LEARNING BOX RELATIONS AS DISTANCES FOR MULTI-MODAL REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal representation learning seeks to capture semantic correspondence beyond pairwise similarity, including intersection, inclusion, and separation. Point embeddings model correspondence through proximity, whereas box embeddings can encode these relations and compose them into more complex structures. Existing box methods, however, primarily rely on volume-based objectives whose coordinatewise products attenuate learning signals as dimensionality or box separation increases. We introduce a new volume-free, distance-based objective that learns box-to-box relations through intersection-aware and directed inclusion losses. Our objective avoids this attenuation and theoretically guarantees box intersection and inclusion. Experiments show that our method reliably learns joint, disjoint, and inclusion relations, achieves the best cross-modal retrieval performance on COCO-1K, COCO-5K, and RedCaps, and shows the closest alignment with human semantic similarity on CxC. On MovieLens-1M, our learned relations compose to answer set-theoretic queries, including combinations not observed during learning. Together, these results establish distance-based box relation learning as a reliable foundation for multimodal correspondence and set-theoretic reasoning. Our code is available at the anonymous repository.

1 INTRODUCTION

Multimodal representations involve richer semantic relations beyond pairwise correspondence: a concept may correspond to multiple concepts, be contained within another, or have no semantic relation. These relations can be naturally expressed through set-theoretic relations such as intersection, inclusion, and separation, allowing multimodal correspondence to be viewed as a relational structure among instances rather than merely pairwise proximity. Point embeddings (Oord et al., 2018; Radford et al., 2021), for example, represent instances as points and focus on proximity through pairwise distance, which limits their ability to express richer cross-modal relations. Accordingly, recent geometric embedding methods (Vendrov et al., 2015; Chun et al., 2021; Chun, 2024; Desai et al., 2023; Ramasinghe et al., 2024) have moved beyond pairwise proximity by encoding specific relational structures into the geometry of the representation, with different approaches focusing on particular types of relations, such as directional inclusion. Box embeddings offer a geometric representation in which intersection, inclusion, and separation can be expressed within a single representation, and these relations can be composed into more complex set-theoretic structures.

However, reliably learning these relations remains a key challenge for box embeddings. Specifically, existing methods have primarily relied on volume-based objectives (Vilnis et al., 2018; Chen et al., 2021b; Xiong et al., 2022; Huang et al., 2023; Geng & Tang, 2024) because volume naturally reflects the intersection and inclusion of boxes. The difficulty arises when two boxes are disjoint: their intersection volume is exactly zero, resulting in zero gradients and no learning signal. As a remedy, volume smoothing (Li et al., 2018; Dasgupta et al., 2020) has been proposed to provide non-zero learning signals for disjoint boxes. However, our analysis shows that the gradients still decay exponentially as the embedding dimension or box separation increases (Theorem 1), while applying log-based optimization leads to gradients that can diverge or vanish (App. A.3). These limitations make volume-based objectives difficult to optimize reliably for learning box relations.

Existing approaches have explored different ways to improve the learning of box relations, with each offering a different trade-off. One direction augments volume-based objectives with geometric

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

distance terms (Liang et al., 2023; Mei et al., 2024) to provide additional learning signals beyond volume. However, as volume remains part of the objective, these methods still inherit the optimization limitations of volume-based objectives. Another direction avoids volume altogether and learns relations through distance, but existing methods have primarily been developed for point-to-box tasks (Ren et al., 2020; Abboud et al., 2020), rather than directly learning box-to-box relations.

To address this, we introduce a new volume-free, distance-based objective that learns box-to-box relations. Our objective combines two complementary geometric signals: an intersection-aware distance for learning joint/disjoint relations, and a directed inclusion distance for asymmetric inclusion relations. By constructing losses from these distances using norms or averages rather than products across dimensions, our objective addresses the dimension- and separation-dependent attenuation inherent in volume-based objectives. The resulting objective reliably learns joint, disjoint, and inclusion relations, allowing box embeddings to realize richer correspondence, including one-to-many relations and differences in information content, and compose these relations to represent multiple conditions that must hold jointly or distinguish conditions that must not hold. Together, we provide a fully distance-based approach to learning box relations and summarize our contributions below.

- We propose new distance-based losses for box embeddings that learn intersection and inclusion relations, theoretically guaranteeing these relations while avoiding the attenuation inherent to volume-based formulations and empirically validating their reliable learning.
- Our method captures richer correspondence, such as one-to-many relations, achieving the best cross-modal retrieval performance across COCO-1K, COCO-5K, CxC, and RedCaps in both I2T and T2I retrieval, with the closest alignment with human semantic similarity.
- Our box representation composes learned relations for set-theoretic operations, accurately distinguishing joint and disjoint items and answering complex queries, including combinations not observed during learning, and achieves the best retrieval performance on MovieLens-1M.

2 Background and Problem Analysis

2.1 Box Representations and Relation Objectives

Box representation. Box embeddings represent each instance as an axis-aligned hyper-rectangle rather than a point in a D-dimensional space, where $D \in \mathbb{N}_{>0}$. Specifically, each instance a is represented by $\mathrm{Box}(a)$ with center $c_a \in \mathbb{R}^D$ and half-widths $\delta_a \in \mathbb{R}^D_{>0}$:

$$\mathrm{Box}(a) = \left\{ x \in \mathbb{R}^D : |x_k - c_{a,k}| \le \delta_{a,k},\ k = 1, \dots, D \right\}. \tag{1}$$

Along dimension k, its interval is $[c_{a,k} - \delta_{a,k}, c_{a,k} + \delta_{a,k}]$, with lower and upper bounds given by $a_{\min,k} = c_{a,k} - \delta_{a,k}$ and $a_{\max,k} = c_{a,k} + \delta_{a,k}$, respectively.

