ROBIN, A TYPE OF CAT: Investigating Hypernymy in Unimodal and Multimodal Models with Contrastive Learning

MSc Thesis *(Afstudeerscriptie)*

written by

Zhirui Chen

under the supervision of **Dr Martha Lewis**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
Aug. 27th, 2025	Prof. dr. Benno van den Berg
	Prof. df. Jelke Bloem
	MSc. Anna Bavaresco

Institute for Logic, Language and Computation

First, are you our sort of a
person?

Sylvia Plath, The Applicant

Abstract

Visual information is commonly assumed to complement distributional semantics in achieving human-like concept understanding, motivating development and evaluation of various vision-language models (VLMs). However, there have been mixed findings on when and how VLMs outperform unimodal LMs. One key challenge lies in the representation of abstract words with low perceptability.

In this work, we focus on contrastive VLMs whose text encoders produce visual-semantic representations for text-only input and have been reported to outperform unimodal LMs across several word-, phrase-, or sentence-level understanding tasks.

We propose a novel approach to the lexical relation hypernymy (IS_A) based on synthetic concepts ("q, a type of p"), and conduct intrinsic evaluation of text encoders of contrastive VLMs accordingly against unimodal counterparts with contrastive learning. We find that contrastive VLMs, though generally outperformed by unimodal sentence transformers possibly due to the absence of unimodal language modeling, achieve competitive performance on traditional hypernymy benchmarks. We further argue that contrastive VLMs hold an inherent advantage on distinguishing hypernymy from one particular distractor relation, coordination (co-hyponymy), and suggest that further research is needed to better complement contrastive VLMs with textual distributional information.

Moreover, we examine the impact of word concreteness on model behaviour on a newly constructed dataset, and argue that abstractness does not necessarily pose a more significant challenge to text encoders of contrastive VLMs than to unimodal LMs. We also highlight the importance of exploring more systematic evaluation protocols for abstract concept representation.

Contents

Chapter 1

Introduction

Perception plays a crucial role in human language learning, and concepts acquire meaning via interaction with the external world [Bis+20]. Despite the contemporary success of distributional semantic models (DSMs) establishing word embeddings from text corpora, distributional semantics [LS23] is criticized as "a 'solipsistic' route to semantics" [Bar16] with no perceptual grounding, affected by the *symbol grounding problem* [Har90], and questioned in terms of human-like language understanding [BK20]. Such discussions lead to growing interest in multimodal distributional semantics [BTB14], integrating complementary distributional and perceptual information with the aim to build more cognitively plausible models of meaning [AVV09]. It is therefore assumed that multimodal models, especially vision-language models (VLMs), could establish more human-like concept representations than unimodal LMs [Bar16; Bis+20].

VLMs integrate multimodal information via fusion strategies from simple concatenation [FL10; BTB11; KB14] to neural mechanisms [Lu+19; Li+19; Che+20b], and are evaluated on semantic benchmarks such as word similarity [FL10; LPB15; PTF21] as well as brain alignment [BF25; Bav+25]. Despite the theoretical advantage, there exist mixed results concerning when and how VLMs outperform their unimodal counterparts, calling for further investigation into VLM mechanisms towards human-like concept representations with complementary information from both modalities. One key challenge to VLMs is abstract words, which are argued to have no [Pai91; Pai13] or indirect [Lou11; Dov14] perceptual grounding, where multimodal fusion is found to be less beneficial [HRK14; Kie+14; PTF21].

Contrastive VLMs such as CLIP [Rad+21] are distinct from multimodal *fusion* models due to their architecture and training objective: there exist separate encoders for the language and vision modalities, jointly trained to maximize the similarity of actual matching image-caption pairs. Due to the cross-modal contrastive objective, linguistic and visual features are integrated via *alignment*, mapped into a shared visual-semantic space, enabling meaningful cross-modal similarity scores. This dual-encoder architecture of contrastive VLMs also allows for independent use of each module. Their text encoders, on one hand, produce sequence-level representations for text-only input in analogy to unimodal LMs such as sentence transformers; on the other hand, they rely more on visual than on textual distributional information being trained with cross-modal contrastive objective, which motivates the comparison of text encoders of contrastive VLMs against unimodal LMs.

While performing poorly on natural language understanding (NLU) benchmarks [HLT22] possibly due to lack of unimodal training, the CLIP text encoder is found to outperform GPT2 [Rad+19] on word-level [RG65; Fin+01; HRK15] as well as sentence-level [Cer+17] similarity benchmarks [WC22b].

Pezzelle, Takmaz, and Fernández [PTF21], however, report mixed results comparing the CLIP text encoder against BERT [Dev+19] on word similarity, suggesting an impact of word concreteness. Chen, Chen, Diao, Wan, and Wang [Che+23] further indicates that CLIP outperforms BERT on a vision-centric task, i.e. to predict whether two captions are from the same image. Yan, Li, Zhu, Lu, Wang, and McAuley [Yan+22] and Hsu, Li, and Yun-Nung [HLY23] compare the CLIP text encoder with PhraseBERT [WTI21] and UCTopic [LSM22], which are unimodal models specialized for phrase-level representations with contrastive learning, where CLIP outperforms its unimodal counterparts when enhanced with domain-aware prompting.

Hypernymy, commonly known as the IS_A or TYPE_OF relation, e.g. (robin, bird) is a hypernymy pair as robin is a type of bird, is a core lexical relation in human concept understanding [Mur04] as well as backbone of semantic hierarchies such as WordNet [Fel00], and therefore used for the evaluation of DSMs [BL11] along with other benchmarks such as word similarity [Fin+01; HRK15]. Baryshnikov and Ryabinin [BR23] even adopts hypernymy to investigate the concept understanding of image synthesis models. Among the hypernymy-related tasks, hypernymy detection is the binary classification between hypernymy pairs such as (robin, bird) and non-hypernymy pairs such as (robin, cat); hypernym discovery [Cam+18] aims to retrieve hypernyms for a given word; graded lexical entailment (GLE) assigns entailment strength scores to word pairs under the assumption that the hypernymy relation is more gradual than binary [Vul+17]. On one hand, GLE successfully reflects *typicality* and *vagueness* in concept categorization [Ham07]; on the other, previous critiques against word similarity evaluation [Far+16] also apply to GLE in this graded setting.

Hypernymy has been used for evaluating word representations ranging from early count-based DSMs [SSS16; Vul+17] as well as the more recent Transformer-based pre-trained LMs such as BERT and GPT2 via prompting [Ett20; HM21; MER21; SVS23]. In particular, Regneri, Abdelhalim, and Laue [RAL24] investigates the representation of hypernymy in BERT qualitatively by analyzing attention matrices of prompts generated with hypernymy versus non-hypernymy pairs, for example, "I like *ravens* and other *animals*" versus "I like *ravens* and other *people*", and observe substantial differences. While hypernymy is mainly used for evaluating unimodal LMs, Liao, Chen, and Du [LCD23] adopt the CLIP text encoder for hypernym discovery with a linear binary classification layer. CLIP outperforms BERT, demonstrating the competitiveness of contrastive VLMs. Both models perform worse on abstract concepts than on concrete concepts, suggesting that abstractness poses a general challenge to both unimodal and multimodal models.

In this work, we focus on text encoders of contrastive VLMs and assess them via hypernymy detection and GLE. Unlike previous work contrasting the CLIP text encoder with GPT2 [Rad+19] or BERT [Dev+19], we ablate the effect of contrastive learning via comparison with sentence transformers [RG19; GYC21], and more directly investigate the contribution of the visual modality. We also explore how word concreteness modulates performance for both types of models. Our research questions are as follows:

RQ1 *How well do contrastive VLMs capture hypernymy compared to unimodal models with contrastive learning Do they exhibit certain advantages due to their visual grounding?* (Chapters 4–5)

RQ2 *Do contrastive VLMs perform worse on abstract or generic words?* (Chapters 5–6)

Our contributions are as follows:

- While previous research on the text encoders of contrastive VLMs mainly focus on comparing CLIP with BERT [Che+23] or GPT2 [WC22b], we perform a direct assessment against sentence transformers with unimodal contrastive learning to examine the effect of visual-semantic training;
- Inspired by the work of Regneri, Abdelhalim, and Laue [RAL24], we propose a novel methodology based on synthetic concepts ("q, a type of p"), defining similarity-based measures to reflect the hypernymy relation, and achieve competitive performance;
- We evaluate 5 multimodal and 7 unimodal models on hypernymy detection and GLE, and qualitatively analyze how contrastive VLMs represent word pairs of different lexical relations;
- We construct a novel dataset, BBC, named after Bolognesi, Burgers, and Caselli [BBC20], to investigate model performance on words of different levels of concreteness and specificity, and find contrastive VLMs to be more robust than previously expected.

We first review related work on visual semantic understanding (Section 2.1) and on hypernymy (Section 2.2); then we propose a novel methodology based on synthetic concept prompts (Chapter 3); in the following chapters we conduct experiments on hypernymy detection (Chapter 4) and on graded lexical entailment (Chapter 5), as well as investigate model performance on word pairs of different levels of concreteness and specificity (Chapter 6); Chapter 7 concludes the work.

Chapter 2

Related work

2.1 Visual-semantic understanding

2.1.1 Multimodal fusion

Distributional semantics [LS23] is a computational approach to word meaning representation [TP10; Len18] and has sparked discussion from the perspective of human cognition [Len+08; Kum20]. Learned from text corpora with no perceptual grounding, its ability to achieve human-like language understanding is questioned [Har90; Bis+20; BK20]. Multimodal distributional semantics is therefore proposed to enhance distributional semantic models (DSMs) with information from the visual [BTB14] or auditory [KC15] modalities.

Feng and Lapata [FL10] propose the first multimodal DSM, representing words based on their distribution via both textual and visual features, improving performance on word similarity [Fin+01] and association [NMS04]. While they learn multimodal information from the same mixed-media corpus, others integrate textual and visual representations which are learnt separately via simple concatenation, obtaining promising performance on a range of semantic tasks including word similarity [BTB11; BTB14; Kie+14], and demonstrating the complementary nature of visual information [Bru+12b; Bru+12a]. More sophisticated multimodal fusion strategies are also explored and shown to outperform concatenation [SL12; SL14]. Lazaridou, Pham, and Baroni [LPB15] extends the predictive skip-gram model [Mik+13b] via jointly predicting visual features for a subset of the vocabulary, thus propagating visual information and improving performance on word similarity benchmarks [BTB14; HRK15].

Apart from static embeddings, contextualized models have been developed to encode multimodal input, typically image-caption pairs, using pre-trained language models such as BERT [Dev+19] as a linguistic backbone. Among these vision-language models (VLMs), ViLBERT [Lu+19] and LXMERT [TB19] adopt a dual-stream architecture encoding text and image input separately before combination, while the single-stream mechanism of VisualBERT [Li+19] and UNITER [Che+20b] conduct early fusion. One specific model, Vokenization [TB20], visually supervised during training, encodes text-only input via automatically retrieving related images. Pezzelle, Takmaz, and Fernández [PTF21] evaluates these multimodal transformers intrinsically, extracting static word embeddings from contextualized multimodal representations with additional visual input. Compared with the unimodal counterpart BERT, they observe increased performance on word similarity evaluation for benchmarks with higher concreteness, i.e. RG-65 [RG65] and MEN [BTB14], but not for more abstract WordSim-353 [Fin+01] and SimLex-999 [HRK15]. Yun, Sun, and Pavlick [YSP21] conduct ablation

studies via training VLMs on text-only data, and find the performance gap between VisualBERT and VideoBERT [Sun+19] variants with vision-language versus text-only pretraining to be marginal. Liu, Yin, Feng, and Zhao [Liu+22] report better spatial common sense knowledge (e.g. a sofa is smaller than a mountain) in the VLM VinVL [Zha+21] than in unimodal models.

Multimodal models are also evaluated on brain alignment as participants read concept words. Anderson, Bruni, Lopopolo, Poesio, and Baroni [And+15] report that textual and visual features correlate better with fMRI activities in linguistic and visual processing areas in the brain, respectively, and that their combination is further more explanatory to conceptual encodings. Bavaresco, Heer Kloots, Pezzelle, and Fernández [Bav+25] investigates human concept processing in two settings, where each concept word either appears in a full sentence that is read by participants and fed to models, or is accompanied by a corresponding image seen by participants and fed to models. Overall, LXMERT [TB19] and VisualBERT [Li+19] correlate with human brain activations more strongly than their unimodal counterparts including BERT. However, the advantage of integrating visual information is not consistent: VisualBERT and MCSE [Zha+22b], a contrastive model with multimodel training, underperform their unimodal counterparts BERT and SimCSE [GYC21] on alignment with both experiential ratings and brain responses [BF25], also casting doubt on whether multimodal models produce more human-like concept representations.

More recently, the triumph of large language models (LLMs) has challenged the necessity of visual grounding for human concept understanding. The empirical evaluation of [Li+24] suggests that unimodal LLMs partially converge towards representations isomorphic to those of vision models. Li, Xu, Dong, Zheng, Liu, Kong, and Sun [Li+23a] probe unimodal LLMs via zero (few)-shot prompting, and demonstrate that the understanding of visual concepts including colour and size emerges as scaling up models in the GPT-based OPT[Zha+22c]-family, which is not the case for embodied concepts such as temperature and mass. While LLMs perform only slightly above chance level on embodied concepts, VLMs including CLIP [Rad+21] and BLIP [Li+22] achieve human-level understanding, in favour of multimodality. Du et al. [Du+25] and Xu, Peng, Nastase, Chodorow, Wu, and Li [Xu+25] investigate object concept understanding in LLMs and multimodal LLMs (MLLMs), suggesting that both capture human conceptual knowledge, while MLLMs have advantages on visual sensory aspects. (For a review on VLM and MLLM architectures, see Wadekar, Chaurasia, Chadha, and Culurciello [Wad+24].)

Overall, while perceptual grounding is argued to be necessary for human-like concept understanding [Har90; Bis+20; BK20], there exist mixed results concerning when and how VLMs outperform their unimodal counterparts, and more investigation into model architectures and multimodal fusion strategies is required to explore how to establish better concept representations with complementary visual information. In this work, we focus on contrastive VLMs (Section 2.1.3).

2.1.2 Concrete vs. abstract words

Word concreteness is a property characterized by the perceptability of its referent [BBC20]. Compared to abstract words corresponding to concepts "neither purely physical nor spatially constrained" [BW05], concrete words are easier to process and remember [Pai91; WH00], and activate overlapping but partly distinct brain systems [Bin+05]. Hill, Korhonen, and Bentz [HKB14] further illustrates that concrete and abstract concepts are organized differently in human concept understanding, where concrete ones are more feature-based and organized according to their semantic similarity. Human concrete ratings are made available in the MRC psycholinguistic database [Col81] as well as the work of Brysbaert,

Warriner, and Kuperman [BWK14], where human concreteness judgments are reported to focus on visual and haptic aspects at the expense of gustatory and auditory ones [LC09; BWK14].

While perception is recognized as necessary for concrete words, it is debatable whether and how abstract words are grounded in human cognition. Paivio's dual-coding theory [Pai91; Pai13] contends that abstract concepts are purely linguistic; Barsalou and Wiemer-Hastings [BW05] argue that abstract concepts are grounded just as concrete ones are, as situational experiences are integrated via mental simulation [Bar99]; others are in favour of an indirect grounding view [Lou11] with language as a cognitive medium [Dov14]. The question then follows whether the assumed advantage of multimodal distributional semantics exists also for abstract concepts, motivating investigation into evaluation of existent VLMs and design of new mechanisms with better abstract concept understanding.

Previous evaluations of multimodal DSMs mainly focus on concrete words [Bru+12a; SL12]. Hill, Reichart, and Korhonen [HRK14] demonstrate that simply concatenating perceptual representations is less beneficial to abstract concepts than to concrete ones, and Kiela, Hill, Korhonen, and Clark [Kie+14] suggest concatenating linguistic and visual representations only for concrete words. Still, Hill and Korhonen [HK14], Takano and Utsumi [TU16], and Utsumi [Uts22] propose new methodologies for establishing multimodal DSMs propagating perceptual information from concrete concepts to abstract ones based on their semantic relatedness, and achieve more robust performance, supporting the indirect grounding view.

For multimodal transformers, it is not clear whether visual information is successfully propagated to enhance abstract concept representations. In the experiments of Pezzelle, Takmaz, and Fernández [PTF21], VLMs outperform their unimodal counterpart BERT only on datasets with higher average concreteness. They further conduct evaluation on subsets containing concrete words only, and find that BERT is consistently outperformed by at least one VLM. This contrast reveals the weakness of evaluated VLMs (ViLBERT, LXMERT, VisualBERT, UNITER, and Vokenization) over abstract words. In this work, we focus on contrastive VLMs, and investigate their performance on abstract words compared to unimodal counterparts with contrastive learning.

2.1.3 Contrastive vision-language models

Contrastive VLMs such as CLIP [Rad+21] have separate text and image encoders, which are jointly trained to maximize the similarity of actual matching image-caption pairs. While other VLMs typically conduct multimodal *fusion* either via simple strategies such as concatenation [FL10; BTB11; KB14] or via neural mechanisms [Lu+19; Li+19; Che+20b], they conduct multimodal *alignment*, mapping representations from the two modalities into a shared visual-semantic space with a contrastive objective, enabling meaningful cross-modal similarity scores for tasks such as cross-modal retrieval [Rad+21]. Such contrastive objectives, which aim to distinguish similar from dissimilar inputs, have also proven effective within single modalities including language [GYC21] and vision [Che+20a; Car+21]. Furthermore, the dual-encoder architecture of contrastive VLMs allows for initialization via pre-trained unimodal models for each module, as well as separate evaluations against counterparts from each modality.

Existent contrastive VLMs adopt different settings for initialization and fine-tuning. For CLIP [Rad+21] and ALIGN [Jia+21], both text and image encoders are trained from scratch. LiT [Zha+22a] uses a pre-trained image encoder frozen during contrastive learning, while the text encoder is trained from scratch. Another model, ALBEF [Li+21], has an extra multimodal fusion encoder, leverages pretrained weights for all encoders, and is trained on extra masked language modeling and image-text matching

bjectives, differing significantly from CLIP-like mechanisms relying solely on contrastive alignment.

With its meaningful cross-modal similarity scores, CLIP has gained popularity for various downstream applications such as zero-shot image classification [Rad+21], as well as other vision-language tasks including Visual Question Answering and Image Captioning [She+21]. It is even evaluated on recognizing emotions in abstract paintings [WT24], ad understanding [BTF24], as well as book/movie genre classification [Bie+22]. Bielawski, Devillers, Van De Cruys, and VanRullen [Bie+22] argue that as CLIP is trained on image-caption pairs "made by human for other humans", it has advantages on dealing with human-centric concepts. Despite its current success, CLIP is reported to behave like a bag-of-words model [Yuk+22], failing to distinguish constituent order and binding in linguistic constructions [Lew+22; CFP23]. Koishigarina, Uselis, and Oh [KUO25] suggest that the individual text and image encoders conduct attribute object binding correctly, whereas the bag-of-words behaviour is due to the insensitivity of the cosine similarity. Kamath, Hessel, and Chang [KHC23], however, investigates this phenomenon with recovery probes to reconstruct captions from their vector representations via the text encoder, and observe failure on more compositional inputs.

On the psycholinguistic task object naming, visual representations produced by the image encoder of CLIP outperform those by computer vision models Bottom-Up [And+18] and the self-supervised Visual Transformer (ViT) [Dos+20] pretrained via DINO [Car+21] on object naming across different image types [Che+24]. Intriguingly, concept representations produced by its text encoder further outperform visual concept representations which are computed by averaging visual exemplar representations, demonstrating the phrase-level concept understanding ability of the CLIP text encoder. Indeed, the text encoders of contrastive VLMs such as CLIP produce phrase- or sentence-level representations for text-only input in analogy to sentence transformers, despite being trained with a cross-modal contrastive learning instead of masked language modeling (MLM) or semantic textual similarity (STS). Compared to unimodal LMs, they have very limited access to distributional patterns in text corpora and benefit more from visual features paired with linguistic captions, which explains the poor performance on natural language understanding (NLU) benchmarks [HLT22].

Several works have evaluated the text encoder of CLIP versus unimodal LMs. As the CLIP text encoder is based on a GPT2[Rad+19]-like architecture [Rad+21], Wolfe and Caliskan [WC22b] compares its semantic representations against those produced by GPT2: first, its contextualized word embeddings do not suffer from high anisotropy as in the case of GPT2 and other unimodal LMs [Eth19], although further ablation studies is required to examine to what extent this results from contrastive learning or multimodality; second, when given single-word inputs, its contextualized embeddings outperform those produced by GPT2 on word similarity benchmarks RG-65 [RG65], WordSim-353 [Fin+01] and SimLex-999 [HRK15], while its sequence-level EOS token embeddings also outperform GPT2 by a smaller margin, in contrast to the findings of Pezzelle, Takmaz, and Fernández [PTF21] where multimodal transformers are outperformed by BERT on WordSim-353 and SimLex-999; third, its sentence-level EOS token embeddings ourperform GPT2 on Semantic Textual Similarity (STS) [Cer+17]. Overall, the text encoder of CLIP establishes semantic representations of remarkable quality despite its small size compared to GPT2.

Chen, Chen, Diao, Wan, and Wang [Che+23] compare the CLIP text encoder with BERT. Despite poor performance on general text understanding including STS, CLIP outperforms BERT on a designed vision-centric task STS-V adapted from STS, i.e. to predict whether two captions are from the same image, demonstrating its advantage in cross-modal association. Apart from BERT, Yan, Li, Zhu, Lu,

Wang, and McAuley [Yan+22] and Hsu, Li, and Yun-Nung [HLY23] compare the CLIP text encoder with two unimodal models specialized for phrase-level representations with contrastive learning, namely PhraseBERT [WTI21] and UCTopic [LSM22]. Hsu, Li, and Yun-Nung [HLY23] report mixed results on phrase understanding, wheras CLIP with the domain-aware prompting strategy proposed by Yan, Li, Zhu, Lu, Wang, and McAuley [Yan+22] outperform the unimodal counterparts.

Research on brain alignment exhibit mixed results: compared to BERT and the contrastive unimodal sentence transformer SimCSE [GYC21], contextualized embeddings produced by the CLIP text encoder are less aligned to human brain activations viewing single word stimuli [BF25]; when concept words are read by human participants and fed to models in full sentences, or when single word concepts are paired with corresponding images which are available to humans and VLMs but not unimodal models, CLIP is more brain-aligned than BERT [Bav+25]. Tikhonov, Bylinina, and Paperno [TBP23] also compare CLIP with BERT-based unimodal LMs, and find that word concreteness contributes significantly to explaining the impact of visual grounding, in accordance with previous findings of Pezzelle, Takmaz, and Fernández [PTF21].

The CLIP text encoder is initialized from scratch and has limited access to textual distributional patterns during training, motivating exploration into the impact of initialization from a pretrained LM or joint training with unimodal objectives. Zhao et al. [Zha+23b] report that SimCSE-like unimodal contrastive learning improves CLIP on cross-modal retrieval, while there is a lack of evidence whether it benefits concept representations of the text encoder as well. Research on initialization from pretrained LMs is more concerned about efficiency, with evaluation limited to cross-modal tasks as well [KF23]. While further investigation is needed, it seems questionable whether distributional semantic information learnt via unimodal training is beneficial to text encoders of contrastive VLMs at all: Zhuang, Fedorenko, and Andreas [ZFA23] train two novel variants of CLIP from scratch on two datasets based on Conceptual-Captions-12M [Cha+21]: a visual + language model on its original image-caption pairs, and a visual + word model via replacing each image-caption pair by multiple image-word pairs, splitting the single words in the original caption and thus ablating co-ocurrence patterns. The visual + language model turns out to significantly underperform its visual + word variant on word learning, calling for more existensive research on how to better integrate complementary visual and distributional information.

With representations in a shared visual-semantic space, the text encoder of CLIP is also used for image synthesis models such as DALL-E 2 [Ram+22], Stable Diffusion [Rom+22], and VQGAN [ERO21], which can in turn assist the analysis of CLIP encoding [RRG22; WC22a] or further enhance its semantic understanding [HLY23]: given a textual prompt, Hsu, Li, and Yun-Nung [HLY23] generates an image via Stable Diffusion with CLIP text encoding, which is then fed to the CLIP image encoder, and the two encodings are combined via concatenation, improving performance on phrase understanding datasets but not consistently. Liu, Yin, Feng, and Zhao [Liu+22] adopt VQGAN-CLIP [Cro+22] for investigating spatial commonsense, where generated images are evaluated both automatically and manually, achieving the most accurate and consistent performance, also illustrating the potential of CLIP concept understanding.

2.2 Hypernymy in distributional semantics

2.2.1 Tasks & datasets

Hypernymy, commonly known as the IS_A or TYPE_OF relation, e.g. robin is a type of bird, is a core lexical relation in human concept understanding [Mur04] and the asymmetric backbone of semantic hierarchies such as WordNet [Fel00]. While evaluation tasks such as word similarity [Far+16] measure how human concept understanding is reflected in distributional semantic models (DSMs) in terms of distance in the semantic space, hypernymy-based evaluation further examines the representation of the semantic hierarchy by distinguishing hypernymy pairs like (robin, bird) from word pairs of other relations, such as (robin, dove) and (robin, feather).

Hypernymy detection is a binary classification task: given a pair (q, p), predict whether q is a type of p. Some task formulations take directionality into account and assign a separate label to cases where p is a type of q [Kie+15]. BLESS [BL11], a benchmark for distributional semantic evaluation spanning 200 distinct target concepts (q) paired with relata corresponding to hypernymy as well as coordination (also known as co-hyponymy, e.g. "robin" and "dove" are both hyponyms of "bird"), meronymy (the part-whole relation, e.g. "robin" has the part "feather") and randomly matched nouns, is often used for hypernymy detection. Alongside variants of BLESS, WBLESS [Wee+14] and BiBLESS [Kie+15], other datasets including Lenci/Benotto [LB12], LEDS [Bar+12], EVALution [San+15], and SHWARTZ [SGD16] have also been developed for hypernymy detection. Another related task is hypernym discovery, i.e. retrieve as many hypernyms as possible for a given query q, with datasets provided by SemEval-2018 Task 9 covering three languages and two specific domains [Cam+18].

While these task formulations treat hypernymy and non-hypernymy as discrete cases, Vulić, Gerz, Kiela, Hill, and Korhonen [Vul+17], however, argues that hypernymy is "more gradual than binary": on one hand, the hypernymy relation tends to be stronger for some word pairs than others, e.g. "robin" is more representative of the "bird" category than "penguin" (*typicality*) [Ros75]; on the other, the boundary between hypernymy and non-hypernymy can be fuzzy (*vagueness* [KP95]), e.g. it is not clear whether "wheelchair", "table", and "bench" can be considered hyponyms of "chair", and to what degree. They define a graded lexical entailment (GLE) task: given a word pair (q, p), predict its entailment strength s, which is then evaluated in terms of correlation with human judgments in their dataset HyperLex.

2.2.2 Previous approaches

Hearst-style patterns

As indicated by the pioneering work of Hearst [Hea92], frequent occurrences of expressions including "p such as q" and "q or any other p" indicate that (q, p) is a hypernymy pair. Such syntagmatic patterns can be automatically extracted and then exploited to enhance hypernymy detection [SJN04]. However, pattern-based methods suffer from the sparsity problem as hypernymy pairs do not necessarily co-occur in Hearst-style patterns [RKN18], obtaining high precision and low recall in the hypernym discovery task. Still, they provide valuable contextual constraints and are used to augment both hypernymy detection [SGD16; RE16; Le+19] and hypernym discovery [SJN04; BB18; HH19].

Unsupervised measures

Early count-based DSMs are evaluated on hypernymy detection via unsupervised measures, following several distributional hypotheses (see Shwartz, Santus, and Schlechtweg [SSS16]). First, a hypernymy pair (q, p) is supposed to be semantically similar and therefore distributionally similar [San+14], and similarity measures such as cosine, Lin [Lin98], and ApSyn [San+16b] are used as hypernymy measures. However, such symmetric measures are insufficient as they cannot even distinguish the directionality of the pair; moreover, coordination pairs also tend to be similar and are therefore very difficult [Wee+14].

