# SYNTAX-PRESERVING HYPERBOLIC VISUAL SEMANTIC EMBEDDINGS

**Genji Ohara[1], Daiki Yoshikawa[1] & Takashi Matsubara[1,2]**
[1]Hokkaido University [2]CyberAgent

## ABSTRACT

Visual Semantic Embeddings (VSE) map images and text into a shared latent space, serving as a core technique for multi-modal applications. While hyperbolic VSE models effectively capture hierarchical data, they often process text as a simple Bag-of-Words, leading to a lack of compositional understanding. Building on recent findings that large language models implicitly acquire syntactic structure, we propose a method to enhance VSE by explicitly learning the syntactic structure of text. We introduce a novel regularization term that preserves parent-child relations from dependency syntax trees as entailment relations within hyperbolic space. Experiments demonstrate that our method outperforms baselines not only on the VL-CheckList benchmark for compositional understanding but also on standard zero-shot tasks. These results confirm that explicitly incorporating syntactic information improves the compositional capabilities of VSE models.

## 1 INTRODUCTION

Visual-Semantic Embeddings (VSE) embed images and text into a common space (Frome et al., 2013), serving as a core technology for tasks such as image retrieval and captioning (Karpathy & Fei-Fei, 2015). While models like CLIP (Radford et al., 2021) achieve remarkable zero-shot performance using contrastive learning, their reliance on Euclidean space poses geometric constraints for hierarchical data. To address this, MERU (Desai et al., 2023) applies hyperbolic space to VSE. By modeling the entailment relation between abstract text and concrete images using entailment cones, MERU achieves hierarchical representations while maintaining high performance.

However, existing VSE models, including MERU, tend to ignore word order and grammatical structure, effectively behaving as Bag-of-Words (BoW) models (Yuksekgonul et al., 2023). For instance, they struggle to distinguish subject-object reversals (e.g., "A cat chases a dog" vs. "A dog chases a cat"), indicating a lack of compositional understanding (Koishigarina et al., 2025). Consequently, performance on benchmarks like Winoground (Thrush et al., 2022) remains poor, limiting applications in complex scenarios such as robotics.

In contrast, large language models (LLM) are known to implicitly acquire syntactic structure through pretraining (Hewitt & Manning, 2019). However, VSE text encoders typically lack the scale and objective to learn such structure spontaneously. Based on these findings, we propose improving VSE compositional understanding by explicitly preserving syntactic structure. Building on MERU, we introduce a regularization term that models parent-child relations from a dependency syntax tree as entailment relations in hyperbolic space.

## 2 RELATED WORK

**CLIP** CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) is a VSE model that achieves high generalization performance through self-supervised learning on a large dataset of image-text pairs. CLIP jointly trains two independent encoders: a text encoder (Transformer (Vaswani et al., 2017)) and an image encoder (ResNet (He et al., 2016) or Vision Transformer (ViT) (Dosovitskiy et al., 2021)). The contrastive loss used in CLIP is defined as a multi-class N-pair loss (InfoNCE loss) (Sohn, 2016), which maximizes the cosine similarity of positive pairs and minimizes the similarity of negative pairs among $N$ image-text pairs in a batch.

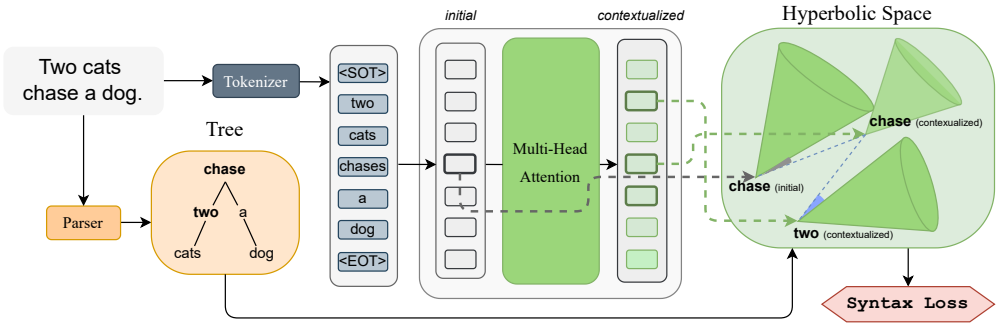


Figure 1: Architecture of our method. The input (initial embeddings) and output (contextualized embeddings) of Multi-Head Attention within the text encoder are used to compute the syntax loss. Focusing on the parent-child relation between "chase" and "two", syntax loss encourages (i) the contextualized embedding of the child "two" to entail the contextualized embedding of the parent "chase", and (ii) the initial embedding of the word "chase" to entail its contextualized embedding.

