Emergent Visual-Semantic Hierarchies in Image-Text Representations

Morris Alper and Hadar Averbuch-Elor

Tel Aviv University

Abstract. While recent vision-and-language models (VLMs) like CLIP are a powerful tool for analyzing text and images in a shared semantic space, they do not explicitly model the hierarchical nature of the set of texts which may describe an image. Conversely, existing multimodal hierarchical representation learning methods require costly training from scratch, failing to leverage the knowledge encoded by state-of-the-art multimodal foundation models. In this work, we study the knowledge of existing foundation models, finding that they exhibit emergent understanding of visual-semantic hierarchies despite not being directly trained for this purpose. We propose the *Radial Embedding* (RE) framework for probing and optimizing hierarchical understanding, and contribute the *HierarCaps* dataset, a benchmark facilitating the study of hierarchical knowledge in image–text representations, constructed automatically via large language models. Our results show that foundation VLMs exhibit zero-shot hierarchical understanding, surpassing the performance of prior models explicitly designed for this purpose. Furthermore, we show that foundation models may be better aligned to hierarchical reasoning via a text-only fine-tuning phase, while retaining pretraining knowledge. Our code, data, and trained models are available at the project page: `https://hierarcaps.github.io/`.

Keywords: Hierarchical representations · Multimodal learning · Vision and language

1 Introduction

The visual world is full of hierarchies. Upon seeing something green flying through the sky, the average person will correctly identify it as a parrot, knowing that it is a special bird which is a type of animal. A trained ornithologist might even recognize that it is a ring-necked parakeet. In general, images and their possible descriptions form a *visual-semantic hierarchy* [57], a directed graph of logical entailment (implication) between items. As shown in Figure 1, a description such as *mammal* may describe many images, while *funny giraffe* describes a smaller subset. Each image is closest to a highly specific description (e.g. *A giraffe's head looking at the camera and making a face.*) which logically entails the corresponding more general descriptions.

The ubiquitous nature of hierarchies in images suggests that an ideal vision-language model (VLM) would be able to understand them similarly to the

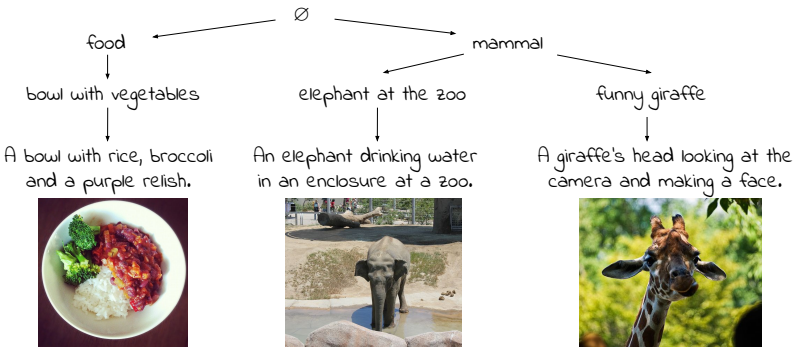

Fig. 1: A single image may be described by many text of varying levels of descriptiveness. While SOTA multimodal foundation models are commonly used to retrieve a single text matching an image, we show that they have learned to model hierarchies. By applying our RE framework to foundation models, we may perform hierarchical image-text matching to place images and captions in the context of a *visual-semantic hierarchy* which encompasses the relative meanings of all possible images and texts. Above, we show a slice of the visual-semantic hierarchy obtained with our method, with $\emptyset$ indicating the root node in the hierarchy and arrows corresponding to the logical entailment relation between general and more specific descriptions.

aforementioned humans. However, state-of-the-art VLMs such as CLIP [47] were trained on tasks like text-image matching which do not explicitly model hierarchical knowledge. Inspired by prior works studying the knowledge acquired by multimodal foundation models [1, 2, 8, 33, 56, 66], we ask the question: Do such VLMs develop an *emergent* understanding of visual-semantic hierarchies? Such understanding could shed light on the inner working of powerful foundation models typically used as black boxes; in addition, prior work has found hierarchical understanding to provide an important inductive bias to improve performance on tasks such as image classification [5,34,44,46,61,62]. We investigate this both in zero-shot as well as fine-tuned settings, to both investigate the presence of this emergent knowledge used as-is as well as its utility as an initialization for transfer learning applied to hierarchical tasks.

In order to probe and optimize models for hierarchical understanding, we must first define the geometry which represents hierarchical entailment in a continuous space. Prior works have represented logical entailment between texts or between images and text as a partial order relation [57]; in particular, the Entailment Cone (EC) framework represents entailment as inclusion in a cone radiating from a point away from the origin [24]. However, these works train models from scratch with EC-based objectives, as well as generally operating in hyperbolic space (while common foundation VLMs operate in Euclidean space). To better model the emergent hierarchical geometry of foundation VLMs, we introduce the Radial Embedding (RE) framework, which relaxes the assumptions of EC to match the existing geometric configuration of pretrained VLM

representations. We show that the RE framework demonstrates the emergent hierarchical knowledge of VLMs. Furthermore, we propose a RE-based contrastive objective for model fine-tuning to enhance hierarchical understanding in VLMs, and show that this outperforms EC-based optimization for these models.

A significant obstacle for studying visual-semantic hierarchies is the lack of ground-truth data, as existing datasets only contain images paired with single or unrelated reference captions. Therefore, we propose the *HierarCaps* dataset, containing 73K images paired with multiple valid texts arranged in a logical hierarchy. To create *HierarCaps*, we leverage large-scale image-caption Internet data and enrich it with hierarchical text generation using a large language model (LLM) and natural language inference (NLI). We also provide a manually-curated 1K-item test set and contribute evaluation metrics for benchmarking visual-semantic hierarchical understanding, quantifying the multimodal hierarchical knowledge which prior works only evaluate qualitatively [20, 57]. We show that *HierarCaps* may supervise VLM fine-tuning with our RE framework, generally boosting hierarchical understanding of pretrained models. Importantly, this can be achieved while still retaining pretrained knowledge – thus performing model *alignment*, parallel to the alignment stages of recent foundation models which unlocks emergent abilities such as instruction following [17, 45].

We validate the effectiveness of our approach on *HierarCaps* and on existing benchmarks for related tasks such as lexical entailment prediction and hierarchical classification. Our evaluation shows that foundation VLMs indeed exhibit emergent hierarchical understanding, significantly outperforming prior models which learn hierarchical representations directly even in the zero-shot setting, with further overall performance enhancement after fine-tuning. We release[1] our code, data, and trained models to the research community to spur future progress on multimodal hierarchical understanding.

