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

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

[1] hyper[3]labs, Berlin, Germany
matin@hyper3labs.com
[2] Hasso Plattner Institute, University of Potsdam, Potsdam, Germany
antonio.rueda-toicen@hpi.de
[3] Faculty of Computers and Data Science, Alexandria University, Alexandria, Egypt
[4] Faculty of Computer Science and Engineering, Alamein International University, New Alamein, Egypt

Abstract. CLIP-like vision-language models (VLMs) trained with contrastive objectives learn strong global image-text representations, but their Euclidean embeddings and global pooling fail to encode relational structure such as part-whole and parent-child relations. Hyperbolic VLMs address this relational representation gap with hierarchy-aware geometry and entailment objectives, and text-conditioned CLIP variants show that caption-, sentence-, and phrase-level queries improve fine-grained alignment. However, these two lines of work remain separate: hyperbolic VLMs use static image and region features, while query-conditioned methods lack hierarchical geometric structure. We present Hyper3-CLIP, a hierarchy-conditioned hyperbolic VLM that combines global, local, and global-local contrastive learning with query-conditioned visual pooling. To train the model, we construct lightweight query hierarchies from text, comprising full captions, sentence fragments, localized part descriptions, and extracted phrases. Each query conditions the pooling of visual patches, and the resulting representations support image-text, whole-part, and parent-child hierarchical entailment losses. Query-conditioned pooling is active only during training: at inference, Hyper3-CLIP is a standard dual encoder and adds no extra computation. Experiments demonstrate improvements in image-text retrieval and hierarchy-sensitive image-caption alignment. Hyper3-CLIP improves R@5 and R@10 retrieval on COCO and Flickr, as well as multi-label classification on VOC and COCO, while remaining competitive on hierarchy metrics. We also audit zero-shot prompt sensitivity under fixed prompt regimes and study the effect of the localized GRIT part budget used during training. Code and pretrained models are available at `https://github.com/Hyper3Labs/hyper3-clip`.

1 Introduction

Vision-language models (VLMs) have become the dominant paradigm for learning transferable multimodal representations. Large-scale contrastive pretraining

methods such as CLIP [20], ALIGN [6], LiT [24], BLIP [10], and SigLIP [23] align images and text in a shared embedding space, enabling strong zero-shot recognition, image-text retrieval, and multimodal reasoning.

Despite their success, most VLMs learn representations in Euclidean space using a single global embedding for each image and caption. While effective for instance-level alignment, this formulation provides limited structure for modeling the hierarchical relationships naturally present in visual scenes and language. Images consist of objects and parts, while captions contain sentences, phrases, and localized descriptions that exist at multiple semantic levels.

Hyperbolic geometry provides a natural representation space for hierarchical data, motivating a broad line of work on hyperbolic representation learning [4, 5, 8, 15, 16] and, more recently, hyperbolic vision-language models. MERU [3] first demonstrated the benefits of hyperbolic embeddings through entailment-based supervision. ATMG [21] later showed that proximity-based contrastive objectives can hinder hierarchical structure learning in hyperbolic spaces and proposed an angle-based alternative to better preserve geometric hierarchy. HyCoCLIP [17] incorporated grounded image-box and text-box pairs for compositional alignment, while UNCHA [7] introduced uncertainty-aware hierarchical alignment for part-to-whole reasoning. More recently, PHyCLIP [22] models hierarchy within individual hyperbolic spaces and compositional semantics via an ℓ_1-product over multiple hyperbolic factors. Although these methods capture hierarchical semantics, they remain constrained to static image and region representations and do not exploit query-conditioned visual features that dynamically adapt to different textual granularities.

In a separate line of work, recent vision-language models have moved beyond global image-text alignment toward grounded and fine-grained supervision. Methods such as ALBEF [11], RegionCLIP [25], GLIP [12], Grounding DINO [14], BLIP-2 [9], and Kosmos-2 [18] incorporate region-level grounding, open-vocabulary localization, or query-driven visual representations, enabling stronger fine-grained alignment between language and visual content. In particular, β-CLIP [26] introduces sentence- and phrase-level queries with query-conditioned attention, producing multiple semantically focused visual embeddings from a single image. While effective for improving granularity, these methods do not explicitly incorporate hierarchical geometric structure for modeling part-whole or multi-granular semantic relations, which motivates the use of hyperbolic representations.

