A Couch Potato is not a Potato on a Couch: Visual Compositionality Prediction using Prompting Strategies and Image Generation for Adequate Image Retrieval

Anonymous ACL submission

Abstract

We explore the role of the visual modality and of vision transformers in predicting the compositionality of English noun compounds. Crucially, we contribute a framework to address the challenge of obtaining adequate images that represent non-compositional compounds (such as *couch potato*), making it relevant for any image-based approach targeting figurative language. Our method uses prompting strategies and diffusion models to generate these images. Comparing and combining our approach with a state-of-the-art text-based approach reveals complementary contributions regarding features as well as degrees of abstractness in compounds.

1 Introduction

Compositionality represents a core concept in linguistics (Partee, 1984): the meaning of complex expressions, such as compounds, phrases and sentences, can be derived from the meanings of their parts. The degree of compositionality however varies; e.g., while the compound *climate change* has a high degree of compositionality, *couch potato* is less so regarding its constituent *potato*, because it does not refer to a potato lying on a couch. For natural language understanding tasks such as summarization, machine translation and retrieval systems, the accurate prediction of compositionally is crucial to ensure precise and reliable results.

The focus of this paper is on predicting degrees of compositionality for English noun compounds. In contrast to state-of-the-art models, which primarily leverage text-based representations to assess the relatedness between compound and constituent meanings (see §2), we explore the contribution of the visual modality, which previously has proven successful across semantic tasks (Bruni et al., 2012; de Deyne et al., 2021; Frank et al., 2021, i.a.). Applying vision models to any task involving non-compositionality however comes with the major challenge of finding appropriate images, because standard image retrieval methods return false positives for non-compositional expressions, e.g., a *couch potato* is actually depicted as a potato (instead of a lazy person) sitting on a couch, cf. `Bing` (left) in Figure 1.

Figure 1: `Bing` (left) and `Vision:Scenario` (right) images of *couch potato*.

The current study suggests a novel way of obtaining "correct" images, which we judge highly valuable for any vision work involving figurative language: We carefully design and compare prompts as input for an image generation model, in order to obtain adequate images for both compositional and non-compositional compounds. The actual compositionality prediction then follows standard routes, i.e., estimating the degree of compositionality via similarity of compound and constituent feature vectors. Evaluation is carried out by measuring the rank correlation between similarity estimates and human ratings. In addition to our main contribution of (i) prompting strategies with increasing contextual description levels to obtain images of non-compositional expressions, we conduct analyses to identify aspects relevant for vision models, including (ii) the role of abstractness, given that abstract concepts are generally more difficult to depict than concrete concepts

(Pezzelle et al., 2021; Tater et al., 2024), and (iii) the role of meaning prototypicality. Finally, (iv) we compare our visual approach against a state-of-the-art text approach, a multimodal approach, and ChatGPT predictions.

2 Related Work

Traditionally, most computational approaches to automatically predict the compositionality of noun compounds have been realized using text-based vector space models (Reddy et al., 2011; Salehi et al., 2015; Schulte im Walde et al., 2016; Cordeiro et al., 2019; Miletić and Schulte im Walde, 2023, i.a.). Few studies addressed compound meaning using multimodal information; Bruni et al. (2014) to identify figurative uses of color terms in adjective–noun phrases, Pezzelle et al. (2016) and Günther et al. (2020) predict compound representations, and Köper and Schulte im Walde (2017) predict the compositionality of German compounds.

3 Gold-Standard Compound Data

Reddy et al. (2011) compiled a compositionality dataset with human ratings for 90 noun–noun compounds, collected via Amazon Mechanical Turk. It contains compounds with varying degrees of compositionality, including compounds where both constituents are literal (e.g., *swimming pool*), only one is literal (e.g., *couch potato*), or neither is literal (e.g., *cloud nine*). Ratings range from 0 (non-compositional) to 5 (highly compositional). We rely on their compound–constituent ratings for 88 compounds,[1] excluding two compounds due to frequency limitations.

4 Our Methodology

Given a compound (e.g., *couch potato*), our task is to assess how related the compound meaning is in relation to the meanings of the constituents, i.e., the modifier (*couch*) and the head (*potato*), by relying on reliable images.