2 Related Work

V&L representation learning. While earlier work in image representation learning focused on models trained on supervised benchmarks such as ImageNet [32], recent approaches have achieved state-of-the-art via contrastive training on web-scale collections of paired images and text [26, 47]. These multimodal representations exhibit various emergent geometric properties which were not explicitly optimized. For instance, vector arithmetic in CLIP's semantic space has been found to correlate with semantic relationships, similar to well-known properties of word embeddings [15, 54, 56]. Liang et al. [36] explore the geometry of text and image embeddings in CLIP space, finding that the modalities are contained in narrow cones separated by a consistent cross-modal gap. Similarly, our work explores the emergent geometry of hierarchies in V&L representations. In addition, our text-only alignment approach bears some similarity to works focusing on deficiencies in textual representations of existing VLMs and how to adapt these models to imbue additional textual understanding [6, 9, 28, 48, 63, 64].

[1] `https://hierarcaps.github.io/`

Hierarchical multimodal reasoning. Images may be described by multiple valid texts of varying granularity levels, forming a hierarchy which may be learned or exploited in various ways. One approach infers concept hierarchies from paired text and images, learning text and image embeddings jointly to implicitly infer the generality of concepts and their logical entailment structure [29, 57, 67]. Other works generate texts or knowledge graphs from images with varying levels of detail and attention to different regions; methods include controllable captioning [10, 19], contrastive captioning [16, 39], dense captioning [27], guided decoding [30], and scene graph generation [68]. Another approach uses existing hierarchical structure in ground-truth annotations to improve predictions for tasks such as image classification [5, 34, 46]; these works use concept hierarchies over categorical image labels, while our benchmark consists free text arranged in entailment hierarchies, not confined to a small set of categories.

We are currently witnessing pioneering efforts in understanding the semantic hierarchies learned by neural VLMs such as CLIP, though current knowledge in this field is still sparse. Xu et al. [61] find that VLMs are more successful at matching fine-grained concepts with images, while underperforming on general concepts (e.g. *leopard* vs. *feline*). Yi et al. [62] show that CLIP may be outperformed on ImageNet classification by a model explicitly trained using concept hierarchies from WordNet. Novack et al. [44] propose a zero-shot classification pipeline with CLIP by leveraging concept hierarchies during inference. Desai et al. [20] introduce neural hyperbolic V&L embeddings with hierarchical image–text matching as one use case, showing qualitative results alone for this task.

Hierarchical embeddings. Hierarchical representations have attracted interest in various fields due to the hierarchical nature of of many types of data. A number of works have approached text (primarily words) as hierarchical for the purpose of learning embeddings to model semantic relationships such as hypernymy [3, 7, 18, 41, 59]. Many works model data with geometric objects that directly represent relations between items [60], notably including the use of hyperbolic manifolds due to their attractive properties for representing hierarchical graphs [50, 53]. Nickel and Kiela [42] introduce hyperbolic word embeddings, and subsequent works refine this idea by exploring different models of hyperbolic space [43], the geometry of entailment relationships [24], and extensions to texts of arbitrary length [22, 65]. Recent work has also applied hyperbolic learning to various tasks in computer vision [4, 21, 23, 35, 40] and multimodal learning [20, 25, 38]. While we take inspiration from these works, particularly the hierarchical nature of text-image data from Desai et al. [20] and the entailment cone framework introduced by Ganea et al. [24], we focus on leveraging the strong knowledge of existing pretrained models such as CLIP which is embedded in Euclidean space, rather than training from scratch in hyperbolic space.

3 Preliminaries

We proceed to cover preliminary mathematical and geometric concepts which provide background for our study of hierarchical knowledge in VLMs. We refer

the reader to Ganea et al. [24] and Desai et al. [20] for further exposition and illustration of the concepts described below.

Notation. Our discussion includes the following notation: $\text{sim}(\mathbf{v}, \mathbf{w}) := \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\| \|\mathbf{w}\|}$ denotes the cosine similarity function between two vectors, and the angle spanned by three points $\mathbf{a}, \mathbf{b}, \mathbf{c}$ in space is denoted by $\angle\mathbf{abc}$. $[x]^+ := \max(0, x)$ is the positive part (ReLU) function.

Geometric Preliminaries. Consider embeddings in embedding space M (e.g. $\mathbb{R}^n$ for Euclidean embeddings). We are interested in embeddings which represent hierarchical relations between items. In particular, embeddings $\mathbf{e}, \mathbf{e}' \in M$ of items x, x' should have a particular geometric configuration if x is entailed by x'. For example, for text embeddings, if $x =$*"animal"* and $x' =$*"cat"* then x is entailed by x' and thus $\mathbf{e}$ and $\mathbf{e}'$ should have a certain relative positioning in embedding space. Denote by $\mathbf{r} \in M$ the *entailment root* (or simply *root*), a special fixed point in space which is the anchor used as a reference point for all entailment relationships. Considering distinct $\mathbf{e}, \mathbf{e}' \in M$ (also distinct from the root), we define the exterior angle $\Xi_{\mathbf{r}}(\mathbf{e}, \mathbf{e}') := \pi - \angle\mathbf{ree}' = \arccos\left(\text{sim}(\mathbf{e} - \mathbf{r}, \mathbf{e}' - \mathbf{e})\right)$, and $d_{\mathbf{r}}(\mathbf{e}) := \|\mathbf{e} - \mathbf{r}\|$, the Euclidean distance of $\mathbf{e}$ from the root.

Entailment Cone Embeddings. Ganea et al. [24] introduce the Entailment Cone (EC) framework for hierarchical representation learning. EC embeddings possess a defined geometric relation between items which is a *partial order*, a desirable mathematical property for modeling logical entailment. For instance, partial orders and logical entailment both obey transitivity (e.g. *black cat* is a type of *cat* which is a type of *animal*, so *black cat* is a type of *animal*). See the supplementary materials for the mathematical definition of partial orders and further explanation of their relevance to entailment hierarchies.