Set relations in box embeddings. The geometry of boxes naturally captures symmetric intersection and asymmetric inclusion as two fundamental set relationships. Given the volume $\mathrm{Vol}(\mathrm{Box}(a)) = \prod_{k=1}^{D} 2\delta_{a,k}$, these relations are quantified by the intersection volume and conditional probability:

$$\mathrm{Vol}(\mathrm{Box}(a) \cap \mathrm{Box}(b)) = \prod_{k=1}^{D} \max(0, m_k), \qquad P(b \mid a) = \frac{\mathrm{Vol}(\mathrm{Box}(a) \cap \mathrm{Box}(b))}{\mathrm{Vol}(\mathrm{Box}(a))}, \tag{2}$$

where $m_k = \min(a_{\max,k}, b_{\max,k}) - \max(a_{\min,k}, b_{\min,k})$ is the signed overlap length along dimension k, and $P(b \mid a)$ is the fraction of $\mathrm{Box}(a)$ covered by $\mathrm{Box}(b)$, equal to 1 iff $\mathrm{Box}(a) \subseteq \mathrm{Box}(b)$.

Smooth volume-based optimization. A hard intersection uses $\max(0, m_k)$ along each dimension, whose derivative is zero throughout the disjoint region $m_k < 0$. To enable gradient-based optimization, Li et al. (2018) replace the hard intersection length with the softplus approximation.

$$f_\tau(m_k) = \frac{1}{\tau}\zeta(\tau m_k), \qquad \lim_{\tau \to \infty} f_\tau(m_k) = \max(0, m_k), \tag{3}$$

where $\zeta(x) = \log(1 + \exp(x))$ and $\tau > 0$ controls the smoothing strength. Dasgupta et al. (2020) instead model box boundaries as Gumbel random variables and approximate smooth expected volumes. For the analysis, we use the following product-form softplus surrogate:

$$\widetilde{\mathrm{Vol}}_\tau(\mathrm{Box}(a), \mathrm{Box}(b)) = \prod_{k=1}^{D} f_\tau(m_k). \tag{4}$$

2.2 Prior Works on Box Embeddings

Volume-based objectives have been widely adopted as a standard framework for learning box-to-box relations (Vilnis et al., 2018; Li et al., 2018; Dasgupta et al., 2020), providing a unified way to learn both joint/disjoint box relations. They have been applied to diverse applications, including probabilistic representation learning, uncertain knowledge graph reasoning, knowledge graph completion, and ontology embedding (Xiong et al., 2022; Chen et al., 2021b; Huang et al., 2023; Geng & Tang, 2024), making them a widely used approach for learning box-to-box relations. Nevertheless, recent studies (Liang et al., 2023; Mei et al., 2024) find that volume-based objectives provide insufficient gradients for learning basic joint/disjoint relations as representation learning increasingly moves to high-dimensional spaces, motivating approaches that move beyond volume-based objectives.

To reliably learn basic joint/disjoint relations, recent methods have proposed complementing volume-based objectives with geometric distances. ContraBox (Liang et al., 2023) uses a point-to-box distance, and OptBox (Mei et al., 2024) applies a box-to-box distance to provide direct signals for learning joint relations between boxes, while none of these distances directly induce disjoint relations, leaving disjoint learning dependent on the volume term. Distance-based alternatives have also been explored for predefined ontology completion (Peng et al., 2022; Jackermeier et al., 2024), but do not address the joint/disjoint transition. Thus, existing approaches still depend on volume-based objectives for learning the basic joint/disjoint relations, leaving this challenge unresolved.

These limitations can lead to failures in learning intended box-to-box relations (Sec. 4.2, 4.4, App. A.4.1). We next further analyze the limitations of volume-based optimization.

2.3 Coordinate-Level Attenuation in Smoothed Volume

To understand the source of these failures, we examine the smoothed-intersection volume (Eq. 4) as a general form of the volume-based objective and analyze how its product structure leads to rapidly decaying gradients. For instances a and b, define the signed gap $g_k(a,b) = |c_{a,k} - c_{b,k}| - \delta_{a,k} - \delta_{b,k}$ and the set of separated dimensions $S = \{k \in \{1,\dots,D\} : g_k(a,b) > 0\}$. If $S \neq \{1,\dots,D\}$, let $\bar{v} = \max_{k \notin S} f_\tau(-g_k(a,b))$ bound the volume factors; when $S = \{1,\dots,D\}$, set $\bar{v} = 1$.

Theorem 1 (Coordinate-level attenuation bound for smoothed intersection volume)**.** *For any separated dimension $j \in S$ and any box-coordinate parameter $\theta \in \{c_{a,j}, c_{b,j}, \delta_{a,j}, \delta_{b,j}\}$, the corresponding derivative of the smoothed intersection volume obeys*

$$\left| \frac{\partial \widetilde{\mathrm{Vol}}_\tau}{\partial \theta} \right| \leq \bar{v}^{\,D-|S|}\, \tau^{-(|S|-1)} \exp\left(-\tau \sum_{k \in S} g_k(a,b) \right). \tag{5}$$

The proof in App. A.1.1 extends to the formulations in Li et al. (2018) and Dasgupta et al. (2020).

Theorem 1 exposes multiplicative attenuation in the smoothed volume from both box separation and dimension. The bound decreases as the accumulated separation $\sum_{k \in S} g_k(a,b)$ grows, and, for $\tau > 1$, as the number of separated dimensions $|S|$ grows. In the fully separated case it becomes $\tau^{-(D-1)}$, decaying exponentially with D. Consistent with the trend predicted by our theoretical bound (Fig. 1 dotted line), we empirically observe that gradients nearly vanish at $D = 16$, remaining negligible for separated boxes with $g_k > 0$ under AdamP (Heo et al., 2020) (see App. A.4.1 for details). Unlike prior analyses of hard-volume variants (Vilnis et al., 2018; Mei et al., 2024), clipped and smoothed objectives (Lees et al., 2020), we directly characterize the smoothed intersection volume, showing that its attenuation arises from the coordinatewise product.