Two other hypotheses have been proposed to better capture the asymmetry of the hypernymy relation. The *distributional inclusion hypothesis* [GD05] contends that contexts of a hyponym q are expected to be included in contexts of its hypernym q: as *a dog barks* entails *an animal barks*, "bark", being a context of "dog", should also be a context of "animal". Several measures have been proposed to quantify the level of context inclusion between a word pair (q, p), including WeedsPrec [WW03], ClarkeDE [Cla09], and APinc [Kot+10]. Some combine multiple measures via geometric means, e.g. cosWeeds [LB12] is the geometric mean of cosine similarity and WeedsPrec, balAPinc [Kot+10] of APinc and Lin similarity, and invCL [LB12]is also a geometric mean based on ClarkeDE. There also exists attempts to refine the distributional inclusion hypothesis, e.g. Pannitto, Salicchi, and Lenci [PSL18] smooths the quantification of inclusion by expanding the range of contexts via neighbours, and Roller, Erk, and Boleda [REB14] report that inclusion only applies selectively to relevant dimensions.

Other measures are based on the *distributional informativeness hypothesis*: hyponyms, being more specific, are also more informative than their hypernyms and therefore likely to occur in less general contexts. The SLQS measure [San+14] evaluates word informativeness via the median entropy of its top contexts; Rimell [Rim14] measures the *topic coherence* of a word also via its top contexts, and then calculates an RCTC (ratio of change in topic coherence) score evaluating the impact of excluding contexts of q from contexts of p, and vice versa. Apart from these methods designed for count-based DSMs, there exists another informativeness measure utilising the visual modality: observing that related images of hypernyms exhibit more variability than their hyponyms [DF11], Kiela, Rimell, Vulic, and Clark [Kie+15] obtain related images of a concept word via *Google Images* and use their CNN-derived visual representations to calculate an image dispersion measure for hypernymy detection.

These unsupervised measures are used for both binary hypernymy detection [SSS16] and graded lexical entailment [Vul+17], and the image dispersion measure of Kiela, Rimell, Vulic, and Clark [Kie+15] outperforms the textual distributional measures on the noun subset of HyperLex [Vul+17], demonstrating the potential of the visual modality. Such unsupervised measures can be further combined, e.g. via the random forest algorithm of Santus, Lenci, Chiu, Lu, and Huang [San+16a].

Supervised approaches

Apart from these unsupervised measures based on distributional hypotheses, attempts have also been made to capture the hypernymy relation via supervision, e.g. to develop classifiers based on the concatenaton [Bar+12], offset [REB14], or even pointwise product [Wee+14] of the two vectors of a word pair via linear regression or support vector machines, achieving better performance. However, such supervised approaches are reported to suffer from *lexical memorization*, as they tend to predict whether p is a "prototypical hypernym" rather than distinguish whether (q, p) is a hypernymy pair [Lev+15].

Beyond early count-based DSMs, predictive models have trained to maximize the probability of observed word-context co-occurrence. From static CBOW embeddings [Mik+13a] to contextualized

models such as BERT [Dev+19], predictive models have gained prominence and are shown to outperform their count-based counterparts on a range of lexical semantic tasks [BDK14]. In the case of hypernymy, the unsupervised inclusion or informativeness measures are no longer applicable due to their lack of interpretability, while diverse supervised approaches have been explored, e.g. Shwartz, Goldberg, and Dagan [SGD16] encode the dependency paths between word pairs via an LSTM network [HS97] in addition to word vector concatenation.

A distinct class of supervised methods aim to capture the hypernymy relation via *projection learning*: Fu, Guo, Qin, Che, Wang, and Liu [Fu+14] argue that the vector offset is insufficient for representing the hierarchical hypernymy relation of a word pair, and propose to map hyponyms to their hypernyms via linear projection instead. Since then, several works have trained transition matrices such that the projections of a word embedding are close to the embeddings of its hypernyms [Yam+16; BB18; Wan+19; Bai+21]. Moreover, Wang and He [WH20] exploit the residual vectors, i.e. the offsets between hyponym projections and hypernyms, as input to classifiers, which substantially alleviates lexical memorization. Meanwhile, Kruszewski, Paperno, and Baroni [KPB15] optimize a mapping function in order to derive Boolean structures between hypernymy pairs.

Specialized embeddings

The hypernymy relation is characterized by its hierarchical structure and often formalized as a partial order. Vendrov, Kiros, Fidler, and Urtasun [Ven+15] introduce an order-preserving semantic space trained on hypernymy pairs from WordNet, which is shown to improve hypernymy detection. Nguyen, Köper, Walde, and Vu [Ngu+17] develop hierarchical embeddings such that hypernymy pairs are more similar than word pairs of other lexical relations in terms of cosine distance, and the magnitude of a hypernym is larger than that of its hyponym in terms of Euclidean norms. Vulić and Mrkšić [VM17] similarly preserve hypernymy via vector norms, while optimizing semantic similarity for both hypernymy and synonymy pairs. While word embeddings are typically based on a Euclidean vector space, attempts have been made to preserve hypernymy via convex cones in hyperbolic spaces, based on the Poincaré ball [NK17; GBH18] or Lorentz model [NK18].

2.2.3 Hypernymy in pre-trained language models

With the emerging popularity of Transformer-based pretrained language models (PLMs) such as BERT [Dev+19] and GPT2 [Rad+19], there is also growing interest in zero-shot evaluation of their semantic knowledge, including lexical semantic understanding [Vul+20; BCS20]. Ettinger [Ett20] examines the predictive capacities of BERT given cloze-style prompts such as "A robin is a [MASK]", and report high accuracy on retrieving canonical hypernyms ("bird" in the case of "robin"). However, the systematicity of such hypernymy knowledge is under question, as performance drops significantly with prompts in plural form ("robins are [MASK]") [Rav+20]. Hanna and Mareček [HM21] experiments with more complex prompts, inspired by Hearst-style patterns, and conduct evaluation on hypernym discovery [Cam+18], where zero-shot BERT proves competitive when compared to supervised approaches. Apart from querying BERT with cloze-style prompts, Shani, Vreeken, and Shahaf [SVS23] probes GPT-based models with a binary question answering scheme ("Is q a type of p?").

Misra, Ettinger, and Rayz [MER21] further extracts the conditional probability of predicting p given the prompt "A(n) q is a(n)" as the *taxonomic verification* score of the pair (q, p) for both BERT-like masked LMs and GPT2-like autoregressive LMs. originally designed for probing typicality

knowledge in PLMs, this measure is shown to outperform previous unsupervised count-based measures on GLE [Vul+17; RDG23]. Also using conditional probabilities as hypernymy measures, Tikhomirov and Loukachevitch [TL24] probe GPT2 with a large number of prompts [Sei+16] and highlight the significance of prompt quality.

In particular, Regneri, Abdelhalim, and Laue [RAL24] investigates the representation of hypernymy in BERT by analyzing attention matrices based on Hearst-style patterns. For example, they generate the positive prompt "I like *ravens* and other *animals*" based on the hypernymy pair (raven, animal), as well as counterfactual prompts using non-hypernymy pairs formed by replacing either q or p with its co-hyponym, such as (raven, crow) and (raven, people). Attention patterns for the three groups of word pairs exhibit substantial differences, and overall attention is lower for positive prompts than for counterfactuals, providing qualitative evidence that BERT successfully captures hypernymy.

While most research probe hypernymy in PLMs via prompting, Liao, Chen, and Du [LCD23] evaluate the CLIP text encoder against on hypernym discovery with a linear binary classification layer. CLIP outperforms BERT, demonstrating the competitiveness of contrastive VLMs. Both models perform worse on abstract concepts than on concrete concepts, suggesting that abstractness poses a general challenge to language models.

Chapter 3

Methodology

3.1 Synthetic concepts

For a word pair (q, p), we generate two synthetic concepts q' and p' for detecting their lexical relation. More specifically, for detecting hypernymy pairs where q is supposed to be the hyponym and p to be the hypernym, e.g. (robin, cat), let q' be "q, a type of p" and p' be "p, such as q" following Hearst patterns [Hea92].

If hypernymy holds for (q, p), i.e. if q is indeed a type of p, then q' ("q, a type of p") easily understood as just q, and p' as p. However, for negative pairs where hypernymy does not hold, such as (robin, tree) and (raven, person), q' and p' presuppose counterfactual class inclusion statements, e.g. "a robin is a tree", and are processed differently. Fischler, Bloom, Childers, Roucos, and Perry Jr [Fis+83] report that such false affirmative sentences produce slower responses than true sentences ("a robin is a bird") as well as substantially more negative averaged ERPs. Counterfactual expressions e.g. "ravens are persons" are also reported to behave differently in terms of BERT attention maps [RAL24]. In this work, we evaluate phrase-level representations of these synthetic concepts, comparing hypernymy against non-hypernymy pairs. Our intuitions are as follows:

For hypernymy pairs, q' ("q, a type of p") would denote the same concept as q, and p' the same as p, reflected in similarity scores of their word/phrase-level embeddings:

$$1 \approx \text{sim}(q', q) > \text{sim}(q', p) \approx \text{sim}(q, p)$$
$$1 \approx \text{sim}(p', p) > \text{sim}(p', q) \approx \text{sim}(p, q)$$

where sim denotes the cosine similarity of embeddings produced by text encoders.

Meanwhile, for non-hypernymy pairs, q' and p' would be synthetic hybrid concepts by blending q and p. On one hand, we could expect $\text{sim}(q', q)$ and $\text{sim}(p', p)$ to be smaller than 1. On the other hand, as q' is stipulated to be "a type of p" due to the appositive, it is expected to be more similar to p than the original q is. The case for p' is similar, as it is stipulated to encompass q. Therefore we expect:

$$1 > \text{sim}(q', q), \text{sim}(q', p) > \text{sim}(q, p)$$
$$1 > \text{sim}(p', p), \text{sim}(p', q) > \text{sim}(p, q)$$

3.2 Measures

With intuitions demonstrated in the last section, we design the following measures reflecting whether (q, p) is a hypernymy pair.

B (baseline) measure We use $\text{sim}(q, p)$, the naïve cosine similarity between q and p, as a baseline, which does not involve the synthetic concepts q' and p'. Note that similarity is a symmetric measure and therefore insufficient for the asymmetric hypernymy relation.

S (similarity) measures We evaluate whether q' is the same concept as q and whether p' is the same as p via their similarity scores. s_q and s_p are defined as follows:

$$s_q(q, p) = \text{sim}(q, q')$$
$$s_p(q, p) = \text{sim}(p, p')$$

s_p and s_q are expected to be higher (closer to 1) for hypernymy pairs and lower for non-hypernymy.

R (ratio) measures Inspired by the RCTC measure [Rim14], we propose the R (ratio) scores comparing the original and synthetic concepts in terms of similarity to their relatum term:

$$r_q(q, p) = \frac{\text{sim}(q, p)}{\text{sim}(q', p)}$$
$$r_p(q, p) = \frac{\text{sim}(p, q)}{\text{sim}(p', q)}$$

r_p and r_q are also expected to be higher (closer to 1) for hypernymy pairs. [1]

C (check) measures Considering the asymmetry of the hypernymy relation, q' and p' are not expected to be a naïve average of q and p. Although it is not clear how the counterfactual hybrid concepts would behave in the case of non-hypernymy, we expect q' to be closer to q than to p and p' to be closer to p than to q for hypernymy pairs. Additional measures are designed for sanity check:

$$c_q(q, p) = \frac{\text{sim}(q', q)}{\text{sim}(q', p)}$$
$$c_p(q, p) = \frac{\text{sim}(p', p)}{\text{sim}(p', q)}$$

The C measures are expected to be larger than 1 for hypernymy pairs.

In our first experiment (see Chapter 4), the synthetic concept p' fails the sanity check as c_p tend to be smaller than 1 in most cases, while q' behaves as expected. We discuss this phenomenon in Appendix A, and use only s_q and r_q as hypernymy measures in the upcoming chapters. Despite its failure in this hypernymy detection scenario, we argue that p', the synthetic concept based on p, may still prove informative for other tasks, and include it here for completeness and future extensibility.

[1] While the numerator, $\text{sim}(q, p)$ might be negative for non-hypernymy pairs, the denominators are much less likely to be negative because of the designed stipulation.

3.3 Models

In this work, we evaluate unimodal and multimodal models with contrastive learning on hypernymy tasks based on their representations of isolated words and phrases. We focus on text encoders of contrastive VLMs (Subsection **??**) alongside unimodal sentence transformers.

Sentence transformers Sentence transformers are typically initialized with a pre-trained language model, e.g. BERT [Dev+19], and then fine-tuned on natural language inference (NLI) or Semantic Textual Similarity (STS) to learn effective sentence embeddings. We evaluate **SimCSE** [GYC21][2], fine-tuned contrastively with self-supervision using dropout; **GTE** [Li+23b] [3], a BERT-based general-purpose text embedding model with multi-stage contrastive learning over diverse datasets; **MiniLM** [Wan+20], **MPNet** [Son+20], and **DistilRoBERTa** variants of SentenceBERT [RG19] [4], which use Siamese networks trained also with contrastive learning; as well as **PhraseBERT** [WTI21][5], fine-tuned contrastively on a phrasal paraphrase dataset.

Contrastive VLMs As for contrastive VLMs, we evaluate the text encoders of the `ViT-B/32` and `ViT-L/14` variants of **CLIP** [Rad+21] [6] with a GPT2-like [Rad+19] architecture, which are jointly trained with the corresponding ViT [Dos+20] variants, as well as **ALIGN** [Jia+21][7] with a BERT-like architecture, jointly trained with an EfficientNet[TL19]-based image encoder using the same loss function while leveraging noisy data. Note that CLIP and ALIGN train both text and image encoders from scratch. **LiT** [Zha+22a] [8] trains the text encoder from scratch alongside a frozen pre-trained image encoder, and outperforms CLIP and ALIGN on ImageNet [Den+09] zero-shot classification.

MCSE Also equipped with a text encoder and an image encoder mapped onto the same space, MCSE [Zha+22b][9] differs from the contrastive VLMs above in terms of initialization, training objective, and training data: its text and image encoders are both initialized with pre-trained models; it extends the contrastive framework of SimCSE with CLIP-like multimodal contrastive learning; its unimodal training data is the same as SimCSE, while its multimodal training data is much smaller. Like LiT, the image encoder of MCSE is also locked. Designed for sentence embedding learning, MCSE is evaluated not on cross-modal tasks but on Semantic Textual Similarity (STS), and outperforms the unimodal SimCSE significantly.

Taxonomic verification Additionally, we use the taxonomic verification method proposed by Misra, Ettinger, and Rayz [MER21] as a baseline, which calculates the conditional probability of producing p as the next word given the prompt "A(n) q is a(n)". According to Renner, Denis, and Gilleron [RDG23], it is the best-performing LM-based approach to GLE [Vul+17].[10]

[2] `https://huggingface.co/princeton-nlp/unsup-simcse-bert-base-uncased`
[3] `https://huggingface.co/thenlper/gte-base` — `gte-small`
[4] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2` — `all-mpnet-base-v2` — `all-distilroberta-v1`
[5] `https://huggingface.co/whaleloops/phrase-bert`
[6] We refer to these two variants with as CLIP-b and CLIP-l. `https://github.com/openai/CLIP`
[7] `https://huggingface.co/kakaobrain/align-base`
[8] We use their `LiT-B16B` model. `https://github.com/google-research/vision_transformer#lit-models`.
[9] `https://huggingface.co/UdS-LSV/mcse-flickr-bert-base-uncased`
[10] Note that Renner, Denis, and Gilleron [RDG23] report results using GPT2-XL, whereas we only experiment with BERT [Dev+19] and GPT2 [Rad+19] due to computational constraints.

Chapter 4

Experiment 1: hypernymy vs. other relations

In this chapter, we perform binary hypernymy detection, distinguishing hypernymy pairs from word pairs of other lexical relations.

4.1 Experimental setup

Dataset BLESS [BL11] is a semantic relation dataset designed for distributional semantic evaluation and commonly used for hypernymy detection (see, inter alia, [SSS16; Kie+15; RKN18; WH20]). We experiment with its noun-noun subset, which contains 14547 (q, p) pairs spanning 200 distinct target concepts (q). All target concepts are unambiguous, concrete, basic-level English nouns consisting of a single word. Apart from 1337 positive hypernymy pairs, each concept is also paired with relata corresponding to other lexical relations including coordination and meronymy, as well as random nouns.

Evaluation For evaluation, we first rank all (q, p) pairs according to the similarity-based measures introduced in Section 3.2. we then conduct a one-sided Welch's t-test [DLL17], and compute average precision (AP) following Shwartz, Santus, and Schlechtweg [SSS16]. Our other metric is MAP, the mean AP value averaged over the 200 concepts, following [LB12]. We also evaluate the models on discriminating hypernymy from each of the negative relations, and provide boxplots in Appendix B.1.

Image generation To qualitatively examine the representation of lexical relations in the multimodal semantic space, we follow White and Cotterell [WC22a] and use the Stable Diffusion model [Rom+22] [1] with CLIP-l encoding for image generation. By looking at generated images with synthetic prompts such as "robin, a type of bird" and "robin, a type of dove", we gain a deeper understanding of the semantic distance between the synthetic q' and the original concepts.

4.2 Results & analysis

Table 4.1 summarizes the performance of all models against different distractor relation types, using proposed measures described in Section 3.2. Images generated for the synthetic concept q' w.r.t. different lexical relations are presented in Figure 4.1.

[1] `https://huggingface.co/CompVis/stable-diffusion-v1-4`

		vs. non		vs. coord		vs. mero		vs. random	
		AP	MAP	AP	MAP	AP	MAP	AP	MAP
	proportion	0.0919	0.0951	0.2727	0.2821	0.3124	0.3626	0.1663	0.1690
b	CLIP-b	0.1741	0.3012	0.3398	0.4384	0.4234	0.6135	0.4210	0.6469
	CLIP-l	0.1585	0.2695	0.2811	0.3926	0.4410	0.6108	0.4583	0.6464
	ALIGN	0.1411	0.2534	0.2322†	0.3342	0.3895	0.5848	0.5722	0.7333
	LiT	0.1889	0.3873	0.3844	0.5314	0.4534	0.7038	0.4146	0.7370
	MCSE	0.1440	0.2215	0.2468†	0.3379	0.3722	0.5193	0.5408	0.5740
	SimCSE	0.1527	0.2351	0.2739†	0.3787	0.3575	0.4983	0.5320	0.5766
	GTE-small	0.2051	0.3223	0.3160	0.4333	0.4109	0.5792	0.7149	0.7675
	GTE-base	0.2512	0.3788	0.3548	0.4771	0.4829	0.6338	**0.7997**	**0.8264**
	MiniLM	0.1813	0.2854	0.2734†	0.3790	0.4091	0.5660	0.7102	0.7476
	MPNet	0.1614	0.2525	0.2303†	0.3313	0.4522	0.5790	0.7295	0.7606
	DistilRoBERTa	0.1516	0.2403	0.2512†	0.3664	0.3956	0.5494	0.5511	0.6036
	PhraseBERT	0.1111	0.1603	0.1816†	0.2395	0.3808†	0.5073	0.6007	0.6545
s_q	CLIP-b	0.2129	0.3230	0.5832	0.7263	0.5984	0.7164	0.2869	0.3924
	CLIP-l	0.2984	0.4699	0.5134	0.7096	0.6253	0.7534	0.4785	0.6424
	ALIGN	0.1903	0.3473	0.4566	0.6907	0.4726	0.6736	0.3120	0.4575
	LiT	0.1152	0.1701	**0.6475**	**0.8109**	0.5075	0.7056	0.1487	0.1971
	MCSE	0.1596	0.2515	0.3511	0.5348	0.4477	0.5765	0.3060	0.4122
	SimCSE	0.2100	0.3835	0.5514	0.7905	0.5040	0.7076	0.3147	0.4665
	GTE-small	0.3815	0.5748	0.5786	0.7657	0.6151	0.7611	0.6095	0.7674
	GTE-base	0.3354	**0.5897**	0.5194	0.7545	0.5983	0.7592	0.5698	0.7733
	MiniLM	0.2795	0.4899	0.4854	0.6870	0.5537	0.6928	0.4635	0.6457
	MPNet	0.2581	0.4754	0.4064	0.6575	0.5670	0.7323	0.5054	0.6530
	DistilRoBERTa	0.2723	0.4033	0.4511	0.6338	0.5612	0.6709	0.4492	0.5510
	PhraseBERT	0.1199	0.1673	0.2231†	0.3052	0.3427†	0.4359	0.4096	0.5626
r_q	CLIP-b	0.2318	0.3776	0.5088	0.6716	0.4593	0.5932	0.4353	0.6015
	CLIP-l	0.1653	0.2919	0.4265	0.5781	0.4202	0.5819	0.2986	0.4658
	ALIGN	0.2869	0.4428	0.5254	0.6698	0.4764	0.6081	0.6198	0.7043
	LiT	0.1490	0.2775	0.5529	0.7679	0.3938	0.5280	0.2297	0.3722
	MCSE	0.1726	0.2637	0.3019†	0.4102	0.3925	0.5414	0.5281	0.5751
	SimCSE	0.2322	0.3391	0.4704	0.5768	0.4552	0.5811	0.4598	0.5633
	GTE-small	0.3394	0.5022	0.4973	0.6463	0.5537	0.6826	0.7258	0.7755
	GTE-base	**0.4268**	0.5611	0.6077	0.7104	0.6318	0.7208	0.7475	0.7907
	MiniLM	0.3134	0.4348	0.4418	0.5649	0.5369	0.6381	0.6488	0.6977
	MPNet	0.3015	0.4343	0.3877†	0.5261	0.5825	0.6713	0.7225	0.7686
	DistilRoBERTa	0.2521	0.3730	0.4154	0.5469	0.5528	0.6628	0.5249	0.6170
	PhraseBERT	0.0992	0.1451	0.1819†	0.2457	0.2793†	0.4245	0.4791	0.5509
veri	BERT	0.2421	0.3525	0.3535	0.4848	**0.7004**	**0.7769**	0.5164	0.6031
	GPT2	0.2255	0.3461	0.2940†	0.4248	0.5508	0.6438	0.7595	0.8093

Table 4.1: Model performance on binary hypernymy detection.
"Proportion" indicates the proportion of positive pairs, i.e., the AP or MAP of a random baseline.
For each model, we report AP and MAP scores evaluating the ability to distinguish hypernymy pairs from other lexical relations in BLESS, based on the measures introduced in Section 3.2. Boldface highlights the best score in each column.
Statistical significance tests (one-sided Welch's t-test) are conducted over the full dataset. †alongside APs denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

Hypernymy: “robin, a type of bird”

Coordination: “robin, a type of dove”

Meronymy: “robin, a type of feather”

Random: “robin, a type of sewerage”

Random: “robin, a type of earring’

Figure 4.1: Generated images for the synthetic concept q' with CLIP-l encoding using Stable Diffusion.

4.2.1 Hypernymy vs. non-hypernymy

We observe that our method outperforms taxonomic verification [MER21], and that s_q and r_q tend to outperform the naïve similarity baseline (b) except with LiT or PhraseBERT, demonstrating the utility of the synthetic concept q'. This indicates that most models successfully capture the hypernymy relation between concepts using the simple template "{q}, `a type of` {p}", although CLIP is sometimes observed to behave like a bag-of-words model [Yuk+22].

GTE-base is the best-performing model both in terms of AP (0.4268 with r_q) and MAP (0.5897 with s_q). While the performance gap between GTE-base and GTE-small is relatively small, GTE surpasses the other models by a significant margin, which can be attributed to its carefully curated training objectives and training data [Li+23b]. Among the other sentence transformers, MiniLM is the best-performing model, achieving an AP of 0.3134 (with r_q) and an MAP of 0.4899 (with s_q) still outperforming the best contrastive VLM combination, i.e. CLIP-l with s_q (AP 0.2984, MAP 0.4699) by a less dramatic margin. CLIP outperforms SimCSE in this hypernymy detection setting, although it is previously reported be yield less brain-aligned word representations than BERT and SimCSE using a different methodology [BF25]. PhraseBERT performs the worst despite being optimized for phrase-level representations. Overall, contrastive VLMs underperform unimodal models. Nevertheless, since their text encoders are trained from scratch, these models have very limited access to distributional patterns in text corpora and rely more on visual features learned during training. Therefore, their current performance is already remarkable.

Interestingly, results show that MCSE is outperformed by SimCSE, in accordance with previous findings that MSE capture less experientual information than SimCSE [BF25]. Such performance drop suggests that additional CLIP-like contrastive learning does not necessarily benefit sentence embedding models. MCSE also underperforms contrastive VLMs, which requires further ablation studies to identify the impact of initialization from a pre-trained LM, text-only contrastive learning, and the small amount of multimodal training data.

4.2.2 Hypernymy vs. coordination

Distinguishing hypernymy from coordination is known to be a challenging task [LB12; Wee+14] where symmetric similarity scores are insufficient. In this setting, our asymmetric s_q and s_q measures consistently outperform naïve cosine similarity, which sometimes underperform a random baseline. The best-performing model turns out to be the contrastive VLM LiT in terms of both AP and MAP. Again, MCSE is outperformed by both SimCSE and contrastive VLMs by a significant margin.

While further analysis is needed, we hypothesize that text encoders of contrastive VLMs hold an inherent advantage in this task due to their visual grounding. As contrastive VLMs are pre-trained with an image-caption matching objective, the word embeddings they produce are aligned with the visual features of images whose captions contain the corresponding word, i.e. their occurrences in the image-caption corpora. For instance, the representation of "robin" is closely associated with exemplars of the visual object "robin". Co-hyponyms and their hypernyms are often distributionally similar in text corpora, making them difficult to distinguish, whereas their occurrences in image-caption corpora can be distinguished in terms of visual features.

Consider the hypernymy pair (robin, bird) and the coordination pair (robin, dove). Exemplars of "robin", can be viewed as a subset of exemplars of "bird", as images depicting robins are depicting birds

at the same time, [2] while exemplars of "dove" constitute another subset of "bird" almost distinct from "robin'. Visual features of these exemplars benefit the understanding of the hypernymy inclusion as well as the distinction between the co-hyponyms.

Intriguingly, while generated images of "robin, a type of bird" consistently depict robins, similar to those generated with just "robin" (see Appendix A), generated images of "robin, a type of dove" tend to depict a hybrid bird with visual features from both robins and doves (see Figure 4.1). This observation aligns with the fact that the s_q score is higher for (robin, bird) than for (robin, dove), supporting our assumption that s_q would be higher for hypernymy pairs.

4.2.3 Hypernymy vs. meronymy

On the hypernymy-meronymy distinction, the best-performing sentence transformer is GTE-base in terms of AP (0.6318 with r_q), and GTE-small in terms of MAP (0.7611 with s_q). The best-performing contrastive VLM is CLIP-l, achieving an AP of 0.6253 and an MAP of 0.7534 with s_q. Multimodal models underperform unimodal models with contrastive learning, and both are outperformed by taxonomic verification with BERT (AP=0.7004, MAP=0.7769) by a more significant margin. Considering the performance gap between GTE and BERT, We hypothesize that this is because BERT has remarkable syntactic abilities [Gol19], sentence embedding models are not sufficiently sensitive to syntactic patterns, which are effective for distingsuishing meronymy from hypernymy [SSS16]. Zhang, Feng, Teng, Liu, and Li [Zha+23a] report that sentence transformers including SimCSE and the MiniLM and MPNet variants of Sentence-BERT [RG19] perform poorly on syntactic understanding.

As for the performance of contrastive VLMs, We observe that r_q tends to underperform the naïve cosine similarity baseline except with ALIGN: for a meronymy pair (q, p), "q, a type of p" is not necessarily significantly more similar to p than q is. We hypothesize that this is due to the distribution of meronymy pairs in the image-caption corpora. Consider "robin" and its meronym "feather": images depicting "robin" also depict feathers and can be paired with captions containing the word "feather". Therefore, exemplars of holonyms and meronyms theoretically form a subset relation in analogy to those of hyponyms and hypernyms, making them more difficult to distinguish relying on visual information alone. Meanwhile, as 'CLIP's training dataset is mainly composed of image annotations made by humans for other humans' [Bie+22], in practice exemplars of "robin" are less likely to be annotated with "feather", "wing", "eye", and "beak", and exemplars of "feather" often depict isolated feathers rather than parts of birds, alleviating this difficulty. Prompted with the synthetic concept "robin, a type of feather", generated images depict robins, feathers or feathers with robin-like colouring inconsistently (Figure 4.1), demonstrating the sophisticated visual understanding of meronymy.