In the EC setting, a relation between two embeddings $\mathbf{e}, \mathbf{e}'$ is defined by $\mathbf{e} \leq \mathbf{e}' \leftrightarrow \mathbf{e}' \in C_\theta(\mathbf{e})$, where $C_\theta(\mathbf{e}) := \{\mathbf{e}' \in M : \Xi_{\mathbf{r}}(\mathbf{e}, \mathbf{e}') \leq \theta\}$ is a convex cone with half-aperture angle θ originating at $\mathbf{e}$ and radiating away from the origin. See Figure 2 for an illustration. Ganea et al. show that this is a partial order if θ is defined as a function $\theta_{\mathbf{r}}(\mathbf{e})$ which varies with the distance of $\mathbf{e}$ from the root $\mathbf{r}$; in Euclidean space, as $\sin\theta_{\mathbf{r}}(\mathbf{e}) = \epsilon / d_{\mathbf{r}}(\mathbf{e})$ for constant $\epsilon > 0$.

In the EC framework the positive excess $\mathcal{L}_{EC} := \left[\Xi_{\mathbf{r}}(\mathbf{e}, \mathbf{e}') - \theta_{\mathbf{r}}(\mathbf{e})\right]^+$ may be used as a margin loss to encourage pairs of inputs with an entailment relation to satisfy the partial order relation [20, 24]. For example, as shown in the left-hand side of Figure 2, if $(\mathbf{e}, \mathbf{e}')$ are embeddings of a pair of items in an entailment relationship (e.g. *animal* and *cat*), then this loss pushes $\mathbf{e}'$ into the entailment cone $C_{\theta_{\mathbf{r}}(\mathbf{e})}(\mathbf{e})$ if it deviates from this cone.

4 Radial Embeddings

We now introduce our Radial Embedding (RE) framework, which relaxes the assumptions of EC and modifies its optimization method for compatibility with the representations learned by pretrained foundation VLMs. Our underlying observation is that foundation VLM representations typically place generic concepts

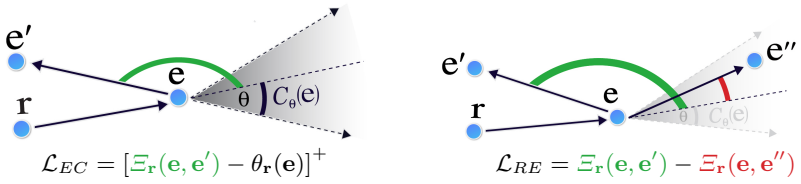

Fig. 2: Illustration of EC and RE optimization. Above we show examples of cases of positive loss under both frameworks. In the EC framework (left), any embedding in the cone $C_{\theta_{\mathbf{r}}(\mathbf{e})}(\mathbf{e})$ represents an item entailed by $\mathbf{e}$ (while $\mathbf{e}'$ is outside this cone). The half-aperture angle $\theta_{\mathbf{r}}(\mathbf{e})$ varies with distance from the root embedding $\mathbf{r}$ to enforce a partial order. During training, the deviation from this cone defines a margin loss. In our proposed RE framework (right), the loss is instead given by the difference between exterior angles of positive and negative examples, and with no dependency on $\theta_{\mathbf{r}}(\mathbf{e})$. In the above case, this loss is positive since the positive item $\mathbf{e}'$ has a larger exterior angle than the negative item $\mathbf{e}''$.

in central locations relative to related, more specific concepts. Intuitively, this agrees with the contrastive objective of models like CLIP which encourages similar concepts to cluster together; for instance, the embedding of *animal* should be close to related, specific concepts such as *dog*, *zebra*, and *fish*; hence, it is expected to be more centrally located. This suggests an emergent hierarchical structure. To uncover (and optionally enhance) this hierarchical structure, we relax assumptions tying entailment to a partial order relation based on entailment cones, as pretrained foundation VLMs have already learned a configuration in high-dimensional Euclidean space that does not necessarily abide by the strong requirements of EC. Instead, we identify the inherent hierarchical geometric configuration of pretrained VLMs, first locating its natural entailment root and then defining functions for hierarchical understanding.

Empty String as Entailment Root To identify an entailment root, we exploit the ability of VLMs to encode generic semantics in underspecified inputs. To this aim, we use the embedding of the empty string $\mathbf{e}_\emptyset$ as the entailment root $\mathbf{r}$, as an empty caption may accompany virtually any image. This allows us to use the pretraining knowledge encapsulated by these models to match their learned embedding configurations. Furthermore, when we perform model alignment (described below) this embedding affects the gradient of the model's weights as part of the loss function, thus optimizing its position.

Measuring Genericness and Entailment with RE. Given the entailment root $\mathbf{r}$ defined above, we adopt the use of the function $d_{\mathbf{r}}(\cdot)$ to measure concept genericness [20]. To measure entailment, we use the exterior angle function $\Xi_{\mathbf{r}}(\cdot, \cdot)$; unlike prior work on entailment cones, we use its value directly without reference to an absolute threshold. This allows us to exploit hierarchical structure in embeddings without making strong assumptions such as the existence of an EC-based partial order.

4.1 RE-Based VLM Alignment

In order to fine-tune pretrained VLMs for enhanced hierarchical understanding, we propose a contrastive objective which separates positive and negative pairs without requiring an absolute point of reference for entailment. We exploit the configuration of hierarchical embeddings and the design of *HierarCaps* (as described in Section 5) by assuming that for each positive pair $(\mathbf{e}, \mathbf{e}')$ there is a negative pair $(\mathbf{e}, \mathbf{e}'')$. For example, $(\mathbf{e}, \mathbf{e}', \mathbf{e}'')$ could be the embeddings of the texts $t =$*"animal"*, $t' =$*"goat"*, and $t'' =$*"portrait"* respectively, as t is entailed by t' and t contradicts t'' (and thus is not entailed by t''). Hence, we define the contrastive RE loss function $\mathcal{L}_{RE} := \Xi_{\mathbf{r}}(\mathbf{e}, \mathbf{e}') - \Xi_{\mathbf{r}}(\mathbf{e}, \mathbf{e}'')$.

Conceptually, $\mathcal{L}_{RE}$ encourages the positive pair $(\mathbf{e}, \mathbf{e}')$ to have a relatively smaller exterior angle than the negative pair $(\mathbf{e}, \mathbf{e}'')$, as illustrated in Figure 2. A direct comparison to $\mathcal{L}_{EC}$ shows significant performance gains for hierarchical image–text retrieval (as illustrated in Section 6.3). In the supplementary material, we also show that $\mathcal{L}_{RE}$ can be derived as the limit of a margin EC loss applied to contrastive pairs as the margin tends to infinity, which empirically outperforms smaller margins in our evaluation.