**Challenges in Compositional Understanding** While VSE models like CLIP excel at recognizing word-level concepts, they struggle with compositional understanding. Yuksekgonul et al. (2023) reported that CLIP often ignores word order and grammatical structure, behaving instead like Bag-of-Words models. Consequently, they face challenges on benchmarks like VL-CheckList (Zhao et al., 2023), which specifically probes whether a model comprehends how individual words combine to form structured meanings across objects, attributes, and relations.

**Hyperbolic space and Hierarchy** While models like CLIP achieve high zero-shot performance, their reliance on Euclidean space poses geometric constraints for hierarchical data (Bridson & Haefliger, 1999). Hyperbolic space is a non-Euclidean space with constant negative curvature. Compared with Euclidean space, it expands exponentially. This property makes it possible to embed tree structures (hierarchical structures), whose number of nodes grows exponentially with depth, with arbitrarily small distortion (Sarkar, 2012).

**Hyperbolic VSE** Desai et al. (2023) proposed MERU, which extends CLIP to hyperbolic space. MERU focuses on the entailment relation in which text (abstract concepts) entails its paired images (concrete instances) (Vendrov et al., 2016) and introduces the entailment cone in hyperbolic space (Ganea et al., 2018). It is trained so that the image embedding is contained inside the cone defined by the embedding of the corresponding text. MERU creates a hierarchical embedding space in which text embeddings are placed near the origin and image embeddings are placed farther outward. HyCoCLIP (Pal et al., 2025) models hierarchies using bounding boxes and inter/intra-modal entailment. By relying solely on text syntax without such region-level annotations, our method addresses a fundamentally different problem setting.

## 3 METHOD

Our method uses MERU (Desai et al., 2023) as the base model and introduces a new regularization term, syntax loss. Figure 1 shows the architecture of our method. Two encoders embed images and texts into hyperbolic space, respectively. We use a Transformer-based encoder as the text encoder. A parser obtains a syntax tree of the input text based on dependency grammar, and computes syntax loss for consistency with the tree by using the input and output of Multi-Head Attention.

The loss function is defined as a weighted sum of contrastive loss, entailment loss, and the newly introduced syntax loss. Given a batch of image-text pairs $B = \{(T_k, I_k)\}_{k=1}^{N}$, where $T_k$ and $I_k$ are the text and image of the k-th pair, the total loss is defined as follows:

$$\mathcal{L}_{\text{ours}}(B) = \mathcal{L}_{\text{cont}}(B) + \lambda \mathcal{L}_{\text{entail}}(B) + \mu \mathcal{L}_{\text{syntax}}(B), \tag{1}$$

where $\lambda$ and $\mu$ are hyperparameters controlling the weights of each loss term.

Table 1: Comparison of zero-shot image classification performance (Top-1 Accuracy).

| | ImageNet | CIFAR-10 | CIFAR-100 | SUN397 | Caltech-101 | STL-10 | Food-101 | CUB | Cars | Aircraft | Pets | Flowers | DTD | EuroSAT | RESISC45 | Country211 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MERU | 32.3 | 69.0 | 41.5 | 44.0 | 66.3 | **90.6** | 37.3 | 5.9 | **5.6** | **2.2** | 31.5 | 14.1 | **19.6** | 32.9 | 36.6 | **5.9** |
| Ours | **32.4** | **72.5** | **42.8** | **44.3** | **69.2** | 89.7 | **51.8** | **8.7** | 4.9 | 1.9 | **37.6** | **15.9** | 19.1 | **35.0** | **37.1** | 4.1 |

**Contrastive Loss & Entailment Loss** The contrastive loss $\mathcal{L}_{cont}$ and entailment loss $\mathcal{L}_{entail}$ are computed in the same manner as in MERU. Details are provided in section A.3. $\mathcal{L}_{cont}$ encourages paired image-text embeddings to be close in hyperbolic space while pushing apart unpaired embeddings. $\mathcal{L}_{entail}$ encourages image embeddings to be contained within the entailment cones defined by their corresponding text embeddings, modeling the entailment relation from text to images.