4.1 Image Acquisition+Representation

To reliably capture the meaning of a word or expression via images, the images are required to accurately represent compositional as well as figurative, non-compositional meanings. Standard strategies to download images, such as Bing[2], however, include false positive images for non-compositional expressions, e.g., a *couch potato* is actually depicted as a potato (instead of a lazy person) sitting on a couch (see examples in Figure 1 and in App. A). We propose a new method for obtaining images that accurately depict non-compositional meanings and may also be highly valuable for figurative expressions in general: We generate images with a text-to-image diffusion transformer (PixArt-Sigma[3]), exploring four prompting strategies to guide image generation[4]:

Word: Prompts consist **solely of the target word (i.e., compound or constituent)**, without any context or modifications.

Sentence: Prompts consist of actual **corpus sentences containing the target word**, extracted from the ENCOW16AX web corpus (Schäfer and Bildhauer, 2012).

Definition: Prompts use definitions of the target words **generated by ChatGPT**.

Scenario: Prompts use diverse, descriptive scenarios involving the target word **generated by ChatGPT**.

For `Word`, we generate 10 images with different seeds. For `Sentence`, we extract 10 sentences per target and generate one image per sentence. For `Definition`, we ask ChatGPT to create 3 definition prompts, and generate one image each; for `Scenario`, we ask ChatGPT to create 25 scenario prompts, and generate one image each[5]. For comparison, we download 10 images per target from Bing, resized to 1024×1024; generated images are directly at this size.

We then extract feature vectors from these images using a vision transformer[6], and create a single representation for each target word by mean-pooling the feature vectors of multiple images of the same word.

4.2 Prediction and Evaluation

We assess the meaning relatedness between a compound and its constituents using cosine

[1] Reddy et al. also collected ratings for the whole compound phrases, but we do not use them.

[2] https://www.bing.com/images

[3] https://huggingface.co/PixArt-alpha/PixArt-Sigma-XL-2-1024-MS; we chose this model after testing various diffusion models.

[4] See examples in App. C.

[5] See instructions in App. B.

[6] https://pytorch.org/vision/main/models/generated/torchvision.models.vit_h_14.html

	Prediction Approach	Mod	Head
	Bing	.345	.232
PixArt	**Word**	-.005	.043
PixArt	**Sentence**	**.506**	.096
PixArt	**Definition**	.414	.288
PixArt	**Scenario**	.457	**.440**
	Skip-gram (T)	.565	.574
	Combined (T+V)	.624	.590
	ChatGPT (direct)	.736	.738

Table 1: Spearman's ρ for model predictions.

similarity between the respective visual representations, where a higher similarity corresponds to a higher degree of compositionality. Our approach predicts two ratings for each target compound: one for the compound–modifier combination and one for the compound–head combination. To assess prediction quality, we compute the Spearman's correlation between the predicted scores and the gold standard ratings provided by Reddy et al. (2011), see §3.

Although our goal is to explore challenges and contributions of the visual modality, and not to optimize performance, we compare our image-based predictions against (i) Word2Vec `Skip-gram`[7] predictions (Mikolov et al., 2013), the state-of-the-art textual approach on our task (Cordeiro et al., 2019; Miletić and Schulte im Walde, 2023), (ii) a weighted combination of the textual and our best visual predictions,[8] and (iii) direct ChatGPT predictions. Table 1 presents the correlation results for visual and textual approaches for compound–modifier and compound–head combinations.

The performance of our visual approaches differs strongly across prompting strategies. `Word` yields weak correlations; `Sentence` provides a strong improvement but only for modifiers, while prompting with more contextualisation (`Definition` and `Scenario`) yields the best results for both constituents. Taken together, the results highlight the challenge of obtaining adequate images of (non-compositional) noun compounds, and reinforce our exploration of prompting strategies. While the text-based approach `Skip-gram` outperforms all individual variants of image-based approaches, it is outperformed by combining text (T) and vision (V) features. `Bing` provides intermediate

[7]Trained on ENCOW16AX web corpus with a window size of 20, minimum count of 5, and 300 dimensions.
[8]See App. D for details.

