Do Vision-Language Models Understand Compound Nouns?

Sonal Kumar[♠*] **Sreyan Ghosh**[♠*] **S Sakshi**[♠] **Utkarsh Tyagi**[♠]
Dinesh Manocha[♠]
[♠]University of Maryland, College Park, USA
{sonalkum, sreyang, utkarsht, dmanocha}@umd.edu

Abstract

Open-vocabulary vision-language models (VLMs) like CLIP, trained using contrastive loss, have emerged as a promising new paradigm for text-to-image retrieval. However, do VLMs understand compound nouns (CNs) (e.g., *lab coat*) as well as they understand nouns (e.g., *lab*)? We curate Compun, a novel benchmark with 400 unique and commonly used CNs, to evaluate the effectiveness of VLMs in interpreting CNs. The Compun benchmark challenges a VLM for text-to-image retrieval where, given a text prompt with a CN, the task is to select the correct image that shows the CN among a pair of distractor images that show the constituent nouns that make up the CN. Next, we perform an in-depth analysis to highlight CLIPs' limited understanding of certain types of CNs. Finally, we present an alternative framework that moves beyond hand-written templates for text prompts widely used by CLIP-like models. We employ a Large Language Model to generate multiple diverse captions that include the CN as an object in the scene described by the caption. Our proposed method improves CN understanding of CLIP by 8.25% on Compun. Code and benchmark are available [1].

1 Introduction

A compound noun (CN) is a noun formed from two or more words combined to create a single noun with a new meaning. A CN usually combines two nouns (noun + noun, e.g., *paper towel*) or an adjective and a noun (adjective + noun, e.g., *full moon*); however, there exist more types, and we show an exhaustive list with examples in Appendix A.3. For the scope of this paper, we focus primarily on the noun + noun type.

Interpreting the meaning of CNs by decoding the implicit semantic relation between their constituent nouns has attracted great interest in Natural Language Processing (NLP) for decades (Wisniewski, 1997; Coil and Shwartz, 2023). This task requires systems to move beyond memorization as CNs are continually emerging, with new combinations frequently appearing (Pinter et al., 2020b). Pre-trained Language Models (PLMs) that are trained on vast amounts of text and acquire broad semantic knowledge in the process have shown impressive performance in interpreting CNs, including unseen CNs (Coil and Shwartz, 2023). The improvements can also be partly attributed to the transformer architecture, which by design computes a word representation as a function of the representation of its context (Coil and Shwartz, 2023).

Though extensively studied in NLP, whether modern vision-language models (VLMs) understand CNs is under-explored. Open-vocabulary VLMs [2] like CLIP (Radford et al., 2021), trained using a contrastive loss between image-caption pairs, have become the go-to choice for image-to-text (zero-shot classification) and text-to-image retrieval (Ray et al., 2023). However, recent work shows that CLIP-like VLM models often act as bag of words and lack understanding of relationships between objects and attributes (Yuksekgonul et al., 2023). For example, prior works explore the failure of VLMs to understand spatial relationships between two objects in the image through the caption (e.g., "left of") (Kamath et al., 2023) or the binding of a verb with its corresponding object (e.g., "running tiger"). To the best of our knowledge, evaluating a VLM's understanding of the semantic relationship between nouns to interpret CNs hasn't been explored in literature.

Main Contributions. We propose Compun, a benchmark with 400 instances that serves as a test

[1] https://github.com/sonalkum/Compun
[*] These authors contributed equally to this work.

[2] Our work only investigates VLMs trained with contrastive loss as they are widely adopted for retrieval tasks (Ray et al., 2023). Investigating other kinds of VLMs (e.g., autoregressive) is beyond the scope of our work.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 519–527

bed to evaluate a VLM's ability to interpret CNs. Each instance in Compun corresponds to a unique compound noun and includes one image representing the compound noun, along with two additional distractor images. These distractor images depict the individual constituent nouns that form the CN (investigating CNs with more than 2 nouns remains part of future work.) (example in Fig. 1). Given the class name (or the CN), the task of a VLM is to retrieve (or select) the correct image among the distractors. We perform a detailed analysis of CLIPs' performance on Compun, providing an in-depth understanding of how well state-of-the-art VLMs interpret CNs. Next, we present a novel framework to improve text-to-image retrieval that moves beyond generic hand-written prompts for text-to-image retrieval. Given a CN, we generate multiple diverse captions using an LLM, where each caption describes a scene with the CN as a key object in it. Finally, the captions are used to construct a custom prompt for text-to-image retrieval. Our proposed method improves CLIP's and OpenCLIP's performance by 8.25% and 2.35%, respectively, on Compun.