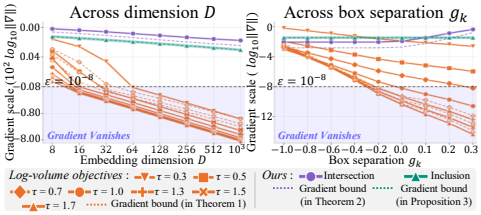

Figure 1: Gradients of the smoothed log-volume versus our distance losses across dimension D and box separation g_k. The "black dashed line" marks AdamP's epsilon ($\epsilon = 10^{-8}$), below which ϵ dominates the update and gradients become negligible.

3 Method

We develop the distance-based box objective motivated by Theorem 1. Our intersection loss learns related boxes to be joint and unrelated boxes to be disjoint (Sec. 3.1), while our inclusion loss

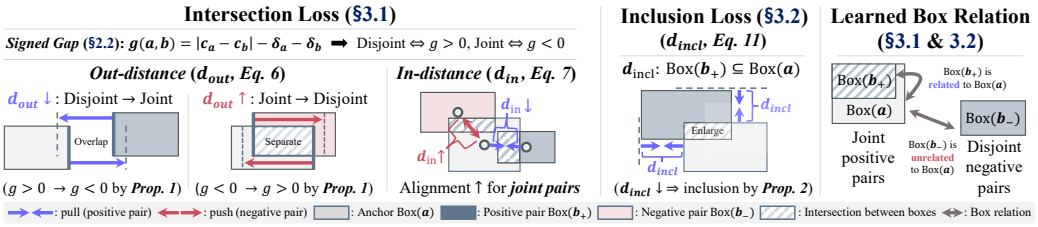

Figure 2: Overview of the proposed distance-based intersection and inclusion losses.

learns directional containment (Sec. 3.2). We then analyze the coordinate gradients of both losses (Sec. 3.4). Fig. 2 provides an overview of the proposed distance-based methods.

3.1 Intersection Loss

The intersection loss is built on two complementary distances. The out-distance measures interval separation, while the in-distance ranks pairs with similar out-distances by their center proximity.

Box out-distance. The out-distance is based on the signed gap g_k defined in Sec. 2 to distinguish whether two boxes intersect or are disjoint by sign, and quantifies the extent of separation for disjoint boxes. The signed gap measures the relation between the facing boundaries of the two boxes, where $g_k(a,b) > 0$ indicates separation and $g_k(a,b) < 0$ indicates overlap. This signed formulation allows a positive gap to be driven toward zero for intersection and a negative gap toward zero for separation. For $g(a,b) = [g_1(a,b), \ldots, g_D(a,b)]$, we define the out-distance as

$$d_{\text{out}}(a,b) = C^{-1} \left\| \phi_\gamma \left(g(a,b) \right) \right\|_2, \qquad \phi_\gamma(x) = \tfrac{1}{\gamma}\zeta(\gamma x) \xrightarrow{\gamma\to\infty} \max(0,x), \tag{6}$$

where ϕ_γ is a smooth approximation via softplus with $\gamma > 0$, $C = \text{sg}(\|c_a - c_b\|_2) + \sqrt{D}\log 2/\gamma$, and sg denotes stop-gradient. The softplus $\phi_\gamma(x)$ provides a smooth approximation to $\max(0,x)$, where the latter exactly characterizes the intersection condition (*i.e.*, $\text{Box}(a) \cap \text{Box}(b) \neq \emptyset$ iff $g_k(a,b) \leq 0$ for all k). This smoothing provides gradients near the transition boundary $g_k(a,b) \approx 0$, where the hard $\max(0,x)$ yields no gradient, enabling optimization across the joint/disjoint boundary (see the optimization bounds in Proposition 1). The parameter γ controls the width of the softplus transition. The normalization C^{-1} bounds the distance between disjoint boxes to $[0,1]$ to match the scale of the in-distance, while stop-gradient prevents a shortcut via increasing the center distance (see App. A.2).

Box in-distance. As the out-distance focuses on the joint/disjoint relation, it provides a coarse ranking signal, while retrieval can further benefit from fine-grained discrimination among pairs with similar separation. We therefore introduce a complementary ranking signal based on the displacement between box centers, called in-distance. We define

$$d_{\text{in}}(a,b) = (1 - d_{\text{out}}(a,b)) \cdot \tanh\left(\|c_a - c_b\|_2 \right), \tag{7}$$

where the $\tanh$ bounds the center displacement to $[0,1)$. Thus, the in-distance refines ranking for closely related boxes, while weakly related boxes are governed by the out-distance.

Intersection-aware box distance. We combine the out- and in-distances to define the distance as

$$d_{\text{inter}}(a,b) = d_{\text{out}}(a,b) + d_{\text{in}}(a,b). \tag{8}$$

In Sec. 4.2, we compare the intersection-aware distance with separately optimized in/out-distances.

Proposition 1 (Out-Distance Optimization Bounds and Certificates)**.** *For every* $\gamma > 0$*,*

$$C^{-1}\phi_\gamma\left(\max_k g_k(a,b) \right) \leq d_{\text{out}}(a,b) \leq \sqrt{D}\, C^{-1}\phi_\gamma\left(\max_k g_k(a,b) \right). \tag{9}$$

Consequently, the following finite-γ *certificates hold.*

(i) $d_{\text{out}}(a,b) \leq (\gamma C)^{-1}\log 2$ *implies* $\text{Box}(a) \cap \text{Box}(b) \neq \emptyset$.
(ii) $d_{\text{out}}(a,b) > \sqrt{D}\,(\gamma C)^{-1}\log 2$ *implies* $\text{Box}(a) \cap \text{Box}(b) = \emptyset$.