4.2.4 Hypernymy vs. random pairs

Since random word pairs are typically dissimilar, symmetric similarity is a strong baseline for distinguishing hypernymy. In fact, in our experiment, the best performance is achieved by GTE-base using naïve cosine similarity, whereas our s_q and r_q measures perform no better. To account for this

[2] In analogy to the *distributional inclusion hypothesis*, it is plausible to propose a *multimodal distributional inclusion hypothesis*, where occurrences of a hyponym can be replaced with its hypernym without falsifying the image-caption match—e.g., any image depicting a robin may also be captioned as "bird". Similarly, following the *distributional informativeness hypothesis*, images of a hypernym tend to be more variable than those of a hyponym—e.g., images of birds encompass not only robins but also doves and ravens—as observed by Deselaers and Ferrari [DF11] and operationalized in the visual generality measure of Kiela, Rimell, Vulic, and Clark [Kie+15].

phenomenon, we also look at generated images via prompting the synthetic concepts for hypernymy. For the random pair (robin, sewerage), generated images tend to depict robins, but differ from images for "robin" in terms of visual background. For (robin, earring), generated images depict earrings, sometimes bird-shaped. We conclude that for some random pairs, the counterfactuality of synthetic concepts causes processing difficulty, and they are represented similarly to the original concept q depite the irrelevant relata p, posing a challenge to our measures specialized for hypernymy.

Considering the importance of asymmetric measures for distinguishing hypernymy from other distractor relation types, it is thus worth exploring the combination of multiple measures [San+16a] for accomplishing a more robust hypernymy detection approach. Inspired by previous unsupervised count-based metrics [Kot+10; LB12] making use of geometric means, we also experiment with combining these measures using their direct multiplication, leading to improved performance, but not consistently (see Appendix D).

4.3 Summary & discussion

In this chapter we investigate how unimodal and multimodal models with contrastive learning represent the hypernymy relation in contrast to other lexical relations. Multimodal models underperform unimodal models on the hypernymy detection task. GTE, a general-purpose text embedding model, performs the best, potentially due to its carefully curated training objectives and data. The performance gap between contrastive VLMs and other sentence transformers is relatively small.

We also examine model performance on distinguishing hypernymy against each distractor relation type, reason about how concepts are represented in the text encoders of contrastive VLMs, and propose multiple hypotheses attempting to explain model performance and what type of information is useful for hypernymy detection. We argue that text encoders of contrastive VLMs hold an inherent advantage for distinguishing coordination pairs, as is reflected in the generated images using the synthetic concepts "q, a type of p" for prompting. We also advocate the combination of multiple measures for more robust hypernymy detection, and experiment with the multiplication of current measures (Appendix D).

Initialized from scratch and pre-trained with cross-modal but not unimodal mapping, contrastive VLMs learn limited information from distributional patterns in text despite the large amount of captions in their training data, which can be a disadvantage. However, previous work suggests that textual distributional information does not necessarily benefit contrastive VLMs. Recall the visual + language and visual + word CLIP variants of Zhuang, Fedorenko, and Andreas [ZFA23], where ablating distributional information with single-word captions improves word learning. More investigation is required towards better integration of the complementary information provided by both modalities.

distributional information However, the typical way to combine them, implemented as Visual + Language models, fails to show benefits over Language-Only models. The Visual + Language (CLIP) models perform significantly worse than the Visual + Word (CLIP) models, indicating that the CLIP architecture is particularly inefficient in associating visual information to single words when full captions are present.

Although SimCSE-like unimodal contrastive learning is reported to enhance CLIP on image-text retrieval [Zha+23b], it is not yet clear whether it enhances the text encoder on representation learning. It is thus desirable to conduct ablation studies comparing MCSE with contrastive VLMs in future

work to identify the impact of initialization from a pre-trained LM, text-only contrastive learning, and the amount of multimodal training data on dual-encoder architectures with cross-modal contrastive learning. Meanwhile, the fact that MCSE underperforms SimCSE suggests that additional CLIP-like contrastive learning does not necessarily benefit sentence embedding models [BF25].

Overall, both unimodal and multimodal contrastive learning prove useful for hypernymy detection. Multimodal models underperform unimodal ones on general hypernymy detection, but outperform them on the hypernymy-coordination distinction in particular. Further analysis suggests that visual information is beneficial for concept understanding, while more research is required to explore how to better integrate the complementary information from the two modalities.

Chapter 5

Experiment 2: graded lexical entailment (GLE)

In this chapter, we evaluate the models on graded lexical entailment (GLE) [Vul+17] to see how their concept understanding correlates with the gradual hypernymy judgment in human cognition, and compare model performance on concrete and abstract words.

5.1 Experimental setup

Dataset & evaluation Motivated by discussions on typicality [Ros75] as well as graded membership [KP95], Vulić, Gerz, Kiela, Hill, and Korhonen [Vul+17] constructed a novel dataset, HyperLex, for GLE, aggregating human judgments on *to what degree is q a type of p* on a 0-6 rating scale. HyperLex focuses on single words and covers different lexical relations in WordNet [Fel00]: apart from hypernymy. coordination, and random pairs, it also includes synonymy, antonymy, and reversed hypernymy pairs. We experiment on its noun-noun subset containing 2163 pairs, 1003 of which are hypernymy pairs according to WordNet. Models are evaluated using Spearman's ρ correlation [Spe61].

Concreteness groups A major source of concepts in HyperLex is the USF norms [NMS04], which provides concreteness scores ranging from 1 to 7. As suggested by Vulić, Gerz, Kiela, Hill, and Korhonen [Vul+17], we classify words with available concreteness scores as concrete or abstract using 4 as a threshold, and obtain four subsets: qc pc (concrete q, concrete p); qc pa (concrete q, abstract p); qa pc (abstract q, concrete p); and qa pa (abstract q, abstract p). We conduct evaluation on each subset and compare model performance across concreteness groups.

5.2 Results & analysis

We report Spearman's ρ in Table 5.1. Over the full dataset, GTE-base achieves the highest correlation score (0.4967) with r_q. Bisides GTE, CLIP-l, ALIGN and MPNet with r_q also surpass the taxonomic verification baseline, an approach known to be competitive on GLE [RDG23][1], demonstrating the effectiveness of our methodology. Similarity-based measures making use of the synthetic concept "q, a

[1]Renner, Denis, and Gilleron [RDG23] report a correlation score of 0.425 achieved by GPT2-XL on the union of noun and verb datasets, whereas we experiment with BERT and GPT2 on the noun dataset only.

	#	all 2163	qc pc 172	qc pa 89	qa pc 119	qa pa 1055
b	CLIP-b	0.1536	0.0946	0.3609	0.4415	$-0.0563^{\dagger}$
	CLIP-l	0.1740	0.1262	0.3307	0.5128	$-0.0159^{\dagger}$
	ALIGN	0.1874	0.1346	0.3842	0.4852	$-0.0415^{\dagger}$
	LiT	0.1342	0.0989	0.3155	0.4021	$0.0038^{\dagger}$
	MCSE	0.1625	0.0978	0.3923	0.3855	$0.0707^{\dagger}$
	SimCSE	0.1946	0.1174	0.4475	0.3233	0.1999
	GTE-small	0.2460	0.1870	0.3834	0.5249	0.2335
	GTE-base	0.3020	0.2654	0.4324	0.5739	0.2351
	MiniLM	0.2209	0.1830	0.3876	0.4398	$0.0302^{\dagger}$
	MPNet	0.2307	0.2019	0.4460	0.4325	$0.1238^{\dagger}$
	DistilRoBERTa	0.1843	0.1474	0.3879	0.5148	$0.0794^{\dagger}$
	PhraseBERT	0.1575	0.0973	0.2980	0.5485	$0.0305^{\dagger}$
s_q	CLIP-b	0.3024	0.3445	0.2198^{*}	0.4528	$0.0633^{\dagger}$
	CLIP-l	0.3583	0.3954	$0.1391^{\dagger}$	0.5413	$0.1286^{\dagger}$
	ALIGN	0.3265	0.3697	$0.1285^{\dagger}$	0.5748	$0.0600^{\dagger}$
	LiT	0.2130	0.2751	$-0.0458^{\dagger}$	0.2901	$0.1259^{\dagger}$
	MCSE	0.2717	0.2709	$0.1077^{\dagger}$	0.5408	0.2282
	SimCSE	0.3702	0.3517	0.3062	0.5954	**0.3189**
	GTE-small	0.4106	0.4532	0.2315^{*}	0.4719	0.3002
	GTE-base	0.4039	0.4575	0.2064^{*}	0.4735	0.2531
	MiniLM	0.4167	0.4424	$0.1308^{\dagger}$	0.5953	0.1956^{*}
	MPNet	0.4267	0.4891	$0.1689^{\dagger}$	0.5845	0.1865^{*}
	DistilRoBERTa	0.3243	0.3530	$0.1704^{\dagger}$	0.6176	$0.0038^{\dagger}$
	PhraseBERT	0.2569	0.2112	0.2900	0.5627	0.2038
r_q	CLIP-b	0.4073	0.4355	0.3417	0.5004	$0.0985^{\dagger}$
	CLIP-l	0.3352	0.3325	0.3774	0.5907	$0.1049^{\dagger}$
	ALIGN	0.4095	0.4307	0.3457	**0.6561**	$0.0904^{\dagger}$
	LiT	0.3072	0.4009	$0.0932^{\dagger}$	0.4144	$-0.0280^{\dagger}$
	MCSE	0.2553	0.2943	0.4215^{*}	0.3483	$0.1391^{\dagger}$
	SimCSE	0.3394	0.2982	0.4366	0.5139	0.2809
	GTE-small	0.4150	0.3986	0.4359	0.6219	0.2641
	GTE-base	**0.4967**	**0.5092**	**0.5225**	0.6488	0.2949
	MiniLM	0.3619	0.3580	0.3836	0.5339	$0.1159^{\dagger}$
	MPNet	0.4179	0.4152	0.4914	0.5571	0.2866
	DistilRoBERTa	0.2856	0.2592	0.3367	0.5817	0.1671^{*}
	PhraseBERT	0.1602	0.0703^{*}	0.3134	0.5402	$0.0988^{\dagger}$
veri	BERT	0.3829	0.4262	0.5024	0.4372	$0.0331^{\dagger}$
	GPT2	0.3536	0.3326	0.4039	0.5842	$0.0964^{\dagger}$

Table 5.1: Model performance on graded lexical entailment.
For each model, we report Spearman's rank correlation between its similarity-based measures and the HyperLex GLE ratings. Boldface highlights the best score in each column.
Statistical significance is assessed using a two-sided permutation test with 10,000 permutations. †denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

type of p" not only exhibit qualitatively different behaviour across different lexical relations (Fig 4.1), but also capture their gradual nature quantitatively.

For instance, the word pairs (rabbit, food) and (scallop, animal), despite being labeled as hypernymy pairs in WordNet, have relatively low GLE ratings. Although "rabbit" is considered a type of food in many cultures, it is more typically considered an animal in a decontextualized setting. In accordance with this fact, generated images with CLIP-l encoding using Stable Diffusion depict live animals when prompted with the isolated word. When prompted with the synthetic concept "rabbit, a type of food", generated images still depict a live rabbit or a rabbit alongside some food instead of depicting rabbit as food per se, likely also due to training data bias. On the other hand, the depiction of "scallop" resembles food rather than live animals, whereas "scallop, a type of animal" resembles sea animals. The image-caption training data provide information about how concepts are typically depicted, allowing contrastive VLMs to learn the corresponding human-like representations, but they also introduce systematic bias [Ham+24] as in the case of unimodal models.

We further analyze model performance across concreteness groups. Results show that GTE-base with r_q is the best-performing combination only when q is concrete. On the "qa pc" subset containing (chemistry, science) and (headache, pain), the contrastive VLM ALIGN performs the best. The case where both q and p are abstract is particularly challenging for all models: while SimCSE maintains a correlation score of 0.3189, correlation scores obtained by contrastive VLMs do not exceed 0.1286 and are not statistically significant (p-value ≥ 0.05), in accordance with the previous statement that abstract words are particularly challenging for multimodal models.

Meanwhile, we argue that previous critique against word similarity evaluation [Far+16] such as subjectivity can extend to graded lexical entailment, especially for abstract words which can be highly ambiguous. To further explore the performance of contrastive VLMs on abstract words, we perform another experiment in Chapter 6, focusing on binary hypernymy detection with coordination as the only distractor relation, i.e. hypernymy versus coordination discrimination.

"rabbit"

"rabbit, a type of food"

"scallop"

"scallop, a type of animal"

Figure 5.1: Generated images for concepts based on (rabbit, food) and (scallop, animal).

Chapter 6

Experiment 3: concreteness & specificity

While abstract words have long been recognized as a challenge to multimodal models, generic words (e.g. bird, furniture) are also grounded less directly via their more specific hyponyms (e.g. robin, rocking chair). In this chapter, we evaluate unimodal and multimodal models with contrastive learning on hypernymy versus coordination discrimination, comparing their performance on concrete versus abstract, as well as specific versus generic word pairs, in order to investigate whether abstractness and genericity are significantly more challenging for multimodal models than unimodal models.

6.1 Experimental setup

Material We adopt concreteness ratings provided by Brysbaert, Warriner, and Kuperman [BWK14], and the corresponding specificity scores computed by Bolognesi, Burgers, and Caselli [BBC20]: for 13518 of the 15030 nouns with available concreteness ratings, they first successfully retrieve a WordNet [Fel00] synset, always using the first sense in the case of polysemy, and then compute the total amount of its direct and indirect hypernyms in the WordNet hierarchy[1]. Both scores range from 1 (most abstract/generic) to 5 (most concrete/specific).

Sampling Following Bolognesi, Burgers, and Caselli [BBC20], we treat each noun q with concreteness and specificity ratings as a WordNet synset. We first randomly sample one of its direct hypernym synsets to form a hypernymy pair (q, h) with the canonical lemma h; then we randomly sample one of the direct hyponym synsets of h that is distinct from q, and obtain a coordination pair (q, c) with the canonical lemma c. [2] (q, h) is discarded if q is the only direct hyponym of h, i.e. if such a coordination pair (q, c) is not available. Such sampling results in 12343 pairs for each relation.

[1] Bolognesi, Burgers, and Caselli [BBC20] provide 3 different specificity measures. Here we use Specificity 3 only, following their Study 3.

[2] For instance, for the word "cat", we retrieve the synset `cat.n.01`, whose direct hypernym is the synset `feline.n.01`. This hypernym synset has two lemmas, namely `feline.n.01.feline` and `feline.n.01.felid`. We use the name of its first lemma, "feline", to form a hypernymy pair (cat, feline). `feline.n.01` has another hyponym `big_cat.n.01`, whose lemmas are `big_cat.n.01.big_cat` and `big_cat.n.01.cat`, so we sample a corresponding coordination pair (cat, big cat).

Dataset In this experiment, we focus on single words following BLESS [BL11] and HyperLex [Vul+17], and conduct evaluation on a subset of the sampled pairs excluding multiword expressions (MWEs). [3] For each word q with a hypernymy pair (q, h) and a coordination (q, c), the two pairs are included in our evaluation if and only if q, h and c are all single words, resulting in 7907 hypernymy pairs and 7907 coordination pairs. We refer to our constructed dataset as $\text{BBC}_{\text{single}}$, named after Bolognesi, Burgers, and Caselli [BBC20]. For each pair (q, p) where p is either a hypernym or a co-hyponym, concreteness and specificity scores are available for q. Given the close distance between q and p in the WordNet hierarchy as guaranteed by our sampling methodology, we believe these scores also work as the concreteness and specificity scores for pair as a whole. Like in Experiment 2, we divide the dataset into concreteness and specificity groups for performance comparison: following Bolognesi, Burgers, and Caselli [BBC20], word pairs fall into 4 quadrants, i.e. conc+ spec+, conc+ spec-, conc- spec+, and conc- spec-, with 3 as a threshold for both dimensions.

Evaluation We adopt two evaluation metrics: The first one is AP; the second one is accuracy, computed via examining whether the score of (q, h) is higher than that of (q, c) for each q, which is actually in analogy to MAP in Experiment 1. Note that due to the balanced number of positive (hypernymy) and negative (coordination) pairs in the dataset, the AP and accuracy (Acc) of a random baseline are both 0.5. We also compute the Spearman's correlation between the concreteness/specificity scores and hypernymy measures produced by models. Additiional boxplots are provided in Appendix B.2.

6.2 Results & analysis

6.2.1 Overall performance

Table 6.1 reports the models' overall performance on hypernymy versus coordination discrimination as well as correlation with concreteness and specificity scores. We first observe that r_q with GTE-base is the best-performing combination under both metrics, in contrast to previous results where s_q with LiT performs the best on the same task using the BLESS dataset. Also, contrastive VLMs tend to underperform sentence transformers with more significant performance gaps. There exist two potential reasons behind this conflict: 1) the target concepts (qs) in BLESS are all unambiguous basic-level concepts such as "robin", "banana" and "sweater" [BL11], whereas the source of target concepts for BBC, namely the Brysbaert, Warriner, and Kuperman [BWK14] concreteness database, aims to include as many entries as possible, thus containing more sophisticated concepts; 2) hypernymy in BBC are limited to direct hyponym-hypernym WordNet pairs, which can be less intuitive for humans [Vul+17]. Some pairs in BBC, such as (kamikaze, fighter), (lactose, disaccharide) and (flathead, scorpaenoid), can be difficult even for humans to process while being well-represented by unimodal models. While we experiment on BBC to investigate the impact of concreteness and specificity on model performance, we admit that it is not an ideal dataset for evaluating human-like concept understanding.

Interestingly, in this experiment MCSE outperforms both SimCSE under both evaluation metrics, contrary to results in our previous experiments as well as the findings of Bavaresco and Fernández [BF25], which calls for more extensive research to compare the performance of SimCSE and MCSE and explore the effect of cross-modal contrastive learning on unimodal sentence embedding models.

[3] For completeness, we also perform an additional experiment using all the sampled pairs including multiword expressions (BBC_{full}), reported in Appendix C for completeness. Results are similar.

measure	model	AP	Acc	$\text{corr}_{\text{conc}}$	$\text{corr}_{\text{spec}}$
b	CLIP-b	0.6094	0.7023	−0.4218	−0.2078
	CLIP-l	0.6275	0.7234	−0.3349	−0.1640
	ALIGN	0.6364	0.7161	−0.0334	−0.0389
	LiT	0.5627	0.6493	−0.4316	−0.1686
	MCSE	0.6259	0.6698	−0.0298	0.0489
	SimCSE	0.6328	0.6670	−0.0290	0.0324
	GTE-small	0.6588	0.7247	−0.0171*	0.0185*
	GTE-base	0.6721	0.7481	−0.0235	0.0679
	MiniLM	0.5482	0.7146	0.0877	0.1160
	MPNet	0.6337	0.7044	0.0175*	0.1855
	DistilRoBERTa	0.5927	0.6423	−0.0232	0.1236
	PhraseBERT	0.5768	0.6221	−0.0068†	0.1636
s_q	CLIP-b	0.6180	0.7290	−0.4492	−0.2165
	CLIP-l	0.6342	**0.7808**	−0.3861	−0.1795
	ALIGN	0.6146	0.7208	−0.1446	−0.1109
	LiT	0.5385	0.5785	−0.3273	−0.1787
	MCSE	0.6246	0.7172	0.0197*	0.0561
	SimCSE	0.6219	0.6748	0.0762	−0.0193*
	GTE-small	0.6877	0.7815	−0.0585	0.0363
	GTE-base	0.6845	0.7936	−0.0707	0.0333
	MiniLM	0.6815	0.7735	−0.0380	0.0626
	MPNet	0.6506	0.7582	0.0614	0.1725
	DistilRoBERTa	0.6166	0.6876	0.0889	0.1422
	PhraseBERT	0.5857	0.6632	−0.1372	0.0095†
r_q	CLIP-b	0.6398	0.7156	0.0214	−0.0213
	CLIP-l	**0.6601**	0.7498	−0.0278	−0.0641
	ALIGN	0.6584	0.7092	0.1923	0.0404
	LiT	0.5638	0.6312	0.1864	0.0143†
	MCSE	0.6575	0.7020	0.0121†	0.0532
	SimCSE	0.6546	0.6961	0.0584	0.0036†
	GTE-small	0.7432	0.8014	−0.0255	−0.0381
	GTE-base	**0.7619**	**0.8246**	−0.0166*	0.0070†
	MiniLM	0.7146	0.7662	0.0684	0.0411
	MPNet	0.7228	0.7899	0.0000	0.1038†
	DistilRoBERTa	0.6307	0.6860	−0.0456	0.0663
	PhraseBERT	0.5828	0.6394	0.0824	0.1503
veri	BERT	0.5433	0.6231	0.0358	0.1310
	GPT2	0.7607	0.8183	0.0001†	0.0379

Table 6.1: Model performance on hypernymy versus coordination discrimination. MWEs excluded. For each model, we report AP and Accuracy scores evaluating the ability to distinguish hypernymy from coordination pairs, based on the measures introduced in Section 3.2. In each column, boldface highlights the best scores obtained by contrastive VLMs and sentence transformers respectively. Statistical significance tests (one-sided Welch's t-test) are conducted over the full dataset, and all differences are statistically significant ($p < 0.01$).

We also report Spearman's rank correlation with concreteness and specificity scores. Statistical significance is assessed using a two-sided permutation test with 10,000 permutations. †denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

6.2.2 Sensitivity to concreteness/specificity

Table 6.1 also demonstrates that s_q with contrastive VLMs CLIP and LiT are particularly sensitive to word concreteness. Their negative correlations reflect that abstract word pairs tend to have higher s_q scores and might therefore be more easily mistaken as hypernymy pairs, as can be observed in boxplots as well (Appendix B.2). In terms of specificity, s_q with contrastive VLMs also exhibit negative correlations, which are less drastic than in the case of concreteness. Meanwhile, sentence transformers MPNet and DistilRoBERTa are also sensitive to specificity, although with a positive correlation: they tend to produce lower s_q scores for more generic pairs. The other measure r_q is less sensitive to concreteness and specificity and yields a higher AP, demonstrating robustness. However, note that s_q with CLIP-l outperforms r_p in terms of accuracy: while this measure tend to produce higher scores, which is a disadvantage when evaluated via AP, it is actually better at distinguishing the hypernymy pair (q, h) from the coordination pair (q, c) in general, and can be considered as better-performing as the tendency to produce higher scores for abstract pairs is factored out in the accuracy metric. [4]

Table 6.2 reports model performance on BBC subsets with high/low concreteness or specificity, respectively, factoring out the confounding effect of comparing across concreteness levels. From the high specificity (spec+) subset to the low specificity (spec-, high genericity) subset, we observe a performance increase in terms of AP and a performance drop in terms of accuracy. This conflict further indicates that evaluation of the understanding of concrete versus abstract or specific versus generic words needs to be handled with caution. We observe a performance drop from the high concreteness (conc+) to the low concreteness (conc-, high abstractness) subset in terms of both AP and accuracy for both unimodal and multimodal models, in accordance with findings of Liao, Chen, and Du [LCD23]. Unlike previous results on GLE (Chapter 5), the comparison of performance drop between contrastive VLMs and sentence transformers is not drastic. Given our former discussions on the subjectivity of the GLE task and caution about evaluation method, we argue from our current findings that abstractness is not necessarily significantly more challenging for contrastive VLMs than for sentence transformers, and that further investigation is required for reaching a definite conclusion.

6.3 Summary & discussion

In this chapter, we construct a novel dataset BBC containing hypernymy pairs and adversarial coordination pairs, with concreteness and specificity scores for the target concept q. As word pairs are sampled from direct hypernymy pairs in WordNet, with a known general misalignment from human intuition [Cao+24], this dataset does not aim to reflect human concept understanding and is only intended for examining model performance on words of different levels of concreteness and specificity.

Word abstractness has long been considered a challenge to multimodal models. In this experiment, we do observe that the s_q measure with contrastive VLMs is more sensitive to word concreteness, but the performance of contrastive VLMs does not decrease drastically on abstract word pairs as previously assumed. Our results demonstrate that abstractness is not significantly more challenging for contrastive VLMs than for sentence transformers. Meanwhile, the impact of specificity is less clear. Overall, we highlight the need to explore more systematic evaluation protocols in order to investigate the role of word concreteness and specificity on model behaviour.

