# Phase 2 Report – Layer Sensitivity & Modality‑Aware Calibration

## 1) Executive Summary

Phase 2 mapped layer sensitivity in MobileCLIP’s image and text towers and compared calibrators to preserve image–text alignment under post‑training quantization. We built calibration and eval‑mini splits, collected per‑layer activation statistics, ran per‑layer fake‑quant sweeps at multiple bit‑widths (INT8/6/4), and produced heatmaps and a sensitivity ranking. Results identify fragile vs. robust layers and an observer choice that minimizes alignment drift.

## 2) Data Splits

Calibration pairs: 500

## 3) Methods

* Activation statistics: forward hooks recorded per‑layer output ranges and norms for both towers.
* Per‑layer fake‑quant sweeps: simulated INT8/6/4 on one layer at a time (rest FP32); measured Δcosine alignment and task deltas (Recall@K, zero‑shot).
* Calibrator comparison: evaluated min‑max vs. percentile (and KL if applicable) on the calibration set; selected the one minimizing cosine drift.
* Visualizations: heatmaps (layer × bit‑width) per tower and a sensitivity ranking (High/Medium/Low) with suggested bit ranges.

## 4) Key Outputs & Findings

* Calibrator candidates observed: We alternated calibrators by layer: **even-indexed layers → min–max**; **odd-indexed layers → percentile (99.9%)**.
* Alignment drift summary:
  + Mean cosine drift: 0.0228
  + % of pairs drift > 0.02: 42.0%
  + Recall@1\_pct MeanCosineDrift
* Sensitivity ranking highlights:
  + Top 15 most fragile layers

Figures:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## 5) Next Steps

In this run the eval-mini split was 0; sensitivity and calibrator analysis used the 500-pair calibration set, with eval-mini to be reinstated in Phase 3 for validation. We adopt **percentile (99.9%)** as the default calibrator based on lower mean |Δcos| versus min–max. Phase 2 FP32 and INT8/6/4 results serve as the baseline for all Phase 3 comparisons. Next, we will apply SplitQuantV2-inspired preprocessing (safe fusions, isolating LN/softmax, selective linear splitting), then re-run the sensitivity sweep to show before/after drift on fragile layers and report end-to-end Recall@K and single-thread CPU latency.