We emphasize that $\mathcal{L}_{RE}$ is fundamentally different from conventional contrastive loss functions. While standard triplet losses also consider triplets of items such as $(\mathbf{e}, \mathbf{e}', \mathbf{e}')$, in the standard case the positive relation between $\mathbf{e}$ and $\mathbf{e}'$ is symmetric [13,51]. By contrast, in our case $\mathbf{e}$ and $\mathbf{e}'$ cannot be swapped, because logical entailment is an asymmetric relation. This is seen in the above example where *"goat"* is a type of animal while *"animal"* is not a type of goat. Hence, the exterior angle terms $\Xi_{\mathbf{r}}(\mathbf{e}, \cdot)$ in the definition of $\mathcal{L}_{RE}$ are asymmetric functions of their arguments, unlike the Euclidean or cosine distance functions typically used in contrastive learning. Moreover, $\mathcal{L}_{RE}$ depends on the learnable root $\mathbf{r}$. This loss pushes items entailed by $\mathbf{e}$ towards the half-line $(\mathbf{r}, \mathbf{e})$, unlike standard contrastive losses which only encourage similar items to be close and dissimilar items to be far apart in representation space.

5 The *HierarCaps* Dataset

While there exist large-scale datasets of paired images and captions, these typically contain a single caption per image or multiple independent reference texts. To evaluate and optimize visual-semantic hierarchical understanding, we propose a new dataset and benchmark, *HierarCaps*, containing images paired with multiple valid texts arranged in a logical hierarchy. To generate this data we use existing image captioning datasets along with a LLM- and NLI-guided hierarchy generation procedure; the train set is produced fully automatically, while the test set is manually curated and corrected to serve as a clean evaluation benchmark.

5.1 Design and Contents

[1] We show a synthetic image in place of the original image from Conceptual Captions due to licensing.

Positive	Negative
animal	*portrait*
goat	*frog*
goat on island	*goats graze at snowy night*
a goat eating leaves of a lemon tree on the island	*goat on island with flowers blooming in the spring*

Fig. 3: Sample item from the *HierarCaps* train set[1]. Ground-truth captions have a four-tiered hierarchical structure. The first tier contains the most generic description matching the image (*animal*), the last contains the most specific description (*a goat eating leaves...*), and each (positive) tier is logically entailed by the following tier. The train set also contains corresponding negative captions; corresponding captions in the same tier logically contradict each other, and each positive caption is implied by both (positive and negative) captions in the following tier.

HierarCaps contains images with paired hierarchical caption data, including a 73K-item train set and 1K-item test set. See Figure 3 for a sample from its train set. Items in *HierarCaps* have a four-tiered hierarchical structure, designed by supplementing ground-truth Internet image captions with a minimal amount of added data to encompass the tasks of lexical and textual entailment. *Lexical entailment* refers to the logical relation between words such as *cat* implying *animal* (as a cat is a type of animal), while *textual entailment* refers to logical implication between full texts, such as *this is a large cat* implying *this is a cat*. As seen in Figure 3, the first two tiers in *HierarCaps* caption hierarchies roughly correspond to lexical entailment, as they usually contain single words or short phrases, while the last two tiers correspond to textual entailment of longer texts. In this way, training or evaluating with *HierarCaps* subsumes hierarchical understanding on the two varieties of entailment, which have both been of significant interest in NLP and multimodal learning.

Each image has four accompanying positive captions (seen on the left of Figure 3). The text in the fourth tier (most specific) is the original caption from an image-caption dataset, while all other texts are LM-generated (described in in Section 5.2), forming a logical entailment hierarchy (e.g. *animal* $\rightarrow$ *goat* $\rightarrow$ *goat on island* $\rightarrow$ *a goat eating leaves of a lemon tree on the island*). In addition, images in the train set also possess four counterfactual negative captions (seen on the right of Figure 3). These are constructed to logically contradict the corresponding positive item in the same tier while also entailing the positive item from the previous tier; for example *frog* in tier 2 of the figure entails *animal*, while contradicting *goat*. When using *HierarCaps* for fine-tuning, the positive and negative from each tier provide the paired contrastive data needed for the RE framework's loss function, as described in Section 4.1.

5.2 Dataset Construction

Modern multimodal learning benefits from large-scale data, particularly an abundance of paired images and textual captions. Unfortunately, in our setting the

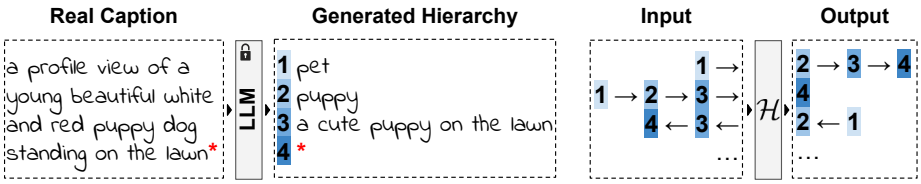

Fig. 4: Dataset construction pipeline, used to create *HierarCaps*. Real image captions from Conceptual Captions are fed to a LLM with various prompts to produce caption hierarchies; above, * indicates the original caption, enriched with a full hierarchy of shorter, logically entailed captions. These are filtered with an NLI model to enforce logical entailment and augmented with few-shot LLM text completion, producing a set of *seed hierarchies*. We then distill these hierarchies into a much smaller language model $\mathcal{H}$ which learns to complete hierarchies in both directions; as described in Section 5.2, $\mathcal{H}$ is used to produce the final hierarchies in *HierarCaps*.

desired data (multiple hierarchically-arranged captions for an image) does not naturally occur alongside Internet images. Images on the Internet most commonly appear with a single caption (or alt text); this data was used for example to produce the Conceptual Captions [52] (CC) dataset. More reference texts per image may be produced with manual annotation, as was done for the Microsoft COCO Captions [11, 37] dataset, but in this case each reference is produced independently and thus they do not form a consistent hierarchical structure. Therefore, we instead build upon these existing image-caption datasets to enrich them with hierarchical texts, using the powerful text generation abilities of modern LLMs and logical understanding of pretrained NLI models.

