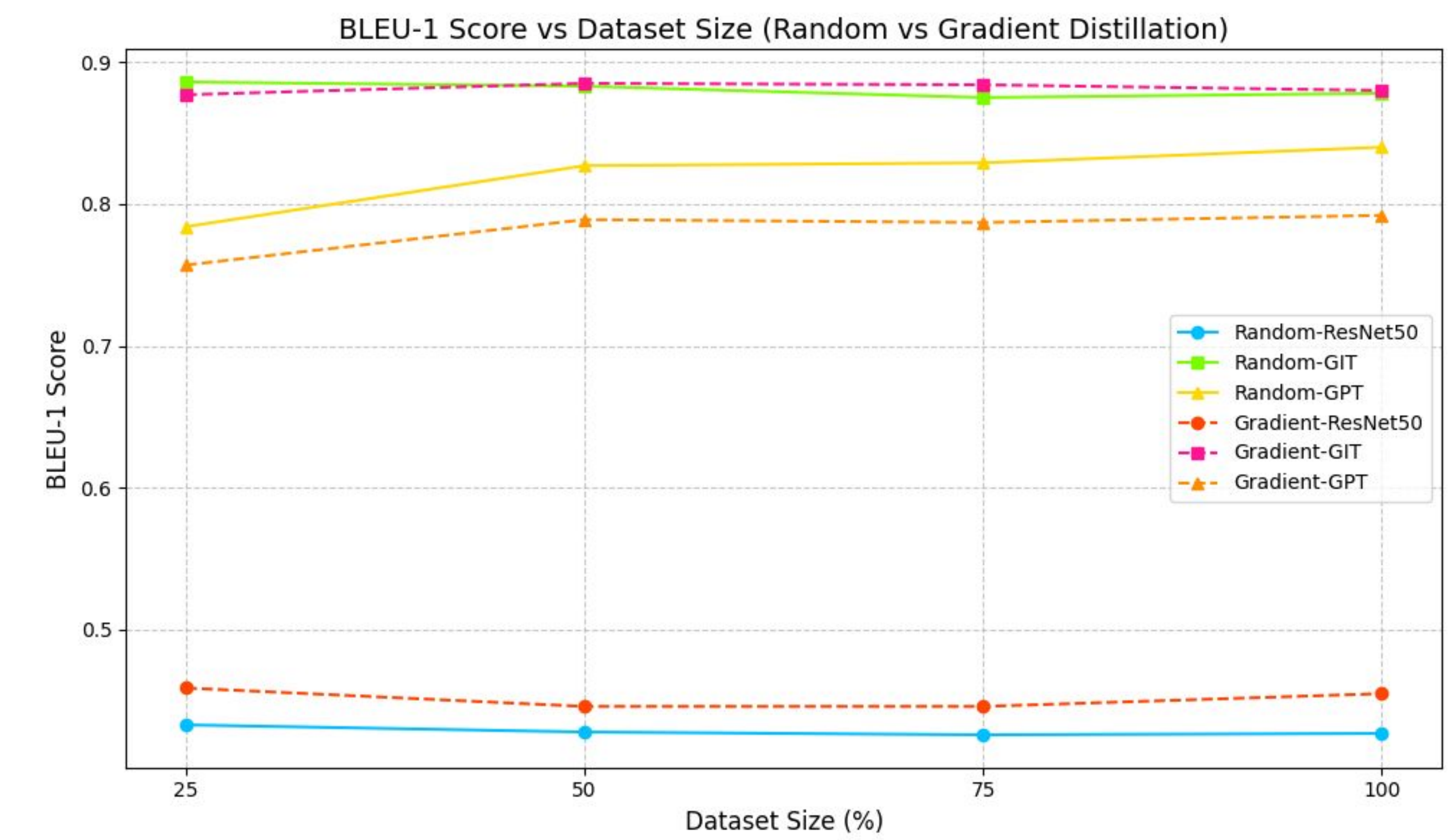

An overview showing the dataset distillation process. Source.