2 Background and Related Work

Interpreting CNs. Cognitive science research has examined human processing of novel noun-noun pairings (Wisniewski, 1997; Costello and Keane, 2000; Connell and Lynott, 2012). Although these pairings can lead to multiple interpretations, typically, one interpretation emerges as the most naturally comprehensible (Costello and Keane, 2000). Early work in interpreting compound nouns has majorly framed the task as a classification task, where each compound noun is classified to a single relation (Kim and Baldwin, 2005; Tratz and Hovy, 2010). However, owing to the ambiguity where a single compound noun can be classified into multiple relations (Shwartz and Dagan, 2018), recent work has adopted paraphrasing for the same task (Kim and Nakov, 2011; Pasca, 2015; Shwartz and Dagan, 2018). The task is again to classify a compound noun into multiple pre-defined templates. Ponkiya et al. (2020) show that PLMs acquire adequate knowledge to understand the semantic relationship between the constituent nouns in a CN during pre-training itself. Following this, a wealth of work employs sequence-to-sequence PLMs (including LLMs) to assess their ability to interpret existing and novel blends of nouns (Shwartz, 2021; Li et al., 2022b; Pinter et al., 2020b).

Contrastive VLMs. Contrastive VLMs include models trained using a contrastive loss between image-caption pairs. CLIP (Radford et al., 2019), a pioneering work in this space, shows that such a model can improve text-to-image and image-to-text retrieval, with applications in zero-shot classification, etc. Following CLIP, a wealth of work focuses on improving different aspects of CLIP, like compositionality (Nayak et al., 2023), and also employ CLIP as a backbone vision encoder for various vision tasks like captioning (Mokady et al., 2021), instruction following (Liu et al., 2023), etc. Our work is inspired by the fact that contrastive VLMs often act as bag of words (Yuksekgonul et al., 2023) and might lack an understanding of the semantic relationship between the constituent nouns to interpret the final CN.

3 Compun Benchmark

The task. Our Compun benchmark serves as a test bed for evaluating a VLM's capability to interpret compound nouns. For evaluation, we focus on the zero-shot text-to-image retrieval task, where given a natural language prompt, the task of a VLM is to retrieve an image that illustrates the image described in the prompt. In the base setting, our prompt just describes a compound noun as *"A photo of a {compound noun}"*. Text-to-image retrieval has been earlier adopted by several works for evaluating compositional understanding (Yuksekgonul et al., 2023; Ray et al., 2023). Inspired by these works, we design the Compun benchmark to challenge a VLM to select the correct image among a set of distractors. More precisely, each instance in the Compun benchmark, attributed to a compound noun, has 3 images, where only one image illustrates the compound noun, while the other two images illustrate the constituent nouns that make up the compound noun (example in Figure 1). All compound nouns in the Compun benchmark have a maximum of two nouns.

Evaluation. We resort to a simple evaluation metric, consistent with prior-art (Thrush et al., 2022) for evaluating a VLM on Compun. Formally, let us denote the image illustrating the compound noun as a positive ($\mathcal{P}$) and the other 2 distractor images as negatives ($\mathcal{N}_1$ and $\mathcal{N}_2$). Thus, given the natural language prompt $\mathcal{C}$ for the compound noun , our evaluation metric $f(\mathcal{C}, \mathcal{P}, \mathcal{N}_1, \mathcal{N}_2)$ is defined as:

$$f(\mathcal{C},\mathcal{P},\mathcal{N}_1,\mathcal{N}_2) = \begin{cases} 1 & \text{if } s(\mathcal{C},\mathcal{P}) > s(\mathcal{C},\mathcal{N}_1) \\ & \text{and } s(\mathcal{C},\mathcal{P}) > s(\mathcal{C},\mathcal{N}_2) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $s(.)$ is the standard cosine similarity, widely used for retrieval.

Data collection and annotation. The Compun benchmark has 400 test instances and a total of 1200 (400×3) images. Each instance in the Compun benchmark is attributed to a unique compound noun (the complete list is provided in Table A.4). We use a combination of compound nouns provided by Levin et al. (2019), Lang et al. (2022), GPT-4 (OpenAI, 2023), and the internet. Next, we discard compound nouns that can have confusing distractors (e.g., *cheesecake*, where it's usually hard to distinguish between a *cheesecake* and any other *cake*). After this step, we filter 400 compound nouns, the most widely used from our list. While a compound noun can have multiple interpretations, we use the more commonly known one. For a compound noun that may have multiple interpretations, we use MTurk to decide the most commonly known one. More details about this study can be found in Appendix A.6. Finally, a group of 4 annotators collects the required 1200 images from various image search engines. All 4 annotators come with extensive vision and language research experience.