	Concrete		Abstract	
	Mod	Head	Mod	Head
Scenario	**.448**	.174	.299	.400
Skip-gram	.439	.220	**.471**	.430

Table 2: Spearman's ρ for `Scenario` and `Skip-gram` predictions for concrete versus abstract compounds.

results, thus emphasizing the deceptive starting point of our study because we know these results incorporate wrong meaning depictions, cf. examples in Figures 1, 4. Finally, `ChatGPT` yields state-of-the-art results, which however come with the usual restriction that we cannot analyze the underlying conditions. Given that Reddy et al. (2011) has been publicly available for years, it is likely part of `ChatGPT`'s training data, requiring caution in interpreting results.

5 Analysis

We conduct a detailed analysis of the image-based approach, focusing on the images and predictions generated by the highest-performing candidate, `Scenario`, with `Skip-gram` included as the textual comparison.

5.1 Abstractness of Compounds

We analyze predictions for concrete and easily perceivable compounds, against abstract and less perceivable compounds, expecting differences in the contributions of visual features (Pezzelle et al., 2021; Tater et al., 2024). First, we collect human concreteness ratings for each compound on a scale from 0 (abstract) to 5 (concrete), following previous work (Brysbaert et al., 2014; Muraki et al., 2023).[9] The 30 compounds with the highest mean ratings are categorized as concrete, and the 30 with the lowest as abstract (see Table 3). Table 2 presents the prediction results. For concrete compounds, `Scenario` and `Skip-gram` perform similarly. In contrast, `Skip-gram` performs noticeably better for abstract compounds, thus aligning with our expectations: the image-based approach performs en par for compounds with clear, recognizable features, such as concrete nouns, which are easier to capture and represent in images. In contrast, abstract compounds, which are harder to visually represent, lead to poorer predictions, and the text-based approach outperforms the image-based one.

[9]We will make these ratings publicly available.

Figure 2: Images of *graveyard shift*, *graveyard*, *shift*.

Figure 3: Images of *engine room*, *engine*, *room*.

5.2 Analysis of Individual Compounds

To assess prediction quality for individual compounds, we rely on Rank Differences (RDs), which compare predicted ranks against corresponding gold ranks by calculating their absolute differences, separately for modifiers and heads (see Table 4), and analyse two examples.

Graveyard Shift refers to *"a work shift taking place from late night to early morning"*, where `Scenario` performs well with low RDs of 4.0 (mod) and 1.0 (head). Figure 2 presents the underlying images. Those of *graveyard* (second row) show graveyards with tombstones, mostly in daylight. In contrast, *shift* (third row) is more abstract and harder to represent; still, the images capture the concept fairly accurately, by depicting people working in various contexts, such as bakers and construction workers. Finally, the images of *graveyard shift* (first row) closely resemble those of *shift*, as they also depict workers in various settings, but with the key distinction of always occurring at night, differentiating them from the daytime scenes associated with *shift*.

The computed visual cosine similarities for *graveyard shift* are .243 for *graveyard* and .753 for *shift*, while the respective gold ratings on the 0–5 range are .380 for *graveyard* and 4.5 for *shift*. The close alignment between the predicted and gold rankings suggests that the visual similarities accurately reflect the semantic contributions of each constituent, resulting in strong predictions for the compound.

Engine Room `Scenario` predicts poor compositionality ratings with high RDs of 16.5 (mod) and 75.5 (head). The underlying images of *room* (Figure 3, third row) are high-quality and accurately depict various types of rooms (e.g., living rooms and conference rooms). In contrast, the images of *engine room* (first row) depict a mix of diverse types of engine rooms with trains and cars.

The visual cosine similarity is .45, while the gold compositionality rating is 5.0, i.e., the maximum value. The captured visual similarity seems reasonable, as images of *engine room* and *room* should intuitively share some features but also exhibit significant differences, given that a prototypical *room* is rather a living or conference than an engine room. Unfortunately, the predicted visual similarity does not align with the compositionality rating, which is also reflected in the high individual RD of 75.5.