[4] Actually, the authors of BLESS suggested factoring out concept-specific effects during evaluation [BL11], which is why we report MAP alongside AP.

		conc+		conc-		spec+		spec-	
		AP	Acc	AP	Acc	AP	Acc	AP	Acc
	#	5021		2886		194		7713	
b	CLIP-b	0.6329	0.7216	0.5990	0.6687	0.6092	0.6856	0.6314	0.7027
	CLIP-l	0.6485	0.7411	0.6153	0.6927	0.6272	0.6907	0.6402	0.7242
	ALIGN	0.6564	0.7391	0.6016	0.6760	0.6362	0.6804	0.6568	0.7170
	LiT	0.5999	0.6909	0.5350	0.5769	0.5620	0.6598	0.5929	0.6490
	MCSE	0.6289	0.6734	0.6217	0.6635	0.6253	0.6804	0.6536	0.6695
	SimCSE	0.6391	0.6772	0.6223	0.6493	0.6322	0.6804	0.6579	0.6667
	GTE-small	0.6684	0.7375	0.6416	0.7024	0.6579	0.7165	0.6871	0.7249
	GTE-base	0.6791	0.7624	0.6603	0.7231	0.6718	0.7474	0.6877	0.7481
	MiniLM	0.6548	0.7256	0.6382	0.6954	0.6467	0.7835	0.7076	0.7128
	MPNet	0.6367	0.7060	0.6290	0.7017	0.6328	0.7062	0.6730	0.7044
	DistilRoBERTa	0.5983	0.6487	0.5830	0.6313	0.5924	0.6649	0.6102	0.6418
	PhraseBERT	0.5680	0.6023	0.5926	0.6566	0.5772	0.5979	0.5733	0.6227
s_q	CLIP-b	0.6537	0.7431	0.6086	0.7044	0.6173	0.7320	0.6719	0.7289
	CLIP-l	**0.6898**	**0.8036**	0.6133	**0.7412**	0.6337	0.7526	0.6691	**0.7815**
	ALIGN	0.6369	0.7411	0.5868	0.6854	0.6140	**0.7835**	0.6412	0.7192
	LiT	0.5458	0.5851	0.5362	0.5669	0.5389†	0.5567	0.5231	0.5790
	MCSE	0.6293	0.7327	0.6166	0.6902	0.6246	0.7113	0.6337	0.7174
	SimCSE	0.6361	0.6865	0.5983	0.6545	0.6225	0.5876	0.6009	0.6770
	GTE-small	0.7105	0.8012	0.6520	0.7471	0.6876	0.7784	0.6957	0.7815
	GTE-base	0.7007	0.8084	0.6626	0.7678	0.6835	**0.8505**	0.7333	0.7922
	MiniLM	0.7021	0.7827	0.6512	0.7574	0.6812	0.7784	0.6994	0.7734
	MPNet	0.6583	0.7598	0.6382	0.7554	0.6505	0.7371	0.6617	0.7587
	DistilRoBERTa	0.6276	0.6961	0.5990	0.6729	0.6166	0.7010	0.6299	0.6873
	PhraseBERT	0.5867	0.6624	0.5871	0.6646	0.5879†	0.5928	0.5300	0.6650
r_q	CLIP-b	0.6416	0.7267	0.6411	0.6961	0.6400	0.6959	0.6438	0.7161
	CLIP-l	0.6625	0.7564	**0.6579**	0.7384	**0.6602**	0.7680	0.6619	0.7494
	ALIGN	0.6811	0.7367	0.6105	0.6615	0.6567	0.7680	**0.7221**	0.7078
	LiT	0.5738	0.6545	0.5395	0.5908	0.5644*	0.6546	0.5536	0.6306
	MCSE	0.6651	0.7122	0.6442	0.6843	0.6565	0.7526	0.6983	0.7008
	SimCSE	0.6668	0.7088	0.6333	0.6739	0.6544	0.6804	0.6634	0.6965
	GTE-small	0.7575	0.8208	0.7171	0.7678	0.7425	0.8144	0.7626	0.8011
	GTE-base	**0.7767**	**0.8423**	**0.7347**	**0.7938**	**0.7619**	0.8454	0.7661	**0.8241**
	MiniLM	0.7215	0.7710	0.7027	0.7561	0.7133	0.7938	0.7646	0.7655
	MPNet	0.7292	0.7911	0.7129	0.7879	0.7213	0.8144	**0.7765**	0.7893
	DistilRoBERTa	0.6428	0.6993	0.6117	0.6629	0.6292	0.7680	0.6897	0.6839
	PhraseBERT	0.5687	0.6190	0.6117	0.6750	0.5836*	0.5979	0.5740	0.6405
veri	BERT	0.5482	0.6280	0.5355	0.6147	0.5447	0.5773	0.5076	0.6243
	GPT2	0.7642	0.8192	0.7548	0.8167	0.7610	0.8247	0.7547	0.8181

Table 6.2: Model performance on the high/low concreteness/specificity subsets. MWEs excluded. In each column, boldface highlights the best scores obtained by contrastive VLMs and sentence transformers respectively.
Statistical significance tests (one-sided Welch's t-test) are conducted for each subset. †alongside APs denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

		conc+ spec+		conc+ spec-		conc- spec+		conc- spec-	
		AP	Acc	AP	Acc	AP	Acc	AP	Acc
	#	179		4842		15		2871	
b	CLIP-b	0.6500	0.6927	0.6324	0.7226	0.5594†	0.6000	0.5997	0.6691
	CLIP-l	0.6482	0.6872	0.6484	0.7431	0.6512†	0.7333	0.6156	0.6924
	ALIGN	0.6648	0.6816	0.6566	0.7412	0.5735†	0.6667	0.6021	0.6761
	LiT	0.6040	0.6704	0.5999	0.6917	0.5414†	0.5333	0.5350	0.5772
	MCSE	0.6638	0.6872	0.6279	0.6729	0.6299†	0.6000	0.6221	0.6639
	SimCSE	0.6772	0.6927	0.6377	0.6766	0.5534†	0.5333	0.6232	0.6499
	GTE-small	0.6973	0.7318	0.6669	0.7377	0.5896†	0.5333	0.6433	0.7032
	GTE-base	0.7066	0.7654	0.6781	0.7623	0.5387†	0.5333	0.6617	0.7241
	MiniLM	0.7209	0.7933	0.6523	0.7230	0.5822†	0.6667	0.6391	0.6956
	MPNet	0.6834	0.7207	0.6349	0.7055	0.5925†	0.5333	0.6297	0.7025
	DistilRoBERTa	0.6143	0.6536	0.5978	0.6485	0.6236†	0.8000	0.5833	0.6304
	PhraseBERT	0.5895	0.6089	0.5674	0.6020	0.4900†	0.4667	0.5942	0.6576
s_q	CLIP-b	0.6790	0.7263	0.6531	0.7437	0.7099†	0.8000	0.6082	0.7039
	CLIP-l	0.6765	0.7598	**0.6904**	**0.8052**	0.6930†	0.6667	0.6127	**0.7416**
	ALIGN	0.6410	**0.7877**	0.6372	0.7394	0.6756†	0.7333	0.5863	0.6851
	LiT	0.5371	0.5810	0.5463	0.5853	0.4518†	0.2667	0.5367	0.5684
	MCSE	0.6301	0.7039	0.6297	0.7338	0.7122†	0.8000	0.6161	0.6897
	SimCSE	0.5931	0.5810	0.6378	0.6904	0.6882†	0.6667	0.5975	0.6545
	GTE-small	0.7087	0.7765	0.7108	0.8021	0.6759†	0.8000	0.6520	0.7468
	GTE-base	0.7500	0.8547	0.6992	0.8067	0.6588†	0.8000	0.6628	0.7677
	MiniLM	0.7017	0.7709	0.7030	0.7831	0.7229†	**0.8667**	0.6508	0.7569
	MPNet	0.6645	0.7263	0.6583	0.7610	0.6984†	**0.8667**	0.6383	0.7548
	DistilRoBERTa	0.6249	0.6872	0.6286	0.6964	0.6758†	**0.8667**	0.5985	0.6719
	PhraseBERT	0.5230*	0.5810	0.5905	0.6654	0.6079†	0.7333	0.5873	0.6642
r_q	CLIP-b	0.6507	0.7095	0.6413	0.7274	0.5884†	0.5333	0.6420	0.6970
	CLIP-l	0.6654	0.7598	0.6625	0.7563	**0.7269**†	**0.8667**	**0.6582**	0.7377
	ALIGN	**0.7355**	**0.7877**	0.6790	0.7348	0.5779†	0.5333	0.6112	0.6621
	LiT	0.5618	0.6816	0.5746	0.6534	0.4522†	0.3333	0.5403	0.5921
	MCSE	0.7026	0.7654	0.6638	0.7102	0.7152†	0.6000	0.6442	0.6848
	SimCSE	0.6710	0.6872	0.6669	0.7096	0.6839†	0.6000	0.6334	0.6743
	GTE-small	0.7717	0.8324	0.7568	0.8203	0.6953*	0.6000	0.7177	0.7687
	GTE-base	**0.7816**	**0.8547**	**0.7767**	**0.8418**	0.6614*	0.7333	**0.7358**	**0.7941**
	MiniLM	0.7705	0.7989	0.7196	0.7710	0.7604*	0.7333	0.7030	0.7562
	MPNet	0.7810	0.8156	0.7270	0.7902	**0.7732***	0.8000	0.7128	0.7879
	DistilRoBERTa	0.6853	0.7654	0.6413	0.6968	0.7353*	0.8000	0.6110	0.6621
	PhraseBERT	0.5806	0.6034	0.5689	0.6196	0.5202†	0.5333	0.6126	0.6757
veri	BERT	0.5121	0.5978	0.5502	0.6291	0.4931†	0.3333	0.5362	0.6162
	GPT2	0.7581	0.8380	0.7647	0.8185	0.7210*	0.6667	0.7550	0.8175

Table 6.3: Model performance on subsets w.r.t. both concreteness and specificity. MWEs excluded. In each column, boldface highlights the best scores obtained by contrastive VLMs and sentence transformers respectively.
Statistical significance tests (one-sided Welch's t-test) are conducted for each subset. †alongside APs denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

Chapter 7

Conclusion & discussion

7.1 Conclusion

This paper investigates the concept understanding of text encoders of contrastive VLMs alongside unimodal models with contrastive learning. We focus on hypernymy, a crucial lexical relation in human semantic memory, and conduct evaluation on both binary hypernymy detection and graded lexical entailment with a novel methodology based on synthetic concepts.

RQ1 *How well do contrastive VLMs capture hypernymy compared to unimodal models with contrastive learning? Do they exhibit certain advantages due to their visual grounding?*
In Experiment 1 (Chapter 4) and Experiment 2 (Chapter 5), we find the best-performing model to be GTE [Li+23b], a state-of-the-art unimodal sentence embedding model, potentially due to its carefully curated training objectives and data. Although contrastive VLMs are outperformed by other unimodal models as well, their performance gap is relatively small. In one particular scenario, namely hypernymy versus coordination discrimination (Section 4.2.2), the contrastive VLM LiT [Zha+22a] performs above all other models by a large margin. We hypothesize that contrastive VLMs hold an inherent advantage in this task due to their visual grounding, with illustrations via text-to-image generation.

RQ2 *Do contrastive VLMs perform worse on abstract or generic words?*
In Experiment 2 (Chapter 5) and Experiment 3 (Chapter 6), we compare model performance on subsets containing concrete and abstract word pairs. On the GLE task, the performance of contrastive VLMs deteriorates significantly on abstract words, while sentence transformers also suffer a degradation in performance. We argue that GLE, like word similarity evaluation, may suffer from problems such as subjectivity [Far+16], especially for abstract word pairs which are highly ambiguous. On the hypernymy versus coordination discrimination task, both demonstrate a comparable decline. While contrastive VLMs exhibit a tendency to produce higher scores for more abstract pairs leading to potential confusion, we argue that with such concept-specific effects factored out in evaluation metrics, abstractness is not necessarily significantly more challenging for text encoders of contrastive VLMs than for sentence transformers. The impact of specificity on both types of models is less clear. We argue for the importance of exploring more systematic evaluation protocols in order to investigate the impact of word concreteness and specificity on model performance.

Our contributions are as follows:

- While previous research on the text encoders of contrastive VLMs mainly focus on comparing CLIP with BERT [Che+23] or GPT2 [WC22b], we perform a direct assessment against sentence transformers with unimodal contrastive learning to examine the effect of visual-semantic training;
- Inspired by the work of Regneri, Abdelhalim, and Laue [RAL24], we propose a novel methodology based on synthetic concepts ("q, a type of p"), defining similarity-based measures to reflect the hypernymy relation, and achieve competitive performance;
- We evaluate 5 multimodal and 7 unimodal models on binary hypernymy detection and graded lexical entailment, and qualitatively analyze how contrastive VLMs represent word pairs of different lexical relations due to their training data;
- We construct a novel dataset, BBC, to investigate model performance on words of different levels of concreteness and specificity, and find contrastive VLMs to be more robust than previously expected.

7.2 Discussion

In this work, we evaluate the text encoders of 3 contrastive VLM models, CLIP [Rad+21], ALIGN [Jia+21], and LiT [Zha+22a], all trained from scratch with an image-caption matching objective. It is possible that initialization from a pre-trained language model or an additional unimodal contrastive learning objective would enhance their linguistic competency, While SimCSE[GYC21]-like unimodal contrastive learning is reported to enhance CLIP on image-text retrieval [Zha+23b], it is not yet clear whether it benefits the text encoder on representation learning.

MCSE [Zha+22b] is an intriguing model for comparison as it has a similar dual-encoder architecture, but is initialized with BERT and is jointly trained on a CLIP-like objective and a SimCSE-like unimodal contrastive objective. Also, note that the amount of multimodal training data for MCSE is significantly smaller than that of contrastive VLMs. In our experiments on BLESS and HyperLex, MCSE is outperformed by both SimCSE and CLIP, discouraging the integration of vision-language contrastive learning and unimodal training objectives. More direct ablation studies are currently lacking.

According to Zhuang, Fedorenko, and Andreas [ZFA23], replacing complete captions in the image-caption training data with single words improves CLIP variants on word learning. They suggest that full captions actually hinders the CLIP architecture from integrating the visual information for single word learning. We thus argue that the visual-semantic concept understanding of contrastive VLMs might differ substantially from word embeddings learnt from distributional patterns in text corpora. While their text encoders achieve remarkable performance in our experiments as well as previous work [PTF21], more exploration is required on the integration of contrastive VLMs and the textual distributional information that empower language models.

7.3 Limitation & future work

In this work we aim to evaluate contrastive VLMs on human concept understanding, and perform experiments on binary hypernymy detection as well as graded lexical entailment. However, our

investigations are restricted to single-word nouns and do not address polysemy, which is highlighted by [Ren+23]: BLESS, the dataset for our first experiment, focuses unambiguous words; HyperLex, the dataset for our second experiment, contains polysemous words, but we do not look into model performance in the case of polysemy; during the construction of the dataset for our third experiment, we link a word to its first WordNet sense by default, overlooking other potential senses.

We report results of Experiment 3 on the complete samples in Appendix C, where multiword expressions (MWEs) are not excluded. In future work, our work can be extended to other parts of speech such as verbs by modifying the templates for generating synthetic concepts, and to other semantic phenomenon, e.g. the lexical relation meronymy [PL25] via "handle, part of some door" and "door, with some handle"; the idiomaticity of MWEs via computing the similarities between "beaver", "eager beaver", and "eager beaver, a type of beaver"; and polysemy, as the expression "q, a type of p" provides minimal context for disambiguation, as can be observed in the (scallop, animal) example (Fig 5.1). Moreover, our synthetic concept "q, a type of p" can be associated to topics such as categorical knowledge editing [PGH24], e.g. if a cobra is a type of dog, then it also barks. It is also possible to produce additional concept representations via image synthesis as in [HLY23; Liu+22].

In Experiment 1, our hypernymy-specialized measures underperform the naïve cosine similarity baseline in hypernymy versus random pair discrimination scenario, and we advocate the combination of multiple measures for a more robust representation of hypernymy. We experiment with the multiplication of current measures, namely $b \cdot s_q$, $b \cdot r_q$, and $r_q \cdot s_q$, and report results in Appendix D. These combinations sometimes yield better performance, but not consistently. In future work, we expect more powerful combination approaches, e.g. via the random forest algorithm [San+16a], to further improve performance, or even combine different models to better integrate information from different modalities.

Bibliography

[And+15] Andrew James Anderson, Elia Bruni, Alessandro Lopopolo, Massimo Poesio, and Marco Baroni. "Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text". In: *NeuroImage* 120 (2015), pages 309–322 (cited on page 7).

[And+18] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-up and top-down attention for image captioning and visual question answering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pages 6077–6086 (cited on page 9).

[AVV09] Mark Andrews, Gabriella Vigliocco, and David Vinson. "Integrating experiential and distributional data to learn semantic representations." In: *Psychological review* 116.3 (2009), page 463 (cited on page 3).

[Bai+21] Yuhang Bai, Richong Zhang, Fanshuang Kong, Junfan Chen, and Yongyi Mao. "Hypernym discovery via a recurrent mapping model". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021, pages 2912–2921 (cited on page 13).

[Bar+12] Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. "Entailment above the word level in distributional semantics". In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012, pages 23–32 (cited on pages 11, 12).

[Bar16] Marco Baroni. "Grounding distributional semantics in the visual world". In: *Language and Linguistics Compass* 10.1 (2016), pages 3–13 (cited on page 3).

[Bar99] Lawrence W Barsalou. "Perceptual symbol systems". In: *Behavioral and brain sciences* 22.4 (1999), pages 577–660 (cited on page 8).

[Bav+25] Anna Bavaresco, Marianne de Heer Kloots, Sandro Pezzelle, and Raquel Fernández. *Modelling Multimodal Integration in Human Concept Processing with Vision-Language Models*. 2025 (cited on pages 3, 7, 10).

[BB18] Gabriel Bernier-Colborne and Caroline Barriere. "CRIM at semeval-2018 task 9: A hybrid approach to hypernym discovery". In: *Proceedings of the 12th international workshop on semantic evaluation*. 2018, pages 725–731 (cited on pages 11, 13).

[BBC20] Marianna Bolognesi, Christian Burgers, and Tommaso Caselli. "On abstraction: decoupling conceptual concreteness and categorical specificity". In: *Cognitive Processing* 21 (2020), pages 365–381 (cited on pages 5, 7, 29, 30).

[BCS20] Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. "Inducing relational knowledge from BERT". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 34. 05. 2020, pages 7456–7463 (cited on page 13).

[BDK14] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pages 238–247 (cited on page 13).

[BF25] Anna Bavaresco and Raquel Fernández. "Experiential Semantic Information and Brain Alignment: Are Multimodal Models Better than Language Models?" In: *arXiv preprint arXiv:2504.00942* (2025) (cited on pages 3, 7, 10, 21, 24, 30).

[Bie+22] Romain Bielawski, Benjamin Devillers, Tim Van De Cruys, and Rufin VanRullen. "When does CLIP generalize better than unimodal models? When judging human-centric concepts". In: *7th Workshop on Representation Learning (Repl4NLP 2022)*. ACL: Association for Computational Linguistics. 2022, pages 29–38 (cited on pages 9, 22).

[Bin+05] Jeffrey R Binder, Chris F Westbury, Kristen A McKiernan, Edward T Possing, and David A Medler. "Distinct brain systems for processing concrete and abstract concepts". In: *Journal of cognitive neuroscience* 17.6 (2005), pages 905–917 (cited on page 7).

[Bis+20] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Yue Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph P. Turian. "Experience Grounds Language". In: *ArXiv* abs/2004.10151 (2020) (cited on pages 3, 6, 7).

[BK20] Emily M Bender and Alexander Koller. "Climbing towards NLU: On meaning, form, and understanding in the age of data". In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. 2020, pages 5185–5198 (cited on pages 3, 6, 7).

[BL11] Marco Baroni and Alessandro Lenci. "How we BLESSed distributional semantic evaluation". In: *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. 2011, pages 1–10 (cited on pages 4, 11, 18, 30, 32, 57).

[BR23] Anton Baryshnikov and Max Ryabinin. "Hypernymy Understanding Evaluation of Text-to-Image Models via WordNet Hierarchy". In: *arXiv preprint arXiv:2310.09247* (2023) (cited on page 4).

[Bru+12a] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. "Distributional semantics in technicolor". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2012, pages 136–145 (cited on pages 6, 8).

[Bru+12b] Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. "Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning". In: *Proceedings of the 20th ACM international conference on Multimedia*. 2012, pages 1219–1228 (cited on page 6).

[BTB11] Elia Bruni, Giang Binh Tran, and Marco Baroni. "Distributional semantics from text and images". In: *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*. 2011, pages 22–32 (cited on pages 3, 6, 8).

[BTB14] Elia Bruni, Nam Khanh Tran, and Marco Baroni. "Multimodal Distributional Semantics". In: *J. Artif. Intell. Res.* 49 (2014), pages 1–47 (cited on pages 3, 6).

[BTF24] Anna Bavaresco, Alberto Testoni, and Raquel Fernández. "Don't buy it! Reassessing the ad understanding abilities of contrastive multimodal models". In: *arXiv preprint arXiv:2405.20846* (2024) (cited on page 9).

[BW05] Lawrence W. Barsalou and Katja Wiemer-Hastings. "Situating Abstract Concepts". In: *Grounding Cognition*. Cambridge University Press, Jan. 2005, pages 129–163. ISBN: 9780521168571 (cited on pages 7, 8).

[BWK14] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. "Concreteness ratings for 40 thousand generally known English word lemmas". In: *Behavior Research Methods* 46 (2014), pages 904–911 (cited on pages 7, 8, 29, 30).

[Cam+18] Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. "SemEval-2018 task 9: Hypernym discovery". In: *Proceedings of the 12th international workshop on semantic evaluation*. 2018, pages 712–724 (cited on pages 4, 11, 13).

[Cao+24] Zhihan Cao, Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. "Misalignment of Semantic Relation Knowledge between WordNet and Human Intuition". In: *arXiv preprint arXiv:2412.02138* (2024) (cited on page 32).

[Car+21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pages 9650–9660 (cited on pages 8, 9).

[Cer+17] Daniel Matthew Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation". In: *International Workshop on Semantic Evaluation*. 2017 (cited on pages 3, 9).

[CFP23] Xinyi Chen, Raquel Fernández, and Sandro Pezzelle. "The BLA benchmark: Investigating basic language abilities of pre-trained multimodal models". In: *arXiv preprint arXiv:2310.15061* (2023) (cited on page 9).

[Cha+21] Soravit Changpinyo, Piyush Kumar Sharma, Nan Ding, and Radu Soricut. "Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pages 3557–3567 (cited on page 10).

[Che+20a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PmLR. 2020, pages 1597–1607 (cited on page 8).

[Che+20b] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. "Uniter: Universal image-text representation learning". In: *European conference on computer vision*. Springer. 2020, pages 104–120 (cited on pages 3, 6, 8).

[Che+23] Zhihong Chen, Guiming Hardy Chen, Shizhe Diao, Xiang Wan, and Benyou Wang. "On the Difference of BERT-style and CLIP-style Text Encoders". In: *Annual Meeting of the Association for Computational Linguistics*. 2023 (cited on pages 4, 5, 9, 36).

[Che+24] Zhirui Chen, Andreas Mädebach, Eleonora Gualdoni, and Gemma Boleda. "On the Use of Language and Vision Models for Cognitive Science: The Case of Naming Norms". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Volume 46. 2024 (cited on page 9).

[Cla09] Daoud Clarke. "Context-theoretic semantics for natural language: an overview". In: *Proceedings of the workshop on geometrical models of natural language semantics*. 2009, pages 112–119 (cited on page 12).

[Col81] Max Coltheart. "The MRC psycholinguistic database". In: *The Quarterly Journal of Experimental Psychology Section A* 33.4 (1981), pages 497–505 (cited on page 7).

[Cro+22] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. "Vqgan-clip: Open domain image generation and editing with natural language guidance". In: *European conference on computer vision*. Springer. 2022, pages 88–105 (cited on page 10).

[Den+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pages 248–255 (cited on page 17).

[Dev+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *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)*. 2019, pages 4171–4186 (cited on pages 4, 6, 13, 17).

[DF11] Thomas Deselaers and Vittorio Ferrari. "Visual and semantic similarity in imagenet". In: *CVPR 2011*. IEEE. 2011, pages 1777–1784 (cited on pages 12, 22).

[DLL17] Marie Delacre, Daniël Lakens, and Christophe Leys. "Why psychologists should by default use Welch's t-test instead of Student's t-test". In: *International Review of Social Psychology* 30.1 (2017), pages 92–101 (cited on page 18).

[Dos+20] 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: *ArXiv* abs/2010.11929 (2020) (cited on pages 9, 17).

[Dov14] Guy Dove. "Thinking in Words: Language as an Embodied Medium of Thought". In: *Topics in Cognitive Science* 6.3 (June 2014), pages 371–389. ISSN: 1756-8765 (cited on pages 3, 8).

[Du+25] Changde Du, Kaicheng Fu, Bincheng Wen, Yi Sun, Jie Peng, Wei Wei, Ying Gao, Shengpei Wang, Chuncheng Zhang, Jinpeng Li, et al. "Human-like object concept representations emerge naturally in multimodal large language models". In: *Nature Machine Intelligence* (2025), pages 1–16 (cited on page 7).

[ERO21] Patrick Esser, Robin Rombach, and Bjorn Ommer. "Taming transformers for high-resolution image synthesis". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pages 12873–12883 (cited on page 10).

[Eth19] Kawin Ethayarajh. "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings". In: *arXiv preprint arXiv:1909.00512* (2019) (cited on page 9).

[Ett20] Allyson Ettinger. "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pages 34–48 (cited on pages 4, 13).

[Far+16] Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. "Problems with evaluation of word embeddings using word similarity tasks". In: *arXiv preprint arXiv:1605.02276* (2016) (cited on pages 4, 11, 27, 35).

[Fel00] Christiane Fellbaum. "WordNet: an electronic lexical database". In: *Language* 76 (2000), page 706 (cited on pages 4, 11, 25, 29).

[Fin+01] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. "Placing search in context: The concept revisited". In: *Proceedings of the 10th international conference on World Wide Web*. 2001, pages 406–414 (cited on pages 3, 4, 6, 9).

[Fis+83] Ira Fischler, Paul A Bloom, Donald G Childers, Salim E Roucos, and Nathan W Perry Jr. "Brain potentials related to stages of sentence verification". In: *Psychophysiology* 20.4 (1983), pages 400–409 (cited on page 15).

[FL10] Yansong Feng and Mirella Lapata. "Visual information in semantic representation". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*. Association for Computational Linguistics. 2010, pages 91–99 (cited on pages 3, 6, 8).

[Fu+14] Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. "Learning semantic hierarchies via word embeddings". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pages 1199–1209 (cited on page 13).

[GBH18] Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. "Hyperbolic Entailment Cones for Learning Hierarchical Embeddings". In: *International Conference on Machine Learning*. 2018 (cited on page 13).

[GD05] Maayan Geffet and Ido Dagan. "The distributional inclusion hypotheses and lexical entailment". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. 2005, pages 107–114 (cited on page 12).

[Gol19] Yoav Goldberg. "Assessing BERT's syntactic abilities". In: *arXiv preprint arXiv:1901.05287* (2019) (cited on page 22).

[GYC21] Tianyu Gao, Xingcheng Yao, and Danqi Chen. "SimCSE: Simple contrastive learning of sentence embeddings". In: *arXiv preprint arXiv:2104.08821* (2021) (cited on pages 4, 7, 8, 10, 17, 36).

[Ham+24] Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. "Identifying implicit social biases in vision-language models". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Volume 7. 2024, pages 547–561 (cited on page 27).

[Ham07] James A Hampton. "Typicality, graded membership, and vagueness". In: *Cognitive science* 31.3 (2007), pages 355–384 (cited on page 4).

[Har90] Stevan Harnad. "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1-3 (1990), pages 335–346 (cited on pages 3, 6, 7).

[Hea92] Marti A. Hearst. "Automatic Acquisition of Hyponyms from Large Text Corpora". In: *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. 1992 (cited on pages 11, 15).

[HH19] William Held and Nizar Habash. "The effectiveness of simple hybrid systems for hypernym discovery". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pages 3362–3367 (cited on page 11).

[HK14] Felix Hill and Anna Korhonen. "Learning Abstract Concept Embeddings from Multi-Modal Data: Since You Probably Can't See What I Mean". In: *Conference on Empirical Methods in Natural Language Processing*. 2014 (cited on page 8).

[HKB14] Felix Hill, Anna Korhonen, and Christian Bentz. "A quantitative empirical analysis of the abstract/concrete distinction". In: *Cognitive science* 38.1 (2014), pages 162–177 (cited on page 7).

[HLT22] Chan-Jan Hsu, Hung-yi Lee, and Yu Tsao. "XdBERT: distilling visual information to BERT from cross-modal systems to improve language understanding". In: *arXiv preprint arXiv:2204.07316* (2022) (cited on pages 3, 9).

[HLY23] Tsung-Yuan Hsu, Chen-An Li, and Chao-Wei Huang Yun-Nung. "Visually-Enhanced Phrase Understanding". In: *Annual Meeting of the Association for Computational Linguistics*. 2023 (cited on pages 4, 9, 10, 37).

[HM21] Michael Hanna and David Mareček. "Analyzing BERT's knowledge of hypernymy via prompting". In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. 2021, pages 275–282 (cited on pages 4, 13).

[HRK14] Felix Hill, Roi Reichart, and Anna Korhonen. "Multi-Modal Models for Concrete and Abstract Concept Meaning". In: *Transactions of the Association for Computational Linguistics* 2 (2014), pages 285–296 (cited on pages 3, 8).

[HRK15] Felix Hill, Roi Reichart, and Anna Korhonen. "Simlex-999: Evaluating semantic models with (genuine) similarity estimation". In: *Computational Linguistics* 41.4 (2015), pages 665–695 (cited on pages 3, 4, 6, 9).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pages 1735–1780 (cited on page 13).

[Jia+21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. "Scaling up visual and vision-language representation learning with noisy text supervision". In: *International conference on machine learning*. PMLR. 2021, pages 4904–4916 (cited on pages 8, 17, 36).

[KB14] Douwe Kiela and Léon Bottou. “Learning image embeddings using convolutional neural networks for improved multi-modal semantics”. In: *Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP)*. 2014, pages 36–45 (cited on pages 3, 8).

[KC15] Douwe Kiela and Stephen Clark. “Multi- and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception”. In: *Conference on Empirical Methods in Natural Language Processing*. 2015 (cited on page 6).

[KF23] Zaid Khan and Yun Fu. “Contrastive alignment of vision to language through parameter-efficient transfer learning”. In: *arXiv preprint arXiv:2303.11866* (2023) (cited on page 10).

[KHC23] Amita Kamath, Jack Hessel, and Kai-Wei Chang. “Text encoders bottleneck compositionality in contrastive vision-language models”. In: *Conference on Empirical Methods in Natural Language Processing*. 2023 (cited on page 9).

[Kie+14] Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. “Improving multi-modal representations using image dispersion: Why less is sometimes more”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2014, pages 835–841 (cited on pages 3, 6, 8).

[Kie+15] Douwe Kiela, Laura Rimell, Ivan Vulic, and Stephen Clark. “Exploiting image generality for lexical entailment detection”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*. ACL; East Stroudsburg, PA. 2015, pages 119–124 (cited on pages 11, 12, 18, 22).

