

CPLIP No-Tuning Benchmark

Aim & Scope

- Aim: test CPLIP “as is” (no tuning).
- Scope: do not train the encoders; use the same data prep and split for all models; use simple heads (zero-shot, linear probe, simple MIL).

1) Data & Labels

- Choose the task (high risk vs low risk).
- Include/exclude rules, where data came from, and the follow-up window (e.g., 36 months).
- Count positives, negatives, and uncertain.

2) Patient-Level Splits (Frozen)

- Make one train/val/test split (or 5-fold CV). Use the same split for all models.
- Keep a similar % of cancer in each split. No patient appears in more than one split.
- Save the random seed and class counts for each split.

3) Data Prep (Same for All Models)

- Tile WSIs: 512×512, 20-30% overlap. Use ROI first.
- Filter bad tiles (too white/black/blurred).
- If you use stain norm, pick one method and keep it.
- Resize to the model size (e.g., 224/256) and normalize.
- Save tiles and a labels file that links tile/WSI/patient to the label.

4) Embeddings (Frozen CPLIP)

- Write down which CPLIP weights/version you used; input size; fp16 or fp32; pooling.
- Set batch size/workers/hardware; set a random seed.
- Compute and save tile embeddings (cache).
- Keep the link from tile → WSI → patient.

5) Heads (No Tuning)

- Zero-shot: one fixed prompt per class; compare image vs text; average tile scores to slide/patient.
- Linear probe: train one linear layer on frozen embeddings (use defaults).
- Simple MIL: mean-pool tile embeddings; logistic/linear layer (use defaults).

6) Repeatability

- Set seeds so runs are repeatable.
- Run at least 3 seeds; report mean ± SD.

7) Test Set Evaluation

- Discrimination: ROC-AUC (and PR-AUC if imbalanced), F1, sensitivity/specificity (set threshold on validation).
- Calibration: Brier score, ECE, and a reliability plot.
- Speed/size: number of parameters, train time, time per slide/tile, peak memory, hardware used.
- Uncertainty: 95% CIs with bootstrap; DeLong test optional.

8) Errors & Subgroups

- Make a confusion matrix and per-class metrics; look at common false positives/false negatives.
- Check groups: site, lesion type, scanner, stain, cohort.
- Write short notes on common failure types.

9) What to Save>Show

- Results table: Model | Head | Params | Train time | Inference time | ROC-AUC | PR-AUC | F1 | Brier | ECE | Seeds (mean \pm SD).
- Figures: ROC and PR curves; calibration plot; a small speed/size chart.
- Configs and seeds: splits.json, preproc.yaml, embed.cfg, metrics.json.

10) No-Tuning Rules

- No hyper-parameter search, no prompt tweaking, no backbone fine-tuning, no trying many stain/tiling options, no test-set peeking.

11) Reproducibility

- Write down software versions (CLIP commit, CUDA, PyTorch).
- Confirm de-identification and that patients are in only one split.
- Keep a log of seeds, configs, hardware, and dates.

Table 1 Patients + follow-up

patient_id	sex	birth_year	start_date	last_followup_date	became_cancer	cancer_date	risk_label_36m
P1001	F	1968	2021-02-10	2023-12-01	1	2022-08-15	Positive
P1002	M	1979	2021-05-20	2024-06-05	0		Negative
P1003	F	1988	2023-01-12	2024-02-10	0		Uncertain
P1004	Other	1959	2020-11-03	2025-01-20	1	2024-12-30	Negative

Table 2 WSI mapping

wsi_id	patient_id	file_path	roi_path	scan_date	magnification	baseline_grade	site
S1001	P1001	D:/WSI/P1001_S1001.tif	D:/ROI/P1001_S1001.geojson	2021-02-10	40x	HGD	tongue
S1002	P1002	D:/WSI/P1002_S2001.svs		2021-05-20	20x	LGD	buccal
S1003	P1003	/data/wsi/P1003_S3001.tif	/data/roi/P1003_S3001.geojson	2023-01-12	40x	HGD	gingiva
S1004	P1004	/data/wsi/P1004_S4001.tif		2020-11-03	40x	LGD	tongue