4 Retrieval with Example Captions

Fig. 1 illustrates our proposed approach. As discussed earlier, the standard approach for text-to-image retrieval using class names is to hand-write several prompt templates (e.g., "a photo of a class name."). We propose an alternative framework – retrieve with example captions. Our framework is zero-shot and requires no further training. Given a compound noun c, we ask an LLM to generate 5 diverse captions, where each caption has the compound noun c as an object in it. The generated captions should have c in diverse settings with diverse adjectives and verbs. We instruct GPT-4 (OpenAI, 2023) with the following prompt to generate the captions:

> Return a list of 5 diverse captions with a compound_noun in a photo. The captions should be a maximum of 10 words and one-liners. All 5 captions should describe the compound noun in diverse settings with different verbs and actions being performed with the compound noun. An example output for "chicken burger": ['Sizzling chicken burger grilling at a lively backyard BBQ.,' 'Chef expertly flipping a juicy chicken burger in a diner.',' Family enjoying homemade chicken burgers on a sunny picnic.', 'Athlete fueling up with a protein-packed chicken burger post-workout.', 'Friends sharing a chicken burger at a vibrant street festival.']. Only return a list of strings and nothing else.

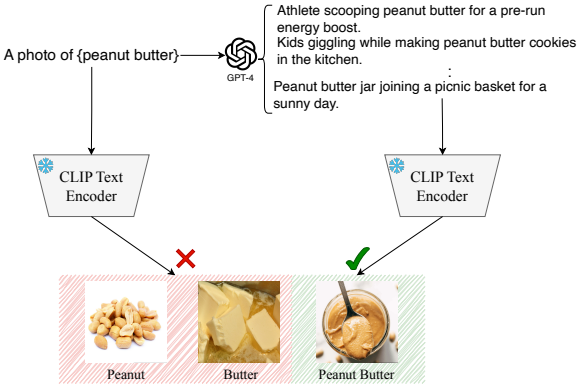

Figure 1: Illustration of our proposed **Retrieval with Captions**. We first generate 5 diverse captions describing 5 diverse scenes, with the compound noun as an object in it. These captions are then used to build 5 custom text prompts for text-to-image retrieval, and the image with the highest mean similarity to all 5 prompts is then selected for retrieval.

and an example output for the CN *"chocolate crocodile"* is as follows:

> ["Pastry chef sculpting a chocolate crocodile with finesse.", "Kids discovering a chocolate crocodile in a candy treasure hunt.", "Artist painting a whimsical chocolate crocodile in a foodie gallery.", "Chocolate crocodile starring in a whimsical patisserie window display.", "Chocolate crocodile sunbathing on a dessert island paradise."]

We then build a prompt for our VLM separately with each of the captions as follows to get 5 final prompts: *"a photo of a {class name}. An example of {compound name} in an image is {caption}"*. Next, we calculate the mean similarity of an image $c \in C$ with the text prompts as follows:

$$\text{Mean Similarity} = \frac{1}{n}\sum_{i=1}^{5} s(c, p_i) \quad (2)$$

where $p_i \in P$ denotes the generated prompts, $s(.)$ is the standard cosine similarity formulation, and $c \in C$ denotes the set of all images available for text-to-image retrieval or every image the text