[Kot+10] Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. “Directional distributional similarity for lexical inference”. In: *Natural Language Engineering* 16.4 (2010), pages 359–389 (cited on pages 12, 23, 70).

[KP95] Hans Kamp and Barbara Partee. “Prototype theory and compositionality”. In: *Cognition* 57.2 (1995), pages 129–191 (cited on pages 11, 25).

[KPB15] German Kruszewski, Denis Paperno, and Marco Baroni. “Deriving boolean structures from distributional vectors”. In: *Transactions of the Association for Computational Linguistics* 3 (2015), pages 375–388 (cited on page 13).

[Kum20] Abhilasha Ashok Kumar. “Semantic memory: A review of methods, models, and current challenges”. In: *Psychonomic Bulletin & Review* 28 (2020), pages 40–80 (cited on page 6).

[KUO25] Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. “CLIP Behaves like a Bag-of-Words Model Cross-modally but not Uni-modally”. In: *arXiv preprint arXiv:2502.03566* (2025) (cited on page 9).

[LB12] Alessandro Lenci and Giulia Benotto. “Identifying hypernyms in distributional semantic spaces”. In: ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. 2012, pages 75–79 (cited on pages 11, 12, 18, 21, 23, 57, 70).

[LC09] Dermot Lynott and Louise Connell. “Modality exclusivity norms for 423 object properties”. In: *Behavior research methods* 41.2 (2009), pages 558–564 (cited on page 8).

[LCD23] Jiayi Liao, Xu Chen, and Lun Du. "Concept Understanding in Large Language Models: An Empirical Study". In: *Tiny Papers @ ICLR*. 2023 (cited on pages 4, 14, 32).

[Le+19] Matt Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. "Inferring Concept Hierarchies from Text Corpora via Hyperbolic Embeddings". In: *Annual Meeting of the Association for Computational Linguistics*. 2019 (cited on page 11).

[Len+08] Alessandro Lenci et al. "Distributional semantics in linguistic and cognitive research". In: *Italian journal of linguistics* 20.1 (2008), pages 1–31 (cited on page 6).

[Len18] Alessandro Lenci. "Distributional models of word meaning". In: *Annual review of Linguistics* 4.1 (2018), pages 151–171 (cited on page 6).

[Lev+15] Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. "Do supervised distributional methods really learn lexical inference relations?" In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2015, pages 970–976 (cited on page 12).

[Lew+22] Martha Lewis, Nihal V Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H Bach, and Ellie Pavlick. "Does clip bind concepts? probing compositionality in large image models". In: *arXiv preprint arXiv:2212.10537* (2022) (cited on page 9).

[Li+19] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. "Visualbert: A simple and performant baseline for vision and language". In: *arXiv preprint arXiv:1908.03557* (2019) (cited on pages 3, 6–8).

[Li+21] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. "Align before fuse: Vision and language representation learning with momentum distillation". In: *Advances in neural information processing systems* 34 (2021), pages 9694–9705 (cited on page 8).

[Li+22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation". In: *International conference on machine learning*. PMLR. 2022, pages 12888–12900 (cited on page 7).

[Li+23a] Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Qi Liu, Lingpeng Kong, and Xu Sun. "Can language models understand physical concepts?" In: *arXiv preprint arXiv:2305.14057* (2023) (cited on page 7).

[Li+23b] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. "Towards general text embeddings with multi-stage contrastive learning". In: *arXiv preprint arXiv:2308.03281* (2023) (cited on pages 17, 21, 35).

[Li+24] Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. "Do vision and language models share concepts? a vector space alignment study". In: *Transactions of the Association for Computational Linguistics* 12 (2024), pages 1232–1249 (cited on page 7).

[Lin98] Dekang Lin. "An Information-Theoretic Definition of Similarity". In: *International Conference on Machine Learning*. 1998 (cited on page 12).

[Liu+22] Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. "Things not written in text: Exploring spatial commonsense from visual signals". In: *arXiv preprint arXiv:2203.08075* (2022) (cited on pages 7, 10, 37).

[Lou11] Max M Louwerse. "Symbol interdependency in symbolic and embodied cognition". In: *Topics in Cognitive Science* 3.2 (2011), pages 273–302 (cited on pages 3, 8).

[LPB15] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. "Combining language and vision with a multimodal skip-gram model". In: *arXiv preprint arXiv:1501.02598* (2015) (cited on pages 3, 6).

[LS23] A. Lenci and M. Sahlgren. *Distributional Semantics*. Studies in Natural Language Processing. Cambridge University Press, 2023. ISBN: 9781107004290 (cited on pages 3, 6).

[LSM22] Jiacheng Li, Jingbo Shang, and Julian McAuley. "UCTopic: Unsupervised contrastive learning for phrase representations and topic mining". In: *arXiv preprint arXiv:2202.13469* (2022) (cited on pages 4, 10).

[Lu+19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks". In: *Advances in neural information processing systems* 32 (2019) (cited on pages 3, 6, 8).

[MER21] Kanishka Misra, Allyson Ettinger, and Julia Rayz. "Do language models learn typicality judgments from text?" In: *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*. 2021 (cited on pages 4, 13, 17, 21).

[Mik+13a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013) (cited on page 12).

[Mik+13b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* 26 (2013) (cited on page 6).

[Mur04] Gregory Murphy. *The big book of concepts*. MIT press, 2004 (cited on pages 4, 11).

[Ngu+17] Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. "Hierarchical embeddings for hypernymy detection and directionality". In: *arXiv preprint arXiv:1707.07273* (2017) (cited on page 13).

[NK17] Maximillian Nickel and Douwe Kiela. "Poincaré embeddings for learning hierarchical representations". In: *Advances in neural information processing systems* 30 (2017) (cited on page 13).

[NK18] Maximilian Nickel and Douwe Kiela. "Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry". In: *ArXiv* abs/1806.03417 (2018) (cited on page 13).

[NMS04] Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. "The University of South Florida free association, rhyme, and word fragment norms". In: *Behavior Research Methods, Instruments, & Computers* 36.3 (2004), pages 402–407 (cited on pages 6, 25).

[Pai13] Allan Paivio. "Dual coding theory, word abstractness, and emotion: A critical review of Kousta et al. (2011)." In: *Journal of Experimental Psychology: General* 142.1 (2013), pages 282–287. ISSN: 0096-3445 (cited on pages 3, 8).

[Pai91] Allan Paivio. “Dual coding theory: Retrospect and current status.” In: *Canadian Journal of Psychology/Revue canadienne de psychologie* 45.3 (1991), page 255 (cited on pages 3, 7, 8).

[PGH24] Derek Powell, Walter Gerych, and Thomas Hartvigsen. “Taxi: Evaluating categorical knowledge editing for language models”. In: *arXiv preprint arXiv:2404.15004* (2024) (cited on page 37).

[PL25] Mattia Proietti and Alessandro Lenci. “The quasi-semantic competence of LLMs: a case study on the part-whole relation”. In: *arXiv preprint arXiv:2504.02395* (2025) (cited on page 37).

[PSL18] Ludovica Pannitto, Lavinia Salicchi, and Alessandro Lenci. “Refining the Distributional Inclusion Hypothesis for Unsupervised Hypernym Identification”. In: *Italian Journal of Computational Linguistics* (2018) (cited on page 12).

[PTF21] Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. “Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation”. In: *Transactions of the Association for Computational Linguistics* 9 (2021), pages 1563–1579 (cited on pages 3, 4, 6, 8–10, 36).

[Rad+19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), page 9 (cited on pages 3, 4, 9, 13, 17).

[Rad+21] 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*. PmLR. 2021, pages 8748–8763 (cited on pages 3, 7–9, 17, 36).

[RAL24] Michaela Regneri, Alhassan Abdelhalim, and Sören Laue. “Detecting conceptual abstraction in LLMs”. In: *arXiv preprint arXiv:2404.15848* (2024) (cited on pages 4, 5, 14, 15, 36).

[Ram+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. “Hierarchical text-conditional image generation with clip latents”. In: *arXiv preprint arXiv:2204.06125* 1.2 (2022), page 3 (cited on page 10).

[Rav+20] Abhilasha Ravichander, Eduard H. Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. “On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT”. In: *STARSEM*. 2020 (cited on page 13).

[RDG23] Joseph Renner, Pascal Denis, and Rémi Gilleron. “WordNet Is All You Need: A Surprisingly Effective Unsupervised Method for Graded Lexical Entailment”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023, pages 9176–9182 (cited on pages 14, 17, 25).

[RE16] Stephen Roller and Katrin Erk. “Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment”. In: *arXiv preprint arXiv:1605.05433* (2016) (cited on page 11).

[REB14] Stephen Roller, Katrin Erk, and Gemma Boleda. "Inclusive yet Selective: Supervised Distributional Hypernymy Detection". In: *International Conference on Computational Linguistics*. 2014 (cited on page 12).

[Ren+23] Joseph Renner, Pascal Denis, Rémi Gilleron, and Angèle Brunellière. "Exploring category structure with contextual language models and lexical semantic networks". In: *arXiv preprint arXiv:2302.06942* (2023) (cited on page 37).

[RG19] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019 (cited on pages 4, 17, 22).

[RG65] Herbert Rubenstein and J. Goodenough. "Contextual correlates of synonymy". In: *Commun. ACM* 8 (1965), pages 627–633 (cited on pages 3, 6, 9).

[Rim14] Laura Rimell. "Distributional Lexical Entailment by Topic Coherence". In: *Conference of the European Chapter of the Association for Computational Linguistics*. 2014 (cited on pages 12, 16).

[RKN18] Stephen Roller, Douwe Kiela, and Maximilian Nickel. "Hearst patterns revisited: Automatic hypernym detection from large text corpora". In: *arXiv preprint arXiv:1806.03191* (2018) (cited on pages 11, 18).

[Rom+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pages 10684–10695 (cited on pages 10, 18).

[Ros75] Eleanor Rosch. "Cognitive representations of semantic categories." In: *Journal of experimental psychology: General* 104.3 (1975), page 192 (cited on pages 11, 25).

[RRG22] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. "DALLE-2 is seeing double: Flaws in word-to-concept mapping in Text2Image models". In: *arXiv preprint arXiv:2210.10606* (2022) (cited on page 10).

[San+14] Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte Im Walde. "Chasing hypernyms in vector spaces with entropy". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers*. 2014, pages 38–42 (cited on page 12).

[San+15] Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. "EVALution 1.0: an Evolving Semantic Dataset for Training and Evaluation of Distributional Semantic Models". In: *LDL@IJCNLP*. 2015 (cited on page 11).

[San+16a] Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. "Nine features in a random forest to learn taxonomical semantic relations". In: *arXiv preprint arXiv:1603.08702* (2016) (cited on pages 12, 23, 37).

[San+16b] Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. "Unsupervised measure of word similarity: How to outperform co-occurrence and vector cosine in vsms". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 30. 1. 2016 (cited on page 12).

[Sei+16] Julian Seitner, Christian Bizer, Kai Eckert, Stefano Faralli, Robert Meusel, Heiko Paulheim, and Simone Paolo Ponzetto. "A large database of hypernymy relations extracted from the web." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pages 360–367 (cited on page 14).

[SGD16] Vered Shwartz, Yoav Goldberg, and Ido Dagan. "Improving hypernymy detection with an integrated path-based and distributional method". In: *arXiv preprint arXiv:1603.06076* (2016) (cited on pages 11, 13).

[She+21] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. "How much can clip benefit vision-and-language tasks?" In: *arXiv preprint arXiv:2107.06383* (2021) (cited on page 9).

[SJN04] Rion Snow, Daniel Jurafsky, and Andrew Ng. "Learning syntactic patterns for automatic hypernym discovery". In: *Advances in neural information processing systems* 17 (2004) (cited on page 11).

[SL12] Carina Silberer and Mirella Lapata. "Grounded models of semantic representation". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics. 2012, pages 1423–1433 (cited on pages 6, 8).

[SL14] Carina Silberer and Mirella Lapata. "Learning grounded meaning representations with autoencoders". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. 2014, pages 721–732 (cited on page 6).

[Son+20] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. "Mpnet: Masked and permuted pre-training for language understanding". In: *Advances in neural information processing systems* 33 (2020), pages 16857–16867 (cited on page 17).

[Spe61] Charles Spearman. "The proof and measurement of association between two things." In: (1961) (cited on page 25).

[SSS16] Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. "Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection". In: *arXiv preprint arXiv:1612.04460* (2016) (cited on pages 4, 12, 18, 22).

[Sun+19] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. "Videobert: A joint model for video and language representation learning". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pages 7464–7473 (cited on page 7).

[SVS23] Chen Shani, Jilles Vreeken, and Dafna Shahaf. "Towards concept-aware large language models". In: *arXiv preprint arXiv:2311.01866* (2023) (cited on pages 4, 13).

[TB19] Hao Tan and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers". In: *arXiv preprint arXiv:1908.07490* (2019) (cited on pages 6, 7).

[TB20] Hao Tan and Mohit Bansal. "Vokenization: Improving language understanding with contextualized, visual-grounded supervision". In: *arXiv preprint arXiv:2010.06775* (2020) (cited on page 6).

[TBP23] Aleksey Tikhonov, Lisa Bylinina, and Denis Paperno. “Leverage points in modality shifts: Comparing language-only and multimodal word representations”. In: *arXiv preprint arXiv:2306.02348* (2023) (cited on page 10).

[TL19] Mingxing Tan and Quoc V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *ArXiv* abs/1905.11946 (2019) (cited on page 17).

[TL24] Mikhail Tikhomirov and Natalia Loukachevitch. “Exploring Prompt-Based Methods for Zero-Shot Hypernym Prediction with Large Language Models”. In: *arXiv preprint arXiv:2401.04515* (2024) (cited on page 14).

[TP10] Peter D Turney and Patrick Pantel. “From frequency to meaning: Vector space models of semantics”. In: *Journal of artificial intelligence research* 37 (2010), pages 141–188 (cited on page 6).

[TU16] Katsumi Takano and Akira Utsumi. “Grounded Distributional Semantics for Abstract Words”. In: *Cognitive Science* (2016) (cited on page 8).

[Uts22] Akira Utsumi. “A test of indirect grounding of abstract concepts using multimodal distributional semantics”. In: *Frontiers in Psychology* 13 (2022) (cited on page 8).

[Ven+15] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. “Order-embeddings of images and language”. In: *arXiv preprint arXiv:1511.06361* (2015) (cited on page 13).

[VM17] Ivan Vulić and Nikola Mrkšić. “Specialising word vectors for lexical entailment”. In: *arXiv preprint arXiv:1710.06371* (2017) (cited on page 13).

[Vul+17] Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. “Hyperlex: A large-scale evaluation of graded lexical entailment”. In: *Computational Linguistics* 43.4 (2017), pages 781–835 (cited on pages 4, 11, 12, 14, 17, 25, 30).

[Vul+20] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. “Probing pretrained language models for lexical semantics”. In: *arXiv preprint arXiv:2010.05731* (2020) (cited on page 13).

[Wad+24] Shakti N Wadekar, Abhishek Chaurasia, Aman Chadha, and Eugenio Culurciello. “The evolution of multimodal model architectures”. In: *arXiv preprint arXiv:2405.17927* (2024) (cited on page 7).

[Wan+19] Chengyu Wang, Yan Fan, Xiaofeng He, and Aoying Zhou. “A Family of Fuzzy Orthogonal Projection Models for Monolingual and Cross-lingual Hypernymy Prediction”. In: *The World Wide Web Conference* (2019) (cited on page 13).

[Wan+20] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. *MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers.* 2020 (cited on page 17).

[WC22a] Jennifer C White and Ryan Cotterell. “Schrödinger’s Bat: Diffusion Models Sometimes Generate Polysemous Words in Superposition”. In: *arXiv preprint arXiv:2211.13095* (2022) (cited on pages 10, 18).

[WC22b] Robert Wolfe and Aylin Caliskan. “Contrastive Visual Semantic Pretraining Magnifies the Semantics of Natural Language Representations”. In: *ArXiv* abs/2203.07511 (2022) (cited on pages 3, 5, 9, 36).

[Wee+14] Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. "Learning to distinguish hypernyms and co-hyponyms". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2014, pages 2249–2259 (cited on pages 11, 12, 21).

[WH00] W. Caroline West and Phillip J. Holcomb. "Imaginal, Semantic, and Surface-Level Processing of Concrete and Abstract Words: An Electrophysiological Investigation". In: *Journal of Cognitive Neuroscience* 12.6 (Nov. 2000), pages 1024–1037. ISSN: 1530-8898 (cited on page 7).

[WH20] Chengyu Wang and Xiaofeng He. "BiRRE: learning bidirectional residual relation embeddings for supervised hypernymy detection". In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. 2020, pages 3630–3640 (cited on pages 13, 18).

[WT24] Hanna-Sophia Widhoelzl and Ece Takmaz. "Decoding Emotions in Abstract Art: Cognitive Plausibility of CLIP in Recognizing Color-Emotion Associations". In: *arXiv preprint arXiv:2405.06319* (2024) (cited on page 9).

[WTI21] Shufan Wang, Laure Thompson, and Mohit Iyyer. "Phrase-BERT: Improved phrase embeddings from BERT with an application to corpus exploration". In: *arXiv preprint arXiv:2109.06304* (2021) (cited on pages 4, 10, 17).

[WW03] Julie Weeds and David Weir. "A general framework for distributional similarity". In: *Proceedings of the 2003 conference on Empirical methods in natural language processing*. 2003, pages 81–88 (cited on page 12).

[Xu+25] Qihui Xu, Yingying Peng, Samuel A Nastase, Martin Chodorow, Minghua Wu, and Ping Li. "Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts". In: *Nature human behaviour* (2025), pages 1–16 (cited on page 7).

[Yam+16] Josuke Yamane, Tomoya Takatani, Hitoshi Yamada, Makoto Miwa, and Yutaka Sasaki. "Distributional hypernym generation by jointly learning clusters and projections". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, pages 1871–1879 (cited on page 13).

[Yan+22] An Yan, Jiacheng Li, Wanrong Zhu, Yujie Lu, William Yang Wang, and Julian McAuley. "Clip also understands text: Prompting clip for phrase understanding". In: *arXiv preprint arXiv:2210.05836* (2022) (cited on pages 4, 9, 10).

[YSP21] Tian Yun, Chen Sun, and Ellie Pavlick. "Does vision-and-language pretraining improve lexical grounding?" In: *arXiv preprint arXiv:2109.10246* (2021) (cited on page 6).

[Yuk+22] 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: *arXiv preprint arXiv:2210.01936* (2022) (cited on pages 9, 21).

[ZFA23] Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. "Visual grounding helps learn word meanings in low-data regimes". In: *arXiv preprint arXiv:2310.13257* (2023) (cited on pages 10, 23, 36).

[Zha+21] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. "Vinvl: Revisiting visual representations in vision-language models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pages 5579–5588 (cited on page 7).

[Zha+22a] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. "Lit: Zero-shot transfer with locked-image text tuning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pages 18123–18133 (cited on pages 8, 17, 35, 36).

[Zha+22b] Miaoran Zhang, Marius Mosbach, David Ifeoluwa Adelani, Michael A Hedderich, and Dietrich Klakow. "MCSE: Multimodal contrastive learning of sentence embeddings". In: *arXiv preprint arXiv:2204.10931* (2022) (cited on pages 7, 17, 36).

[Zha+22c] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. "Opt: Open pre-trained transformer language models". In: *arXiv preprint arXiv:2205.01068* (2022) (cited on page 7).

[Zha+23a] Yan Zhang, Zhaopeng Feng, Zhiyang Teng, Zuozhu Liu, and Haizhou Li. "How well do text embedding models understand syntax?" In: *arXiv preprint arXiv:2311.07996* (2023) (cited on page 22).

[Zha+23b] Mengjie Zhao, Junya Ono, Zhi-Wei Zhong, Chieh-Hsin Lai, Yuhta Takida, Naoki Murata, Wei-Hsiang Liao, Takashi Shibuya, Hiromi Wakaki, and Yuki Mitsufuji. "On the Language Encoder of Contrastive Cross-modal Models". In: *Annual Meeting of the Association for Computational Linguistics*. 2023 (cited on pages 10, 23, 36).

Appendix A

Synthetic concepts

We devise two synthetic concepts, q' and p', for investigating different models' representation of the lexical relation between the pair (q, p). In particular, we hypothesize that for hypernymy pairs where p is the hypernym of q, e.g. (robin, bird), q' (q, a type of p) would be highly similar to q, and p' (p, such as q) would be similar to p. Compared to a naïve average of q and p, q' is expected to be closer to q than to p, while p' is expected to be closer to p than to q. To examine these hypotheses, we introduce two measures, c_q and c_p, as a sanity check, with the assumption that both values would surpass 1 in the case of hypernymy (Section 3.2).

Fig A.2 and Fig A.3 summarize the distribution of c_q and c_p on the dataset BLESS in our first experiment (Chapter 4). We observe that for all models except PhraseBERT, the interquartile range of c_q lies entirely above 1, providing descriptive support for our hypothesis concerning q'. However, the synthetic concept p' fails the sanity check with the median of the c_p measure falling below 1 for most models, contradicting with our former hypothesis. Exploration via text-to-image generation also demonstrates that images generated with synthetic prompts "q, a type of p" and "p, such as q" both tend to depict the specific concept q rather than the more generic hypernym p (Fig A.1).

This phenomenon can be explained in terms of their in natural language: "bird, such as robin", although semantically similar to the hypernym "bird", is more distributionally constrained as it is unlikely to occur in contexts which do not also apply to the hyponym "robin"; if a caption contains the specialization "such as robin", the corresponding image would also depict a robin rather than an arbitrary type of bird with high probability. Therefore, the expression "p, such as q" is distributionally more similar to q. Consequently, only measures based on q' are used as hypernymy measures in our experiments, while measures based on p' are discarded.

Despite its failure in this hypernymy detection scenario, we argue that this observed phenomenon is hypernymy-specific, and that p', the synthetic concept based on p, may still prove informative for lexical relations. For example, the synthetic concept p' specialized for meronymy, "p, with some q", is more similar to p than to q: "door, with some handle" differ from the meronym "handle" w.r.t. distribution in text corpora, and images depicting "door, with some handle" are more likely to depict a complete door rather than the local region around the handle, while images depicting "handle" typically focus on the handle. In our preliminary experiment on BLESS using the synthetic concept p', "p, with some q", s_p and r_p successfully distinguish meronymy from the other lexical relations.

q: “robin”

p: “bird”

q': “robin, a type of bird”

p': “bird, such as robin”

q': “robin, a type of cat”

Figure A.1: Generated images for concepts based on (robin, bird) and (robin, cat).

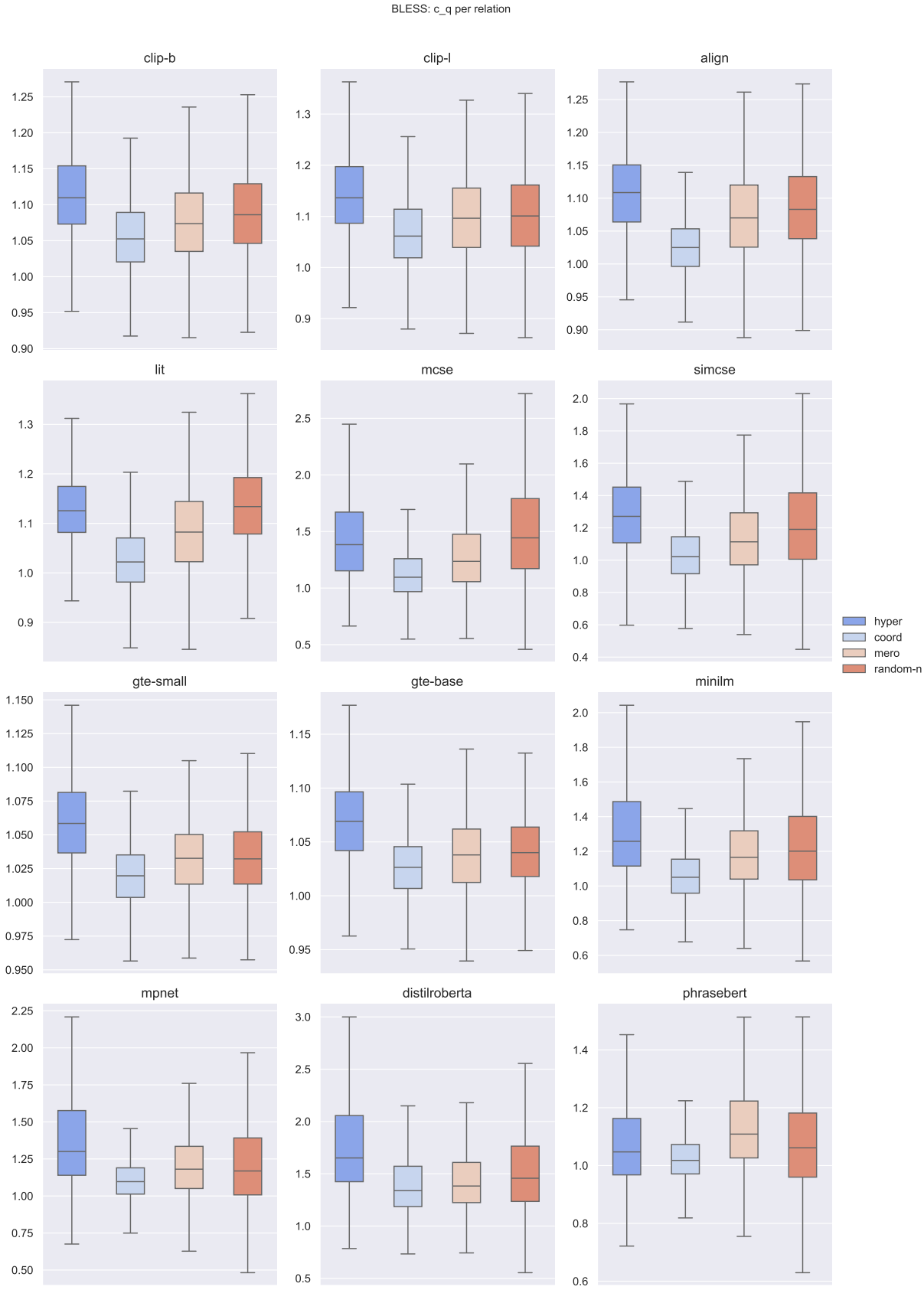

Figure A.2: Boxplots summarizing the distribution of c_q produced by different models on BLESS.