We observe that the image-based approach, which relies solely on visual similarity, performs well when shared visual features align with the semantic contributions of constituents to the compound's meaning. However, it struggles in cases where visual similarity does not accurately capture these contributions, thus highlighting the limitations of using visual features alone when predicting compositionality.

6 Conclusion

This study explored the contribution of the visual modality to the prediction of compositionality for English noun–noun compounds, focusing on the challenge of obtaining adequate images, especially for non-compositional compounds, by providing prompting strategies for generative models with increasing contextual description levels. We further analyzed especially challenging sub-cases, such as abstract targets and meaning prototypicality, as well as complementary distributions of visual and textual information.

Limitations

The image-based approach relies heavily on the quality and availability of relevant, accurate images for the compounds. While image generation can address some of these challenges, it comes with significant resource demands (GPU) and can be time-consuming, which may hinder scalability, especially when generating large numbers of images for many compounds. Additionally, while the approach performs well for concrete compounds, it struggles with abstract compounds and those that are difficult to visualize.

Ethics Statement

We see no ethical issues related to this work. All experiments involving human participants were voluntary, with fair compensation (12 Euros per hour), and participants were fully informed about data usage. We did not collect any information that can link the participants to the data. All modeling experiments were conducted using open-source libraries, which received proper citations. All relevant information (including created artifacts, used packages, information for reproducibility, etc.) can be found in (PLACEHOLDER for GitHub repository, will be added upon paper acceptance).

References

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. 2012. Using Image Analysis to improve Computational Representations of Word Meaning. In *Proceedings of the 20th Anniversary ACM Multimedia*, Nara, Japan.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 64:904–911.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Simon de Deyne, Danielle J. Navarro, Guillem Collell, and Andrew Perfors. 2021. Visual and Affective Multimodal Models of Word Meaning in Language and Mind. *Cognitive Science*, 45.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, online.

Fritz Günther, Marco Alessandro Petillia, and Marco Marelli. 2020. Semantic transparency is not invisibility: A computational model of perceptually-grounded conceptual combination in word processing. *Journal of Memory and Language*, 112.

Maximilian Köper and Sabine Schulte im Walde. 2017. Complex verbs are different: Exploring the visual modality in multi-modal models to predict compositionality. In *Proceedings of the 13th Workshop on Multiword Expressions*, pages 200–206, Valencia, Spain.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Filip Miletić and Sabine Schulte im Walde. 2023. A systematic search for compound semantics in pretrained BERT architectures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1499–1512, Dubrovnik, Croatia.

Emiko J. Muraki, Summer Abdalla, Marc Brysbaert, and Penny M. Pexman. 2023. Concreteness ratings for 62,000 English multiword expressions. *Behavior Research Methods*, 5:2522–2531.

Barbara H. Partee. 1984. Compositionality. In Fred Landman and Frank Veltman, editors, *Varieties of Formal Semantics: Proceedings of the 4th Amsterdam Colloquium*, pages 281–311. Foris Publications.

Sandro Pezzelle, Ravi Shekhar, and Raffaella Bernardi. 2016. Building a bagpipe with a bag and a pipe: Exploring conceptual combination in vision. In *Proceedings of the 5th Workshop on Vision and Language*, pages 60–64, Berlin, Germany. Association for Computational Linguistics.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies*, pages 977–983, Denver, Colorado, USA.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. 2016. The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany.

Tarun Tater, Sabine Schulte im Walde, and Diego Frassinelli. 2024. Unveiling the mystery of visual attributes of concrete and abstract concepts: Variability, nearest neighbors, and challenging categories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21581–21597, Miami, Floria, USA.

A Bing versus Vision:Scenario

Figure 4 provides further examples of images of non-compositional compounds, comparing the extraction via `Bing` (on the left) against image generation using the `Vision:Scenario` prompting method (on the right), also see Figure 1.