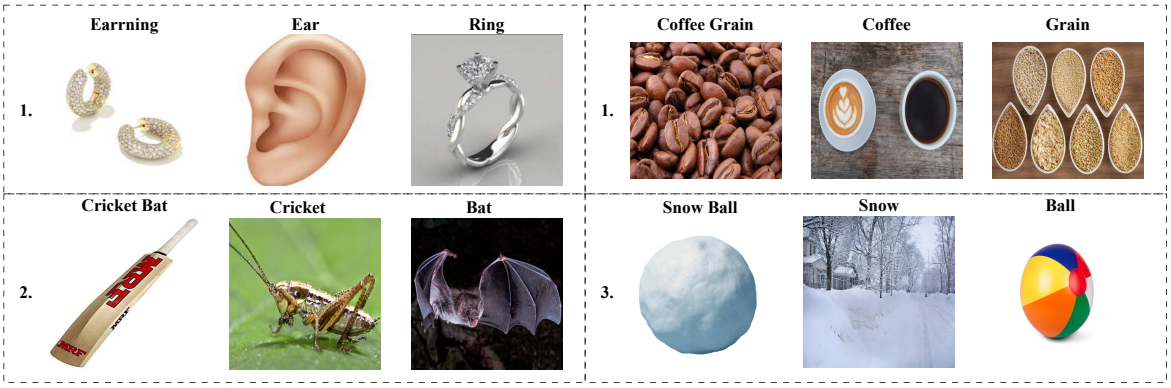

Figure 2: Illustration of 3 types of CNs used in our study: Either Noun, Both Nouns and None. A brief explanation of the 3 types is provided in Section 6. **1. (left)** An example of Either Noun, where *earring* looks like an ordinary *ring* but not like an *ear*, and the noun *ear* just acts as an attribute that modifies the visual of a *ring* to an *earring*. **1. (right)** An example of Either Noun, where *coffee grain* looks like an ordinary grain but is modified by the noun *coffee*, which acts as an attribute. **2.** An example of None, where a *cricket bat* looks completely different from both *cricket* and *bat*. **3.** An example of Both Nouns, where a *snow ball* looks both like *snow* and *ball*.

has to be compared with. Finally, we choose the image with the highest mean cosine similarity.

Our core hypothesis builds on existing work in using language as an internal representation for visual recognition, which creates an interpretable bottleneck for computer vision tasks (Menon and Vondrick, 2023; Pratt et al., 2023). Instead of querying a VLM with just the compound noun, employing language enables us to compare to any words flexibly. Since interpreting compound nouns is easier when provided with proper context in example sentences, getting exposed to diverse keywords through examples makes the image with the compound noun a strongly activating image while the distractors are lowly activating. Taking an example from Fig. 1, keywords like *player* and *wooden* obtained through diverse captions make the original image more activating than its distractors. Our proposed method also improves performance on benchmark text-to-image retrieval datasets, and we provide additional results in Appendix A.2.

5 Experiments and Results

Baselines. For our baselines, we employ the original CLIP (Radford et al., 2019), OpenCLIP (Ilharco et al., 2021), ALIGN (Jia et al., 2021), ALBEF (Li et al., 2021), BLIP (Li et al., 2022a) and MetaCLIP (Zhai et al., 2023). All these methods are trained with contrastive learning on image-text pairs. We also employ CLIP *w/ desc* (Menon and Vondrick, 2023), which adds image descriptors to the prompt for retrieval. Finally, we also ablate with CLIP *rev.* where we switch the order of nouns in the compound noun. We prompt GPT-4 with a temperature of 0.1 and top-p of 1.

Results. Table 1 compares the performance of various VLMs on the Compun benchmark. We also perform a human evaluation on our benchmark using MTurk. While OpenCLIP achieves the highest performance with simple template prompts, our method improves OpenCLIP performance by 2.35%. Similarly, our method improves CLIP performance by 8.24%. CLIP *rev.* leads to a 37.25% drop in performance over CLIP, indicating that CLIP has some understanding of the semantic relationship between the nouns. In the next section, we make important conclusions regarding CLIP's limited understanding of *attributed compound nouns*.

Model	Text-to-Image Acc.
Human	96.25
Random	33.33
ALBEF (Li et al., 2021)	80.55
BLIP (Li et al., 2022a)	79.85
MetaCLIP (Xu et al., 2023)	81.35
CLIP (Radford et al., 2019)	78.25
CLIP *rev.*	41.00
CLIP *w/ desc* (Menon and Vondrick, 2023)	81.15
CLIP *w/ examples (ours)*	86.50
OpenCLIP (Ilharco et al., 2021)	83.90
OpenCLIP *w/ examples (ours)*	86.25

Table 1: Comparison of our proposed version of CLIP with other baselines on the Compun benchmark. Our proposed method outperforms CLIP by 8.25% and OpenCLIP by 2.35%.

6 Results Analysis

To perform an in-depth analysis of the results, we first perform an MTurk study to divide all CNs in Compun into 3 main categories as illustrated in Fig. 2: **1.**CNs that clearly show either of their constituent nouns in the picture. In this case, one noun

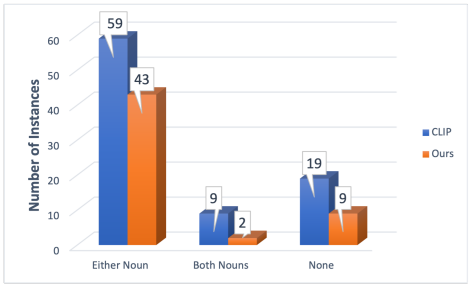

Figure 3: Count of misclassified instances by CLIP on COMPUN for three settings, either, both, and none. Section 6 describes these settings. CLIP is more likely to retrieve a negative when the positive image shows either constituent noun, highlighting CLIP's limited understanding of attributed CNs.

