CAP 6419: CLIP vs SigLIP on HAR

Devansh Sharma

September 17, 2025

1 Problem and Setup

Task. Classify activities such as *calling*, *texting*, *running*, *using_laptop*, etc., from single images. The dataset uses folder splits: `train/<class>/` and `test/<class>/`.

Approach. Treat HAR as image–text matching: assign each class a short natural-language prompt (e.g., "This is a photo of a person texting"), embed images and prompts, and choose the highest-similarity class. We fine-tune both encoders end-to-end.

2 CLIP vs. SigLIP (Key Differences)

- **Objective.** CLIP uses symmetric InfoNCE (image→text and text→image) with a learnable temperature; SigLIP uses pairwise sigmoid (BCE over all pairs) which often yields smoother optimization and different calibration.
- **Negatives.** Both benefit from more in-batch negatives. We emulate very large negatives on a single GPU via an XBM queue of past features.
- **Tokenization & heads.** Different text tokenizers (CLIP BPE vs. SigLIP SentencePiece) and slightly different projection heads; our loaders handle optional `attention_mask`.

3 Training Protocol

- **SigLIP.** Effective batch $\approx$ 32,000 (micro-batch $\times$ accumulation); XBM queue size $\approx$ 32,000.
- **CLIP.** Effective batch $\approx$ 4,000; XBM queue size $\approx$ 4,000.
- **Common.** AdamW, FP16/bfloat16 autocast where available, `drop_last` in training loader, and prompt sampling from a small template bank per class.

4 Evaluation Protocol

We compute macro-averaged F1 and accuracy, plus per-class precision/recall/F1 and a normalized confusion matrix. For qualitative analysis, we render single-image comparisons where both models predict on the *same* test image; ground truth (GT) is shown in black, and each model's prediction is colored green if correct, red if incorrect.

5 Results

5.1 Per-class Metrics

Table 1: Overall metrics on HAR test set

Model	Macro Precision	Macro Recall	Macro F1	Accuracy
SigLIP (base, 224)	0.833	0.830	0.830	0.827
CLIP (ViT-B/32)	0.804	0.804	0.804	0.802