Parts (i) and (ii) are equivalent when $D = 1$*. For any pair of boxes, the hard limit recovers the intersection relation exactly,* $\text{Box}(a) \cap \text{Box}(b) \neq \emptyset \iff \lim_{\gamma\to\infty} d_{\text{out}}(a,b) = 0$, $\text{Box}(a) \cap \text{Box}(b) = \emptyset \iff 0 < \lim_{\gamma\to\infty} d_{\text{out}}(a,b) < 1$. *See App. A.1.3 for the proof.*

Intersection Loss. We train the intersection-aware box distance with pairwise binary cross-entropy, defining the single-pair loss as $\ell_{\text{intersection}}(a, b; y) = -\log \sigma\big(y\alpha\big(\beta - d_{\text{inter}}(a, b)\big)\big)$. Given a batch of paired boxes $\{(a_i, b_i)\}_{i=1}^{B}$, we treat the same-index pairs as positive pairs with $y_{in} = +1$ and different-index pairs as negative pairs with $y_{in} = -1$. The batch objective is

$$\mathcal{L}_{\text{intersection}} = \frac{1}{B^2} \sum_{i=1}^{B} \sum_{n=1}^{B} \ell_{\text{intersection}}(a_i, b_n; y_{in}). \tag{10}$$

Here $\alpha, \beta > 0$ are learnable sharpness and soft-margin parameters, following Chun et al. (2021).

3.2 Inclusion Loss

The inclusion loss models the directed relation $\text{Box}(a) \subseteq \text{Box}(b)$ when such a directional relation is present. We define the containment violation in dimension k as

$$\nu_k(a, b) = \max(0, b_{\min,k} - a_{\min,k}) + \max(0, a_{\max,k} - b_{\max,k}), \tag{11}$$

which is zero when the k-th interval of $\text{Box}(a)$ lies inside the corresponding interval of $\text{Box}(b)$. Applying the strictly increasing map $\psi(x) = x/(1+x)$ from $[0, \infty)$ to $[0, 1)$, the inclusion distance is $d_{\text{inclusion}}(a, b) = \ell_{\text{inclusion}}(a, b) = \frac{1}{D} \sum_{k=1}^{D} \psi(\nu_k(a, b))$. For box pairs $\{(a_i, b_i)\}_{i=1}^{B}$, we minimize:

$$\mathcal{L}_{\text{inclusion}} = \frac{1}{B} \sum_{i=1}^{B} \ell_{\text{inclusion}}(a_i, b_i). \tag{12}$$

Proposition 2 (Exact Containment and Directed Triangle Inequality)**.** *For any boxes* $\text{Box}(a)$*,* $\text{Box}(b)$*, and* $\text{Box}(u)$*,* $d_{\text{inclusion}}(a, b) = 0$ *if and only if* $\text{Box}(a) \subseteq \text{Box}(b)$*, and* $d_{\text{inclusion}}(a, u) \leq d_{\text{inclusion}}(a, b) + d_{\text{inclusion}}(b, u)$*. Hence, the zero-distance relation is transitive. See App. A.1.4.*

3.3 Training Objective

The overall training objective combines the intersection loss and inclusion loss:

$$\mathcal{L} = \mathcal{L}_{\text{intersection}} + \lambda \mathcal{L}_{\text{inclusion}}, \tag{13}$$

where λ is the inclusion loss weight, while the intersection loss smoothing parameter γ is selected via the ablation study in Sec. 4.5. In Sec. 4, we set the inclusion direction following prior works (Desai et al., 2023; Ramasinghe et al., 2024; Pal et al., 2025) as $\text{Box}(\text{image}) \subseteq \text{Box}(\text{text})$.

3.4 Gradient Analysis of the Distance-Based Losses

Unlike the smoothed volume objective in Theorem 1, which causes gradient decay by multiplying coordinate factors, our distances aggregate coordinate separation and containment violations using a norm and an average, respectively. We quantify and analyze their coordinate gradients below.

All derivatives are with respect to box coordinates, with $\text{sg}' = 0$. For the intersection loss with $\alpha, \gamma > 0$, fix a disjoint pair, let $j \in \arg\max_k g_k(a, b)$, and choose $\theta \in \{c_{a,j}, c_{b,j}, \delta_{a,j}, \delta_{b,j}\}$. For $y \in \{-1, +1\}$, write $r = \tanh(\|c_a - c_b\|_2)$ and $p = \sigma(y\alpha(\beta - d_{\text{inter}}(a, b)))$.

Theorem 2 (Gradient Bound for the Pairwise Intersection Loss)**.** *Under these conditions,*

$$\left| \frac{\partial \ell_{\text{intersection}}(a, b; y)}{\partial \theta} \right| > \frac{\alpha(1-p)(1-r)}{2\sqrt{D}\, C} > 0. \tag{14}$$

The bound scales as D^{-1}*, with no product across dimensions. See App. A.1.2 for the proof.*

For the inclusion loss, fix $(a, b) = (a_i, b_i)$, choose j with $\nu_j(a, b) > 0$, and let $\theta \in \{\delta_{a_i,j}, \delta_{b_i,j}\}$, assuming $b_{\min,j} \neq a_{\min,j}$ and $a_{\max,j} \neq b_{\max,j}$.

Proposition 3 (Gradient Bound for the Pairwise Inclusion Loss)**.** *Under these conditions,*

$$\left| \frac{\partial \ell_{\text{inclusion}}(a, b)}{\partial \theta} \right| \geq \frac{1}{D\,(1 + \nu_j(a, b))^2} > 0. \tag{15}$$

The bound scales as D^{-1}*. See App. A.1.5 for the proof.*

These theoretical results and empirical analysis in Fig. 1 show that our losses maintain non-vanishing gradients, with at most polynomial decay in D rather than exponential decay as in Theorem 1.