Our dataset construction pipeline is illustrated in Figure 4. We first automatically create seed data using captions from the train set of CC, using the SOTA LLM Llama 2 [55] for text generation along with heavy NLI filtering to enforce logical entailment. We engineer prompts to produce the desired output for each tier; these prompts are reproduced in the supplementary material. While this process is inefficient and only succeeds in a minority of cases, we augment them with more synthetic hierarchies using in-context completion with Llama 2. We then distill this core of examples into a smaller language model $\mathcal{H}$ (using the encoder-decoder model Flan-T5 [14]) that learns to generate a valid positive hierarchy given an input detailed caption. We then run it on all input captions (CC train captions for our train set, COCO validation captions for our test set). $\mathcal{H}$ is trained to complete caption hierarchies in both directions (specific to general and general to specific); on the train set, after producing the positive hierarchies we run $\mathcal{H}$ in the general→specific direction to deliberately hallucinate negative captions. All outputs are filtered with NLI to ensure the correct entailment and contradiction relations between texts.

Along with 73K automatically-generated train items, we manually review and correct 1K test items as a clean evaluation benchmark. When performing

this test set curation, we found 75% of automatically-generated items to be fully valid; remaining items typically contained minor inaccuracies. As we show below, training on our fully automatically-generated train set provides an empirical performance boost despite the presence of such noise. See the supplementary material for further dataset construction details, including models used, inference settings, examples of inaccuracies produced by automatic generation, and additional technical details.

6 Experiments

Models Considered. We focus on dual encoder VLMs (i.e. paired text and image encoders). We evaluate OpenAI and LAION implementations of CLIP [12,47] which we refer to as *CLIP* and *OpenCLIP* respectively, and we test an open implementation of ALIGN [26]. We test various model sizes; see the supplementary material for checkpoints and numbers of parameters.

Full Alignment Framework. As seen in Figure 3, each caption hierarchy in the *HierarCaps* train set consists of four positive captions (P_1, P_2, P_3, P_4) and four negative captions (N_1, N_2, N_3, N_4); we use caption triplets (P_i, P_{i+1}, N_i) for tiers $i \in \{1, 2, 3\}$ to calculate the RE loss $\mathcal{L}_{RE}$ as defined in Section 4.1. We aggregate this loss over tiers 1–3 along with hard example mining within minibatches to compute aggregate RE loss $\tilde{\mathcal{L}}_{RE}$. We also add an additional loss term that discourages fine-tuned text embeddings from deviating from their original values. Given fine-tuned embeddings $\{\mathbf{e}_i\}_{i=1,\cdots,8}$ of the image's ground-truth caption texts, and corresponding original (pretrained, not fine-tuned) embeddings $\{\mathbf{e}_i^*\}_{i=1,\cdots,8}$, we define the regularization loss $\mathcal{L}_{reg} := -\frac{1}{8}\sum_{i=1}^{8} \text{sim}(\mathbf{e}_i, \mathbf{e}_i^*)$, i.e. the mean cosine similarity loss between pretrained and fine-tuned text representations. The total loss for fine-tuning is given by $\mathcal{L}_{total} := \lambda_{RE}\tilde{\mathcal{L}}_{RE} + \lambda_{reg}\mathcal{L}_{reg}$. For all fine-tuned models, the image encoder is frozen and we only update weights of the text encoder, training for a single epoch on the *HierarCaps* train data. Following the pretraining procedure of the models under consideration, we unit-normalize all embeddings before calculating losses (i.e. we constrain embeddings to the surface of the unit sphere). Loss calculation details and training hyperparameters are detailed in the supplementary material.

6.1 Test Datasets and Metrics

We evaluate all models both zero-shot and after alignment (fine-tuning) on the *HierarCaps* train set. We propose evaluation metrics on *HierarCaps* to quantify multimodal hierarchical understanding, contrasting with the purely qualitative evaluations provided in prior works [20,57]. We complement this by evaluating on various external datasets measuring hierarchical and multimodal understanding in other settings. We describe our evaluation procedure below.

HierarCaps. We perform hierarchical image–text matching on *HierarCaps* using the methodology of Desai et al. [20] by selecting 50 equally-spaced points

between the root node and the closest text embedding to the given image; at each point we retrieve the text closest to the image (via cosine similarity) within the given radius of the image. We calculate standard retrieval metrics (precision and recall) relative to the four ground-truth captions for each image. In addition, we calculate the order-aware metric τ_d to check that ground-truth texts are correctly ordered by their distances from the root node (i.e. the embedding of the empty string); for a given image, this equals the Kendall correlation between $d_{\mathbf{r}}(\mathbf{e})$ for each ground-truth text's embedding $\mathbf{e}$, and their ground-truth order. In particular, this equals 1 if and only if all ground-truth texts are ordered correctly by these distances. Note that unlike the retrieval metrics previously described, this metric measures to what extent the textual embedding space agrees with the expected (ground truth) hierarchical structure. These are calculated relative to the manually-curated test set; for qualitative results we perform retrieval over an expanded set of candidate texts (see supp.) to yield more extensive hierarchies.

HyperLex. The HyperLex dataset [58] is a benchmark for lexical entailment understanding. We use the standard metric of Spearman correlation with ground-truth entailment scores, using the exterior angle $\Xi_{\mathbf{r}}(\mathbf{e}, \mathbf{e}')$ between embeddings of word pairs as the predicted entailment score. As in prior works, we report this score on all items (ρ_{all}) as well as restricted to nouns (ρ_N). See the supplementary material for details on prompts used.

BREEDS. The BREEDS dataset [49] contains a subset of images from ImageNet annotated with two-tiered hierarchical labels (e.g. *stringed instrument* $\rightarrow$ *banjo*); we use the subsets[2] of BREEDS selected by Novack et al. [44] and test the ability of our models to predict both the coarse and fine-grained labels for a given image (in the correct order) out of all labels of all granularities. In order to predict ordered pairs of labels, we consider all pairs of distinct labels (x, y) where the embedding of x is closer to the entailment root, and calculate the score $c_x + c_y + C \cdot \Xi_{\mathbf{r}}(x, y)$, where c_x is the CLIP similarity between x and the image (and similarly c_y), C is a constant chosen by cross-validation, and $\Xi_{\mathbf{r}}(x, y)$ is the exterior angle between the embeddings of texts x and y. We predict the highest scoring $k \in \{1, 5\}$ pairs of labels and report recall at k.

Standard multimodal benchmarks. We evaluate on various standard cross-modal tasks to test whether pretraining knowledge is preserved. Results for cross-modal retrieval on MS-COCO [11, 37] and zero-shot image classification on CIFAR-10 and CIFAR-100 [31] are given here; results on additional datasets are given in the supplementary material.