To bridge this gap, we propose **Hyper3-CLIP**[5], a hierarchy-conditioned hyperbolic vision-language model that combines query-conditioned visual pooling with hyperbolic hierarchical representation learning. Hyper3-CLIP constructs lightweight textual hierarchies from captions, sentence fragments, localized descriptions, and extracted phrases; each textual query attends to image patch tokens to produce a query-specific visual embedding, which is projected into hyperbolic space alongside its corresponding text representation. The resulting query-level representations are integrated with grounded image-box supervision

[5] Pretrained model: `https://huggingface.co/hyper3labs/hyper3-clip`

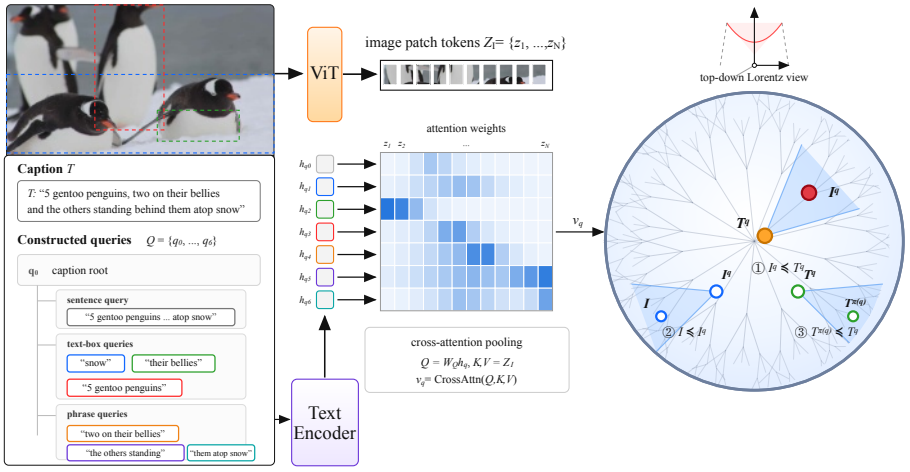

Fig. 1: Hyper3-CLIP overview. The method assumes an image along with a whole-image caption and localized image boxes with corresponding text boxes. We define caption, sentence, text-box, and phrase query nodes. Each text query attends over full-image ViT patch tokens to form a query-conditioned visual node for the hierarchy losses, while the stored image boxes and text boxes remain available for grounded local alignment.

through hyperbolic contrastive and entailment learning, enabling explicit modeling of both semantic hierarchy and fine-grained visual composition. Through extensive experiments on image-text retrieval benchmarks, we demonstrate that combining hyperbolic hierarchical representations with query-conditioned visual alignment improves image-text retrieval and hierarchy-sensitive image-caption alignment.

Our contribution is the integration of these two lines of work rather than a new objective family or pooling mechanism. Query decomposition and text-conditioned attention pooling follow β-CLIP [26], and the base contrastive and grounded box-level entailment objective is inherited from UNCHA [7]. What is new is their parent-linked hyperbolic combination: each query carries a parent link and instantiates its own visual node in the shared hyperbolic space, and these nodes are supervised through three query-level entailment relations. Query-conditioned pooling is active only during training; at inference, Hyper3-CLIP is a standard dual encoder, and the pooling module adds no inference-time computation.

2 Method

Hyper3-CLIP extends a grounded hyperbolic training setup for image-caption and image-box/text-box pairs. Each training sample consists of a full image-

caption pair (I, T) and grounded box-level pairs $\{(I_j^{\text{box}}, T_j^{\text{box}})\}_{j=1}^{P}$, where I_j^{box} is an image box (a localized crop) and T_j^{box} is the corresponding text box, typically a noun or phrase from the caption. The inherited objective supplies the contrastive and entailment losses over these full-image and localized nodes. Hyper3-CLIP maintains this base structure and adds query-conditioned visual nodes derived from the full-image patch tokens. Appendix A gives the Lorentz geometry and inherited HyCoCLIP/UNCHA loss terms. Relative to this inherited base, Hyper3-CLIP adds three components: the hierarchical query construction (Sec. 2.1), the query-conditioned visual pooling module adapted from β-CLIP [26] (Sec. 2.2), and the query-level entailment relations added to the training objective (Sec. 2.3).

2.1 Hierarchical Query Construction

For each training sample, Hyper3-CLIP builds an additional query set $\mathcal{Q}$ from the sample text annotations: the caption and localized text boxes. The caption is the root query. Sentence queries are sentence-level fragments of the caption; text-box queries reuse the localized text boxes; phrase queries are lightweight phrases extracted from the caption. Unlike text-box queries, phrase queries have no image-box annotation.

Each query stores a parent $\pi(q)$ and loss weight $w(q)$. Table 1 summarizes the query types and parent links. The weights encode a simple reliability prior: sentence fragments are direct caption spans and receive full weight, text boxes are grounded but shorter and less contextual, extracted phrases are the weakest textual units, and the caption root has zero query weight because the full image-caption pair is already supervised by the inherited global objective.