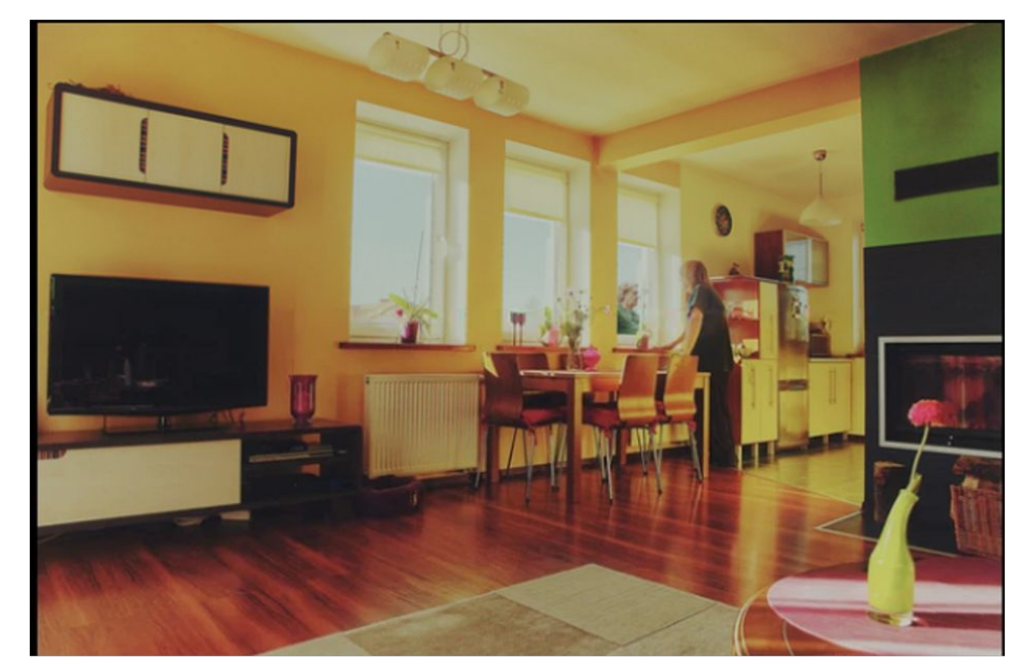
### *Comparison of Image Captioning Architectures*

| Feature | ResNet-50 + LSTM | GIT Model | CLIP + GPT-2 |
|---|---|---|---|
| **Image Encoder** | ResNet-50 (pretrained CNN) | Transformer-based Vision Encoder (ViT-like, BERT-style pretraining) | CLIP Pretrained Vision Encoder |
| **Text Decoder** | LSTM RNN | Multimodal Transformer Decoder | GPT-2 Transformer Decoder |
| **Image-Text Connection** | Image features fed as initial LSTM input | Unified vision-language model | Linear projection maps image features to GPT-2 space |
| **Training Type** | Train LSTM on top of fixed ResNet-50 features | Fine-tune full model (encoder + decoder) | Fine-tune GPT-2 with learnable image token |
| **Pretraining Used** | ResNet-50 ImageNet pretrained | Pretrained on large image-text datasets | CLIP pretrained on image-text pairs + GPT-2 pretrained |
| **Strength** | Simple and fast; Good with small data | Strong multimodal understanding | Leverages strong language modeling capabilities |
| **Weakness** | LSTM less powerful than Transformers | Heavy model; More compute needed | Needs strong image embedding quality |



*Sample Test Image*

- **Resnet-50:** image of a fireplace in a table with a table and a fireplace and a
- **GIT:** a woman stands in the dining area at the table
- **CLIP**: A window and home with yellow

### *Insights*

- Model choice (GIT≫GPT≫ResNet) dominates performance.
- Most gains happen before 75% dataset size — more data after that gives diminishing returns.
- Model architecture matters more than dataset distillation or size.

### *References:*

- Ruonan Yu, Songhua Liu, X.W.: Dataset distillation: A comprehensive review (2023)
- Raphael Gontijo Lopes, T.S.: Data-free knowledge distillation for deep neural net works(2017)
- S. Bengio, D.E.: Show and tell: A neural image caption generator (2015)