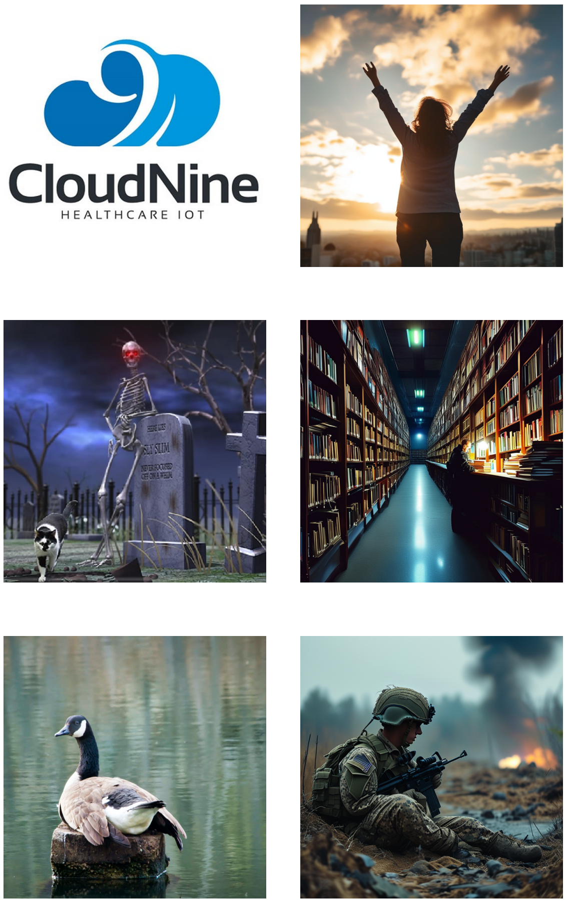

Figure 4: `Bing` (left) and `Vision:Scenario` (right) images of *cloud nine* (top), *graveyard shift* (mid) and *sitting duck* (bottom).

B Prompt Generation Using ChatGPT

This appendix describes the procedure for generating `Definition` and `Scenario` prompts for text-to-image models using `ChatGPT`. The process consists of three phases, carried out separately for each of the two prompting strategies:

- **Preparation Phase:** `ChatGPT` is introduced to the task, including the goal of generating prompts that accurately reflect the meanings of compounds and their constituents. Prompts are described as detailed descriptions of the intended image, formatted in CSV without headers or numbering for easy copying.
- **Instruction Phase:** `ChatGPT` receives guidelines for each strategy. For `Definition`, it creates three prompts based directly on the noun definitions. For `Scenario`, it generates 25 diverse prompts capturing real-world scenarios related to the target word, ensuring a broad diversity of representations.
- **Query Phase:** `ChatGPT` generates prompts for each target compound in sequence (modifier, head, compound) based on the provided instructions, outputting them in CSV format for further use.

C Prompt Examples

We present examples of prompts for the compound *couch potato* and its constituents, *couch* and *potato*, using the `Sentence`, `Definition`, and `Scenario` prompting strategies. On paper acceptance we will provide all prompts.

C.1 Sentence

For the `Sentence` prompting strategy, we gather sentences containing the target word from the ENCOW16AX web corpus. Below are three examples for *couch potato*, *couch*, and *potato*.

***Couch Potato*:**

1. *"if we do not give them opportunity for play and exercise, they can become couch_potato just like some people."*
2. *"that be not to say there be anything inherently wrong with spend a certain amount with couch_potato fare, be it TV, movie, music, or game."*
3. *"be sit and watch the (unknown) athlete at the Olympics make you feel like a couch_potato?"*

***Couch*:**

1. *"they be sit side-by-side on the couch."*
2. *"while they sit down on the couch, one of those cheap spiky contraption with polished wooden arm, his wife make them a cup of tea."*
3. *"there be another bed in the front room and a couch that be a (unknown)."*

***Potato*:**

1. *"while the canned food sector generally be see growth - and benefit from strong brand such as (unknown) - this have not help canned potato."*
2. *"despite the (unknown) ' accident, in @card@ Belarus be still a net exporter of meat, milk, egg, flour, and potato to other former Soviet republic, although its export be routinely test for radioactive contamination."*
3. *"and most of us know a rise when we see one, as we know the African marigold we plant beside the potato and bean in our garden, and the maple, elm, cedar, and pine that shade our backyard and line our street."*