Table 1: Query types used to build the hierarchy. The caption root provides parent context; sentence, text-box, and phrase queries receive nonzero entailment weights.

Query	Source and parent	$w(q)$
Caption root	caption; none	0.0
Sentence	caption sentence; caption	1.0
Text box	text box; caption	0.75
Phrase	caption phrase; sentence/caption	0.5

Text boxes continue to supervise their paired image boxes through the inherited objective. Hyper3-CLIP also uses each text box to condition pooling over full-image patch tokens.

2.2 Query-Conditioned Visual Pooling

Query-conditioned visual pooling converts each text query into a separate visual node for its owner image.

Let $Z_I \in \mathbb{R}^{N \times d_v}$ denote the full-image patch tokens produced by the image encoder after removing the class token, and let $h_q \in \mathbb{R}^{d_t}$ be the text-encoder representation of query q. A learned linear map $W_Q \in \mathbb{R}^{d_v \times d_t}$ projects h_q to the visual-token dimension, after which it attends over the image patch tokens:

$$\widetilde{v}_q = \mathrm{MHA}(Q = W_Q h_q,\ K = Z_I,\ V = Z_I)\,, \qquad v_q = \widetilde{v}_q + \mathrm{MLP}(\mathrm{LN}(\widetilde{v}_q))\,.$$

The cross-attention block uses 8 heads, and the MLP has hidden dimension $4d_v$ with a GELU activation. This text-conditioned attention pooling follows β-CLIP [26]; Hyper3-CLIP uses it to instantiate hyperbolic query nodes rather than Euclidean embeddings. The resulting representation v_q is the visual node for query q. It is pooled from the full image, while the text query guides the pooling. Finally, v_q passes through the shared image projection and Lorentz lift, while h_q passes through the shared text projection and lift; we denote the resulting hyperbolic embeddings by I^q and T^q, respectively. Query-conditioned pooling is used only during training, and evaluation uses the standard global embeddings of the dual encoder.

2.3 Training Objective

The training loss combines contrastive alignment with hyperbolic entailment and uncertainty calibration. The contrastive term $\mathcal{L}^{\mathrm{con}}_{\mathrm{UNCHA}}$ keeps the UNCHA global, local, and global-local alignments. The base entailment term $\mathcal{L}^{\mathrm{ent}}_{\mathrm{UNCHA}}$ keeps the inherited relations among I, T, and the grounded box-level pairs $\{(I_j^{box}, T_j^{box})\}_{j=1}^{P}$.

For each non-root query $q \in \mathcal{Q}$ with $w(q) > 0$, the query-conditioned nodes add three entailment relations:

$$(1)\ I^q \preceq T^q, \qquad (2)\ I \preceq I^q, \qquad (3)\ T^{\pi(q)} \preceq T^q.$$

Following MERU [3], $E(a \preceq b)$ denotes the hyperbolic entailment-cone violation for an ordered pair. Node b roots a cone of more specific points, and node a is penalized when it lies outside that cone. Relation (1) is the query-level analogue of image-to-text entailment. For relation (2), the ordering is defined by semantic specificity rather than by the computational origin of the representations. Query-conditioned pooling removes image content unrelated to the query, so I^q represents a broader query-specific concept, whereas I retains the complete scene and its additional context. Thus, we treat the full image as more specific than the query-conditioned visual node, consistent with the inherited relation $I \preceq I^{box}$, where the complete scene is more specific than a localized part. Relation (3) preserves the query hierarchy on the text side: a parent query is treated as more specific than its child because it contains additional contextual information.

The query terms are summed over non-root queries with reliability weights $w(q)$ and calibrated with the same uncertainty mechanism as the inherited UNCHA objective to form $\mathcal{L}^{\mathrm{ent}}_{\mathrm{query}}$. The full objective is

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

3 Experiments

GRIT data. Hyper3-CLIP is trained on GRIT [18] using all available localized parts up to a cap of 5 parts per image. Figure 2 studies this max-parts setting and shows that it retains 99.38% of localized part instances in the processed training data.

Query construction. Queries are constructed by rule-based string processing, without a learned parser. For each image, candidates are added in a fixed order: the caption root, then sentence queries, then text-box queries, then extracted phrases. Every candidate is normalized by collapsing whitespace runs to single spaces and is accepted only if it is at least 3 characters long and its case-insensitive text has not already been accepted for that image. Sentence queries split the caption at whitespace following ".", "!", "?", or ";" and at line breaks, keeping the first 5 candidates; their parent is the caption. Text-box queries reuse the localized text boxes, also with the caption as parent. Phrase extraction splits the caption at commas, semicolons, colons, and parentheses, and at the connector words *and*, *with*, *near*, *beside*, *behind*, *under*, *above*, *around*, and *next to* (matched case-insensitively); each resulting fragment is reduced to its alphanumeric words, allowing internal hyphens and apostrophes and discarding other punctuation. A fragment of 2–8 words yields one phrase; a longer fragment is cut into windows of at most 6 words starting every 4 words, keeping windows with at least 2 words; fragments with fewer than 2 words are discarded. Text-box and phrase queries share a budget of 30 accepted queries per image, filled by text boxes first; each accepted phrase takes the first accepted sentence query as parent, or the caption when no sentence query exists. Finally, at most 6 queries per image are kept, truncating in construction order.