6.2 Results and Discussion

Quantitative results are displayed in Table 1. Across tasks and datasets, we see that our RE framework demonstrates that models like CLIP possess hierarchical understanding in the zero-shot regime, while our fine-tuning generally further enhances performance across model types and sizes. In addition, Figure

[2] living17, nonliving26, entity13, entity30

	HierarCaps			HyperLex		BREEDS		COCO		C10	C100
Model	P	R	τ_d	ρ_{all}	ρ_N	R@1	R@5	R@1	R@5	acc	acc
CLIP^{B}	0.14	0.36	0.89	0.44	0.48	0.22	0.50	0.30	0.55	0.90	**0.66**
$\text{CLIP}^{\text{B}}_{\text{FT}}$	**0.15**	**0.47**	**0.99**	**0.51**	**0.55**	**0.24**	**0.55**	**0.31**	**0.56**	0.90	0.65
CLIP^{L}	**0.16**	0.37	0.88	0.42	0.44	0.29	0.61	0.36	0.60	0.96	0.78
$\text{CLIP}^{\text{L}}_{\text{FT}}$	0.15	**0.44**	**0.97**	**0.50**	**0.54**	**0.32**	**0.65**	0.36	**0.61**	0.96	**0.79**
$\text{OpenCLIP}^{\text{B}}$	**0.16**	0.33	0.87	0.34	0.37	0.23	0.50	0.39	0.65	**0.94**	0.76
$\text{OpenCLIP}^{\text{B}}_{\text{FT}}$	0.14	**0.40**	**0.98**	**0.49**	**0.55**	**0.25**	**0.56**	0.39	0.65	0.93	**0.77**
$\text{OpenCLIP}^{\text{H}}$	0.16	0.32	0.83	0.06	0.03	0.23	0.50	0.48	0.73	0.98	0.86
$\text{OpenCLIP}^{\text{H}}_{\text{FT}}$	0.16	**0.36**	**0.97**	**0.37**	**0.39**	**0.31**	**0.65**	0.48	0.73	0.98	0.86
ALIGN	0.16	0.36	0.89	0.35	0.37	0.21	0.52	0.42	0.67	0.78	0.53
ALIGN_{FT}	0.16	**0.42**	**0.96**	**0.44**	**0.47**	**0.23**	**0.58**	0.42	0.67	0.78	0.53

Table 1: Results for hierarchical text–image matching on *HierarCaps* (test set), existing hierarchical understanding benchmarks, and standard text→image retrieval on COCO (val set). Best results are in **bold**; ties are underlined. Superscript letters indicate model size. C10 and C100 indicate CIFAR-10 and 100 respectively. P and R are precision and recall, and τ_d is the order-aware metric defined in Section 6.1. ρ_{all} and ρ_N are Spearman correlation between predicted and ground-truth values for all items and for nouns. R@k refers to recall at k and acc refers to categorical accuracy. As seen above, fine-tuning mostly improves hierarchical understanding without significantly impacting standard cross-modal task performance.

5 shows qualitative results of hierarchical image–text matching with CLIP on *HierarCaps*. We see that CLIP already shows the emergent ability to perform hierarchical retrieval, although its quality is further improved by model alignment. We note that this improvement is most evident for more general terms, consistent with prior work observing that foundation VLMs may struggle with matching highly general terms to images [44,61]. In the supplementary material, we further show that this improvement is seen even when controlling for text length.

As seen in the cross-modal retrieval results for COCO in Table 1 (and as shown on additional datasets and tasks in the supplementary material), our fine-tuning has a negligible effect on standard multimodal tasks. This supports our fine-tuning being a type of model alignment, bringing latent knowledge of hierarchies to the surface without significantly impacting pretrained knowledge.

Additionally, our results on existing hierarchical understanding datasets in Table 1 show that VLMs can be probed and aligned with the RE framework for more general hierarchical understanding, beyond hierarchical image–text matching and the specific format of *HierarCaps*. Results on BREEDS show the applicability of our methodology to other multimodal datasets. Results on HyperLex demonstrate the surprising finding that VLMs such as CLIP with alignment applied perform competitively to methods using dedicated embeddings for the

CLIP	+alignment	CLIP	+alignment
fun	*food animal*	*two*	*cat*
top	*vegetables*	*sleep*	*cats*
...	...	...	...
a close up of a plate of food with broccoli	*A raw piece of broccoli with something growing from it.*	*cats sleeping with a remote*	*Couple of cats sleeping on opposite ends of the couch*
A worm sits on top of a piece of broccoli.	*A worm sits on top of a piece of broccoli.*	*Two cats sleeping with a remote control near each of them.*	*Two cats sleeping with a remote control near each of them.*

Fig. 5: Qualitative hierarchical text-image matching results on *HierarCaps* (test set). The left column of each table shows hierarchical text-image matching applied using pretrained CLIP-Large, while the right column shows results on the same images after alignment (fine-tuning). The results above are abridged; for full results, see the supplementary material.

unimodal (text-only) task of lexical entailment prediction. For example, our best ρ_{all} is 0.51 versus 0.69 for LEAR [59] and 0.59 for DOA-E [3], both of which were trained explicitly on lexical entailment data; see the supplementary material for an extensive comparison to prior methods. While lexical entailment prediction is a non-obvious task for VLMs as it is purely unimodal (text-only), we hypothesize that paired image-text data naturally provides supervision for learning concept hierarchies (supported by prior works discussed in Section 2), and that this knowledge is unlocked by our alignment procedure.

We also provide a comparison to MERU [20], a recent hierarchical understanding model which is trained from scratch (rather than building on top of an existing pretrained foundation model). We use the strongest MERU variant, a hyperbolically-trained dual encoder model with a ViT-L backbone, and compare our RE results on those of the comparable CLIP-L model (reported in Table 1). For standard multimodal metrics, we reproduce MERU's reported 0.32 R@5 on COCO text→image (vs. our 0.61 for fine-tuned CLIP-L). We also perform hierarchical retrieval on *HierarCaps* (using the hyperbolic inference procedure of Desai et al. to match MERU's learned geometry), yielding precision 0.11, recall 0.11, and τ_d 0.79, far underperforming our results. As seen in Figure 6, these metrics reflect weaker performance on both semantic text-image matching and hierarchical knowledge, emphasizing the importance of strong foundation VLMs for downstream tasks such as ours.

