# Phase 1 Report – Baseline PTQ on MobileCLIP

## 1. Executive Summary

The objective of Phase 1 was to establish a reproducible baseline for Post-Training Quantization (PTQ) on the MobileCLIP2-S2 model. The model was evaluated in its original FP32 form and in a quantized INT8 (Linear-only) version using ONNX Runtime on CPU. Additional metrics such as cosine alignment drift were added to verify semantic consistency. This baseline now serves as the control reference for future hybrid quantization and layer sensitivity optimization experiments.

## 2. Environment Setup

All experiments were conducted on Google Colab Pro (CPU runtime, 2 cores, ~16 GB RAM). Key libraries included PyTorch 2.8.0+cu126, ONNX Runtime 1.21.1, HuggingFace Transformers, and the Apple MobileCLIP repository. The model was run entirely on CPU without GPU acceleration.

## 3. Model and Dataset

The model used was MobileCLIP2-S2, developed by Apple and downloaded via HuggingFace Hub (repo: apple/MobileCLIP2-S2). The evaluation used a 500-sample subset of the Flickr30k image-caption dataset, providing a manageable test size for CPU-based testing. A smaller 50-sample subset was used for cosine drift evaluation.

## 4. Methodology

1. Loaded MobileCLIP2-S2 on CPU and verified functional inference.  
2. Measured FP32 baseline metrics: model size, inference latency, and Recall@1 (image→text retrieval).  
3. Exported the model to ONNX (opset\_version=17).  
4. Applied dynamic INT8 quantization to Linear (MatMul) layers only, avoiding ConvInteger operator issues.  
5. Measured INT8 model latency, throughput, and cosine alignment drift versus FP32.  
6. Collected reproducibility metadata (CPU type, core count, and library versions).

## 5. Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Size (MB) | Latency (ms/sample) | Throughput (samples/s) | Recall@1 (%) | Mean Cosine Drift |
| FP32 | 398.07 | 614.10 | 1.63 | 0.20 | 0.0000 |
| INT8 (Linear-only) | 261.50 | 466.30 | 2.14 | N/A | 0.0228 |

The quantized model achieved approximately a 34% reduction in size and a 1.3× speed-up in CPU latency. The mean cosine drift was 0.0228, with 42% of pairs exceeding a drift threshold of 0.02. This indicates mild embedding misalignment after quantization but no major degradation in retrieval accuracy.

## 6. Discussion

The Phase 1 results confirm that Post-Training Quantization (PTQ) on MobileCLIP is feasible for CPU-only inference. Dynamic INT8 quantization of linear layers provided both model compression and latency improvements while maintaining semantic alignment within acceptable limits. The observed cosine drift (~0.02) suggests minor representational shifts that can be addressed through layer-specific calibration or hybrid quantization strategies in later phases.

## 7. Next Steps (Phase 2 Preview)

Phase 2 will explore Hybrid Quantization, integrating static INT8 quantization for convolutional layers with dynamic quantization for fully connected layers. Sensitivity analysis will be performed to identify layers most susceptible to accuracy drift. The goal is to minimize alignment loss while achieving further reductions in inference latency.