Hyper3-CLIP: Hierarchy-Conditioned Hyperbolic Vision-Language Training

A hyperbolic CLIP that learns from every part of the caption.

Matin Mahmood[1] **Antonio Rueda-Toicen**[2] **Mohamed ElBassat**[3] **Seifeldin Elkerdany**[4] **Weixing Wang**[2] **Gerard de Melo**[2]

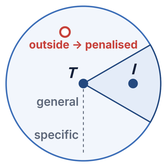
[1]hyper[3]labs, Berlin · [2]Hasso Plattner Institute, University of Potsdam · [3]Alexandria University · [4]Alamein International University

ECCV 2026 · Beyond Euclidean Workshop · Malmö · ORAL

BACKGROUND · 1

Why hyperbolic space?

Hyperbolic **volume grows exponentially with radius**, supporting branching hierarchies. Images are trained to lie inside their captions' **entailment cones**.



Each node *b* roots an entailment cone of more-specific points.

$a \preceq b$: *a* lies inside the cone of *b*; a node outside the cone is penalised.

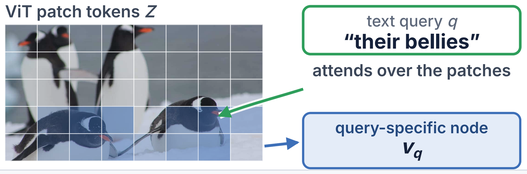
T = caption node, *I* = image node.

Prior work: $I \preceq T$ (MERU, HyCoCLIP, UNCHA)

BACKGROUND · 2

One query, one visual view

A single image vector must match the whole caption, so **"their bellies"** averages away. β-CLIP: each text query **attends over the patches** and pools its own vector.



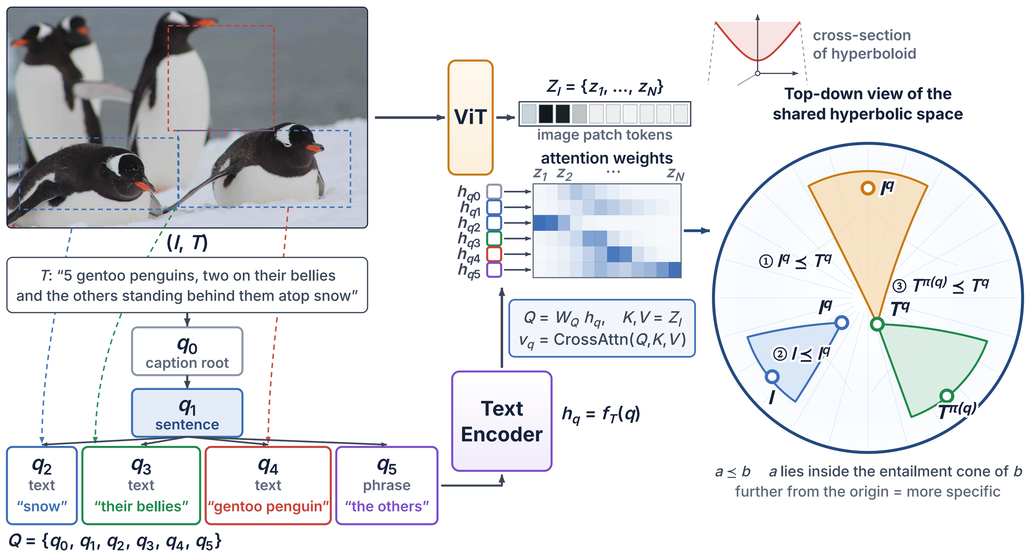
HYPER3-CLIP

Every query becomes a node

From each caption we build a lightweight **query hierarchy**. Each query pools its own visual node; **parent links** tie all nodes together in one hyperbolic space.

QUERY TYPE	PARENT	LOSS WEIGHT
Caption root	none	0.0
Sentence	caption	1.0
Text box (grounded)	caption	0.75
Phrase	sentence	0.5

Example: q_0 caption → q_1 sentence → phrases.



TRAINING OBJECTIVE

Three entailment relations per query (①②③ above)

Node *b* roots a cone of **more-specific** points; node *a* is penalised when it lies outside that cone.

① $I^q \preceq T^q$

Query-level analogue of image-to-text entailment: the visual node lies inside the cone of its query text.

② $I \preceq I^q$

Pooling removes content unrelated to *q*, so I^q is the broader concept; the full image keeps the whole scene and is **more specific** (as $I \preceq I^{box}$).

③ $T^{\pi(q)} \preceq T^q$

$\pi(q)$ = parent of *q*. A parent query carries **more context** than its child, so it is treated as more specific.

$$\mathcal{L} = \mathcal{L}^{con}_{UNCHA} + \lambda_{ent} \left(\mathcal{L}^{ent}_{UNCHA} + \mathcal{L}^{ent}_{query} \right)$$

Query terms are summed over non-root queries with weights $w(q)$. Pooling is **training-only**: inference is a plain CLIP dual encoder.