Table 1: Cross-modal retrieval performance on COCO-1K, 5K, and CxC. Ours (Decoupled) independently optimizes the out- and in-distances with separate contrastive objectives, while Ours (Intersection) uses our intersection loss, and Ours (Full) additionally incorporates our inclusion loss.

	COCO-1K			COCO-5K			CxC		
Method	R@1 (I2T)	R@1 (T2I)	RSUM	R@1 (I2T)	R@1 (T2I)	RSUM	R@1 (I2T)	R@1 (T2I)	RSUM
InfoNCE	80.50	67.26	532.12	59.64	45.67	442.67	61.24	47.63	452.94
PCME	79.84	67.32	530.88	59.48	45.81	442.64	61.24	47.73	453.14
PCME++	81.68	69.24	536.79	61.92	47.96	452.33	63.32	49.91	461.67
SmoothBox	68.08	54.90	484.92	44.50	33.56	370.46	46.48	35.64	384.71
GumbelBox	58.58	52.71	462.45	34.70	30.58	335.55	36.60	32.73	350.93
OptBox	11.57	1.86	114.72	3.29	0.51	40.67	4.08	0.68	50.59
ContraBox	63.02	50.03	469.42	38.23	27.49	339.09	40.55	29.69	355.67
Ours (Decoupled)	80.00	67.64	531.70	59.82	46.00	443.70	61.34	48.10	454.31
Ours (Intersection)	81.18	68.65	535.55	62.82	47.44	451.24	64.36	49.34	461.29
Ours (Full)	**82.70**	**69.98**	**538.98**	**63.78**	**49.40**	**457.97**	**65.44**	**51.26**	**467.66**
Δ over best baseline	+1.02	+0.74	+2.19	+1.86	+1.44	+5.64	+2.12	+1.35	+5.99

Table 2: Cross-modal retrieval performance on RedCaps (11M).

Method	R@1 (I2T)	R@5 (I2T)	R@10 (I2T)	R@1 (T2I)	R@5 (T2I)	R@10 (T2I)	RSUM
InfoNCE	43.98	66.14	73.32	43.54	64.98	72.72	364.68
PCME	45.02	65.74	73.68	43.30	65.42	72.54	365.70
PCME++	43.44	65.09	72.61	41.36	63.83	71.96	358.29
SmoothBox	25.88	48.13	57.00	24.07	47.16	56.30	258.54
GumbelBox	24.17	46.38	55.59	22.42	44.85	54.83	248.24
OptBox	23.91	45.75	55.52	21.94	44.66	54.23	246.01
ContraBox	19.17	38.40	47.76	17.96	37.70	47.35	208.34
Ours (Full)	**45.52**	**66.52**	**73.78**	**43.96**	**65.65**	**73.30**	**368.73**
Δ over best baseline	+0.50	+0.38	+0.10	+0.42	+0.23	+0.58	+3.03

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation. To assess the quality of learned multimodal representations and the box-to-box relations they encode, we use cross-modal retrieval, a representative task commonly used to evaluate multi-modal representation learning. We train and evaluate models on MS COCO (Lin et al., 2014) and RedCaps (Desai et al., 2021), two widely used cross-modal retrieval benchmarks. For COCO, models are trained on the training split and evaluated on the test split under the COCO-1K and COCO-5K settings. We additionally evaluate on CxC (Parekh et al., 2021), an extension of MS COCO with human semantic similarity judgments, used only for evaluation. For RedCaps, models are trained on its training split and evaluated on its test split. We measure Recall@K for $K \in \{1, 5, 10\}$ in both Image-to-Text (I2T) and Text-to-Image (T2I), and report RSUM as the sum of recalls across both directions (Chun et al., 2021). Full results are provided in App. A.4.

Baselines. We compare point, box-based, and probabilistic methods. As a point baseline, we train a vector embedding with the standard CLIP objective using InfoNCE (Oord et al., 2018). For box-based baselines, we use SmoothBox (Li et al., 2018), GumbelBox (Dasgupta et al., 2020), ContraBox (Liang et al., 2023), and OptBox (Mei et al., 2024). We also consider probabilistic embeddings, PCME (Chun et al., 2021) and PCME++ (Chun, 2024), which represent instances as probability distributions and learn cross-modal relations through probabilistic distance. For retrieval, we use each method's main loss distance, including the intersection-aware distance (in Sec. 3.1) for ours.

Architecture, training, and model selection. All methods share the same backbone, projection head, and optimization setup. Encoders are initialized from pre-trained CLIP ViT-B/32 (Radford et al., 2021), and the projection head is a linear projection to $D = 1024$ followed by generalized pooling (GPO) (Chen et al., 2021a). Following PCME++ (Chun, 2024), probabilistic and box-based methods use two parallel heads to produce the center and per-dimension scale, corresponding to the standard deviation and box half-width, respectively; the point baseline uses only the center head. We train with AdamP for 25 epochs with batch size $B = 128$ on COCO, and 32 epochs with $B = 1024$ on RedCaps, selecting the best model by validation RSUM (Chun, 2024). See App. A.5.4 for details.