Architecture. Following prior hyperbolic VLMs [7,17,22], we use a CLIP-style dual-encoder architecture with a ViT image encoder and a Transformer text encoder. The main Hyper3-CLIP result uses a ViT-B/16 image backbone, the CLIP-B/32 text-encoder architecture, a 512-dimensional multimodal embedding, Lorentz projection, and random initialization for both towers. ViT-S/16 appears in the published baselines and controlled ablations, but not as a main Hyper3-CLIP result.

Evaluation. All downstream results are zero-shot evaluations of frozen checkpoints. The cross-attention pooling module is dropped at inference: every evaluation scores images and text with the standard global dual-encoder embeddings, so the pooling module adds no computation at inference time. We report image-text retrieval, ImageNet hierarchy metrics, 16-dataset classification, and VOC/COCO multi-label classification.

3.1 Zero-Shot Retrieval and Hierarchy

In zero-shot retrieval, the model must find the matching caption for an image, and vice versa, using only its learned multimodal alignment. Using COCO [13] and Flickr30K [19], we measure Recall@K for image-to-text and text-to-image

Table 2: Zero-shot image-text retrieval and ImageNet hierarchy evaluation for the ViT-B/16 setting. Published baseline values are from UNCHA Table 2 [7]; Hyper3-CLIP is evaluated with the same retrieval and hierarchy metrics. Paired retrieval columns report R@5/R@10. T and I denote text retrieval and image retrieval. Higher is better except TIE and LCA. Bold marks the best value in each column, with ties bolded.

Backbone	Model	Text retrieval				Image retrieval				Hierarchy				
		COCO		Flickr		COCO		Flickr						
		R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10	TIE↓	LCA↓	J↑	H-P↑	H-R↑
ViT-B/16	CLIP	71.4	81.5	93.6	96.9	57.4	68.5	83.5	89.9	3.60	2.21	0.79	0.85	0.85
	MERU	72.3	82.0	93.5	96.2	57.4	68.6	84.0	90.0	3.63	2.22	0.78	0.85	0.85
	ATMG	62.9	74.0	85.1	92.2	51.2	62.6	78.0	85.3	4.19	2.48	0.75	0.83	0.83
	HyCoCLIP	72.0	82.0	92.6	95.4	58.4	69.3	84.9	90.3	3.17	2.05	0.81	0.87	0.87
	UNCHA	72.7	82.7	91.4	95.9	60.0	71.0	84.9	91.2	**2.94**	**1.96**	**0.83**	**0.88**	**0.88**
	Hyper3-CLIP	**75.7**	**84.3**	**95.4**	**97.6**	**63.0**	**73.2**	**86.4**	**91.4**	3.16	2.08	0.82	0.87	**0.88**

Table 3: Zero-shot classification on the 16-dataset CLIP evaluation suite. Values are mean-per-class accuracy. Food-101, CUB, and Flowers, marked with *, use the common Photo prompt audited in Table 6; the remaining datasets use the baseline evaluation prompts. Bold marks the best value in each column.

	General						Fine-grained						Misc.				
Model	IN	C10	C100	SUN	Cal.	STL	Food*	CUB*	Cars	Air.	Pets	Flwr.*	DTD	Euro.	RES.	C211	Avg.
HyCoCLIP-B/16	45.80	88.80	60.10	57.20	81.30	95.00	61.14	17.76	11.60	3.70	56.80	28.79	29.40	35.80	45.60	**6.50**	45.33
UNCHA-B/16	**48.80**	90.40	63.20	57.70	**83.90**	95.70	64.10	18.74	14.00	3.80	57.10	30.75	**30.30**	**41.30**	**52.70**	6.10	47.41
PHyCLIP-B/16	44.31	89.33	59.05	55.32	76.35	94.84	62.40	17.97	10.89	3.24	54.18	28.27	25.50	36.29	48.22	5.56	44.48
Hyper3-CLIP-B/16	46.98	**91.14**	**65.64**	**59.39**	83.42	**96.56**	**67.31**	**19.36**	**16.12**	**4.33**	**62.57**	**39.66**	29.41	28.52	48.94	6.06	**47.84**

retrieval. We also report ImageNet hierarchy metrics to test whether the hyperbolic embedding preserves class-label structure.