6.3 Ablations

To justify the effectiveness of our framework relative to the existing EC framework (described in Section 3), we align CLIP-Base while replacing the loss $\mathcal{L}_{RE}$

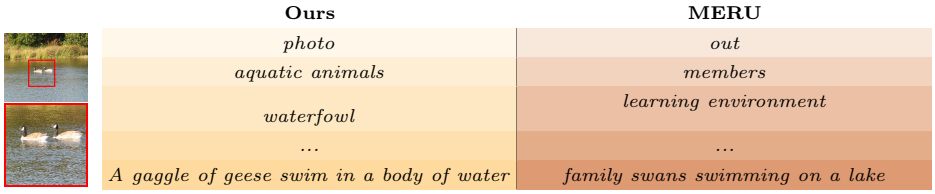

Fig. 6: Comparison to MERU [20]. We show an image from the *HierarCaps* test set and a zoomed-in view bordered in red, along with our (aligned CLIP-L) hierarchical text-image matching results and those of MERU (ViT-L). As seen above, our fine-tuned model yields more semantically plausible results with a clear hierarchical structure.

in fine-tuning with an EC loss. Following Ganea et al. [24], we apply this to both positive and negative examples by using the loss $\mathcal{L}_{EC} = [\pm(\Xi_{\mathbf{r}}(\mathbf{e}, \mathbf{e}') - \theta_{\mathbf{r}}(\mathbf{e}))]^+$, where the inner sign is negative for negative examples. Here the half-aperture angle defined by $\sin\theta_{\mathbf{r}}(\mathbf{e}) = \min(1, 0.05/d_{\mathbf{r}}(\mathbf{e}))$ For hierarchical image–text matching on *HierarCaps*, this yields precision 0.13 and recall 0.30, significantly underperforming our RE alignment results as seen in Table 1.

In the supplementary material, we further ablate each key element of our framework to show their necessity, including the use of pretrained weights followed by fine-tuning with our loss; the use of a learnable root initialized with the empty string embedding (rather than the fixed roots used in prior works [20,24]); the use of regularization loss, and hard negative mining; and the structure of our training data (four-tiered, with positive and negative items). We also show that learning the position of the root vector alone (while keeping other embeddings fixed) is insufficient, and we compare to additional EC framework variants.

7 Conclusion

In our work, we study the emergent capability of foundation VLMs to understand visual-semantic hierarchies, proposing the Radial Embedding (RE) framework to probe and optimize them for hierarchical understanding. We present the *HierarCaps* dataset and benchmark for hierarchical text-image matching, along with evaluation metrics for this task. We show that these models demonstrate emergent hierarchical understanding, outperforming prior methods designed explicitly for representing hierarchical structures even when used zero-shot, with an additional overall performance boost after fine-tuning.

Regarding limitations, we focus on the representation learning setting using dual encoder VLMs, while other architectures might require a different analysis. Additionally, while we assign a single linear hierarchy of texts to an image, the semantic hierarchy of texts describing an image naturally form a branching structure, as they may focus on different visual details. Future work could explicitly model this branching structure to automate organization of image collections similarly to the manually-organized branching hierarchical structure of widely-used online image repositories such as Wikimedia Commons.

Acknowledgements

We thank Yotam Elor, Roi Livni, Guy Tevet, Chen Dudai, and Rinon Gal for providing helpful feedback. This work was partially supported by ISF (grant number 2510/23).

References

1. Alper, M., Averbuch-Elor, H.: Kiki or bouba? sound symbolism in vision-and-language models. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2023)
2. Alper, M., Fiman, M., Averbuch-Elor, H.: Is bert blind? exploring the effect of vision-and-language pretraining on visual language understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
3. Athiwaratkun, B., Wilson, A.G.: Hierarchical density order embeddings. arXiv preprint arXiv:1804.09843 (2018)
4. Atigh, M.G., Schoep, J., Acar, E., Van Noord, N., Mettes, P.: Hyperbolic image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4453–4462 (2022)
5. Bertinetto, L., Mueller, R., Tertikas, K., Samangooei, S., Lord, N.A.: Making better mistakes: Leveraging class hierarchies with deep networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12506–12515 (2020)
6. Carlsson, F., Eisen, P., Rekathati, F., Sahlgren, M.: Cross-lingual and multilingual clip. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 6848–6854 (2022)
7. Chang, H.S., Wang, Z., Vilnis, L., McCallum, A.: Distributional inclusion vector embedding for unsupervised hypernymy detection. arXiv preprint arXiv:1710.00880 (2017)
8. Chefer, H., Lang, O., Geva, M., Polosukhin, V., Shocher, A., Irani, M., Mosseri, I., Wolf, L.: The hidden language of diffusion models. arXiv preprint arXiv:2306.00966 (2023)
9. Chen, G., Hou, L., Chen, Y., Dai, W., Shang, L., Jiang, X., Liu, Q., Pan, J., Wang, W.: mclip: Multilingual clip via cross-lingual transfer. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 13028–13043 (2023)
10. Chen, L., Jiang, Z., Xiao, J., Liu, W.: Human-like controllable image captioning with verb-specific semantic roles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16846–16856 (June 2021)
11. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
12. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
13. Chuang, C.Y., Robinson, J., Lin, Y.C., Torralba, A., Jegelka, S.: Debiased contrastive learning. Advances in neural information processing systems **33**, 8765–8775 (2020)

14. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
15. Couairon, G., Douze, M., Cord, M., Schwenk, H.: Embedding arithmetic of multimodal queries for image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4950–4958 (2022)
16. Dai, B., Lin, D.: Contrastive learning for image captioning. Advances in Neural Information Processing Systems **30** (2017)
17. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
18. Dash, S., Chowdhury, M.F.M., Gliozzo, A., Mihindukulasooriya, N., Fauceglia, N.R.: Hypernym detection using strict partial order networks. Proceedings of the AAAI Conference on Artificial Intelligence **34**(05), 7626–7633 (Apr 2020). https://doi.org/10.1609/aaai.v34i05.6263
19. Deng, C., Ding, N., Tan, M., Wu, Q.: Length-controllable image captioning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 712–729. Springer (2020)
20. Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., Vedantam, S.R.: Hyperbolic image-text representations. In: International Conference on Machine Learning. pp. 7694–7731. PMLR (2023)
21. Dhall, A., Makarova, A., Ganea, O., Pavllo, D., Greeff, M., Krause, A.: Hierarchical image classification using entailment cone embeddings. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 836–837 (2020)
22. Dhingra, B., Shallue, C.J., Norouzi, M., Dai, A.M., Dahl, G.E.: Embedding text in hyperbolic spaces. arXiv preprint arXiv:1806.04313 (2018)
23. Ermolov, A., Mirvakhabova, L., Khrulkov, V., Sebe, N., Oseledets, I.: Hyperbolic vision transformers: Combining improvements in metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7409–7419 (2022)
24. Ganea, O., Bécigneul, G., Hofmann, T.: Hyperbolic entailment cones for learning hierarchical embeddings. In: International Conference on Machine Learning. pp. 1646–1655. PMLR (2018)
25. Hong, J., Hayder, Z., Han, J., Fang, P., Harandi, M., Petersson, L.: Hyperbolic audio-visual zero-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7873–7883 (2023)
26. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
27. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4565–4574 (2016)
28. Kamath, A., Hessel, J., Chang, K.W.: Text encoders are performance bottlenecks in contrastive vision-language models. arXiv preprint arXiv:2305.14897 (2023)
29. Kiela, D., Rimell, L., Vulic, I., Clark, S.: Exploiting image generality for lexical entailment detection. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015). pp. 119–124. ACL; East Stroudsburg, PA (2015)