**Syntax Loss** $\mathcal{L}_{\text{syntax}}$ consists of two terms:

$$\mathcal{L}_{\text{syntax}}(B) = \sum_{(T,I)\in B} l_{\text{syntax}}(T), \tag{2}$$

$$l_{\text{syntax}}(T) = \sum_{(p,c)\in\mathcal{T}(S)} l_{\text{entail}}(\boldsymbol{e}_c^{\text{cont}}, \boldsymbol{e}_p^{\text{cont}}) + \sum_{w\in S} l_{\text{entail}}(\boldsymbol{e}_w^{\text{init}}, \boldsymbol{e}_w^{\text{cont}}). \tag{3}$$

where $T$ is the text (sequence of words), $w$ is a word in the text, $\boldsymbol{e}_w^{\text{init}}$ is the initial embedding of word $w$, $\boldsymbol{e}_w^{\text{cont}}$ is the contextualized embedding of word $w$, and $\mathcal{S}(T)$ represents the set of parent-child pairs $(p, c)$ in the syntax tree of sentence $T$. The initial embedding $e_w^{init}$ corresponds to the input of the Multi-Head Attention in the text encoder and represents the latent meaning of the word $w$ itself, independent of context. The contextualized embedding $\boldsymbol{e}_w^{\text{cont}}$ corresponds to the output of the Multi-Head Attention and represents the meaning of the word incorporating contextual information. $l_{entail}(x, y)$ is the entailment loss defined in section A.3, which encourages the embedding $y$ to be contained within the entailment cone defined by the embedding $x$.

The first term of eq. (3) aims to preserve parent-child relations in the syntax tree as entailment relations in hyperbolic space. Child nodes represent abstract concepts as parts of a sentence, and parent nodes represent more concrete concepts as wholes that integrate those parts. For example, in the sentence "Two cats chase a dog", the phrases "two cats" and "a dog" are child nodes of the node "chase", but the sentence "Two cats chase a dog" is clearly a sentence about "two cats" and also about "a dog". To model this relation in hyperbolic space, we encourage the contextualized embeddings of a parent node to be contained by the contextualized embeddings of its child node, that is, to lie inside the entailment cone.

The second term of eq. (3) aims to control the process by which each word embedding acquires information from context. The initial embeddings represent abstract concepts that include all possible meanings, and the contextualized embeddings represent concrete concepts whose meaning is restricted by a specific context. For example, without context, the word "chase" can mean abstract and polysemous senses such as "someone chases something", but in the sentence "Two cats chase a dog" it is restricted to a specific action with a single meaning. To model this relation in hyperbolic space, we encourage the contextualized embeddings to be entailed by their own initial embeddings.

## 4 EXPERIMENTS

**Experimental Setup** For pretraining, we used the GRIT dataset (Peng et al., 2023). As a shared architecture for MERU and our method, we used ViT-S/16 (Chen et al., 2021) for the image encoder and a Transformer-based encoder for the text encoder, and we adopted AdamW (Loshchilov & Hutter, 2019) as the optimizer. We used spaCy (en_core_web_sm) (Honnibal, 2017) as the parser.

**Evaluation: Zero-shot Classification** First, we measured Top-1 accuracy on a standard benchmark of 16 datasets including ImageNet (Deng et al., 2009). Zero-shot classification tasks test

Table 2: Comparison of zero-shot retrieval performance (Recall@K).

| | Text → Image | | | | Image → Text | | | |
|---|---|---|---|---|---|---|---|---|
| | COCO | | Flickr | | COCO | | Flickr | |
| | R@5 | R@10 | R@5 | R@10 | R@5 | R@10 | R@5 | R@10 |
| MERU | **47.59** | 59.15 | 74.72 | **83.64** | 65.40 | 75.40 | 86.70 | 92.00 |
| Ours | 47.42 | **59.16** | **75.34** | 83.54 | **66.14** | **75.46** | **87.60** | **94.00** |

Table 3: Performance Comparison on VL-CheckList.

| | **Object** | | | | | | **Attribute** | | | | | **Relation** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Location | | | Size | | | | | | | | | |
| | Center | Mid | Margin | Large | Medium | Small | Size | Material | State | Action | Color | Spatial | Action |
| MERU | 61.7 | 60.7 | 59.2 | 65.3 | 57.6 | 55.9 | **50.9** | 63.0 | 44.5 | 49.7 | 65.5 | **35.0** | **55.9** |
| Ours | **67.6** | **63.3** | **62.0** | **68.8** | **63.7** | **63.8** | 50.6 | **69.4** | **48.4** | **50.7** | **67.4** | 34.4 | 54.2 |

a model's ability to accurately assign labels to images from novel categories without relying on any dataset-specific training data. This evaluates how well the learned visual-semantic embeddings generalize to broad, unseen concepts. The results (table 1) show that our method improves Top-1 accuracy on 11 of the 16 datasets compared with MERU, with a large improvement of 14.5 points on Food-101. These results show that explicitly providing syntactic information to VSE models contributes to better zero-shot image classification performance.