The results are given in Table 2. Hyper3-CLIP improves over UNCHA across all eight considered COCO/Flickr retrieval settings. On hierarchy metrics, it does not improve TIE or LCA, but remains competitive with UNCHA.

3.2 Zero-Shot Classification

In zero-shot classification, the model assigns labels to images from categories without using dataset-specific training data. This evaluates how well the learned visual-semantic embeddings generalize to broad, unseen concepts. We measure mean-per-class accuracy on a standard 16-dataset zero-shot classification benchmark that includes ImageNet [2].

Under the stated prompt protocol, Hyper3-CLIP obtains the best 16-dataset average, reaching 47.84 compared with 47.41 for UNCHA. The gains are concentrated on CIFAR-10, CIFAR-100, SUN397, STL-10, Food-101, Flowers, Cars, Aircraft, and Pets, whereas weaknesses remain for Country211, DTD, and EuroSAT. Section 4.2 audits the three datasets evaluated with the common Photo prompt and reports their corresponding official-prompt results separately.

Table 4: Zero-shot multi-label classification on VOC and COCO under a single evaluator. Values are mAP; Avg. is the mean of VOC and COCO. Higher is better, and bold marks the best value in each column.

Backbone	Model	Multi-label classification		
		VOC	COCO	Avg.
ViT-B/16	CLIP	68.55	44.89	56.72
	MERU	71.29	45.33	58.31
	HyCoCLIP	74.89	50.73	62.81
	UNCHA	75.51	50.93	63.22
	PHyCLIP	73.29	49.64	61.47
	Hyper3-CLIP	**78.20**	**54.28**	**66.24**

3.3 Zero-Shot Multi-Label Classification

For VOC and COCO multi-label classification, each image can contain multiple valid object categories. We score each class independently and report mean Average Precision (mAP), following the framing used by UNCHA.

Under a shared evaluator, Hyper3-CLIP improves over UNCHA by 2.69 mAP on VOC and 3.35 mAP on COCO. Because the VOC/COCO values reported in UNCHA Table 5 use a different protocol, we do not mix those numbers into the main comparison.

4 Ablation Study

We further conduct a series of analyses, including ablating the query-conditioned objective to determine which relations are responsible for the hierarchy-sensitive gains and whether those gains preserve standard retrieval performance.

4.1 Query-Conditioned Objective Components

The full objective augments the grounded contrastive loss with three query-level entailment relations. These relations align the query-conditioned visual node with its query text, $I^q \preceq T^q$; treat the full-image node as more specific than the query-conditioned visual node, $I \preceq I^q$; and treat the construction-parent text node as more specific than its child query, $T^{\pi(q)} \preceq T^q$. To isolate their effects, each variant is first trained with only the base objective for 80k steps, then continues training to 100k steps with the query entailment terms specified in Table 5. Data, batch size, optimizer schedule, and seed are fixed; the control rows keep all query entailment terms enabled and modify only the named control.

We evaluate retrieval and hierarchy-sensitive image-caption alignment. For hierarchy evaluation, following the image-caption structure of HierarCaps [1], positives pair an image with captions from its hierarchy and negatives pair the image with captions sampled from other fine-grained nodes. Pairs are ranked by the model's hyperbolic entailment value, $p(a \preceq b) = \max(1 - 2\phi(a, b)/\pi, 0)$, where ϕ is the exterior cone angle between the image and text embeddings. We report

Average Precision (AP) and AUROC over 4,000 positive and 100,000 negative pairs from 1,000 images. Avg. R@10 is the mean of the four COCO/Flickr text- and image-retrieval R@10 values.

Table 5: Component ablation for the query-conditioned objective. A checkmark indicates that the corresponding query loss from Sec. 2.3 is enabled. The control rows keep all three query losses enabled while replacing text-conditioned visual pooling, reversing the direction of the visual whole-to-query relation, or corrupting text-parent links. ΔAP and ΔR@10 are measured relative to the no-query baseline.

Variant	$I^q \preceq T^q$	$I \preceq I^q$	$T^{\pi(q)} \preceq T^q$	Hierarchy AP	ΔAP	AUROC	ΔR@10
No query losses	–	–	–	76.21	0.00	98.34	0.00
Query image-text only	✓	–	–	75.34	−0.87	98.24	−0.46
Visual hierarchy only	–	✓	–	71.85	−4.36	98.01	−0.77
Text hierarchy only	–	–	✓	78.86	+2.65	98.44	−0.09
Full objective	✓	✓	✓	**79.85**	**+3.64**	**98.55**	+0.06
Mean-pooled visual nodes	✓	✓	✓	75.97	−0.25	98.32	**+0.15**
Reversed visual order ($I^q \preceq I$)	✓	✓	✓	76.06	−0.15	98.29	−0.09
Shuffled text parents	✓	✓	✓	77.22	+1.00	98.40	−0.93

In the results in Table 5, we observe that the full objective achieves the highest caption-hierarchy AP, improving over the no-query baseline by 3.64 points while leaving retrieval essentially unchanged. Among the single-relation variants, the text parent-to-query term is the strongest and recovers 73% of the full AP gain (2.65/3.64). The query image-text and visual whole-to-query terms do not improve AP alone, but the full objective improves over the text-only variant by 0.99 AP; their effect is therefore visible in the jointly trained objective rather than as separate single-term gains.