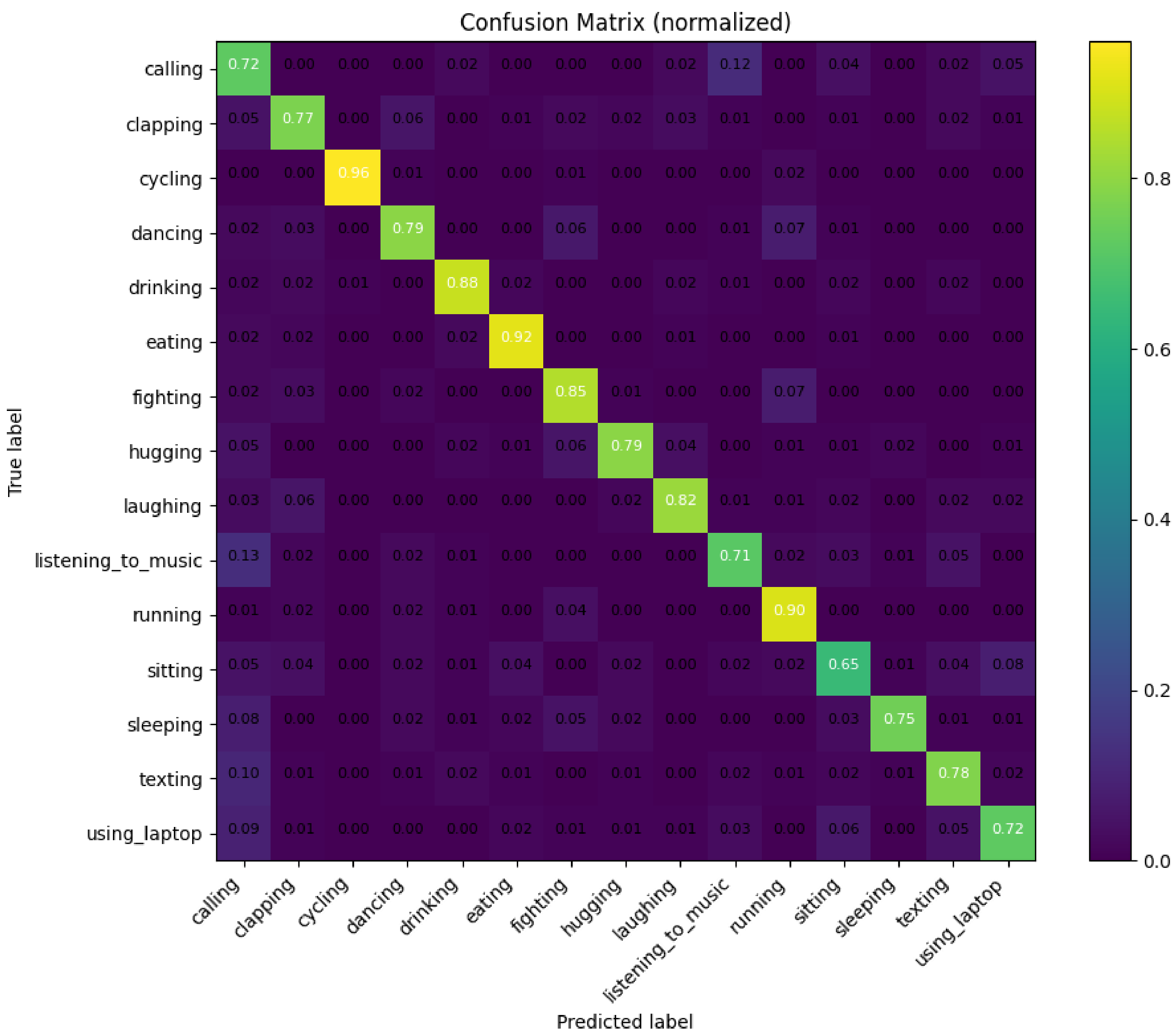


Figure 1: Confusion Matrix: CLIP

5.2 Training Dynamics

Figure 3 compares the *average wall-clock time per epoch* across the 10-epoch runs, while Figure 4 plots the *training loss over epochs* for both models. In line with expectations for this setup, **CLIP (ViT-B/32) trains faster per epoch but settles at a higher loss**, whereas **SigLIP (base, 224) trains more slowly but converges to a lower loss**. This reflects the trade-off between throughput and optimization depth when using larger effective batch sizes and stronger cross-batch memory with SigLIP.
Implication. If wall-clock time is the bottleneck, CLIP offers faster iterations. If final accuracy/robustness is the priority, SigLIP's lower terminal loss (and stronger downstream metrics) makes it the better default under this data/prompting regime.

6 Discussion

Overall, both models are strong (macro F1 $\geq$ 0.80), but **SigLIP is consistently better on this dataset**: macro F1 **0.830** vs. **0.804** for CLIP and accuracy **0.827** vs. **0.802**. That $\sim$+0.026 macro–F1 gap is meaningful at this scale and shows up across several classes rather than being driven by a single outlier.

