Modal Training Summary

We evaluated several data-access approaches for training, and Modal Volumes proved to be the most effective within the Modal environment. Using this workflow reduced the time required to train for 10 epochs from roughly 10 hours to about 30 minutes. Because reproducibility and modularity are central to our pipeline, each experiment's configuration file records the dataset version and any applied subsetting and all tunable hyperparameters, making it easy to re-run experiments after updates or rapidly adjust parameters. Checkpoints are stored in a dedicated directory for later reuse, and logs are captured through Weights & Biases, enabling real-time monitoring and quick cancellation of unpromising runs to save compute.

After generating dataset_v1, we ran a series of controlled experiments to understand how data subsetting and augmentation influence performance (using baseline model ResNet50 with PubMedBERT). Subsetting the dataset did not improve results, while light augmentation provided small but consistent gains and more aggressive augmentation tended to reduce performance.

Our multimodal architecture combines visual and textual backbones, with a classifier operating on fused embeddings (either through concatenation or weighted sum). To establish a baseline, we trained multiple vision models (including EfficientNet-B0, DenseNet121, ResNet50, EfficientNet-B1, and ViT-B/16) paired with the same text encoder (PubMedBERT) under identical training settings. The models performed similarly overall. Since our deployment strategy requires running inference on local user devices, we selected EfficientNet-B0 for its favorable accuracy-to-compute ratio. We then compared several text encoders and sequence lengths. SapBERT (CLS, max_length=256) emerged as one of the strongest options, performing on par with SapBERT at 512 tokens using either CLS or mean pooling. With this setup, the model achieved a validation F1 of approximately 0.61 after 10 epochs.

Before scaling training to more epochs, we examined modality masking. Fully masking text caused a significant drop in performance, while fully masking images had only a modest effect. This indicates that the model is strongly dependent on text and that the vision backbone remains underpowered. To address this imbalance, we experimented with higher dropout, probabilistic text masking, increased learning rates for the vision model, and auxiliary per-modality losses. None of these interventions produced significant improvements.

Our next steps include refining the fusion and training strategy (experiments that will help mitigate overreliance on text and experiments with other loss function options, fusion strategies, etc.) and developing dataset_v2, which will apply stricter text filtering to minimize potential data leakage. We also plan to train models for a higher number of epochs once the final version is established.