The controls further support the role of the proposed hierarchy construction. Replacing text-conditioned visual pooling with mean-pooled visual nodes removes the AP gain. Reversing the visual hierarchy eliminates the full objective's hierarchy gain: AP falls from 79.85 to 76.06, slightly below the no-query baseline of 76.21, while average retrieval remains near baseline. This matched control confirms that the hierarchy improvement under the Table 5 protocol depends on the proposed $I \preceq I^q$ direction rather than merely adding a visual-hierarchy constraint. Shuffling the text-parent links substantially lowers AP by 2.63 points relative to the full objective. Together, these results show that query-conditioned training improves hierarchy-sensitive image-caption alignment while preserving retrieval performance.

4.2 Prompt Sensitivity and Zero-Shot Protocol

The zero-shot classification results in Table 3 are most sensitive to prompt choice on Food-101, CUB, and Flowers102. We therefore evaluate these three datasets under both the *Official* prompt template and a shared *Photo* prompt template for every model. The exact prompt strings are:

Prompt regime	Templates
Official Food-101	`food : {}.; food porn : {}.`
Official CUB	`bird pics : {}.; birding : {}.; birds : {}.; bird photography : {}.`
Official Flowers102	`flowers : {}.`
Photo	`a photo of a {}.`

The remaining 13 datasets in the 16-dataset suite are unchanged across the reported averages.

Table 6: Prompt sensitivity on the three zero-shot classification datasets whose inherited prompts were unstable. The Photo prompt template uses the same CLIP-style prompt family for every model and every dataset in this audit. The 16-dataset average changes only the three prompt-sensitive columns; all other zero-shot columns are held fixed. Values are mean-per-class accuracy, and bold marks the best value in each column.

	Food/CUB/Flowers avg.			16-dataset avg.		
Model	Official	Photo prompt	Δ	Official	Photo prompt	Δ
HyCoCLIP-B/16	33.09	35.90	+2.81	44.82	45.33	+0.51
UNCHA-B/16	33.90	37.86	+3.96	46.69	47.41	+0.72
Hyper3-CLIP-B/16	3.26	**42.11**	**+38.85**	40.55	**47.84**	**+7.29**

The common Photo prompt template improves the recorded HyCoCLIP and UNCHA three-dataset averages relative to the official template, showing that these columns are prompt-sensitive even for released baselines. Hyper3-CLIP performance collapses on the Official prompt template for these datasets, but it recovers when the Photo prompt template is used. This suggests that short query phrases such as `flower` can interact poorly with the short Official prompt prefixes used for these datasets, possibly because training supervises the text encoder with compositional multi-word queries rather than terse class-name templates.

4.3 Parts per Image

The processed GRIT data exhibits a long tail in the number of localized parts per image. The left panel of Fig. 2 summarizes an exact scan of 2,051 shards with 13,149,251 examples and 25,185,017 localized parts. The mean is 1.9153 parts per example, and the maximum observed example has 20 parts. Although 40.68% of examples have exactly one part and 79.23% have only one or two parts, the tail introduces a small number of examples with many local boxes.

This tail can also affect HyCoCLIP, which trains on localized GRIT boxes. It is more consequential for UNCHA and Hyper3-CLIP training because each retained part can contribute not only a local image-text correspondence, but also part-whole, uncertainty-weighted, and query-conditioned hierarchy terms. If the additional boxes on long-tail examples are marginal or noisy, retaining

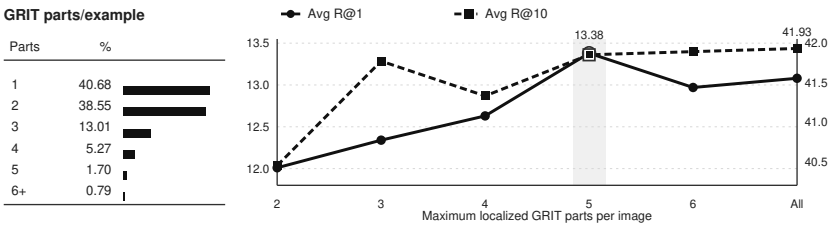

Fig. 2: Left: GRIT parts per example. Right: 10k proxy sweep over the cap on localized parts per image. Solid circles: Avg R@1; dashed squares: Avg R@10.

more localized parts per image can therefore add noisy constraints while also increasing the number of local embeddings and loss terms contributed by a single example. We therefore sweep caps of 2 to 6 localized parts per image, plus an unbounded all-parts condition, and report the 10k proxy retrieval results in the right panel of Fig. 2.