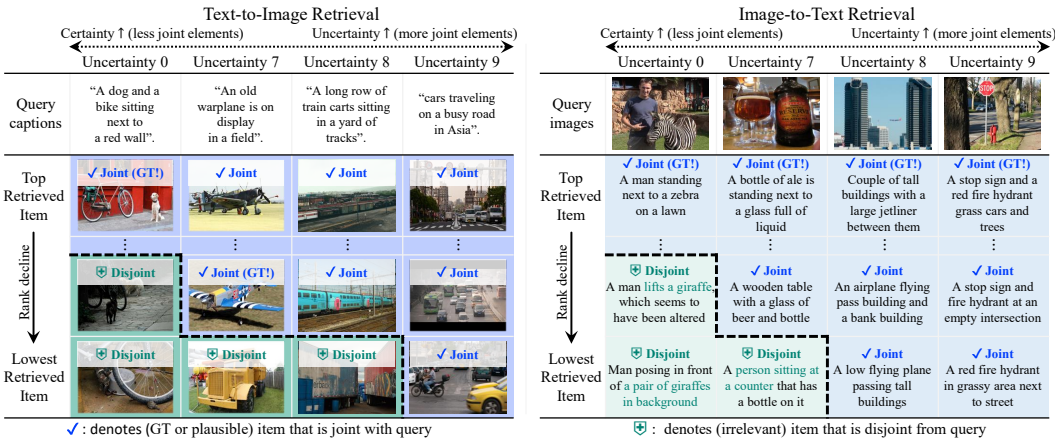

Figure 3: Qualitative retrieval results for image and caption queries on COCO-5k across varying levels of matching uncertainty. Uncertainty is measured by the number of intersecting text boxes for an image query, and image boxes for a caption query, with more intersections indicating higher uncertainty. The three highest uncertainty levels (7, 8, and 9) are shown alongside the lowest level.

4.2 Image–Text Retrieval on COCO, CxC, and RedCaps

Cross-modal retrieval performance on COCO and CxC. We compare our model against point, probabilistic, and box-based baselines in Tab. 1. Ours (Full) achieves the best performance across all metrics, outperforming the strongest baseline on every dataset (see full results in App. A.4.3). Specifically, the advantage over the best baseline increases from COCO-1K to 5K, demonstrating the effectiveness of our learned box relations for ranking larger retrieval candidate sets. The largest improvement is observed on CxC, which provides additional human judgments of semantic similarity, suggesting that our learned box relations better capture varying degrees of semantic similarity. Qualitatively, as shown in Fig. 3, relevant pairs with varying degrees of semantic similarity are represented by joint box relations, reflecting the ability of box geometry to preserve graded relevance rather than treating relevance as a binary relation. Overall, our method consistently achieves the best performance across datasets and metrics, while best aligning with human semantic similarity.

Generalization to Larger-Scale Data with RedCaps. We examine whether our approach scales to substantially larger training data while maintaining its performance advantage, using RedCaps, which contains approximately 11 million training image-caption pairs, compared with approximately 0.57 million pairs in MS COCO. As shown in Tab. 2, PCME++, the strongest baseline on MS COCO (Tab. 1), falls behind the point baseline (InfoNCE) on RedCaps in both R@1 and RSUM, illustrating that method performance can vary with training data scale. Box-based baselines remain below both the point and probabilistic baselines across all metrics, indicating their limited effectiveness for retrieval. In contrast, our method achieves the best performance across all metrics on both benchmarks. Together, these results show that our approach generalizes across training data scales while maintaining its advantage over other methods.

Analysis of Joint/Disjoint Relation Learning. On COCO-1K, we compare box-based methods for joint/disjoint relation learning, while evaluating retrieval performance when these relations are well learned. For each box baseline, we ablate its soft-volume objective and select the setting that maximizes joint relations for positive pairs and disjoint relations for negative pairs (Tab. A7). We evaluate the learned relations using P-Joint/N-Joint, the fractions of joint positive/negative pairs, where higher P-Joint and lower N-Joint indicate better relation learning. As shown in Tab. A7, volume-based methods (SmoothBox and GumbelBox) fail to learn reliable joint/disjoint relations, and their best settings do not translate to reliable retrieval. For example, GumbelBox achieves only 37.6%/36.7% P-Joint/N-Joint while its RSUM drops from 462.4 to 3.2. Hybrid methods, designed to address volume-based limitations, show mixed results: OptBox fails to reliably learn box relations, while ContraBox improves relation learning over volume-based methods but remains far from ideal, with 53.6% P-Joint and 0.2% N-Joint, and still underperforms the point baseline (InfoNCE) in retrieval. In contrast, Ours (Full) achieves 93.3% P-Joint and 9.9% N-Joint, achieving the strongest joint/disjoint relation learning and retrieval performance. These results demonstrate that our method enables reliable learning of box relations and improves cross-modal retrieval.

Effect of Loss Design. We compare three loss designs for retrieval. Ours (Intersection) uses our intersection loss (Eq. 10), which jointly learns d_{out} for joint/disjoint relations and d_{in} for proximity-based ranking. In contrast, Ours (Decoupled) optimizes d_{in} and d_{out} independently with separate contrastive objectives, providing concurrent learning signals. Ours (Full) further incorporates our inclusion loss (Eq. 12) to capture asymmetric cross-modal relations (image $\subseteq$ text). In Tab. 1, comparing Ours (Intersection) and Ours (Decoupled), the former consistently achieves higher scores across all metrics (*e.g.*, RSUM of 461.29 vs. 454.31), demonstrating the benefit of structured relation-to-proximity learning over independent optimization of d_{in} and d_{out}. Furthermore, Ours (Full) consistently outperforms Ours (Intersection) across all metrics (*e.g.*, RSUM of 467.66 vs. 461.29), indicating that modeling asymmetric cross-modal relations provides complementary information beyond intersection relations. Overall, these results highlight the complementary benefits of structured relation-to-proximity learning and asymmetric relation modeling for cross-modal retrieval.

4.3 Analysis of Cross-Modal Matching Uncertainty via Box Relations

Point embeddings represent each image and text with a single point, making it difficult to geometrically express cross-modal matching uncertainty, such as one-to-many correspondences. Our box-based representation instead expresses uncertainty through box relations. We quantify image-to-text uncertainty by the number of text boxes intersecting an image box, and text-to-image uncertainty analogously by the number of image boxes intersecting a text box, where higher values indicate greater matching ambiguity. We divide the uncertainty values into 10 equal-width bins and report R@1 for each bin to examine how retrieval changes with matching uncertainty.