ZERO-SHOT RETRIEVAL · VIT-B/16

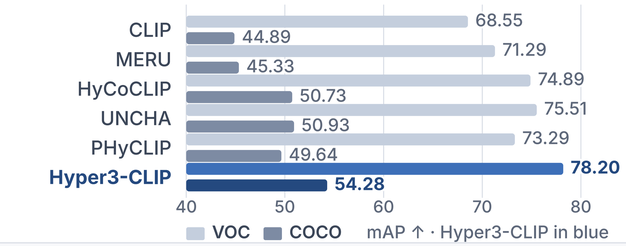
Higher recall on COCO and Flickr

Retrieval R@5 / R@10	CLIP	UNCHA	Hyper3-CLIP	Δ vs UNCHA
COCO text	71.4 / 81.5	72.7 / 82.7	**75.7 / 84.3**	+3.0 / +1.6
Flickr text	93.6 / 96.9	91.4 / 95.9	**95.4 / 97.6**	+4.0 / +1.7
COCO image	57.4 / 68.5	60.0 / 71.0	**63.0 / 73.2**	+3.0 / +2.2
Flickr image	83.5 / 89.9	84.9 / 91.2	**86.4 / 91.4**	+1.5 / +0.2

Hierarchy trade-off: on ImageNet TIE / LCA (lower is better) UNCHA stays best: 2.94 / 1.96 vs our 3.16 / 2.08.

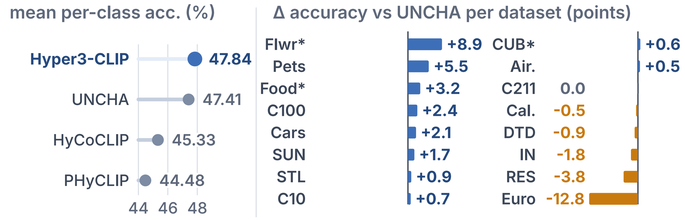
ZERO-SHOT MULTI-LABEL · MAP

Multi-label gains on VOC and COCO



ZERO-SHOT CLASSIFICATION · 16 DATASETS

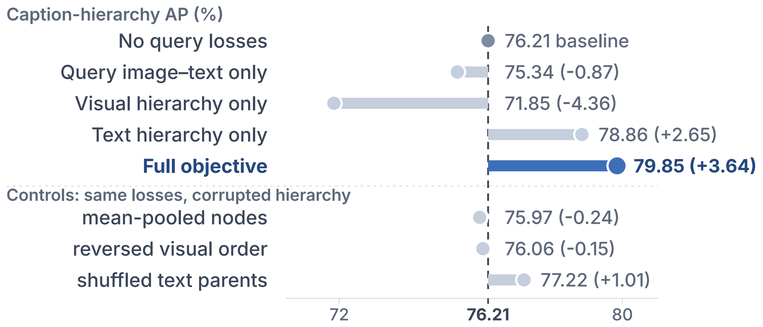
Highest 16-dataset average



Mean per-class accuracy. * = photo prompt (see caveat below).

ABLATION · CAPTION-HIERARCHY ENTAILMENT

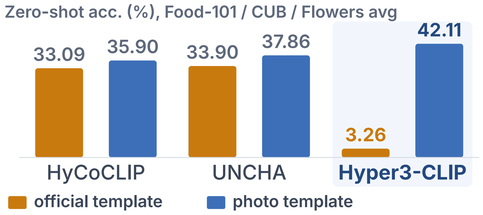
Hierarchy links drive the gain



CAVEAT · ZERO-SHOT CLASSIFICATION

Prompt sensitivity

Zero-shot acc. (%), Food-101 / CUB / Flowers avg

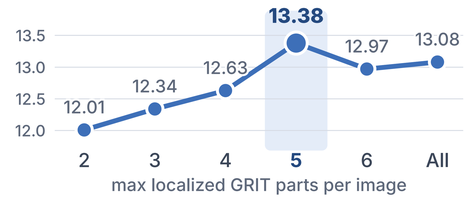


Official: `food : {}. · bird pics : {}. · flowers : {}.`
Photo: `a photo of a {}.`
Trained on multi-word queries, Hyper3-CLIP needs a sentence-like prompt.

DATA · CAP ON GROUNDED PARTS PER IMAGE

Five grounded parts suffice

Avg R@1 on a 10k-image proxy



R@10 plateaus too (41.85 at 5 vs 41.93 all); 79% of images have ≤ 2 parts.

SUMMARY

Take-aways and links

- Query-conditioned **hyperbolic nodes** improve retrieval and multi-label transfer.
- Correct **parent links** and query pooling raise caption-hierarchy AP.
- Inference = **plain CLIP**.
- **Next:** query nodes at inference: late-interaction entailment.

Paper arxiv.org/abs/2608.29313 · **Code** github.com/Hyper3Labs/hyper3-clip · **Model** hf.co/hyper3labs/hyper3-clip

matin@hyper3labs.com
antonio.rueda-toicen@hpi.de