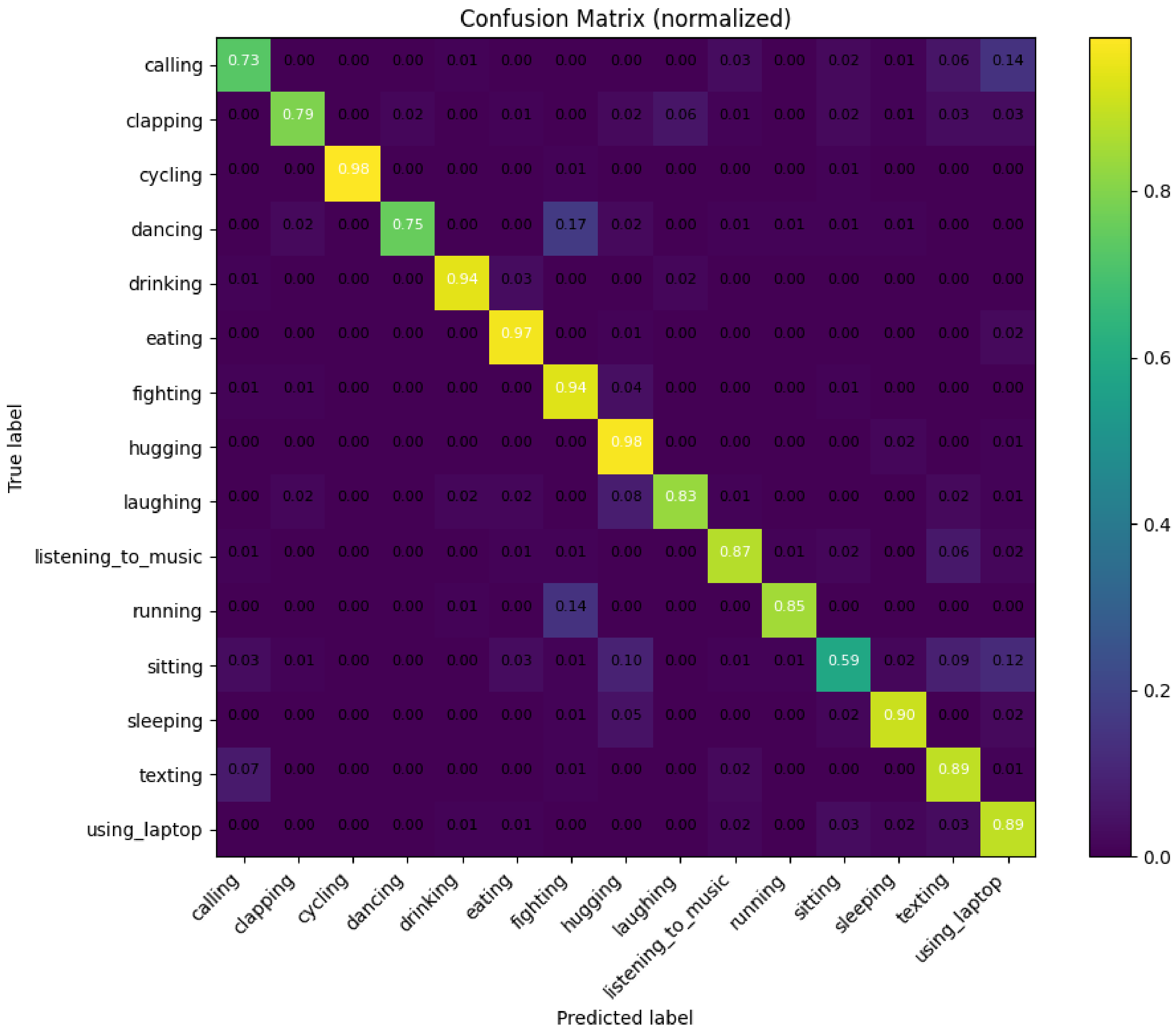


Figure 2: Confusion Matrix: SigLIP

Where SigLIP pulls ahead — object/device–centric actions

The largest class–level gains favor SigLIP on categories anchored by explicit *hand–held devices* or strong *scene layout* cues:

- **using_laptop:** 0.841 vs. 0.675 (**+0.166** for SigLIP) — biggest gap. Likely due to clearer text–image alignment on specific objects (screen/keyboard, hands on keys).
- **texting:** 0.857 vs. 0.794 (**+0.063**) — small object near hands/face; SigLIP better at phone cues under pose/occlusion.
- **calling:** 0.857 vs. 0.810 (**+0.048**) — consistent device advantage (phone to ear).
- **sleeping:** 0.857 vs. 0.810 (**+0.048**) — distinct global layout (prone posture, bedding) exploited more reliably.

Where CLIP does better — interaction/motion cues

CLIP shows smaller but real edges when the signal relies more on *interpersonal dynamics* or *subtle motion/pose* rather than explicit objects:

- **fighting:** 0.841 vs. 0.810 (**+0.032** for CLIP) — cluttered multi–person scenes; CLIP's symmetric InfoNCE/temperature may bias toward human–configuration patterns.