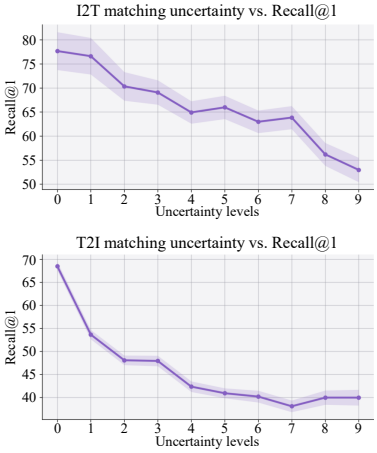

Figure 4: Relationship between cross-modal matching uncertainty and R@1 for I2T and T2I retrieval on COCO-5K.

Qualitative examples in Fig. 3 and App. A.4.6 show that higher matching uncertainty yields more candidates as plausible as the ground-truth match, reflecting one-to-many relations. As shown in Fig. 4, this one-to-many ambiguity is reflected in retrieval performance, with R@1 decreasing as the matching uncertainty level increases in both I2T and T2I retrieval, indicating that multiple plausible candidates make the correct match less distinctive. These results show that our box representation naturally captures one-to-many cross-modal correspondences through geometric relations.

4.4 Set-Theoretic Recommendation Beyond Pairwise Retrieval

Beyond pairwise retrieval, we evaluate whether learned box relations can be combined to retrieve items that satisfy multiple conditions simultaneously. We follow the set-theoretic recommendation setup of Dasgupta et al. (2025) on MovieLens-1M, where a query is defined over a user u, one or more genre attributes a, and an item i. Training supervises only two relations, whether a user u prefers an item i and whether an item i carries an attribute a, so their compositions (*e.g.*, $u \cap a$) are not directly supervised during training. We test three query types built from these relations: $u \cap a$ asks for an item that user u prefers and that has genre a, $u \cap a_1 \cap a_2$ requires two genres at once, and $u \cap a_1 \cap \neg a_2$ excludes the second (*e.g.*, a thriller that is not a drama). Each query is answered by filtering items with the set relations it requires, then ranking the items by each method's own score. As shown in Tab. 3, our method outperforms the best box baseline across all query types. The baselines achieve high P-joint on the supervised relations $u \cap i$ and $a \cap i$, but also assign high N-joint on $u \cap i$, indicating poor separation of positive and negative relations (Tab. A2). Moreover, their P-joint drops sharply on unsupervised compositions (*e.g.*, $u \cap a_1 \cap a_2$), while ours remains high. This P-joint degradation lowers retrieval performance, whereas our method maintains high P-joint on these compositions and achieves reliable set-theoretic retrieval. See App. A.4.2 for full results.

4.5 Ablation Study

Here, we ablate the smoothing parameter γ of the intersection loss and the inclusion-loss weight λ.

Table 3: Set-theoretic recommendation on MovieLens-1M. The relations $u \cap i$ and $a \cap i$ are observed during training, while $u \cap a$, $u \cap a_1 \cap a_2$, and $u \cap a_1 \cap \neg a_2$ are unseen during training. GumbelBox corresponds to GEOBox of Dasgupta et al. (2025). See full results in App. A.4.2.

Methods	R@10			P-joint (↑)			
	$u \cap a$	$u \cap a_1 \cap a_2$	$u \cap a_1 \cap \neg a_2$	$u \cap i$	$a \cap i$	$u \cap a$	$u \cap a_1 \cap a_2$
SmoothBox (Li et al., 2018)	2.5	3.3	0.5	98.7	100.0	24.7	5.8
GumbelBox (Dasgupta et al., 2025)	11.0	16.9	8.6	100.0	100.0	58.1	31.4
OptBox (Mei et al., 2024)	11.2	27.5	11.8	100.0	100.0	58.0	31.2
ContraBox (Liang et al., 2023)	6.1	21.2	4.3	100.0	100.0	54.5	29.3
Ours (Full)	**24.3**	**40.8**	**19.2**	98.7	97.1	86.3	73.6

Table 4: Ablation study of the intersection loss parameter γ. Purple shading indicates γ selection criterion.

	γ	P-Joint % (↑)	N-Joint % (↓)
P-Joint ↑	5	100.0	54.5
	10	99.7	17.2
	15	98.7	11.1
	20	95.6	3.46
	25	67.2	0.52
	30	66.1	0.43
	40	56.1	0.41
N-Joint ↓	60	7.70	0.03

Table 5: Ablation study of the inclusion loss weight λ on COCO-5K. The selection criterion follows Tab. 4.

Method	P-Joint % (↑)	N-Joint % (↓)	R@1 I2T	R@1 T2I
Intersection Loss	95.6%	3.46%	62.82	47.44
+0.2 Inclusion	87.5%	3.69%	+1.20	+2.11
+0.4 Inclusion	87.5%	4.04%	+0.84	+1.88
+0.6 Inclusion	94.0%	8.48%	+0.96	+1.96
+0.8 Inclusion	96.1%	18.6%	+1.66	+1.89
+1.0 Inclusion	98.3%	31.9%	+0.36	+1.48

Smoothing parameter γ (Tab. 4). Proposition 1 characterizes the out-distance of the intersection loss, guaranteeing intersection when $d_{\text{out}}(a, b) \leq (\gamma C)^{-1} \log 2$ and disjointness when $d_{\text{out}}(a, b) > \sqrt{D}\,(\gamma C)^{-1} \log 2$. Both thresholds scale as γ^{-1}. Small γ makes disjointness harder, while large γ makes intersection harder. At $\gamma = 5$, 54.5% of negative pairs remain joint, whereas at $\gamma = 60$, only 7.70% of positive pairs remain joint. The tested $\gamma = 20$ satisfies P-Joint $\geq 90\%$ and N-Joint $\leq 10\%$, with 95.6% and 3.46%, respectively. We use $\gamma = 20$ throughout our experiments.