Five parts gives the best average R@1 while retaining 99.38% of localized part instances and leaving 99.20% of examples untruncated. Using all parts marginally improves Avg R@10 but does not improve Avg R@1. This suggests that the remaining long-tail parts add limited retrieval benefit while increasing the number of per-example embeddings and hierarchy losses.

5 Conclusion

We presented Hyper3-CLIP, a hyperbolic vision-language model that introduces query-conditioned visual nodes into grounded image-text training. From each caption, Hyper3-CLIP constructs a lightweight query hierarchy spanning the full caption, sentence fragments, localized text boxes, and extracted phrases. Each query pools full-image patch tokens through query-conditioned cross-attention, producing a visual node that is embedded in the shared hyperbolic space alongside its query text. These nodes support three hierarchy-aware entailment relations: between each query's visual node and its query text, between the whole image and its query nodes, and between parent and child texts. Standard global, local, and global-local contrastive alignment is retained, and since the pooling module is used only during training, inference keeps the standard dual-encoder cost of CLIP.

Zero-shot evaluations support this design. In the ViT-B setting, Hyper3-CLIP improves R@5 and R@10 retrieval on COCO and Flickr while remaining competitive on ImageNet hierarchy metrics. Under the stated prompt and evaluator protocols, it also records the highest reported 16-dataset zero-shot classification average and VOC/COCO multi-label mAP. Controlled ablations show that the full query-conditioned objective raises caption-hierarchy entailment AP without reducing average retrieval. They also show that correct text-parent links and query-conditioned visual pooling are important for this improvement. The prompt-sensitivity and parts-per-image analyses provide two practical diagnos-

tics for this setting: zero-shot classification can depend sharply on prompt wording, and a moderate cap on localized parts can retain nearly all grounded supervision while avoiding long-tail training overhead.

Overall, these results suggest that conditioning visual representations on a textual query hierarchy is an effective way to inject fine-grained, multi-granular structure into hyperbolic vision-language training. Future work could explore using query-conditioned visual nodes during inference, as well as richer query hierarchies and broader compositional evaluations of hierarchy-aware multimodal models.

A Base Hyperbolic Objective Details

This appendix gives the geometric and objective background inherited by Hyper3-CLIP. The main method builds on these components and adds query-conditioned visual nodes.

A.1 Hyperbolic Space and the Lorentz Model

Hyperbolic space has exponential volume growth, making it suitable for embeddings with tree-like or part-whole structure. Following MERU, HyCoCLIP, and UNCHA, the model represents image and text embeddings on the Lorentz manifold. With curvature magnitude $c > 0$, the d-dimensional Lorentz model is

$$\mathbb{L}_c^d = \left\{ x = (x_0, \bar{x}) \in \mathbb{R}^{d+1} : \langle x, x \rangle_L = -1/c,\ x_0 > 0 \right\},$$

where the Lorentz inner product is

$$\langle x, y \rangle_L = -x_0 y_0 + \bar{x}^\top \bar{y}.$$

The geodesic distance between two points is

$$d_{\mathbb{L}}(x, y) = \frac{1}{\sqrt{c}} \operatorname{arcosh}(-c \langle x, y \rangle_L).$$

Image and text encoders produce Euclidean features that are projected to this manifold, and contrastive retrieval uses $-d_{\mathbb{L}}(x, y)$ as the similarity score.

Hyperbolic entailment uses cones rooted at more general nodes. For an ordered pair $a \preceq b$, the node a is treated as more specific and is penalized if it falls outside the cone of b. Let $E(a \preceq b)$ denote this cone penalty, computed from the exterior angle between a and the cone rooted at b. This notation is used in the method section for both inherited part-whole relations and the added query-conditioned relations.