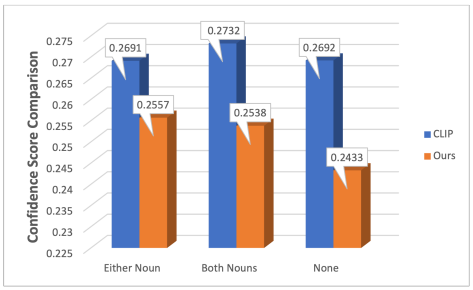

Figure 4: Average CLIP similarity scores for correct predictions on COMPUN on three unique settings, either, both, and none. Section 6 describes these settings. High scores on the COMPUN benchmark are superficial, and CLIP often wins by low similarity scores.

acts as an attribute to the other, changing its visual minimally, but is not visible itself (e.g., *coffee grain*). **2.**CNs that clearly show both the constituent nouns in the picture. This is the same as **1.**, but both nouns are visible (e.g., *snow ball*) and **3.** CNs that show none of the constituent nouns in the picture and make up a completely new CN (e.g., *cricket bat*). The 3 settings have 199, 106, and 95 instances in COMPUN. Fig. 3 compares the number of incorrectly predicted instances in COMPUN across these 3 categories. CLIP makes the highest number of mistakes in the first category, which also indicates CLIPs' limited understanding of such CNs, which can also be interpreted as attributed CNs. Such CNs are also emerging in nature (Coil and Shwartz, 2023), and correctly interpreting them is a long-standing problem in NLP. On the other hand, CLIP makes the least mistakes in the third type, indicating that CLIP might have acquired adequate knowledge about unique objects in its pre-training stage. Fig. 4 shows that results on the COMPUN benchmark are superficially low – similarity scores for correct predictions are $\approx 25_{+\text{-}2\%}$ (in contrast to ImageNet retrieval with ≈82%). We perform retrieval, treating the entire benchmark images as negatives, and achieve a score of 12%, a 66.25% drop.

7 Conclusion

This paper presents the first study of VLMs in interpreting CNs. We curate COMPUN, a novel benchmark with 400 unique CNs, and show that CLIP has a limited understanding of CNs where one of the two constituent nouns acts as an attribute to the other. Next, we present a novel approach that moves beyond generic template-based prompts and leverages an LLM to generate diverse captions describing scenes with the CN as an object in the scene. Our proposed method improves the performance of CLIP on COMPUN significantly.

Limitations and Future Work

We list down some potential limitations of our work:

1. COMPUN focuses on this sole definition of CN interpretation – Can VLMs distinguish between a CN and its constituent nouns? COMPUN does not consist of emerging CNs like the NYTWIT dataset (Pinter et al., 2020a). This dataset proposed CNs where humans created entirely new CNs using editing nouns corresponding to entirely new concepts. These CNs are particularly challenging for even modern LLMs to interpret and require strong reasoning abilities over context (Coil and Shwartz, 2023). However, after a preliminary analysis, we hypothesize that most of the CNs in Pinter et al. (2020a) are rare, and VLMs might not have come across these CNs or concepts from their pre-training stage. We want to explore this as part of future work as this brings entirely new challenges to VLMs, including complex reasoning abilities.

2. COMPUN, like other text-to-image retrieval benchmarks, would benefit from better evaluation metrics. Though our metrics are inspired by prior art, as shown in Section 6, results on COMPUN are superficial, and VLMs can perform well even with low confidence scores (corresponding to low activations when the VLM sees the CN). Additionally, evaluating COMPUN with the entire benchmark as negatives makes it difficult to gain an understanding of where and how VLMs go wrong in interpreting CNs. Thus, as part of future

work, we would like to explore better evaluation metrics and benchmark design.

3. We evaluate COMPUN only on contrastive trained VLMs as we try to study CN interpretation through the lens of text-to-image retrieval, and contrastive VLMs fit well to the retrieval task. As part of future work, we would like to study how well other types of VLMs, like auto-regressive (Liu et al., 2023) VLMs, interpret CNs.

References

Jordan Coil and Vered Shwartz. 2023. From chocolate bunny to chocolate crocodile: Do language models understand noun compounds? *arXiv preprint arXiv:2305.10568*.