**Evaluation: Zero-shot Retrieval** Using COCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015), we measured Recall@K in both directions: text to image and image to text. In zero-shot retrieval, the model is tasked with finding the most relevant image from a large database given a text query, or vice-versa, relying solely on its pre-trained multimodal alignment. The results (table 2) show that our method improves performance on many metrics compared with MERU, indicating that introducing syntactic information also contributes to better zero-shot image retrieval.

**Evaluation: Compositional Understanding** VL-CheckList (Zhao et al., 2023) is a benchmark for evaluating compositional understanding of images and language, and it consists of three categories: Object, Attribute, and Relation. These tasks specifically probe whether a model comprehends how individual words combine to form complex, structured meanings, rather than just recognizing isolated concepts as a "Bag-of-Words". The results (table 3) show that our method outperforms MERU in 10 of 13 subcategories. Performance improves especially in the Object and Attribute categories, while it decreases in the Relation category. This is likely because our method focuses only on parent-child relations in syntax trees based on dependency grammar and does not distinguish relation types (for example, subject-predicate relations versus object-predicate relations).

## 5 Conclusion

We proposed a hyperbolic VSE method to improve compositional understanding by preserving syntactic structure. Building on MERU, we formulated a syntax loss that models syntactic dependencies and embedding transitions as entailment relations in hyperbolic space. Experiments show that our method outperforms the baseline on zero-shot classification and the VL-CheckList benchmark, confirming the value of explicit syntactic information. However, gains in the Relation category remain limited. Future work will address this by considering relation types in syntax trees.

### Acknowledgments

This study was partly supported by JST BOOST (JPMJBY24H0) and CREST (JPMJCR24Q5), and partly achieved through the use of SQUID at D3 Center, Osaka University.

## REFERENCES


Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 1999.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9640–9649, October 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7694–7731. PMLR, 7 2023. URL https://proceedings.mlr.press/v202/desai23a.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1646–1655. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ganea18a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

M Honnibal. Spacy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, 2017. URL https://cir.nii.ac.jp/crid/1370021390573874949.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. Clip behaves like a bag-of-words model cross-modally but not uni-modally, 2025. URL https://arxiv.org/abs/2502.03566.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models, 2025. URL https://arxiv.org/abs/2410.06912.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 7 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In Marc van Kreveld and Bettina Speckmann (eds.), *Graph Drawing*, pp. 355–366, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-25878-7.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248, June 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *International Conference on Learning Representations*, 2016.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 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*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations, 2023. URL https://arxiv.org/abs/2207.00221.

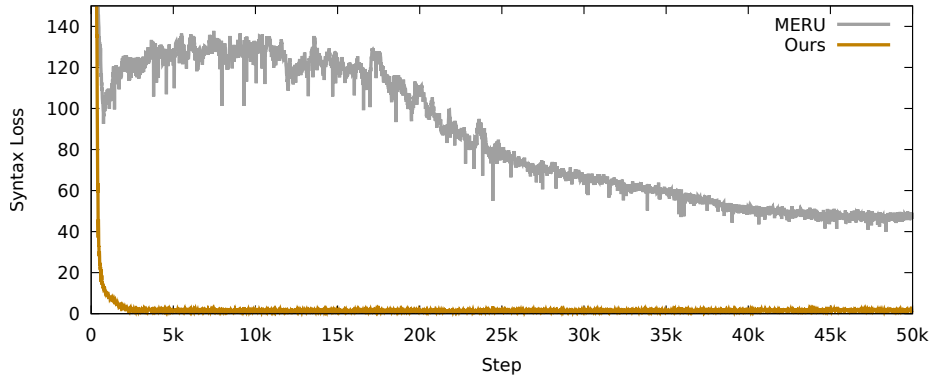


Figure 2: Trends of syntax loss during training for MERU (gray) and our method (blue). Ours shows a rapid decrease in the early stages and maintains a low value. MERU also shows some decrease, suggesting implicit acquisition of some syntactic information.