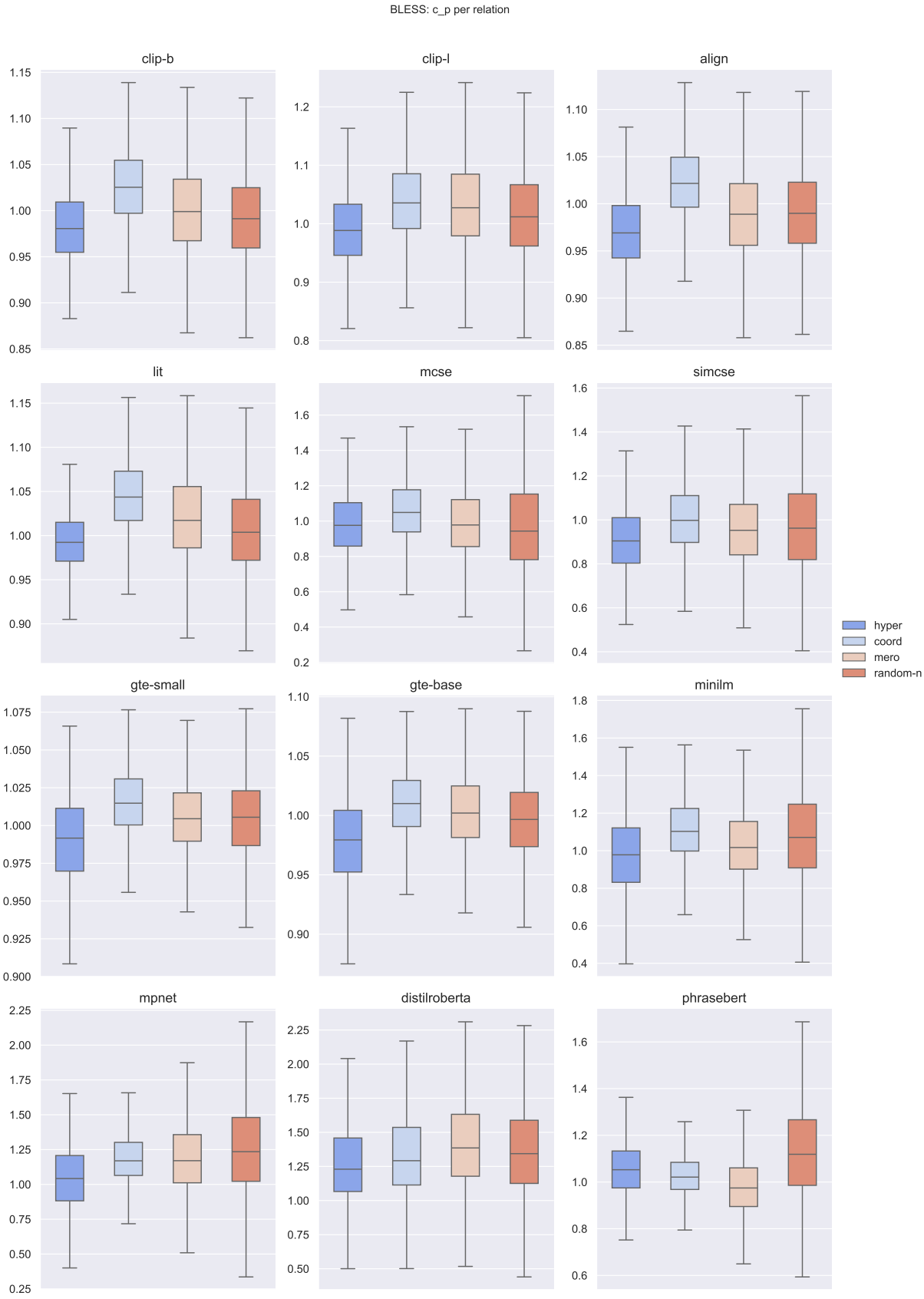

Figure A.3: Boxplots summarizing the distribution of c_p produced by different models on BLESS.

Appendix B

Boxplots

Following Baroni and Lenci [BL11] and Lenci and Benotto [LB12], we produce boxplots summarizing the distribution of different measures for each lexical relation. While their original setting includes a z-normalization procedure for each target concept to factor out concept-specific effects, e.g. the tendency of more frequent concepts to receive higher scores, our boxplots visualize the raw measures with no additional preprocessing.

B.1 Distribution per relation

Fig B.1 summarizes the distribution of b, namely the naïve cosine similarity between q and p, produced by different models across lexical relations in BLESS. We observe that the boxes for hypernymy and for coordination substantially overlap, indicating limited separability. This is confirmed by a one-sided Welch's t-test, where only CLIP, LiT, and GTE exhibit statistically significant differences between the two relations (Table 4.1). Distinguishing hypernymy from meronymy is another challenge for cosine similarity. Still, it serves as an effective baseline for distinguishing hypernymy from random pairs, which are semantically dissimilar.

With our specialized hypernymy measures s_q and r_q (Fig B.2–B.3), hypernymy can be more easily distinguished from coordination as well as meronymy. However, with contrastive VLMs these measures tend to be surprisingly high for random pairs. We argue that for random pairs, the representation of q' might be dominated by the more plausible component q due to the extreme counterfactuality of the synthetic concept "q, a type of p", leading to potential confusion with hypernymy. We thus suggest that combining these measures may contribute to more robust hypernymy detection, and report results of a simple combination method via multiplication in Appendix D.

B.2 Distribution w.r.t. concreteness/specificity

We also provide boxplots summarizing the distribution of s_q and r_q on BBC w.r.t. different lexical relations (hyper vs. coord) as well as different levels of concreteness and specificity. From Fig B.4 we observe that contrastive VLMs tend to produce higher s_q scores for both hypernymy and coordination pairs if the word pair has low concreteness, i.e. if the pair is abstract, making it difficult to distinguish concrete hypernymy pairs from abstract coordination pairs. A similar but weaker tendency exists for specificity (Fig B.5). Meanwhile, the measure r_q does not appear to be sensitive to concreteness and specificity as s_q is (Figures B.6-B.7).

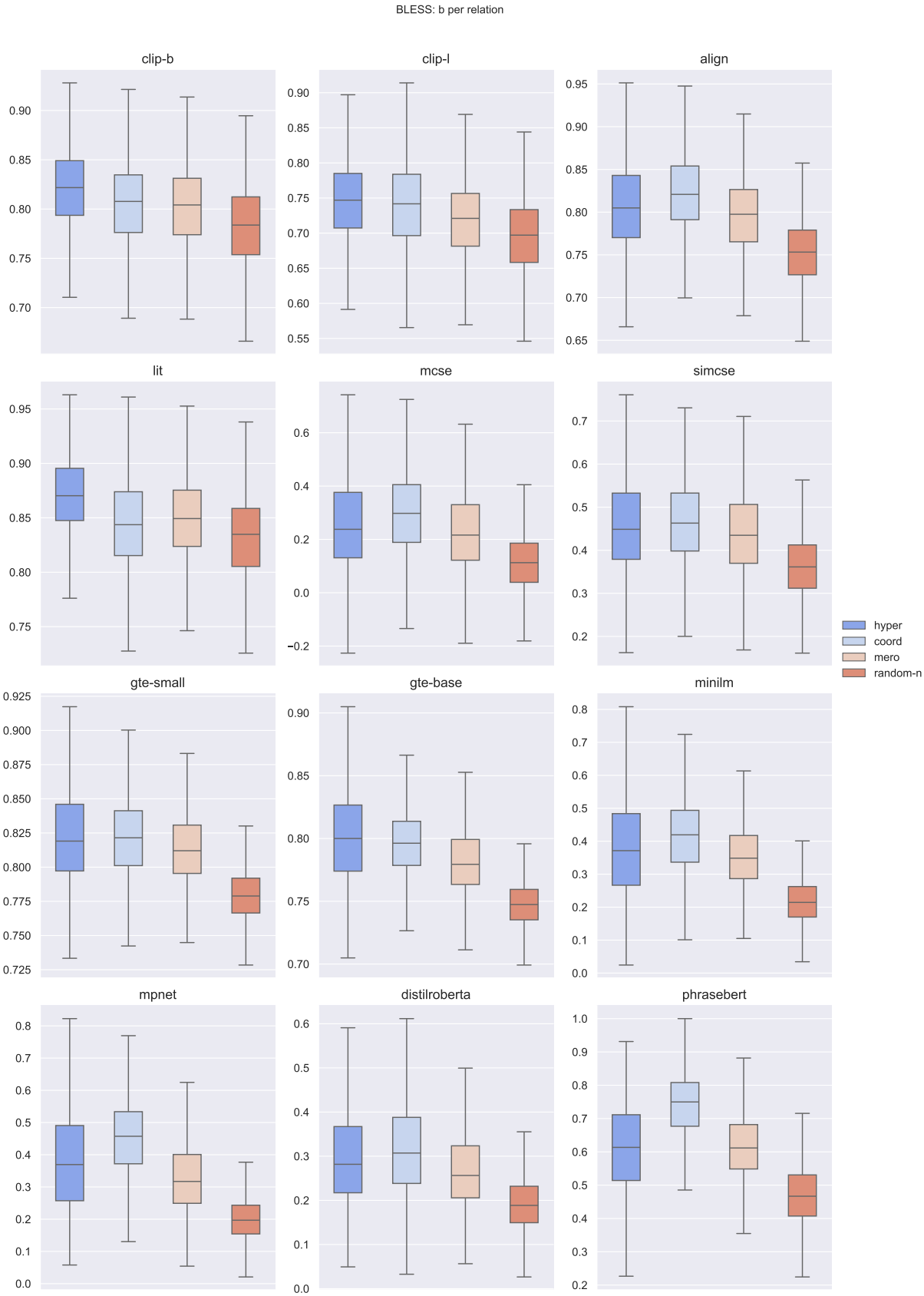

Figure B.1: Boxplots summarizing the distribution of naïve cosine similarity scores produced by different models across lexical relations in BLESS.

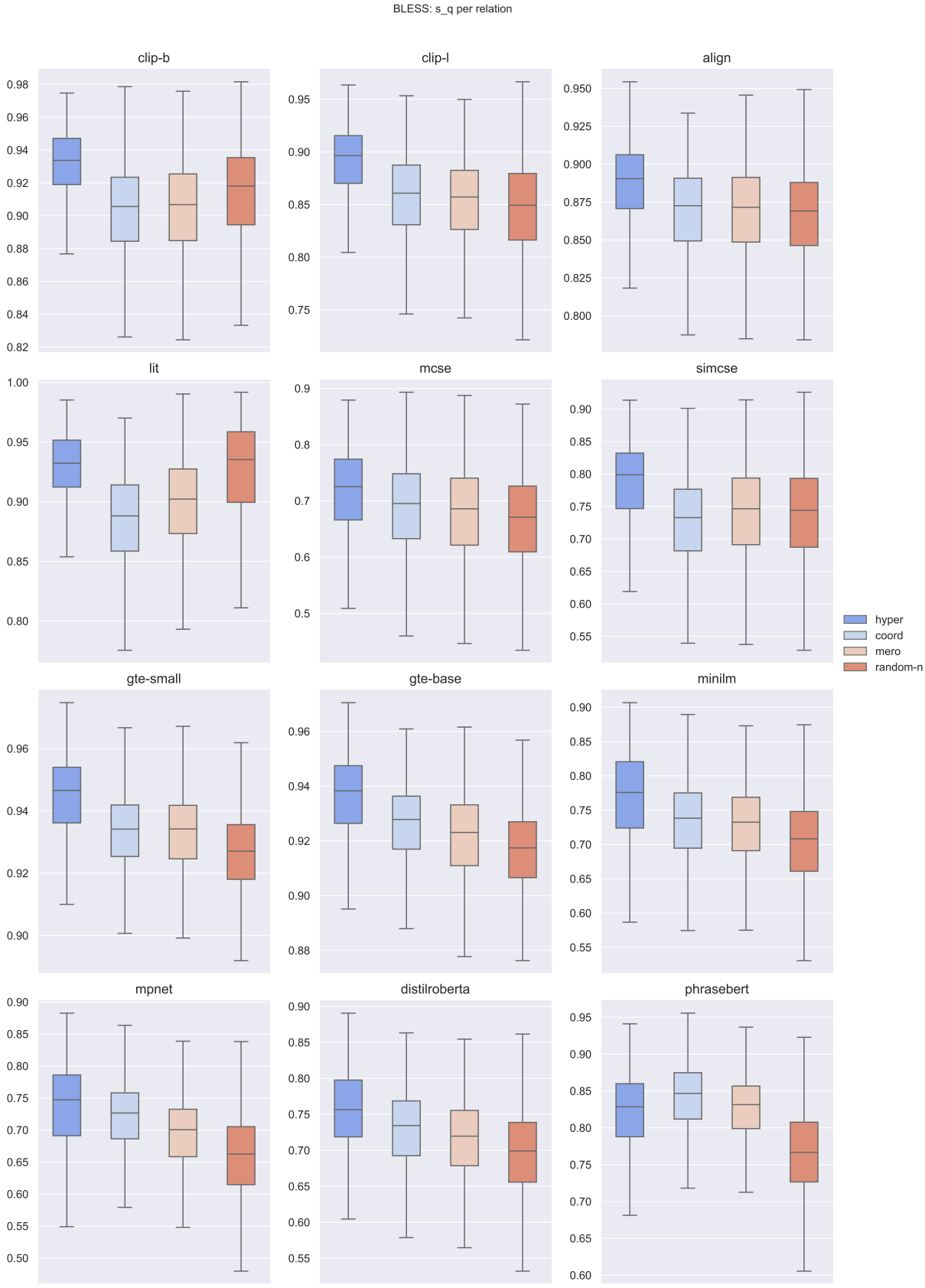

Figure B.2: Boxplots summarizing the distribution of s_q produced by different models on BLESS.

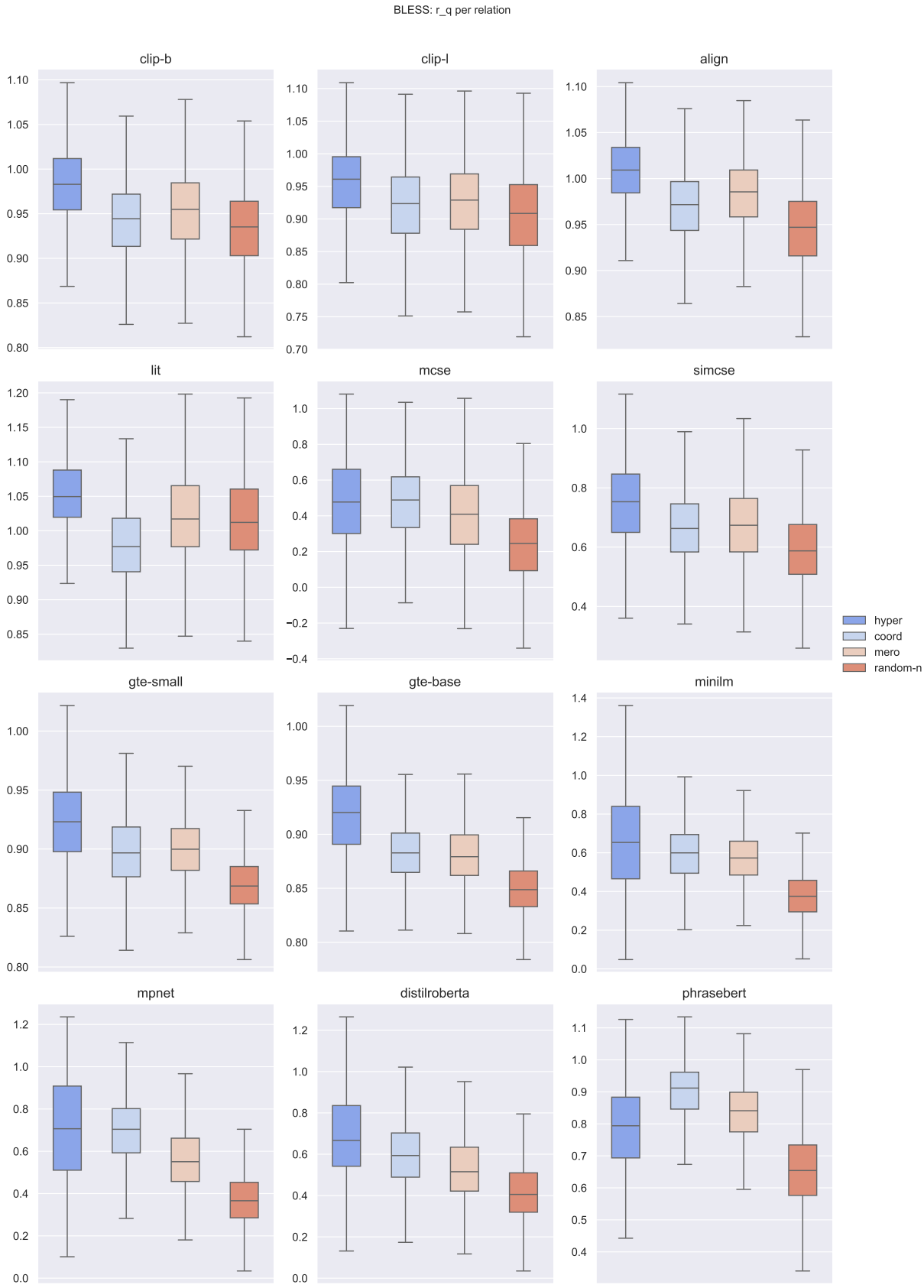

Figure B.3: Boxplots summarizing the distribution of r_q produced by different models on BLESS.

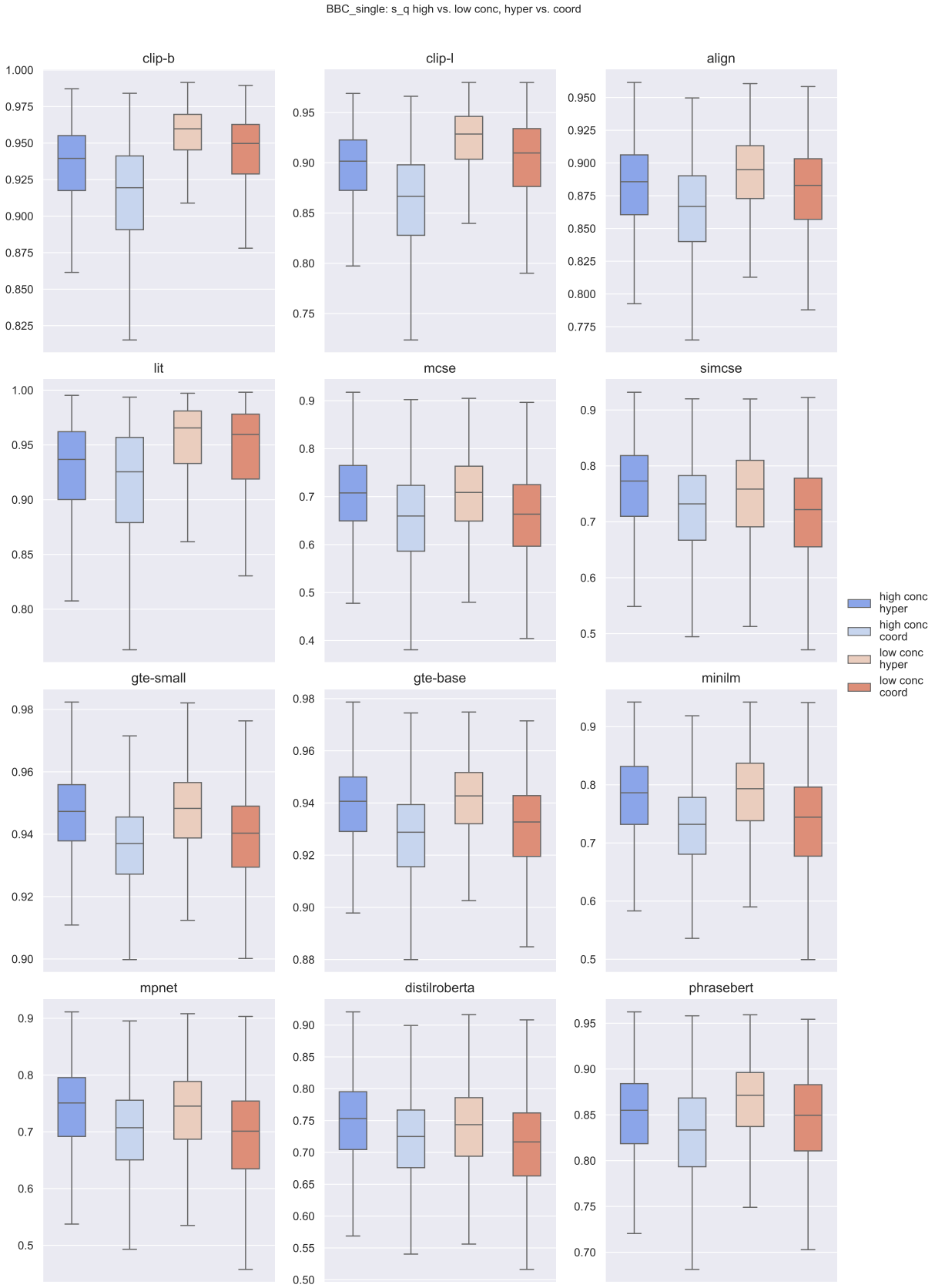

Figure B.4: Boxplots summarizing the distribution of s_q on concrete and abstract pairs on BBC. MWEs excluded.

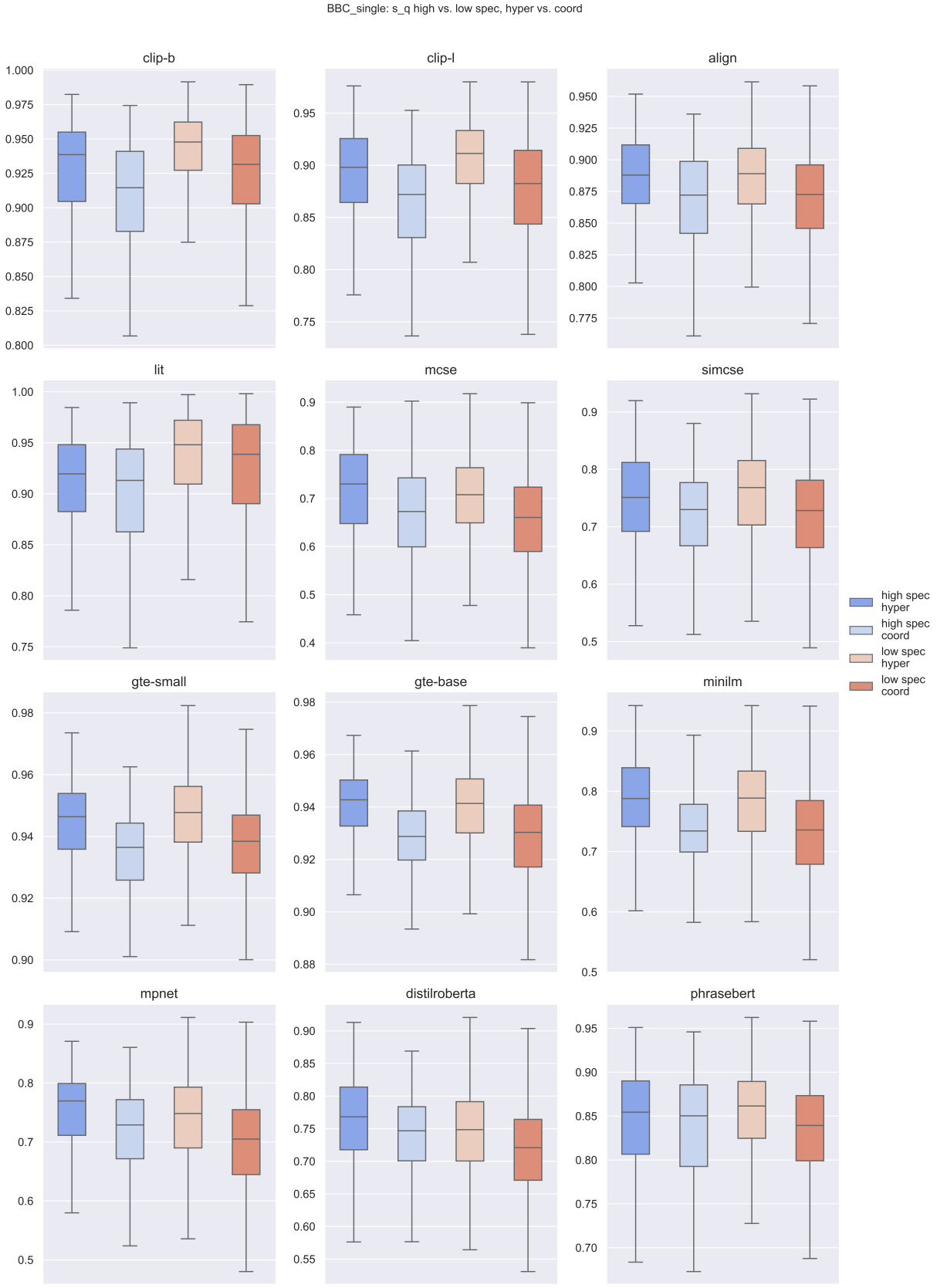

Figure B.5: Boxplots summarizing the distribution of s_q on specific and generic pairs on BBC. MWEs excluded.

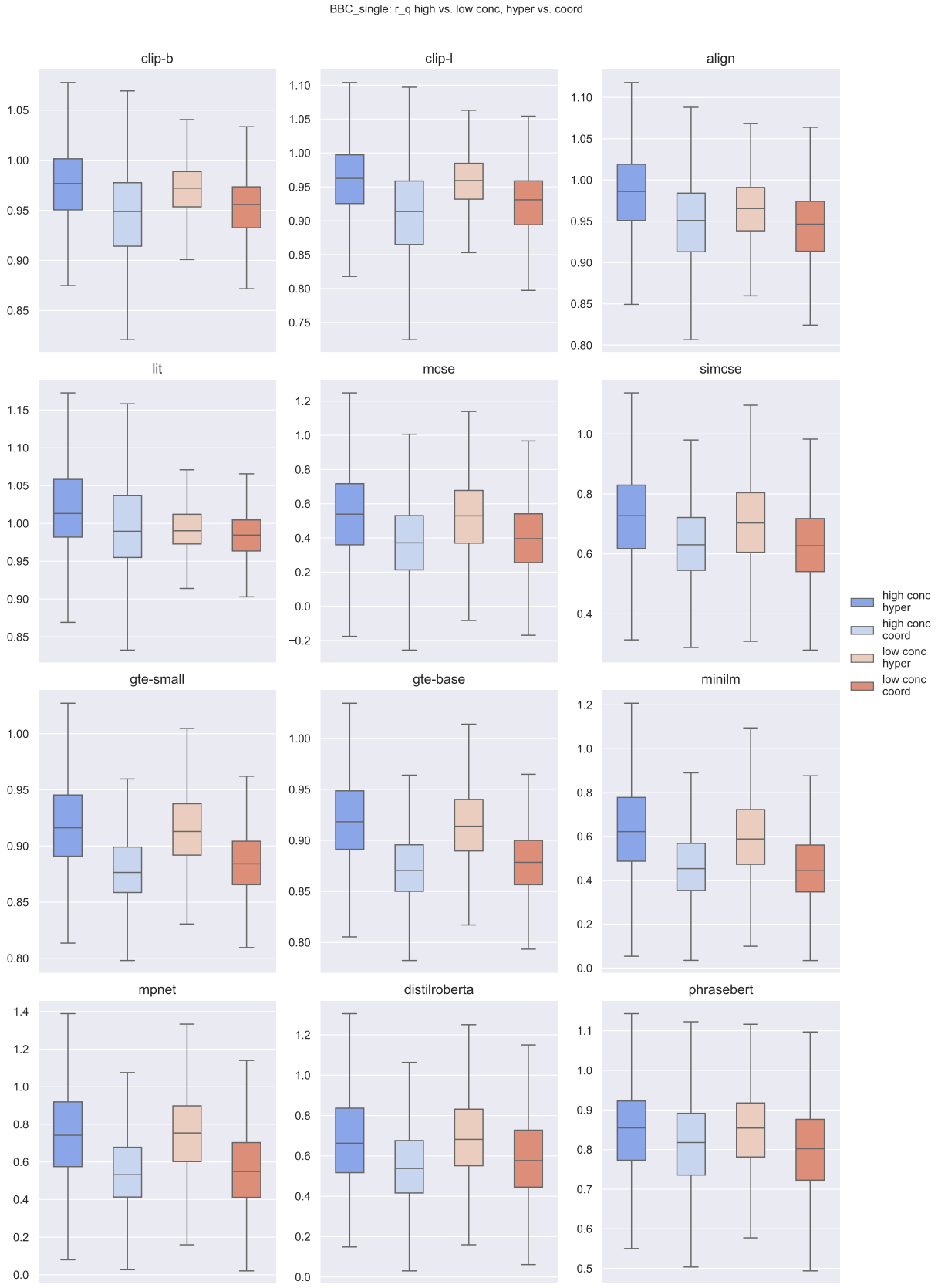

Figure B.6: Boxplots summarizing the distribution of r_q on concrete and abstract pairs in BBC. MWEs excluded.