Louise Connell and Dermot Lynott. 2012. Flexible shortcuts: Linguistic distributional information affects both shallow and deep conceptual processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Fintan J Costello and Mark T Keane. 2000. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science*, 24(2):299–349.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

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

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. "what's 'up' with vision-language models? investigating their struggle to understand spatial relations.". In *EMNLP*.

Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of noun compounds using wordnet similarity. In *International Conference on Natural Language Processing*, pages 945–956. Springer.

Su Nam Kim and Preslav Nakov. 2011. Large-scale noun compound interpretation using bootstrapping and the web as a corpus. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 648–658, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Inga Lang, Lonneke van der Plas, Malvina Nissim, Albert Gatt, et al. 2022. Visually grounded interpretation of noun-noun compounds in english. In *Proceedings of the Workshop on Cognitive Modelling and Computational Linguistics (CMCL'22)*. Association for Computational Linguistics.

Beth Levin, Lelia Glass, and Dan Jurafsky. 2019. Systematicity in the semantics of noun compounds: The role of artifacts vs. natural kinds. *Linguistics*, 57(3):429–471.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Siyan Li, Riley Carlson, and Christopher Potts. 2022b. Systematicity in GPT-3's interpretation of novel English noun compounds. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 717–728, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Sachit Menon and Carl Vondrick. 2023. Visual classification via description from large language models. *ICLR*.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nihal V. Nayak, Peilin Yu, and Stephen Bach. 2023. Learning to compose soft prompts for compositional zero-shot learning. In *The Eleventh International Conference on Learning Representations*.

OpenAI. 2023. Gpt-4 technical report.

Marius Pasca. 2015. Interpreting compound noun phrases using web search queries. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 335–344.

Yuval Pinter, Cassandra L. Jacobs, and Max Bittker. 2020a. NYTWIT: A dataset of novel words in the New York Times. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6509–6515, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuval Pinter, Cassandra L. Jacobs, and Jacob Eisenstein. 2020b. Will it unblend? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1525–1535, Online. Association for Computational Linguistics.

Girishkumar Ponkiya, Rudra Murthy, Pushpak Bhattacharyya, and Girish Palshikar. 2020. Looking inside noun compounds: Unsupervised prepositional and free paraphrasing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4313–4323, Online. Association for Computational Linguistics.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A Plummer, Ranjay Krishna, and Kate Saenko. 2023. Cola: A benchmark for compositional text-to-image retrieval. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Vered Shwartz. 2021. A long hard look at mwes in the age of language models. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*.

Vered Shwartz and Ido Dagan. 2018. Paraphrase to explicate: Revealing implicit noun-compound relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1200–1211, Melbourne, Australia. Association for Computational Linguistics.

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

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687.

Edward J Wisniewski. 1997. When concepts combine. *Psychonomic bulletin & review*, 4:167–183.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023. Demystifying clip data.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

A Additional Details

A.1 Hyper-parameter tuning for number of example captions

Table 2 compares the results of our proposed method for a varying number of captions. While performance monotonically increases with an increasing number of diverse captions, the performance plateaus at 5 captions.

#Exemplars	1	2	3	4	5	6	7
Accuracy	79.55	79.90	81.30	83.55	86.50	86.50	**86.55**

Table 2: Effect of number of exemplars

A.2 ImageNet accuracy with Retrieval with Captions

To prove the efficacy of our proposed approach in a more general setting, we perform zero-shot classification with ImageNet with example captions for each class. Table 3 compares the performance of baseline template-based retrieval with CLIP with our proposed method. Our proposed method outperforms generic template-based retrieval by x% on ImageNet.

Model	Accuracy
CLIP	71.58
CLIP *w/ desc* (Menon and Vondrick, 2023)	75.00
CLIP *w/ examples (ours)*	**76.85**

Table 3: Accuracy Comparison on ImageNet Dataset

A.3 Types of Compound Nouns

There are three forms for compound nouns:

1. open or spaced - space between words (tennis shoe)

2. hyphenated - hyphen between words (six-pack)

3. closed or solid - no space or hyphen between words (bedroom)

Table 4 shows some examples of compound nouns of different forms.