Table 2: Per-class accuracy (recall per class)

Class	SigLIP Acc.	CLIP Acc.	Δ (CLIP−SigLIP)	Support
calling	0.857	0.810	-0.048	126
clapping	0.857	0.857	+0.000	126
cycling	0.905	0.905	+0.000	126
dancing	0.817	0.794	-0.024	126
drinking	0.817	0.786	-0.032	126
eating	0.825	0.833	+0.008	126
fighting	0.810	0.841	+0.032	126
hugging	0.730	0.722	-0.008	126
laughing	0.817	0.833	+0.016	126
listening to music	0.817	0.833	+0.016	126
running	0.905	0.889	-0.016	126
sitting	0.786	0.794	+0.008	126
sleeping	0.857	0.810	-0.048	126
texting	0.857	0.794	-0.063	126
using laptop	0.841	0.675	-0.167	126

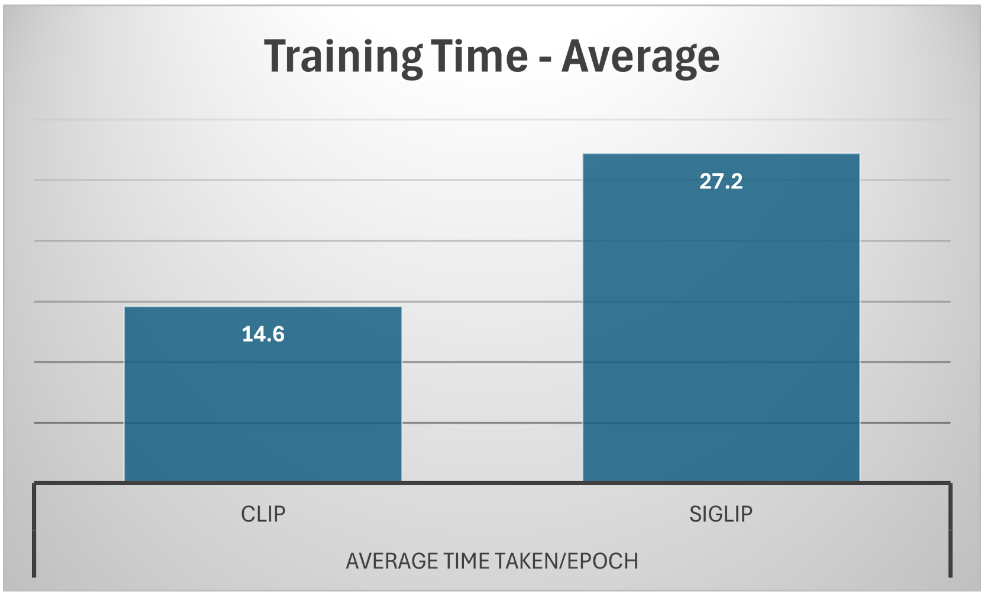


Figure 3: Average time per epoch (mean over 10 epochs). Lower is better. CLIP is faster per epoch; SigLIP is slower.

- **eating, laughing, listening_to_music, sitting:** CLIP leads modestly (+0.008 to +0.016) — fine gesture/expression dominated classes.

What the gaps imply about training choices

- **Negatives matter.** SigLIP trained with ∼32k effective batch + large XBM; CLIP used ∼4k. More negatives help separate near–neighbor prompts (*texting* vs. *using_laptop*). *Action:* grow CLIP's XBM and/or accumulation (target 8–16k) to close device–centric gaps.
- **Loss shape & calibration.** SigLIP's pairwise BCE does not force a single softmax winner and can reward multiple prompt matches; CLIP's InfoNCE can be peakier and brittle on subtle class boundaries. *Action:* for CLIP, tune temperature schedule/clamping and add harder prompt negatives; for SigLIP, consider class–balanced `pos_weight` if headroom remains.
- **Prompt sensitivity.** The prompt bank seems to benefit device classes more ("holding a phone", "typing on a laptop"). *Action:* run per–class prompt ablations; keep the best per class. For CLIP, try longer paraphrases that explicitly anchor the object.