Inclusion loss weight λ (Tab. 5). The directional containment relation induced by our inclusion loss consistently benefits retrieval. Every tested nonzero weight improves R@1 over the intersection-only baseline in both directions, with gains of 0.36–1.66 for I2T and 1.48–2.11 for T2I. As λ increases, P-Joint rises from 87.5% at $\lambda = 0.2$ to 98.3% at $\lambda = 1.0$, but N-Joint also rises from 3.69% to 31.9%. Among the tested values, $\lambda = 0.6$ provides the best balance, with P-Joint $= 94.0\%$ and N-Joint $= 8.48\%$, and is used throughout our experiments.

5 Conclusion

Existing box embedding methods rely on volume-based objectives that suffer from vanishing learning signals as the embedding dimension increases or the initial box separation is large. We address this limitation with volume-free, distance-based losses that directly learn intersection and inclusion relations, avoiding volume-induced attenuation. Our method achieves the best retrieval performance across COCO-1K, COCO-5K, CxC, and RedCaps. On set-theoretic recommendation, it achieves the best retrieval performance on MovieLens-1M and composes learned box relations to answer joint/disjoint queries, including combinations not observed during training. These results demonstrate that our method can reliably learn box representations, opening box embeddings to broader applications. For example, our representation could support condition-aware search and retrieval-augmented generation, where multiple semantic constraints need to be jointly satisfied. It could also provide a structured concept space for medical applications, such as organizing diseases from broad categories to fine-grained subtypes. Together, these directions point toward geometric representations that move beyond pairwise similarity to support structured and compositional understanding.

AI USE STATEMENT

In this work, we used generative AI tools for grammar checking and improving the readability of text. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for language editing. Sections revised with suggestions from generative AI were additionally reviewed by four authors. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ralph Abboud, Ismail Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. Boxe: A box embedding model for knowledge base completion. *Advances in neural information processing systems*, 33:9649–9661, 2020.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15789–15798, 2021a.

Xuelu Chen, Michael Boratko, Muhao Chen, Shib Sankar Dasgupta, Xiang Lorraine Li, and Andrew McCallum. Probabilistic box embeddings for uncertain knowledge graph reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 882–893, 2021b.

Sanghyuk Chun. Improved probabilistic image-text representations. In *International Conference on Learning Representations*, volume 2024, pp. 39787–39815, 2024.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8411–8420. IEEE, 2021.

Shib Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Li, and Andrew McCallum. Improving local identifiability in probabilistic box embeddings. *Advances in Neural Information Processing Systems*, 33:182–192, 2020.

Shib Dasgupta, Michael Boratko, and Andrew McCallum. A geometric approach to personalized recommendation with set-theoretic constraints using box embeddings. *arXiv preprint arXiv:2502.10875*, 2025.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731. PMLR, 2023.

Yuchong Geng and Ao Tang. A concept-centric approach to multi-modality learning. *arXiv preprint arXiv:2412.13847*, 2024.

Google Research Datasets. Genre annotations for MovieLens movies curated from Wikidata (Genre2Movies). `https://github.com/google-research-datasets/genre2movies`, 2024. Accessed: 2026-09-16.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pp. 173–182, 2017.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 639–648, 2020.

Byeongho Heo, Sanghyuk Chun, Seong Joon Oh, Dongyoon Han, Sangdoo Yun, Gyuwan Kim, Youngjung Uh, and Jung-Woo Ha. Adamp: Slowing down the slowdown for momentum optimizers on scale-invariant weights. *arXiv preprint arXiv:2006.08217*, 2020.

Zijie Huang, Daheng Wang, Binxuan Huang, Chenwei Zhang, Jingbo Shang, Yan Liang, Zhengyang Wang, Xian Li, Christos Faloutsos, Yizhou Sun, et al. Concept2box: Joint geometric embeddings for learning two-view knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10105–10118, 2023.

Mathias Jackermeier, Jiaoyan Chen, and Ian Horrocks. Dual box embeddings for the description logic el++. In *Proceedings of the ACM web conference 2024*, pp. 2250–2258, 2024.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alyssa Lees, Chris Welty, Shubin Zhao, Jacek Korycki, and Sara Mc Carthy. Embedding semantic taxonomies. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1279–1291, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.110. URL https://aclanthology.org/2020.coling-main.110/.

Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. Smoothing the geometry of probabilistic box embeddings. In *International conference on learning representations*, 2018.

Tingting Liang, Yuanqing Zhang, Qianhui Di, Congying Xia, Youhuizi Li, and Yuyu Yin. Contrastive box embedding for collaborative reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 38–47, 2023.

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

Lang Mei, Jiaxin Mao, and Ji-Rong Wen. Optimizing probabilistic box embeddings with distance measures. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 5088–5100. IEEE, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Avik Pal, Max Van Spengler, Guido D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 87371–87399, 2025.

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2855–2870, 2021.

Xi Peng, Zhenwei Tang, Maxat Kulmanov, Kexin Niu, and Robert Hoehndorf. Description logic el++ embeddings with intersectional closure. *arXiv preprint arXiv:2202.14018*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27253–27262. IEEE, 2024.

Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. *arXiv preprint arXiv:2002.05969*, 2020.

Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. Neural collaborative filtering vs. matrix factorization revisited. In *Proceedings of the 14th ACM conference on recommender systems*, pp. 240–248, 2020.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 263–272, 2018.

Bo Xiong, Nico Potyka, Trung-Kien Tran, Mojtaba Nayyeri, and Steffen Staab. Faithful embeddings for el++ knowledge bases. In *International semantic web conference*, pp. 22–38. Springer, 2022.