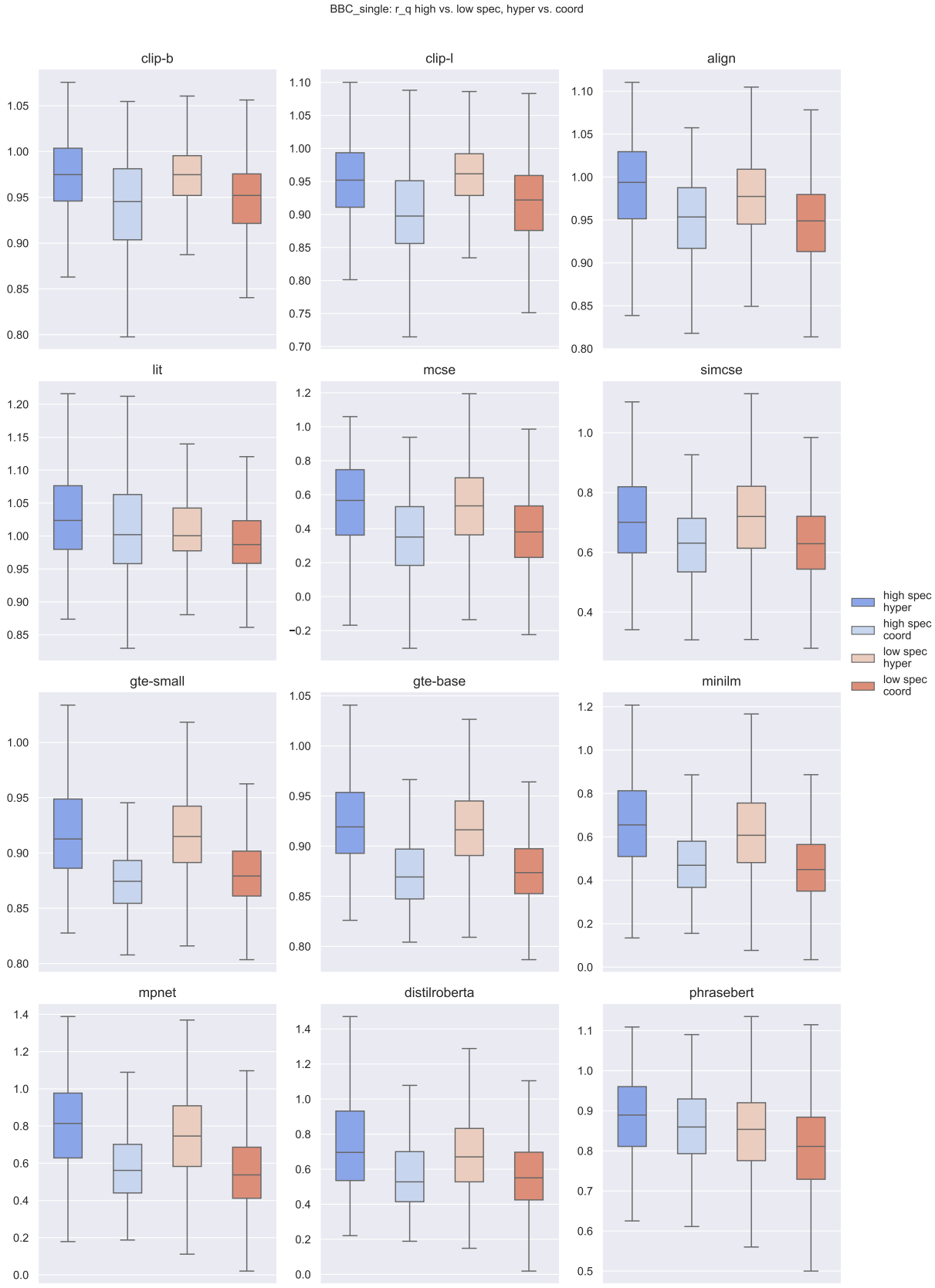

Figure B.7: Boxplots summarizing the distribution of r_q on specific and generic pairs in BBC. MWEs excluded.

While findings on s_q cast doubts on contrastive VLMs' representation of abstract words, as is reflected from the performance gap between contrastive VLMs and sentence transformers in terms of AP, this performance gap is much less drastic in terms of accuracy as concept-specific effects concerning concreteness are factored out (see Table 6.1). Comparing model performance on concrete and abstract subsets further demonstrates that both contrastive VLMs and sentence transformers exhibit a performance drop from concrete to abstract words (Table 6.2). With discussions on the choice of evaluation metrics, we argue from our current findings that abstractness and genericity are not necessarily significantly more challenging for contrastive VLMs than for sentence transformers, and that further investigation is required for reaching a reasonable conclusion.

Appendix C

Experiment 3: including multiword expressions

In Experiment 3, we sample 12343 hypernymy pairs (q, h) and 12343 adversarial coordination pairs (q, c). Focusing on single words, we then report evaluation results on the $\text{BBC}_{\text{single}}$ subset consisting of 7907 pairs per relation. For completeness, here we report results over the full BBC_{full} dataset where multiword expressions are not excluded. Table C.1 reports the models' overall performance as well as correlation with concreteness and specificity scores on the full dataset, and Tables C.2-C.3 report performance on subsets according to concreteness and specificity. It can be seen that r_q with GTE-base is still the best-performing combination overall, and that s_q with CLIP and LiT are sensitive to word concreteness. Comparison between MCSE and SimCSE yields mixed results. As the overall patterns closely mirror those reported in Chapter 6, we provide no further analysis here.

measure	model	AP	Acc	$\text{corr}_{\text{conc}}$	$\text{corr}_{\text{spec}}$
b	CLIP-b	0.6069	0.7094	−0.4124	−0.2072
	CLIP-l	0.6244	0.7319	−0.3091	−0.1468
	ALIGN	0.6260	0.7136	$0.0095^{\dagger}$	$-0.0058^{\dagger}$
	LiT	0.5695	0.6652	−0.4170	−0.1724
	MCSE	0.6047	0.6577	$0.0120^{\dagger}$	0.0619
	SimCSE	0.6125	0.6588	$0.0059^{\dagger}$	0.0439
	GTE-small	0.6438	0.7263	0.0185	0.0349
	GTE-base	0.6578	0.7501	0.0245	0.0897
	MiniLM	0.6274	0.7069	0.1231	0.1300
	MPNet	0.6140	0.6901	0.0813	0.2078
	DistilRoBERTa	0.5823	0.6375	0.0190	0.1258
	PhraseBERT	0.5616	0.6044	0.0691	0.2010
s_q	CLIP-b	0.6192	0.7361	−0.4396	−0.2173
	CLIP-l	0.6389	**0.7830**	−0.3493	−0.1620
	ALIGN	0.6150	0.7209	−0.1226	−0.0869
	LiT	0.5432	0.5974	−0.3088	−0.1745
	MCSE	0.6202	0.7203	0.0177	0.0541
	SimCSE	0.6260	0.6991	0.0516	−0.0313
	GTE-small	0.6912	0.7893	−0.0563	0.0372
	GTE-base	0.6894	0.8005	−0.0434	0.0558
	MiniLM	0.6807	0.7778	−0.0351	0.0681
	MPNet	0.6484	0.7622	0.0728	0.1778
	DistilRoBERTa	0.6206	0.7043	0.0677	0.1302
	PhraseBERT	0.5814	0.6588	−0.1201	0.0172
r_q	CLIP-b	0.6442	0.7181	-0.0137^{*}	−0.0221
	CLIP-l	**0.6616**	0.7475	−0.0700	−0.0678
	ALIGN	0.6593	0.7067	0.1607	0.0300
	LiT	0.5721	0.6476	0.1653	0.0003
	MCSE	0.6396	0.6926	0.0302	0.0539
	SimCSE	0.6450	0.7001	0.0499	−0.0028
	GTE-small	0.7315	0.7932	−0.0304	−0.0343
	GTE-base	**0.7539**	**0.8205**	$-0.0070^{\dagger}$	0.0165
	MiniLM	0.6981	0.7559	0.0729	0.0542
	MPNet	0.7003	0.7682	0.0299	0.1220
	DistilRoBERTa	0.6281	0.6855	−0.0319	0.0617
	PhraseBERT	0.5689	0.6205	0.1088	0.1686
veri	BERT	0.5578	0.6404	0.0433	0.1134
	GPT2	0.7406	0.7889	−0.0248	0.0151^{*}

Table C.1: Model performance on hypernymy versus coordination discrimination. MWEs included. For each model, we report AP and Accuracy scores evaluating the ability to distinguish hypernymy from coordination pairs, based on the measures introduced in Section 3.2. In each column, boldface highlights the best scores obtained by contrastive VLMs and sentence transformers respectively. Statistical significance tests (one-sided Welch's t-test) are conducted over the full dataset, and all differences are statistically significant ($p < 0.01$).

We also report Spearman's rank correlation with concreteness and specificity scores. Statistical significance is assessed using a two-sided permutation test with 10,000 permutations. †denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

		conc+		conc-		spec+		spec-	
		AP	Acc	AP	Acc	AP	Acc	AP	Acc
	#	8405		3938		364		11979	
b	CLIP-b	0.6239	0.7255	0.5997	0.6750	0.6265	0.7115	0.6067	0.7093
	CLIP-l	0.6391	0.7459	0.6149	0.7021	0.6351	0.7198	0.6243	0.7323
	ALIGN	0.6390	0.7322	0.5977	0.6739	0.6512	0.7115	0.6255	0.7137
	LiT	0.5985	0.6990	0.5438	0.5932	0.5963	0.6648	0.5688	0.6652
	MCSE	0.6034	0.6613	0.6090	0.6501	0.6132	0.6648	0.6045	0.6575
	SimCSE	0.6118	0.6632	0.6151	0.6493	0.6160	0.6676	0.6124	0.6585
	GTE-small	0.6483	0.7366	0.6352	0.7044	0.6632	0.7033	0.6433	0.7270
	GTE-base	0.6605	0.7615	0.6522	0.7257	0.6708	0.7473	0.6576	0.7501
	MiniLM	0.6286	0.7137	0.6289	0.6922	0.6497	0.7280	0.6270	0.7062
	MPNet	0.6136	0.6926	0.6170	0.6849	0.6267	0.6703	0.6141	0.6907
	DistilRoBERTa	0.5849	0.6420	0.5776	0.6280	0.5731	0.6401	0.5829	0.6374
	PhraseBERT	0.5513	0.5858	0.5852	0.6440	0.5507	0.5962	0.5623	0.6046
s_q	CLIP-b	0.6471	0.7469	0.6134	0.7131	0.6426	0.7253	0.6189	0.7365
	CLIP-l	**0.6848**	**0.8007**	0.6162	**0.7450**	0.6654	**0.7665**	0.6384	**0.7835**
	ALIGN	0.6337	0.7379	0.5864	0.6846	0.6195	0.7390	0.6149	0.7203
	LiT	0.5512	0.6077	0.5390	0.5754	0.5463	0.6016	0.5433	0.5973
	MCSE	0.6227	0.7299	0.6151	0.6998	0.6175	0.7170	0.6205	0.7204
	SimCSE	0.6370	0.7095	0.6039	0.6770	0.5872	0.6126	0.6273	0.7017
	GTE-small	0.7103	0.8055	0.6549	0.7547	0.6807	0.7500	0.6916	0.7905
	GTE-base	0.7042	0.8150	0.6643	0.7697	0.7144	0.8187	0.6888	0.8000
	MiniLM	0.6953	0.7847	0.6547	0.7631	0.6804	0.7610	0.6808	0.7783
	MPNet	0.6564	0.7661	0.6331	0.7539	0.6463	0.7225	0.6486	0.7634
	DistilRoBERTa	0.6299	0.7141	0.6024	0.6833	0.6071	0.6841	0.6214	0.7049
	PhraseBERT	0.5808	0.6581	0.5849	0.6602	0.5332*	0.5989	0.5834	0.6606
r_q	CLIP-b	0.6488	0.7268	0.6347	0.6993	0.6700	0.7060	0.6435	0.7184
	CLIP-l	0.6643	0.7525	**0.6574**	0.7369	0.6657	0.7527	**0.6616**	0.7474
	ALIGN	0.6774	0.7286	0.6127	0.6600	**0.6973**	0.7363	0.6582	0.7058
	LiT	0.5805	0.6702	0.5475	0.5993	0.5497	0.6236	0.5730	0.6483
	MCSE	0.6425	0.7002	0.6347	0.6765	0.6627	0.7099	0.6390	0.6921
	SimCSE	0.6499	0.7066	0.6353	0.6861	0.6287	0.6648	0.6456	0.7011
	GTE-small	0.7410	0.8074	0.7112	0.7628	0.7354	0.7775	0.7313	0.7936
	GTE-base	**0.7641**	**0.8338**	**0.7320**	**0.7920**	**0.7466**	**0.8324**	**0.7543**	**0.8201**
	MiniLM	0.6997	0.7601	0.6970	0.7466	0.7154	0.7473	0.6978	0.7562
	MPNet	0.7000	0.7682	0.7019	0.7682	0.7179	0.7527	0.6999	0.7687
	DistilRoBERTa	0.6361	0.6954	0.6128	0.6651	0.6497	0.7060	0.6275	0.6846
	PhraseBERT	0.5558	0.6029	0.6038	0.6582	0.5602	0.5797	0.5697	0.6218
veri	BERT	0.5655	0.6482	0.5419	0.6237	0.5304	0.6291	0.5590	0.6407
	GPT2	0.7413	0.7860	0.7393	0.7953	0.7197	0.7720	0.7414	0.7895

Table C.2: Model performance on the high/low concreteness/specificity subsets. MWEs included. In each column, boldface highlights the best scores obtained by contrastive VLMs and sentence transformers respectively.
Statistical significance tests (one-sided Welch's t-test) are conducted for each subset. †alongside APs denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

		conc+ spec+		conc+ spec-		conc- spec+		conc- spec-	
		AP	Acc	AP	Acc	AP	Acc	AP	Acc
	#	345		8060		19		3919	
b	CLIP-b	0.6375	0.7159	0.6235	0.7259	0.5660†	0.6316	0.6002	0.6752
	CLIP-l	0.6408	0.7217	0.6392	0.7469	0.6350†	0.6842	0.6153	0.7022
	ALIGN	0.6556	0.7130	0.6387	0.7330	0.5886†	0.6842	0.5980	0.6739
	LiT	0.6047	0.6754	0.5982	0.7000	0.5263†	0.4737	0.5439	0.5938
	MCSE	0.6174	0.6667	0.6030	0.6610	0.6222†	0.6316	0.6094	0.6502
	SimCSE	0.6230	0.6725	0.6113	0.6628	0.5722†	0.5789	0.6159	0.6497
	GTE-small	0.6682	0.7101	0.6474	0.7377	0.5948†	0.5789	0.6356	0.7050
	GTE-base	0.6808	0.7565	0.6599	0.7617	0.5598†	0.5789	0.6532	0.7265
	MiniLM	0.6563	0.7333	0.6275	0.7129	0.5539†	0.6316	0.6298	0.6925
	MPNet	0.6293	0.6754	0.6134	0.6933	0.5990†	0.5789	0.6174	0.6854
	DistilRoBERTa	0.5744	0.6319	0.5857	0.6424	0.6136†	**0.7895**	0.5780	0.6272
	PhraseBERT	0.5565	0.6000	0.5513	0.5852	0.5094†	0.5263	0.5862	0.6446
s_q	CLIP-b	0.6519	0.7246	0.6471	0.7479	0.6516†	0.7368	0.6132	0.7129
	CLIP-l	0.6748	**0.7739**	**0.6852**	**0.8019**	0.6744†	0.6316	0.6158	**0.7456**
	ALIGN	0.6210	0.7420	0.6346	0.7377	0.6352†	0.6842	0.5862	0.6846
	LiT	0.5603	0.6174	0.5511	0.6073	0.4490†	0.3158	0.5395	0.5767
	MCSE	0.6148	0.7130	0.6234	0.7306	0.6875†	0.7895	0.6147	0.6994
	SimCSE	0.5848	0.6116	0.6392	0.7136	0.6506†	0.6316	0.6035	0.6772
	GTE-small	0.6870	0.7478	0.7113	0.8079	0.6707†	**0.7895**	0.6550	0.7545
	GTE-base	0.7291	0.8232	0.7034	0.8146	0.6095†	0.7368	0.6647	0.7698
	MiniLM	0.6819	0.7594	0.6964	0.7857	0.7026†	**0.7895**	0.6545	0.7629
	MPNet	0.6482	0.7188	0.6569	0.7681	0.6794†	**0.7895**	0.6332	0.7538
	DistilRoBERTa	0.6028	0.6783	0.6316	0.7156	0.6791†	**0.7895**	0.6021	0.6828
	PhraseBERT	0.5303*	0.5913	0.5839	0.6609	0.6135†	0.7368	0.5849	0.6599
r_q	CLIP-b	0.6767	0.7159	0.6477	0.7273	0.5576†	0.5263	0.6357	0.7002
	CLIP-l	0.6713	0.7478	0.6642	0.7527	**0.6941**†	**0.8421**	**0.6580**	0.7364
	ALIGN	**0.7045**	0.7449	0.6764	0.7279	0.5796†	0.5789	0.6132	0.6604
	LiT	0.5552	0.6377	0.5818	0.6716	0.4492†	0.3684	0.5482	0.6004
	MCSE	0.6646	0.7130	0.6418	0.6996	0.6998†	0.6316	0.6347	0.6767
	SimCSE	0.6300	0.6696	0.6508	0.7082	0.6888†	0.5789	0.6355	0.6867
	GTE-small	0.7385	0.7855	0.7410	0.8083	0.7161*	0.6316	0.7115	0.7635
	GTE-base	**0.7546**	**0.8377**	**0.7646**	**0.8336**	0.6886*	0.7368	**0.7328**	**0.7923**
	MiniLM	0.7168	0.7507	0.6991	0.7605	0.7349*	0.6842	0.6973	0.7471
	MPNet	0.7163	0.7536	0.6994	0.7689	**0.7556***	0.7368	0.7018	0.7683
	DistilRoBERTa	0.6454	0.7014	0.6359	0.6948	0.7216*	**0.7895**	0.6124	0.6645
	PhraseBERT	0.5624	0.5797	0.5560	0.6038	0.5403†	0.5789	0.6043	0.6586
veri	BERT	0.5327	0.6435	0.5677	0.6484	0.5220†	0.3684	0.5424	0.6249
	GPT2	0.7203	0.7768	0.7426	0.7864	0.7259*	0.6842	0.7394	0.7959

Table C.3: Model performance on subsets w.r.t. both concreteness and specificity. MWEs included. In each column, boldface highlights the best scores obtained by contrastive VLMs and sentence transformers respectively.
Statistical significance tests (one-sided Welch's t-test) are conducted for each subset. †denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

Appendix D

Combining measures

In Experiment 1, we observe that our hypernymy measures s_q and r_q effectively distinguish hypernymy from coordination or meronymy pairs, but underperform naïve cosine similarity when it comes to random pairs. We then suggest that combining multiple measures may help accomplish a more robust hypernymy detection approach against different distractor relation types. Inspired by previous unsupervised count-based metrics [Kot+10; LB12] making use of geometric means, we experiment with three composite measures, $b \cdot s_q$, $b \cdot r_q$, and $s_q \cdot r_q$, using the direct multiplication of our previous measures b, s_q and r_q.

Table D.1 and Table D.2 report results on BLESS and HyperLex, the two distributional semantic benchmarks. Overall, combining measures via simple multiplication achieves improved or comparable performance on both hypernymy detection and GLE. The symmetric cosine similarity is especially beneficial for distinguishing hypernymy from random pairs, as expected. However, composite measures do not obtain consistent gains on GLE across concreteness groups with smaller amounts of data. Note that the best-performing combination on the concrete subset (qc pc) is $s_q \cdot r_q$ with LiT, demonstrating the potential of concept representations of concrete words via contrastive VLMs. On our newly-constructed dataset BBC, combining measures sometimes but not always leads to improved performance (see Tables D.4-D.8).

		vs. non		vs. coord		vs. mero		vs. random	
		AP	MAP	AP	MAP	AP	MAP	AP	MAP
	proportion	0.0919	0.0951	0.2727	0.2821	0.3124	0.3626	0.1663	0.1690
	individual	0.4268	**0.5897**	0.6475	0.8109	0.6318	0.7611	0.7997	0.8264
$b \cdot s_q$	CLIP-b	0.2165	0.3498	0.4275	0.5522	0.5057	0.6834	0.4283	0.5880
	CLIP-l	0.2131	0.3440	0.3447	0.4764	0.5286	0.6838	0.5324	0.6968
	ALIGN	0.1701	0.3089	0.2733†	0.3991	0.4391	0.6397	0.5905	0.7389
	LiT	0.1861	0.3528	0.5261	0.6819	0.5287	**0.7678**	0.2778	0.4804
	MCSE	0.1516	0.2316	0.2568†	0.3564	0.3886	0.5263	0.5404	0.5757
	SimCSE	0.1879	0.2921	0.3301	0.4655	0.4041	0.5461	0.5574	0.6028
	GTE-small	0.2543	0.3713	0.3700	0.4887	0.4748	0.6176	0.7696	0.8063
	GTE-base	0.3011	0.4275	0.4045	0.5305	0.5517	0.6641	0.8111	**0.8463**
	MiniLM	0.2106	0.3200	0.3041†	0.4207	0.4516	0.5854	0.7324	0.7606
	MPNet	0.1804	0.2877	0.2487†	0.3677	0.4887	0.6043	0.7441	0.7771
	DistilRoBERTa	0.1712	0.2672	0.2699†	0.3954	0.4315	0.5740	0.5876	0.6300
	PhraseBERT	0.1136	0.1618	0.1854†	0.2450	0.3856†	0.5000	0.6036	0.6597
$b \cdot r_q$	CLIP-b	0.2316	0.3771	0.4396	0.5585	0.4696	0.6286	0.4945	0.6792
	CLIP-l	0.1771	0.3019	0.3325	0.4682	0.4589	0.6068	0.4085	0.5925
	ALIGN	0.2016	0.3542	0.3131	0.4577	0.4554	0.6183	0.6580	0.7775
	LiT	0.1894	0.3853	0.5235	0.7046	0.4482	0.6310	0.3157	0.5460
	MCSE	0.1548	0.2361	0.2663†	0.3662	0.3824	0.5297	0.5338	0.5710
	SimCSE	0.1938	0.2853	0.3423	0.4540	0.4006	0.5400	0.5645	0.6073
	GTE-small	0.2673	0.4084	0.3936	0.5298	0.4786	0.6342	0.7597	0.7993
	GTE-base	0.3533	0.4903	0.4778	0.5984	0.5711	0.6885	**0.8206**	0.8358
	MiniLM	0.2332	0.3508	0.3352	0.4553	0.4625	0.6946	0.7263	0.7560
	MPNet	0.2126	0.3191	0.2829†	0.3980	0.5135	0.6233	0.7633	0.7887
	DistilRoBERTa	0.1955	0.2981	0.3005†	0.4298	0.4668	0.6030	0.6052	0.6498
	PhraseBERT	0.1066	0.1551	0.1804	0.2411	0.3427†	0.4759	0.5771	0.6218
$s_q \cdot r_q$	CLIP-b	0.3084	0.4563	0.6406	0.7707	0.5906	0.6994	0.4462	0.5949
	CLIP-l	0.2697	0.4150	0.5103	0.6600	0.5741	0.7000	0.4382	0.6092
	ALIGN	0.3169	0.4686	0.5578	0.6926	0.5241	0.6453	0.5855	0.6875
	LiT	0.1663	0.2965	**0.7524**	**0.8787**	0.4908	0.6436	0.2153	0.3451
	MCSE	0.1788	0.2680	0.3087†	0.4260	0.4129	0.5447	0.5097	0.5630
	SimCSE	0.2445	0.3750	0.5106	0.6491	0.4844	0.6258	0.4353	0.5523
	GTE-small	0.3879	0.5264	0.5409	0.6755	0.5992	0.7031	0.7510	0.7953
	GTE-base	**0.4620**	0.5865	0.6182	0.7310	**0.6662**	0.7362	0.7534	0.8049
	MiniLM	0.3376	0.4547	0.4628	0.5891	0.5612	0.6511	0.6599	0.6991
	MPNet	0.3230	0.4574	0.4074	0.5515	0.6023	0.6887	0.7210	0.7698
	DistilRoBERTa	0.2805	0.4059	0.4317	0.5797	0.5808	0.6795	0.5684	0.6423
	PhraseBERT	0.1052	0.1494	0.1893†	0.2547	0.3053†	0.4241	0.4880	0.5641

Table D.1: Model performance on binary hypernymy detection.
"Proportion" indicates the proportion of positive pairs, i.e., the AP or MAP of a random baseline.
"Individual" indicates the best performance achieved by an individual measure (b, s_q or r_q).
For each model, we report AP and MAP scores evaluating the ability to distinguish hypernymy pairs from other lexical relations in BLESS, based on the measures introduced in Section 3.2. Boldface highlights the best score in each column.
Statistical significance tests (one-sided Welch's t-test) are conducted over the full dataset. †alongside APs denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

		all	qc pc	qc pa	qa pc	qa pa
	#	2163	172	89	119	1055
	individual	0.4967	0.5092	**0.5225**	**0.6561**	0.3189
$b \cdot s_q$	CLIP-b	0.2238	0.2029	0.3867	0.4815	$-0.0264^{\dagger}$
	CLIP-l	0.2632	0.2473	0.3004	0.5597	$0.0553^{\dagger}$
	ALIGN	0.2773	0.2563	0.3502	0.5724	$-0.0085^{\dagger}$
	LiT	0.2170	0.2440	0.2297^{*}	0.3674	$0.1001^{\dagger}$
	MCSE	0.1966	0.1375	0.3909	0.4380	$0.1125^{\dagger}$
	SimCSE	0.2770	0.2083	0.4724	0.4436	0.2638
	GTE-small	0.3253	0.2859	0.3835	0.5472	0.2952
	GTE-base	0.3799	0.3674	0.4010	0.6069	0.2646
	MiniLM	0.2957	0.2720	0.3846	0.5020	$0.0890^{\dagger}$
	MPNet	0.3016	0.2906	0.4573	0.4786	0.1594^{*}
	DistilRoBERTa	0.2303	0.1999	0.4035	0.5712	$0.0435^{\dagger}$
	PhraseBERT	0.1965	0.1360	0.3094	0.5933	$0.0896^{\dagger}$
$b \cdot r_q$	CLIP-b	0.2959	0.2717	0.3853	0.4936	$0.0082^{\dagger}$
	CLIP-l	0.2621	0.2274	0.3954	0.5646	$0.0470^{\dagger}$
	ALIGN	0.3042	0.2783	0.3790	0.5770	$0.0219^{\dagger}$
	LiT	0.2819	0.3319	0.2779	0.4514	$-0.0333^{\dagger}$
	MCSE	0.2022	0.1414	0.4052	0.4340	$0.0994^{\dagger}$
	SimCSE	0.2678	0.2019	0.4633	0.4274	0.2388
	GTE-small	0.3364	0.2939	0.4348	0.5888	0.2685
	GTE-base	0.4114	0.3993	0.4873	0.6327	0.2741
	MiniLM	0.2895	0.2655	0.4017	0.4930	$0.0640^{\dagger}$
	MPNet	0.3183	0.2999	0.4918	0.5025	0.1936^{*}
	DistilRoBERTa	0.2388	0.2080	0.3861	0.5674	$0.1068^{\dagger}$
	PhraseBERT	0.1586	0.0837	0.3071	0.5512	$0.0510^{\dagger}$
$s_q \cdot r_q$	CLIP-b	0.4332	0.4730	0.3307	0.5351	$0.0793^{\dagger}$
	CLIP-l	0.3980	0.4134	0.3436	0.6261	$0.1390^{\dagger}$
	ALIGN	0.4278	0.4614	0.2907	0.6414	$0.0779^{\dagger}$
	LiT	0.4374	**0.5482**	$0.1118^{\dagger}$	0.4968	$0.0050^{\dagger}$
	MCSE	0.2746	0.2314	0.3954	0.5346	0.1698^{*}
	SimCSE	0.3682	0.3335	0.4252	0.5486	0.3164
	GTE-small	0.4549	0.4604	0.3966	0.6064	**0.3242**
	GTE-base	**0.5154**	0.5421	0.4525	0.6550	0.2943
	MiniLM	0.4079	0.4145	0.3619	0.5734	$0.1437^{\dagger}$
	MPNet	0.4604	0.4776	0.4695	0.5803	0.2953
	DistilRoBERTa	0.3184	0.3003	0.3341	0.6234	$0.1187^{\dagger}$
	PhraseBERT	0.1989	0.1144	0.3167	0.5908	0.1635^{*}

Table D.2: Model performance on graded lexical entailment.
"Individual" indicates the best performance achieved by an individual measure (b, s_q or r_q).
For each model, we report Spearman's rank correlation between its similarity-based measures and the HyperLex GLE ratings. Boldface highlights the best score in each column.
Statistical significance is assessed using a two-sided permutation test with 10,000 permutations. †denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

measure	model	AP	Acc	$\text{corr}_{\text{conc}}$	$\text{corr}_{\text{spec}}$
individual	(multimodal)	0.6601	0.7808		
individual	(unimodal)	**0.7619**	0.8246		
$b \cdot s_q$	CLIP-b	0.6200	0.7357	−0.4749	−0.2300
	CLIP-l	0.6395	0.7705	−0.3862	−0.1862
	ALIGN	0.6469	0.7430	−0.0841	−0.0730
	LiT	0.5526	0.6292	−0.4269	−0.1917
	MCSE	0.6323	0.6829	$-0.0261^{\dagger}$	0.0527
	SimCSE	0.6465	0.6942	0.0068	0.0186^{*}
	GTE-small	0.6811	0.7553	−0.0300	0.0352
	GTE-base	0.6910	0.7784	−0.0393	0.0770
	MiniLM	0.6631	0.7382	0.0714	0.1212
	MPNet	0.6421	0.7245	0.0310	0.2010
	DistilRoBERTa	0.6025	0.6621	$-0.0009^{\dagger}$	0.1413
	PhraseBERT	0.5842	0.6416	−0.0418	0.1352
$b \cdot r_q$	CLIP-b	0.6429	0.7199	−0.2655	−0.1426
	CLIP-l	0.6605	0.7616	−0.2317	−0.1340
	ALIGN	0.6606	0.7345	0.0650	$-0.0077^{\dagger}$
	LiT	0.5820	0.6547	−0.1390	−0.0877
	MCSE	0.6400	0.6891	$-0.0087^{\dagger}$	0.0516
	SimCSE	0.6501	0.6941	$0.0132^{\dagger}$	0.0192^{*}
	GTE-small	0.7052	0.7744	−0.0236	$-0.0112^{\dagger}$
	GTE-base	0.7226	0.8025	−0.0221	0.0391
	MiniLM	0.6816	0.7471	0.0806	0.0828
	MPNet	0.6751	0.7546	$0.0100^{\dagger}$	0.1539
	DistilRoBERTa	0.6167	0.6740	−0.0367	0.1032
	PhraseBERT	0.5817	0.6330	0.0303	0.1625
$s_q \cdot r_q$	CLIP-b	0.6759	0.7450	−0.2019	−0.1147
	CLIP-l	**0.6944**	**0.7855**	−0.2074	−0.1296
	ALIGN	0.6662	0.7313	0.0546	−0.0216
	LiT	0.5711	0.6365	-0.0188^{*}	−0.0836
	MCSE	0.6569	0.7090	$0.0105^{\dagger}$	0.0557
	SimCSE	0.6523	0.6975	0.0683	$-0.0035^{\dagger}$
	GTE-small	0.7478	0.8066	−0.0353	$-0.0128^{\dagger}$
	GTE-base	0.7616	**0.8280**	−0.0324	0.0238
	MiniLM	0.7183	0.7731	0.0503	0.0543
	MPNet	0.7202	0.7941	0.0166^{*}	0.1298
	DistilRoBERTa	0.6408	0.6996	-0.0169^{*}	0.0913
	PhraseBERT	0.5932	0.6526	$0.0149^{\dagger}$	0.1136

Table D.3: Model performance on hypernymy versus coordination discrimination. MWEs excluded. For each model, we report AP and Accuracy scores evaluating the ability to distinguish hypernymy from coordination pairs, based on the measures introduced in Section 3.2. "Individual" indicates the best performance achieved by an individual measure (b, s_q or r_q). In each column, boldface highlights the best scores obtained by contrastive VLMs and sentence transformers respectively. Statistical significance tests (one-sided Welch's t-test) are conducted over the full dataset, and all differences are statistically significant ($p < 0.01$).