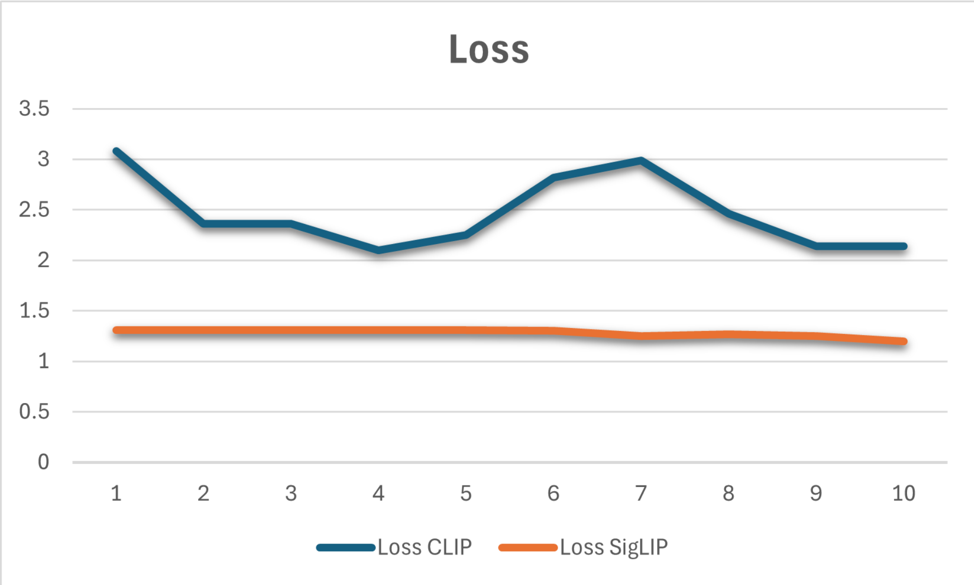


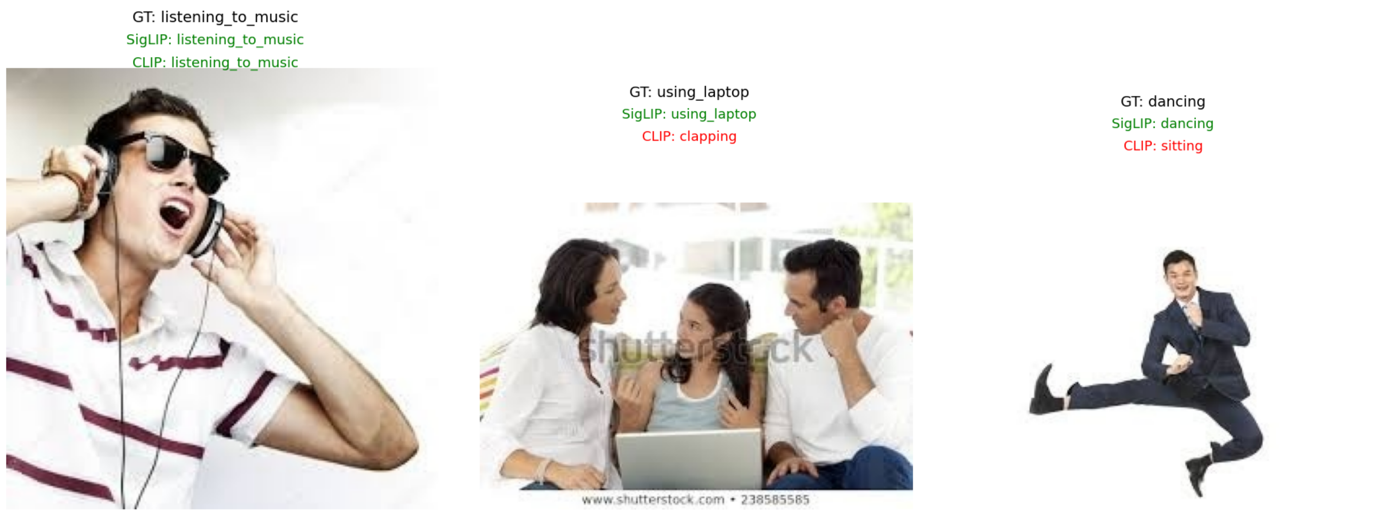
Figure 4: Training loss vs. epoch (both models). CLIP converges quicker in wall-clock but plateaus at a higher loss; SigLIP descends further to a lower terminal loss.

Bottom line

If choosing one today, **SigLIP is the safer default for this HAR dataset**: it is better on average and materially better on the most error–prone, object–defined actions (*using_laptop, texting, calling*). **CLIP is competitive** and slightly better on interaction–heavy categories (*fighting* and small edges on *laughing/listening/sitting*), and should narrow the gap with more negatives and class–specific prompt/augmentation tuning. Until then, **ship SigLIP as the baseline**.

7 Visualizations

You can find more qualitative examples below under, where correct predictions by models are show in green and incorrect in red.



(a) Example A (b) Example B (c) Example C

Figure 5: Produced by `compare_viz.py`.