1. noun	+	noun	bus stop	Is this the bus stop for the number 12 bus?
			fire-fly	In the tropics you can see fire-flies at night.
			football	Shall we play footballtoday?
2. adjective	+	noun	full moon	I always feel crazy at full moon.
			blackboard	Clean the blackboardplease.
			software	I can't install this softwareon my PC.
3. verb(-ing)	+	noun	breakfast	We always eat breakfast at 8am.
			washing machine	Put the clothes in the red washing machine.
			swimming pool	What a beautiful swimming pool!
4. noun	+	verb(-ing)	sunrise	I like to get up at sunrise.
			haircut	You need a haircut.
			train-spotting	His hobby is train-spotting.
5. verb	+	preposition	check-out	Please remember that check-out is at 12 noon.
6. noun	+	prepositional phrase	mother-in-law	My mother-in-law lives with us.
7. preposition	+	noun	underworld	Do you think the police accept money from the underworld?
8. noun	+	adjective	truckful	We need 10 truckfuls of bricks.

Table 4: Types of Compound Nouns

A.4 List of Compound Nouns in Compun

Table 5 lists down all CNs in the Compun benchmark.

A.5 Visual examples of various categories of instances in Compun

Fig. 2 illustrates 3 types of CNs used in our study: Either Noun, Both Nouns and None. A brief explanation of the 3 types is provided in Section 6.

A.6 MTurk Study

Our institution's Institutional Review Board(IRB) has granted approval for the data collection. We will release our benchmark under the CC-BY-NC 4.0 License, which is freely available for research purposes.

Initial Annotator recruitment. We first performed a pilot run amongst 10 English-speaking MTurk annotators to test their intelligibility in identifying a set of 10 images with 10 different but commonly used compound nouns. From this study, we finally recruited 3 annotators. Our institution's Institutional Review Board (IRB) has approved this study.

Removing confusing instances. Post collection of all instances in Compun, the first step was to remove confusing instances where humans found it extremely difficult to distinguish between the image of a compound noun and its constituent nouns. The annotators were just asked a binary answer, i.e., confusing or not, after showing some 5 examples of confusing (e.g., cheesecake) and not confusing instances (e.g., cricket bat) to each. Finally, only instances with a majority vote of confusion amongst the 3 annotators were removed.

Human Evaluation on Compun. Finally, we perform a human evaluation of our benchmark Compun with 3 different annotators. Each annotator evaluates Compun once, and the final reported scores in Table 1 are an average of scores of all 3 annotators with the proposed evaluation metric in Equation 1.

B Extra Details

Model Parameters: We use CLIP-ViT-L/14 for all our experiments which have ≈ 673M parameters with 24 and 8 encoder and decoder layers, respectively, and 16 attention heads per layer.

Compute Infrastructure: All our experiments are conducted on a single NVIDIA A100 GPU.

Implementation Software and Packages: We implement all our models in PyTorch [3] and use the HuggingFace [4] implementations of CLIP. We also use the following repositories for running the baselines: ALBEF (Li et al., 2021)[5], BLIP (Li et al., 2022a)[6], CLIP (Radford et al., 2019)[7], MetaCLIP (Xu et al., 2023)[8], OpenCLIP (Ilharco et al., 2021)[9], CLIP w/ Descriptors(Menon and Vondrick, 2023)[10]. All softwares and packages are open source and are available for academic use, and were used only for academic purpose.

Image Curation: We use multiple websites to curate the images for our Compun benchmark. Some of the these websites are:
1. $https://pixabay.com$
2. $https://www.pinterest.com$
3. $https://www.wikipedia.org$
4. $https://www.istockphoto.com$
5. $https://www.britannica.com$

[3] https://pytorch.org/
[4] https://huggingface.co/
[5] https://github.com/salesforce/ALBEF
[6] https://github.com/salesforce/BLIP
[7] https://github.com/openai/CLIP
[8] https://github.com/facebookresearch/MetaCLIP
[9] https://github.com/mlfoundations/open_clip
[10] https://github.com/sachit-menon/classify_by_
_release/tree/master/descriptors