A.2 Inherited HyCoCLIP and UNCHA Losses

Let $g_I(\cdot)$ and $g_T(\cdot)$ denote image and text encoders after projection to the Lorentz manifold. For a batch, let $\mathcal{L}_c^*(A, B)$ denote the directional contrastive loss that matches embeddings of type A to their paired embeddings of type B and treats other batch elements as negatives. HyCoCLIP extends global image-caption alignment with grounded image-box and text-box supervision. In the notation of this paper, the inherited contrastive groups are

$$\mathcal{L}_{\text{global}}^{\text{con}} = \mathcal{L}_c^*(I, T) + \mathcal{L}_c^*(T, I),$$

$$\mathcal{L}_{\text{local}}^{\text{con}} = \mathcal{L}_c^*(I^{box}, T^{box}) + \mathcal{L}_c^*(T^{box}, I^{box}),$$

$$\mathcal{L}_{\text{global-local}}^{\text{con}} = \mathcal{L}_c^*(I^{box}, T) + \mathcal{L}_c^*(T^{box}, I).$$

UNCHA preserves these groups and modulates the global-local terms using part uncertainty, so parts estimated to be more representative of the full sample receive stronger alignment.

The inherited entailment objective applies $E(a \preceq b)$ to four HyCoCLIP relations:

$$I \preceq T, \qquad I^{box} \preceq T^{box}, \qquad I \preceq I^{box}, \qquad T \preceq T^{box}.$$

The first two relations are cross-modal image-text entailments at global and local granularity. The last two are part-whole entailments within each modality: the full image or caption contains scene context that is absent from its image box or text box, so the full node is treated as more specific.

UNCHA keeps these relations but introduces uncertainty-guided calibration for the part-whole terms. Its uncertainty estimate is derived from hyperbolic radius and is used to reduce the influence of ambiguous or weakly representative parts while preserving the same underlying compositional ordering. The inherited base objective is written compactly as

$$\mathcal{L}_{\text{UNCHA}} = \mathcal{L}^{\text{con}}_{\text{UNCHA}} + \lambda_{\text{ent}} \mathcal{L}^{\text{ent}}_{\text{UNCHA}}.$$

Hyper3-CLIP retains this base objective, while adding the query-conditioned entailment terms described in Sec. 2.3.

References

1. Alper, M., Averbuch-Elor, H.: Emergent visual-semantic hierarchies in image-text representations. In: Computer Vision – ECCV 2024. pp. 220–238. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72943-0_13
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009). https://doi.org/10.1109/CVPR.2009.5206848
3. Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., Vedantam, S.R.: Hyperbolic image-text representations. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 7694–7731. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/desai23a.html
4. Ganea, O., Bécigneul, G., Hofmann, T.: Hyperbolic entailment cones for learning hierarchical embeddings. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 1646–1655. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/ganea18a.html
5. He, N., Liu, J., Zhang, B., Bui, N., Maatouk, A., Yang, M., King, I., Weber, M., Ying, R.: Position: Beyond euclidean – foundation models should embrace non-euclidean geometries (2025), https://arxiv.org/abs/2504.08896
6. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of

Machine Learning Research, vol. 139, pp. 4904–4916. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/jia21b.html
7. Kim, H., Jang, J.H., Kim, J.J., Chun, S.Y.: Uncertainty-guided compositional alignment with part-to-whole semantic representativeness in hyperbolic vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 36861–36870 (June 2026)
8. Law, M., Liao, R., Snell, J., Zemel, R.: Lorentzian distance learning for hyperbolic representations. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 3672–3681. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/law19a.html
9. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 19730–19742. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/li23q.html
10. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/li22n.html
11. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 9694–9705. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/505259756244493872b7709a8a01b536-Paper.pdf
12. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10965–10975 (June 2022)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-10602-1_48
14. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: Computer Vision – ECCV 2024. pp. 38–55. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-031-72970-6_3
15. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. In: Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/59dfa2df42d9e3d41f5b02bfc32229dd-Paper.pdf
16. Nickel, M., Kiela, D.: Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th Inter-

national Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 3779–3788. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/nickel18a.html
17. Pal, A., van Spengler, M., di Melendugno, G.M.D., Flaborea, A., Galasso, F., Mettes, P.: Compositional entailment learning for hyperbolic vision-language models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=3i13Gev2hV
18. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world (2023), https://arxiv.org/abs/2306.14824
19. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2641–2649 (2015)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
21. Ramasinghe, S., Shevchenko, V., Avraham, G., Thalaiyasingam, A.: Accept the modality gap: An exploration in the hyperbolic space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27263–27272 (June 2024)
22. Yoshikawa, D., Matsubara, T.: PHyCLIP: ℓ_1-product of hyperbolic factors unifies hierarchy and compositionality in vision-language representation learning. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=I3Ct1eDmVI
23. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11975–11986 (October 2023)
24. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: LiT: Zero-shot transfer with locked-image text tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18102–18112 (June 2022)
25. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., Gao, J.: RegionCLIP: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16793–16803 (June 2022)
26. Zohra, F., Zhao, C., Itani, H., Ghanem, B.: b-CLIP: Text-conditioned contrastive learning for multi-granular vision-language alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 680–689 (June 2026)