We also report Spearman's rank correlation with concreteness and specificity scores. Statistical significance is assessed using a two-sided permutation test with 10,000 permutations. †denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

		conc+		conc-		spec+		spec-	
		AP	Acc	AP	Acc	AP	Acc	AP	Acc
	#	5021		2886		194		7713	
individual (multimodal)		0.6898	**.8036**	0.6579	0.7412	0.6602	0.7835	**0.7221**	0.7815
individual (unimodal)		**0.7767**	0.8423	**0.7345**	0.7938	**0.7619**	**0.8505**	**0.7765**	0.8241
$b \cdot s_q$	CLIP-b	0.6537	0.7566	0.6106	0.6992	0.6198	0.6856	0.6441	0.7369
	CLIP-l	0.6757	0.7933	0.6240	0.7308	0.6393	0.7320	0.6512	0.7714
	ALIGN	0.6724	0.7642	0.6084	0.7062	0.6474	0.6804	0.6393	0.7446
	LiT	0.5832	0.6632	0.5345	0.5700	0.5521	0.6186	0.5933	0.6295
	MCSE	0.6362	0.6881	0.6265	0.6739	0.6317	0.6959	0.6594	0.6826
	SimCSE	0.6557	0.7056	0.6310	0.6743	0.6463	0.6598	0.6564	0.6951
	GTE-small	0.6939	0.7678	0.6589	0.7335	0.6803	0.7526	0.7060	0.7553
	GTE-base	0.6997	0.7925	0.6766	0.7540	0.6907	0.8041	0.7104	0.7778
	MiniLM	0.6713	0.7506	0.6497	0.7166	0.6618	0.7990	0.7163	0.7367
	MPNet	0.6457	0.7279	0.6362	0.7186	0.6416	0.7216	0.6678	0.7246
	DistilRoBERTa	0.6085	0.6676	0.5913	0.6525	0.6024	0.6598	0.6077	0.6621
	PhraseBERT	0.5781	0.6272	0.5953	0.6667	0.5853	0.6031	0.5584	0.6429
$b \cdot r_q$	CLIP-b	0.6597	0.7383	0.6286	0.6878	0.6425	0.7062	0.6650	0.7202
	CLIP-l	0.6747	0.7749	0.6488	0.7384	0.6605	0.7320	0.6679	0.7623
	ALIGN	0.6838	0.7584	0.6141	0.6930	0.6599	0.6907	0.6927	0.7356
	LiT	0.6047	0.6907	0.5447	0.5922	0.5823	0.6495	0.5793	0.6549
	MCSE	0.6448	0.6969	0.6323	0.6757	0.6393	0.6753	0.6712	0.6895
	SimCSE	0.6588	0.7060	0.6347	0.6733	0.6497	0.6804	0.6630	0.6944
	GTE-small	0.7165	0.7897	0.6855	0.7477	0.7045	0.7887	0.7257	0.7740
	GTE-base	0.7327	0.8192	0.7047	0.7734	0.7226	0.8299	0.7260	0.8018
	MiniLM	0.6880	0.7564	0.6717	0.7308	0.6802	0.7990	0.7333	0.7458
	MPNet	0.6787	0.7600	0.6691	0.7453	0.6739	0.7577	0.7257	0.7546
	DistilRoBERTa	0.6238	0.6819	0.6043	0.6601	0.6161	0.7062	0.6448	0.6731
	PhraseBERT	0.5708	0.6144	0.6022	0.6653	0.5821	0.5979	0.5815	0.6339
$s_q \cdot r_q$	CLIP-b	0.6902	0.7594	0.6066	0.7200	0.6760	0.7268	0.6768	0.7455
	CLIP-l	**0.7154**	0.8000	**0.6719**	**0.7602**	**0.6947**	0.7732	0.6840	**0.7858**
	ALIGN	0.6929	0.7588	0.6116	0.6833	0.6655	**0.7990**	0.6953	0.7295
	LiT	0.5813	0.6584	0.5488	0.5984	0.5714	0.6546	0.5627	0.6361
	MCSE	0.6648	0.7206	0.6425	0.6888	0.6559	0.7577	0.6957	0.7078
	SimCSE	0.6660	0.7102	0.6286	0.6753	0.6524	0.6546	0.6497	0.6986
	GTE-small	0.7639	0.8257	0.7182	0.7734	0.7473	0.8144	0.7667	0.8064
	GTE-base	0.7766	**0.8441**	0.7344	**0.8001**	0.7615	0.8351	0.7715	**0.8278**
	MiniLM	0.7277	0.7795	0.7009	0.7620	0.7173	0.7887	0.7574	0.7727
	MPNet	0.7274	0.7965	0.7077	0.7900	0.7190	0.8041	0.7656	0.7939
	DistilRoBERTa	0.6542	0.7120	0.6183	0.6781	0.6395	0.7835	0.6939	0.6975
	PhraseBERT	0.5831	0.6375	0.6116	0.6788	0.5949*	0.6082	0.5503	0.6537

Table D.4: Model performance on the high/low concreteness/specificity subsets. MWEs excluded. "Individual" indicates the best performance achieved by an individual measure (b, s_q or r_q). In each column, boldface highlights the best scores obtained by contrastive VLMs and sentence transformers respectively.
Statistical significance tests (one-sided Welch's t-test) are conducted for each subset. †alongside APs denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

		conc+ spec+		conc+ spec-		conc- spec+		conc- spec-	
		AP	Acc	AP	Acc	AP	Acc	AP	Acc
	#	179		4842		15		2871	
individual (multimodal)		**0.7355**	0.7877	0.6904	**0.8052**	0.7269†	**0.8667**	0.6582	0.7416
individual (unimodal)		0.7816	**0.8547**	**0.7767**	0.8418	**0.7732***	**0.8667**	**0.7358**	0.7941
$b \cdot s_q$	CLIP-b	0.6646	0.6983	0.6533	0.7588	0.5760†	0.5333	0.6111	0.7001
	CLIP-l	0.6679	0.7318	0.6760	0.7955	0.6551†	0.7333	0.6242	0.7308
	ALIGN	0.6486	0.6816	0.6740	0.7672	0.5904†	0.6667	0.6089	0.7064
	LiT	0.6105	0.6480	0.5825	0.6638	0.5074†	0.2667	0.5345	0.5716
	MCSE	0.6692	0.7039	0.6353	0.6875	0.6554†	0.6000	0.6268	0.6743
	SimCSE	0.6733	0.6648	0.6551	0.7071	0.5738†	0.6000	0.6319	0.6747
	GTE-small	0.7184	0.7709	0.6927	0.7677	0.6141†	0.5333	0.6596	0.7346
	GTE-base	0.7339	0.8268	0.6986	0.7912	0.5625†	0.5333	0.6781	0.7551
	MiniLM	0.7311	0.8156	0.6691	0.7482	0.5855†	0.6000	0.6506	0.7172
	MPNet	0.6783	0.7318	0.6445	0.7278	0.6034†	0.6000	0.6369	0.7193
	DistilRoBERTa	0.6102	0.6480	0.6085	0.6683	0.6127†	0.8000	0.5916	0.6517
	PhraseBERT	0.5739	0.6145	0.5788	0.6276	0.4997†	0.4667	0.5968	0.6677
$b \cdot r_q$	CLIP-b	0.6833	0.7151	0.6589	0.7392	0.5792†	0.6000	0.6295	0.6883
	CLIP-l	0.6751	0.7207	0.6748	0.7770	0.6892†	**0.8667**	0.6488	0.7377
	ALIGN	0.7040	0.6927	0.6833	0.7608	0.5796†	0.6667	0.6147	0.6931
	LiT	0.5889	0.6704	0.6059	0.6914	0.4879†	0.4000	0.5451	0.5932
	MCSE	0.6791	0.6917	0.6438	0.6970	0.6402†	0.4667	0.6325	0.6768
	SimCSE	0.6787	0.6872	0.6581	0.7067	0.5890†	0.6000	0.6354	0.6736
	GTE-small	0.7349	0.8045	0.7155	0.7891	0.6208†	0.6000	0.6862	0.7485
	GTE-base	0.7457	0.8492	0.7323	0.8181	0.5670†	0.6000	0.7062	0.7743
	MiniLM	0.7427	0.8101	0.6858	0.7544	0.6418†	0.6667	0.6724	0.7311
	MPNet	0.7324	0.7654	0.6765	0.7598	0.6849†	0.6667	0.6694	0.7457
	DistilRoBERTa	0.6466	0.6983	0.6231	0.6813	0.6795†	0.8000	0.6044	0.6594
	PhraseBERT	0.5965	0.6145	0.5701	0.6144	0.4978†	0.4000	0.6036	0.6667
$s_q \cdot r_q$	CLIP-b	0.6871	0.7318	0.6904	0.7604	0.6959†	0.6667	0.6609	0.7203
	CLIP-l	0.6895	0.7765	**0.7166**	0.8009	**0.7310**†	0.7333	**0.6716**	**0.7604**
	ALIGN	0.7115	**0.8156**	0.6925	0.7567	0.6404†	0.6000	0.6122	0.6837
	LiT	0.5698	0.6704	0.5820	0.6580	0.4474†	0.4667	0.5494	0.5991
	MCSE	0.7003	0.7654	0.6636	0.7189	0.7223†	0.6667	0.6424	0.6890
	SimCSE	0.6560	0.6592	0.6666	0.7121	0.6759†	0.6000	0.6285	0.6757
	GTE-small	0.7787	0.8324	0.7634	0.8255	0.6683†	0.6000	0.7188	0.7743
	GTE-base	**0.7930**	0.8492	0.7762	**0.8439**	0.6231*	0.6667	0.7356	**0.8008**
	MiniLM	0.7607	0.7933	0.7264	0.7790	0.7369*	0.7330	0.7011	0.7621
	MPNet	0.7685	0.7989	0.7259	0.7964	0.7626*	**0.8667**	0.7076	0.7896
	DistilRoBERTa	0.6868	0.7821	0.6531	0.7094	0.7667*	0.8000	0.6175	0.6775
	PhraseBERT	0.5561	0.6145	0.5848	0.6384	0.5215†	0.5333	0.6128	0.6796

Table D.5: Model performance on subsets w.r.t. both concreteness and specificity. MWEs excluded. "Individual" indicates the best performance achieved by an individual measure (b, s_q or r_q). In each column, boldface highlights the best scores obtained by contrastive VLMs and sentence transformers respectively.

Statistical significance tests (one-sided Welch's t-test) are conducted for each subset. †denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

measure	model	AP	Acc	$\text{corr}_{\text{conc}}$	$\text{corr}_{\text{spec}}$
individual	(multimodal)	0.6616	**0.7830**		
individual	(unimodal)	0.7539	0.8205		
$b \cdot s_q$	CLIP-b	0.6189	0.7401	−0.4639	−0.2304
	CLIP-l	0.6396	0.7744	−0.3522	−0.1657
	ALIGN	0.6380	0.7373	−0.0450	−0.0395
	LiT	0.5592	0.6472	−0.4046	−0.1924
	MCSE	0.6145	0.6744	$0.0101^{\dagger}$	0.0632
	SimCSE	0.6321	0.6953	0.0249	0.0213
	GTE-small	0.6702	0.7586	$0.0017^{\dagger}$	0.0479
	GTE-base	0.6813	0.7822	$0.0099^{\dagger}$	0.1003
	MiniLM	0.6454	0.7330	0.1034	0.1330
	MPNet	0.6157	0.7136	0.0890	0.2199
	DistilRoBERTa	0.5941	0.6602	0.0330	0.1391
	PhraseBERT	0.5726	0.6287	0.0210	0.1650
$b \cdot r_q$	CLIP-b	0.6406	0.7249	−0.2706	−0.1399
	CLIP-l	0.6572	0.7612	−0.2282	−0.1221
	ALIGN	0.6553	0.7300	0.0793	$0.0085^{\dagger}$
	LiT	0.5888	0.6689	−0.1355	−0.0978
	MCSE	0.6197	0.6745	0.0228	0.0594
	SimCSE	0.6332	0.6093	0.0278	0.0219
	GTE-small	0.6912	0.7714	$-0.0067^{\dagger}$	$-0.0001^{\dagger}$
	GTE-base	0.7104	0.8032	$0.0089^{\dagger}$	0.0557
	MiniLM	0.6618	0.7382	0.1020	0.0970
	MPNet	0.6540	0.7386	0.0598	0.1743
	DistilRoBERTa	0.6082	0.6711	$-0.0056^{\dagger}$	0.1013
	PhraseBERT	0.5666	0.6151	0.0877	0.1930
$s_q \cdot r_q$	CLIP-b	0.6739	0.7453	−0.2219	−0.1180
	CLIP-l	**0.6921**	0.7801	−0.2096	−0.1202
	ALIGN	0.6666	0.7219	0.0470	−0.0166
	LiT	0.5848	0.6556	−0.0475	−0.1007
	MCSE	0.6419	0.7033	0.0252	0.0557
	SimCSE	0.6470	0.7079	0.0544	$-0.0123^{\dagger}$
	GTE-small	0.7391	0.8004	−0.0375	$-0.0099^{\dagger}$
	GTE-base	**0.7564**	**0.8259**	-0.0160^{*}	0.0373
	MiniLM	0.7040	0.7641	0.0548	0.0654
	MPNet	0.7016	0.7753	0.0443	0.1450
	DistilRoBERTa	0.6382	0.7017	$-0.0098^{\dagger}$	0.0834
	PhraseBERT	0.5824	0.6352	0.0388	0.1279

Table D.6: Model performance on hypernymy versus coordination discrimination. MWEs included. For each model, we report AP and Accuracy scores evaluating the ability to distinguish hypernymy from coordination pairs, based on the measures introduced in Section 3.2. "Individual" indicates the best performance achieved by an individual measure (b, s_q or r_q). In each column, boldface highlights the best scores obtained by contrastive VLMs and sentence transformers respectively. Statistical significance tests (one-sided Welch's t-test) are conducted over the full dataset, and all differences are statistically significant ($p < 0.01$).
We also report Spearman's rank correlation with concreteness and specificity scores. Statistical significance is assessed using a two-sided permutation test with 10,000 permutations. †denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.

		conc+		conc-		spec+		spec-	
		AP	Acc	AP	Acc	AP	Acc	AP	Acc
	#	8405		3938		364		11979	
individual (multimodal)		0.6848	**0.8007**	0.6574	0.7450	**0.6973**	**0.7665**	0.6616	**0.7835**
individual (unimodal)		0.7641	0.8338	**0.7320**	0.7920	0.7466	**0.8324**	0.7543	0.8201
$b \cdot s_q$	CLIP-b	0.6438	0.7553	0.6132	0.7077	0.6341	0.7170	0.6188	0.7408
	CLIP-l	0.6668	0.7914	0.6247	0.7382	0.6524	0.7445	0.6394	0.7754
	ALIGN	0.6563	0.7549	0.6047	0.6998	0.6447	0.7060	0.6388	0.7383
	LiT	0.5851	0.6742	0.5403	0.5896	0.6019	0.6484	0.5585	0.6472
	MCSE	0.6140	0.6792	0.6165	0.6640	0.6248	0.6786	0.6143	0.6743
	SimCSE	0.6342	0.7010	0.6285	0.6831	0.6198	0.6566	0.6325	0.6965
	GTE-small	0.6778	0.7687	0.6540	0.7369	0.6806	0.7363	0.6699	0.7592
	GTE-base	0.6862	0.7921	0.6707	0.7610	0.6963	0.7830	0.6810	0.7822
	MiniLM	0.6482	0.7413	0.6425	0.7151	0.6672	0.7555	0.6451	0.7323
	MPNet	0.6268	0.7165	0.6252	0.7075	0.6322	0.6978	0.6259	0.7141
	DistilRoBERTa	0.5975	0.6645	0.5870	0.6511	0.5750	0.6401	0.5949	0.6608
	PhraseBERT	0.5645	0.6145	0.5894	0.6590	0.5499	0.5934	0.5738	0.6298
$b \cdot r_q$	CLIP-b	0.6544	0.7386	0.6262	0.6958	0.6710	0.7253	0.6397	0.7249
	CLIP-l	0.6678	0.7714	0.6473	0.7395	0.6588	0.7473	0.6574	0.7617
	ALIGN	0.6723	0.7503	0.6137	0.6869	0.6881	0.7088	0.6544	0.7307
	LiT	0.6054	0.6992	0.5553	0.6041	0.5804	0.6401	0.5892	0.6698
	MCSE	0.6199	0.6802	0.6209	0.6623	0.6327	0.6676	0.6194	0.6747
	SimCSE	0.6343	0.6949	0.6316	0.6803	0.6258	0.6676	0.6336	0.6910
	GTE-small	0.6975	0.7832	0.6785	0.7463	0.6999	0.7582	0.6909	0.7719
	GTE-base	0.7160	0.8152	0.6986	0.7776	0.7098	0.8077	0.7106	0.8031
	MiniLM	0.6628	0.7447	0.6637	0.7245	0.6801	0.7445	0.6615	0.7380
	MPNet	0.6534	0.7409	0.6570	0.7336	0.6738	0.7170	0.6539	0.7392
	DistilRoBERTa	0.6118	0.6766	0.6006	0.6595	0.6032	0.6786	0.6086	0.6709
	PhraseBERT	0.5552	0.5964	0.5946	0.6549	0.5580	0.5714	0.5674	0.6164
$s_q \cdot r_q$	CLIP-b	0.6875	0.7559	0.6556	0.7227	0.6855	0.7280	0.6736	0.7458
	CLIP-l	**0.7091**	0.7893	**0.6708**	**0.7605**	0.6892	0.7610	**0.6923**	0.7807
	ALIGN	0.6881	0.7422	0.6130	0.6785	0.6779	0.7527	0.6663	0.7209
	LiT	0.5947	0.6758	0.5598	0.6125	0.5657	0.6538	0.5856	0.6556
	MCSE	0.6453	0.7127	0.6350	0.6833	0.6620	0.7253	0.6414	0.7026
	SimCSE	0.6543	0.7151	0.6317	0.6927	0.6243	0.6511	0.6477	0.7097
	GTE-small	0.7510	0.8139	0.7132	0.7715	0.7394	0.7720	0.7391	0.8012
	GTE-base	**0.7680**	**0.8389**	0.7310	**0.7981**	**0.7580**	0.8187	**0.7564**	**0.8261**
	MiniLM	0.7080	0.7693	0.6961	0.7529	0.7175	0.7500	0.7038	0.7645
	MPNet	0.7043	0.7767	0.6966	0.7722	0.7197	0.7473	0.7013	0.7761
	DistilRoBERTa	0.6475	0.7102	0.6194	0.6836	0.6537	0.7225	0.6380	0.7011
	PhraseBERT	0.5727	0.6227	0.6043	0.6618	0.5551	0.5907	0.5837	0.6365

Table D.7: Model performance on the high/low concreteness/specificity subsets. MWEs included. "Individual" indicates the best performance achieved by an individual measure (b, s_q or r_q). In each column, boldface highlights the best scores obtained by contrastive VLMs and sentence transformers respectively.
Statistical significance tests (one-sided Welch's t-test) are conducted for each subset. All results are statistically significant (p-value < 0.01).

		conc+ spec+		conc+ spec-		conc- spec+		conc- spec-	
		AP	Acc	AP	Acc	AP	Acc	AP	Acc
	#	345		8060		19		3919	
individual (multimodal)		**0.7045**	**0.7739**	0.6852	**0.8019**	0.6941†	**0.8421**	0.6580	0.7456
individual (unimodal)		0.7546	**0.8377**	0.7646	0.8336	**0.7556***	**0.7895**	**0.7328**	0.7923
$b \cdot s_q$	CLIP-b	0.6456	0.7246	0.6439	0.7566	0.5697†	0.5789	0.6137	0.7083
	CLIP-l	0.6632	0.7478	0.6671	0.7933	0.6337†	0.6842	0.6250	0.7385
	ALIGN	0.6504	0.7101	0.6569	0.7568	0.5819†	0.6316	0.6050	0.7002
	LiT	0.6159	0.6667	0.5841	0.6746	0.4987†	0.3158	0.5404	0.6910
	MCSE	0.6274	0.6812	0.6137	0.6792	0.6694†	0.6316	0.6167	0.6642
	SimCSE	0.6252	0.6580	0.6347	0.7029	0.5988†	0.6316	0.6292	0.6833
	GTE-small	0.6860	0.7449	0.6775	0.7697	0.6196†	0.5789	0.6545	0.7377
	GTE-base	0.7095	0.7942	0.6855	0.7921	0.5764†	0.5789	0.6718	0.7619
	MiniLM	0.6736	0.7652	0.6474	0.7403	0.5840†	0.5789	0.6433	0.7157
	MPNet	0.6347	0.7014	0.6268	0.7171	0.6147†	0.6316	0.6257	0.7078
	DistilRoBERTa	0.5756	0.6319	0.5987	0.6659	0.6157†	**0.7895**	0.5873	0.6504
	PhraseBERT	0.5572	0.5971	0.5655	0.6153	0.5268†	0.5263	0.5903	0.6596
$b \cdot r_q$	CLIP-b	0.6833	0.7304	0.6532	0.7390	0.5723†	0.6316	0.6269	0.6961
	CLIP-l	0.6647	0.7420	0.6682	0.7727	0.6721†	**0.8421**	0.6477	0.7390
	ALIGN	0.6955	0.7101	0.6716	0.7520	0.5821†	0.6842	0.6141	0.6869
	LiT	0.5881	0.6522	0.6065	0.7012	0.4765†	0.4211	0.5558	0.6050
	MCSE	0.6367	0.6754	0.6194	0.6804	0.6409†	0.5263	0.6211	0.6629
	SimCSE	0.6309	0.6696	0.6345	0.6960	0.6245†	0.6316	0.6323	0.6805
	GTE-small	0.7035	0.7652	0.6972	0.7840	0.6312†	0.6316	0.6789	0.7469
	GTE-base	0.7201	0.8174	0.7161	0.8151	0.5855†	0.6316	0.6996	0.7783
	MiniLM	0.6838	0.7507	0.6621	0.7444	0.6299†	0.6316	0.6643	0.7249
	MPNet	0.6743	0.7188	0.6530	0.7418	0.6877†	0.6842	0.6571	0.7339
	DistilRoBERTa	0.6034	0.6725	0.6125	0.6768	0.6723†	**0.7895**	0.6008	0.6588
	PhraseBERT	0.5633	0.5768	0.5552	0.5973	0.5173†	0.4737	0.5955	0.6558
$s_q \cdot r_q$	CLIP-b	0.6965	0.7304	0.6870	0.7569	0.6492†	0.6842	0.6563	0.7229
	CLIP-l	0.6980	0.7623	**0.7096**	0.7904	0.6854†	0.7368	**0.6710**	**0.7607**
	ALIGN	0.6878	0.7623	0.6882	0.7413	0.6240†	0.5789	0.6134	0.6790
	LiT	0.5722	0.6609	0.5959	0.6764	0.4401†	0.5263	0.5605	0.6129
	MCSE	0.6618	0.7275	0.6447	0.7120	**0.7304**†	0.6842	0.6348	0.6833
	SimCSE	0.6243	0.6551	0.6556	0.7176	0.6639†	0.5789	0.6318	0.6933
	GTE-small	0.7444	0.7797	0.7512	0.8154	0.6807†	0.6316	0.7136	0.7721
	GTE-base	**0.7701**	0.8261	**0.7681**	**0.8395**	0.6545*	0.6842	0.7319	**0.7987**
	MiniLM	0.7180	0.7536	0.7077	0.7700	0.7182*	0.6842	0.6963	0.7533
	MPNet	0.7181	0.7449	0.7040	0.7780	0.7440*	**0.7895**	0.6964	0.7721
	DistilRoBERTa	0.6471	0.7188	0.6478	0.7098	0.7391*	**0.7895**	0.6188	0.6831
	PhraseBERT	0.5566	0.5913	0.5739	0.6241	0.5470†	0.5789	0.6049	0.6622

Table D.8: Model performance on subsets w.r.t. both concreteness and specificity. MWEs included. "Individual" indicates the best performance achieved by an individual measure (b, s_q or r_q). In each column, boldface highlights the best scores obtained by contrastive VLMs and sentence transformers respectively.
Statistical significance tests (one-sided Welch's t-test) are conducted for each subset. †denotes p-value ≥ 0.05, and * denotes $0.01 \leq p$-value < 0.05.