snow ball	courtyard	mountain bike	building stone	seat belt	pocketknife	teaspoon	spray paint
sun roof	bomb squad	curtain rail	bookshelf	golf cart	freight train	herb scissors	goldfish
steam train	space station	sandpaper	castle gate	pasta tongs	tailcoat	cassette tape	ice cap
raincoat	thumb pin	fruitcake	earwig	snow boot	pasteboard	shell pearl	fur coat
copper wire	billboard	birdhouse	zebra crossing	eardrum	clotheshorse	trash can	Gas station
firefly	eyeball	streetlight	peanut butter	nutmeg mill	lemon peel	marble corridor	soup ladle
windshield	Coffee grain	fishbowl	chocolate crocodile	mountain goat	watershed	popcorn ball	Cotton ball
duckpin	wastebasket	catfish	hand brake	sugar pea	cement mixer	potato salad	floodlight
pig farm	sand castle	farm machine	bullet train	Tea cup	Wallflower	Ice skate	Web page
ice scoop	eggshell	scoop strainer	splatter screen	motorcycle	clam knife	fishtail	beach house
blade guard	shoe brush	crossbow	toothbrush	fireman	dogwood	Computer mouse	swordfish
meat market	steam iron	football	aircraft engine	handbag	pasta salad	dirt bike	farmhouse
tennis shoes	houseboat	coconut haystack	tailbone	Woodshop	deck chair	fingernail	corn dog
skyscraper	metal spatula	ice tongs	oil pump	saucepan	prison dining	water meter	flagpole
food mill	horsefly	bookstore	streetcar	bedroom	key chain	pepper spray	fishhook
rubber band	Garage door	alphabet soup	Bathroom sink	Toothpaste	egg ring	paint brush	corn salad
sugarcane	headband	lipstick	Hairband	Ice hockey	silkworm	bike rack	clothesline
garlic bread	bow tie	skateboard	palm tree	seahorse	Candy cane	golf ball	cow pasture
ladybug	snowball	forehand	headdress	wiretap	Cupboard	dove necklace	chocolate macaroons
oven mitt	spaceship	toast tongs	ballpark	bedsheet	pinwheel	face mask	pancake pen
stone wall	sunfish	yardstick	dishwasher	footnote	Snack house	chocolate chips	earring
dog house	shoe rack	shellfish	tumbleweed	meat hammer	snow cone	trophy case	dish rack
panini spatula	corner spoon	Fish tank	telephone cord	ponytail	oven tongs	wine bar	rolling pin
rattlesnake	gas guzzler	almond biscotti	honeycomb	fingerprint	paper clip	Kitchen sink	cricket bat
robot arm	cigarette butt	coca leaf	oil thermometer	coconut tree	church bell	cartwheel	graveyard
jet engine	paper towel	bankbook	bread knife	tablespoon	eyelid	Waterwheel	toothpick
dump truck	station wagon	hair brush	penknife	key card	grasshopper	seaweed	banana tree
greenhouse	pasta rake	firearm	bus stop	duckbill	waterspout	pigtail	flower obsidian
shoe box	shoelace	roller coaster	drugstore	pumpkin gutter	gumball	car charger	coffee table
Cotton bud	water tank	headline	honeybee	starfish	Pool Table	courthouse	toadstool
crab cake	wheelchair	toilet paper	fountain pen	teapot	moonstone	watermelon	caveman
car tyre	wind turbine	rubber spatula	rib cage	fire truck	Wheelhouse	baseball bat	paperback
piggybank	garbage man	silicon chips	jellyfish	School bus	relay station	ice cream	soapstone
handcuff	sunflower	wristwatch	firewood	banknote	cattail	lime juicer	basketball court
drumstick	alarm box	bourbon balls	campfire	cowboy	straining scoop	chain armor	paddle wheel
grill surface thermometer	spaghetti spoon	candy bar	keyboard	kneecap	footprint	car door	waistcoat
pasta scoop	chalkboard	candlestick	peanut	seafood	skullcap	bulldog	Sugar plum
ring finger	jungle gym	buttercup	handshake	Hair spray	thunderhead	Fish net	rubber glove
Ice cube	horseshoe	shuttle cork	headlamp	headphone	space shuttle	cocktail spoon	snow man
street lamp	steel drum	Beach resort	Ground beef	exhaust fan	fruit fly	barman	tennis court
seafood scissors	powder brush	car factory	manhole cover	plastic bag	jellybean	backbone	police van
carpet	earthworm	snowbird	strawberry	doorbell	earbuds	tar pit	dish brush
horse cart	Cupcake	riverbank	ink pot	water butt	car phone	pancake	rainbow
headlight	M & M cookies	bolt cutter	eggplant	boat house	oyster knife	elevator shaft	Coffee mug
railroad	gunpowder	shoe shop	wallpaper	horse earrings	laser tag	tapeworm	tree house
pizza wheel	Keyhole	Kitchen counter	butterfly	bullseye	mailbox	avocado tool	sheep dog
grapefruit knife	necklace	pizza lady	letterhead	arrowhead	eyeglasses	earphone	shoe horn
armchair	glasshouse	fish spatula	elephant ear	suitcase	exercise bike	vanilla bean	cassette recorder
butter knife	Table cloth	honey dipper	dustpan	paper cup	sunspot	hornbill	lighthouse
food court	hand grip	fruitcup	watercolor	pinecone	lab coat	seashell	piston ring

Table 5: List of Compound Nouns in CompUN