# A Appendix

## A.1 Training: Trend of Syntax Loss

We compared the trends of syntax loss when training MERU and our method (fig. 2). Our method shows a rapid decrease in syntax loss in the early stage of training and maintains a stable low value until the late stage. This indicates that our method effectively preserves the syntactic structure of text. MERU also shows some decrease in syntax loss, which is consistent with the observation that large language models acquire syntactic structure internally (Hewitt & Manning (2019)). However, the decrease is limited compared with our method, which indicates the effectiveness of explicitly providing syntactic information.

## A.2 Hyperbolic Space: Lorentz Model

Known models of hyperbolic space include the Poincaré ball model and the Lorentz (Minkowski) model. The Lorentz model (or Minkowski model) is numerically stable and is widely used in applications to VSE. The Lorentz model defines the n-dimensional hyperbolic space $\mathcal{L}^n$ as a hyperboloid in an $(n, 1)$-Minkowski space (an $\mathbb{R}^{n+1}$ space consisting of n-dimensional space $\boldsymbol{x}_{space}$ and 1-dimensional time $x_{time}$). The Lorentz inner product and Lorentz distance in Minkowski space are defined as follows:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}} = \langle \boldsymbol{x}_{space}, \boldsymbol{y}_{space} \rangle - x_{time} y_{time}, \tag{4}$$

$$d_{\mathcal{L}}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{c}} \cosh^{-1}(-\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}}). \tag{5}$$

## A.3 Implementation Details

**Contrastive Loss** $\mathcal{L}_{\text{cont}}$ in eq. (1) is computed on the Lorentz model as in MERU:

$$\mathcal{L}_{\text{cont}}(B) = \mathcal{L}_{\text{text}}(B) + \mathcal{L}_{\text{image}}(B), \tag{6}$$

$$\mathcal{L}_{\text{text}} = -\sum_{(T,I)\in B} \log \frac{\exp(-d_{\mathcal{L}}(\boldsymbol{e}_T, \boldsymbol{e}_I)/\tau)}{\sum_{(T',I')\in B} \exp(-d_{\mathcal{L}}(\boldsymbol{e}_T, \boldsymbol{e}_{I'})/\tau)} \tag{7}$$

$$\mathcal{L}_{\text{image}} = -\sum_{(T,I)\in B} \log \frac{\exp(-d_{\mathcal{L}}(\boldsymbol{e}_T, \boldsymbol{e}_I)/\tau)}{\sum_{(T',I')\in B} \exp(-d_{\mathcal{L}}(\boldsymbol{e}_{T'}, \boldsymbol{e}_I)/\tau)} \tag{8}$$

$$\tag{9}$$

**Entailment Loss** $\mathcal{L}_{entail}$ in eq. (1) is computed on the Lorentz model as in MERU:

$$\mathcal{L}_{\text{entail}}(B) = \sum_{(\boldsymbol{x},\boldsymbol{y})\in B} l_{\text{entail}}(\boldsymbol{x},\boldsymbol{y}), \tag{10}$$

$$l_{\text{entail}}(\boldsymbol{x},\boldsymbol{y}) = \max(0, \text{ext}(\boldsymbol{x},\boldsymbol{y}) - \text{aper}(\boldsymbol{x})), \tag{11}$$

where $B$ is the set of positive pairs (text embedding $x$ and image embedding $y$) in a batch and $\text{aper}(x)$ is the half-aperture angle of the cone defined by text embedding $x$. It is designed such that more abstract concepts (closer to the origin) have larger half-aperture angles:

$$\text{aper}(x) = \sin^{-1}\left(\frac{2K}{\sqrt{c}||x_{space}||}\right), \tag{12}$$

where $K$ is a hyperparameter for cone size, and $c$ is the curvature of the hyperbolic space. $ext(x, y)$ is the angle representing how far the image embedding $y$ deviates from the axis of the text embedding $x$'s cone:

$$\text{ext}(x,y) = \cos^{-1}\left(\frac{-x_{time}y_{time} + c\langle x,y\rangle_{\mathcal{L}}}{||x_{space}||\sqrt{\left(c\langle x,y\rangle_{\mathcal{L}}\right)^2 - 1}}\right). \tag{13}$$