30. Kornblith, S., Li, L., Wang, Z., Nguyen, T.: Guiding image captioning models toward more specific captions. arXiv preprint arXiv:2307.16686 (2023)
31. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
32. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
33. Lee, S., Zhang, Y., Wu, S., Wu, J.: Language-informed visual concept learning. arXiv preprint arXiv:2312.03587 (2023)
34. Li, A., Luo, T., Lu, Z., Xiang, T., Wang, L.: Large-scale few-shot learning: Knowledge transfer with class hierarchy. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 7212–7220 (2019)
35. Li, L., Zhang, Y., Wang, S.: The euclidean space is evil: Hyperbolic attribute editing for few-shot image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22714–22724 (2023)
36. Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 17612–17625. Curran Associates, Inc. (2022)
37. 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: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
38. Long, T., van Noord, N.: Cross-modal scalable hierarchical clustering in hyperbolic space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16655–16664 (2023)
39. Luo, R., Price, B., Cohen, S., Shakhnarovich, G.: Discriminability objective for training descriptive captions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6964–6974 (2018)
40. Mettes, P., Atigh, M.G., Keller-Ressel, M., Gu, J., Yeung, S.: Hyperbolic deep learning in computer vision: A survey. arXiv preprint arXiv:2305.06611 (2023)
41. Nguyen, K.A., Köper, M., Walde, S.S.i., Vu, N.T.: Hierarchical embeddings for hypernymy detection and directionality. arXiv preprint arXiv:1707.07273 (2017)
42. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. Advances in neural information processing systems **30** (2017)
43. Nickel, M., Kiela, D.: Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In: International conference on machine learning. pp. 3779–3788. PMLR (2018)
44. Novack, Z., McAuley, J., Lipton, Z.C., Garg, S.: Chils: Zero-shot image classification with hierarchical label sets. In: International Conference on Machine Learning. pp. 26342–26362. PMLR (2023)
45. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
46. Phoo, C.P., Hariharan, B.: Coarsely-labeled data for better few-shot transfer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9052–9061 (2021)

47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
48. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:2004.09813 (2020)
49. Santurkar, S., Tsipras, D., Madry, A.: Breeds: Benchmarks for subpopulation shift. arXiv preprint arXiv:2008.04859 (2020)
50. Sarkar, R.: Low distortion delaunay embedding of trees in hyperbolic plane. In: International symposium on graph drawing. pp. 355–366. Springer (2011)
51. Shah, A., Sra, S., Chellappa, R., Cherian, A.: Max-margin contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 8220–8230 (2022)
52. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
53. Suzuki, R., Takahama, R., Onoda, S.: Hyperbolic disk embeddings for directed acyclic graphs. In: International Conference on Machine Learning. pp. 6066–6075. PMLR (2019)
54. Tewel, Y., Shalev, Y., Schwartz, I., Wolf, L.: Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17918–17928 (2022)
55. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023)
56. Trager, M., Perera, P., Zancato, L., Achille, A., Bhatia, P., Soatto, S.: Linear spaces of meanings: Compositional structures in vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15395–15404 (October 2023)
57. Vendrov, I., Kiros, R., Fidler, S., Urtasun, R.: Order-embeddings of images and language. In: Bengio, Y., LeCun, Y. (eds.) ICLR (2016)
58. Vulić, I., Gerz, D., Kiela, D., Hill, F., Korhonen, A.: Hyperlex: A large-scale evaluation of graded lexical entailment. Computational Linguistics **43**(4), 781–835 (2017)
59. Vulić, I., Mrkšić, N.: Specialising word vectors for lexical entailment. arXiv preprint arXiv:1710.06371 (2017)
60. Xiong, B., Nayyeri, M., Jin, M., He, Y., Cochez, M., Pan, S., Staab, S.: Geometric relational embeddings: A survey. arXiv preprint arXiv:2304.11949 (2023)
61. Xu, Z., Zhu, Y., Deng, T., Mittal, A., Chen, Y., Wang, M., Favaro, P., Tighe, J., Modolo, D.: Challenges of zero-shot recognition with vision-language models: Granularity and correctness. arXiv preprint arXiv:2306.16048 (2023)
62. Yi, K., Shen, X., Gou, Y., Elhoseiny, M.: Exploring hierarchical graph representation for large-scale zero-shot image classification. ECCV (2022)

63. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: The Eleventh International Conference on Learning Representations (2022)
64. 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. pp. 18123–18133 (2022)
65. Zhang, C., Gao, J.: Hype-han: Hyperbolic hierarchical attention network for semantic embedding. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. pp. 3990–3996 (2021)
66. Zhang, C., Van Durme, B., Li, Z., Stengel-Eskin, E.: Visual commonsense in pretrained unimodal and multimodal models. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5321–5335. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.390, https://aclanthology.org/2022.naacl-main.390
67. Zhang, H., Hu, Z., Deng, Y., Sachan, M., Yan, Z., Xing, E.: Learning concept taxonomies from multi-modal data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1791–1801. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1169
68. Zhu, G., Zhang, L., Jiang, Y., Dang, Y., Hou, H., Shen, P., Feng, M., Zhao, X., Miao, Q., Shah, S.A.A., et al.: Scene graph generation: A comprehensive survey. arXiv preprint arXiv:2201.00443 (2022)