C.2 Definition

The following are three `Definition` prompts, each generated by ChatGPT, based on the noun definitions of *couch potato*, *couch*, and *potato.*

***Couch Potato*:**

1. *"A person who spends a significant amount of time sitting or lying down, typically watching television or engaging in sedentary activities."*
2. *"A term describing someone who leads a sedentary lifestyle, preferring indoor activities such as watching TV or playing video games."*
3. *"An informal term for a person who is inactive or lazy, often spending leisure time on a couch or sofa."*

***Couch*:**

1. *"A piece of furniture designed for seating two or more people, typically with a back and armrests."*

2. *"A long upholstered piece of furniture for reclining or sitting, often found in living rooms or lounges."*
3. *"A sofa or settee, usually with cushions and upholstered arms and back, used for relaxation or casual seating."*

Potato:

1. *"An edible tuber that is a staple food in many cultures, typically underground and harvested from the Solanum tuberosum plant."*
2. *"A starchy vegetable with a variety of culinary uses, such as boiling, baking, frying, or mashing."*
3. *"The plant itself, Solanum tuberosum, which belongs to the nightshade family and produces tubers that vary in size, shape, and color."*

C.3 Scenario

Finally, we present three `Scenario` prompts, each generated by ChatGPT, based on real-world scenarios related to *couch potato*, *couch*, and *potato*, respectively. These examples aim to cover a wide range of contexts in which the target words may appear.

Couch Potato:

1. *"A couch potato binge-watching their favorite TV series, surrounded by cushions and blankets."*
2. *"A person on the couch, flipping through a photo album or scrapbook."*
3. *"A person lounging on a couch with a bowl of popcorn, absorbed in a movie marathon."*

Couch:

1. *"A vintage leather couch with tufted upholstery, adding a touch of elegance to a study."*
2. *"A cozy reading nook with a couch by the window, bathed in natural sunlight."*
3. *"A modular couch with interchangeable pieces, allowing for easy customization and rearrangement."*

Potato:

1. *"A beautifully plated baked potato topped with melting butter and dollops of sour cream."*
2. *"A farmer harvesting potatoes in a sun-lit field, with rows of potato plants in the background."*
3. *"A close-up of potato peelings on a kitchen countertop, with a peeler and scattered peels."*

D Combining Textual and Visual Predictions

We conduct an experiment to explore how different contributions of text-based and image-based predictions interact with each other. Specifically, we compute a weighted combination of the individual predictions (cosine similarities) from `Scenario` and `SkipGram`:

$$\texttt{Combined} = \alpha * \texttt{SkipGram} + (1 - \alpha) * \texttt{Scenario}$$

We vary α from 0 to 1 in increments of 0.1. When $\alpha = 0$, the predictions correspond entirely to `Scenario`, while $\alpha = 1$ results in purely `SkipGram`-based predictions.

The results are shown in Figure 5, where we present the modifier, head and mean correlations across α values. The results indicate that combining text-based and vision-based predictions provides an improvement over the individual predictions. While this outcome aligns with expectations, given that `SkipGram` performs better than `Scenario` individually, we also find that `Combined` surpasses `SkipGram` for α values between 0.5 and 0.9. Performance peaks at $\alpha = 0.7$, yielding modifier and head correlations of .624 and .590, respectively. These results suggest that leveraging both modalities provides a meaningful advantage over relying solely on one.

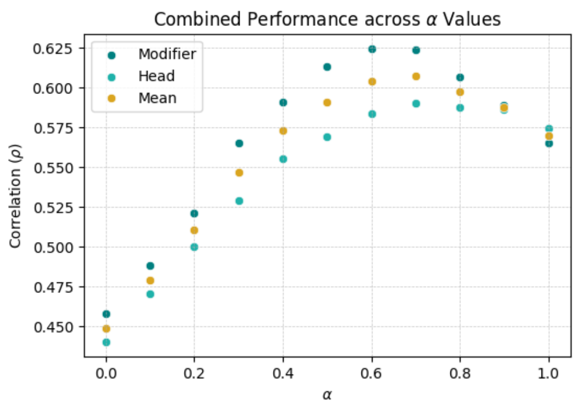

Figure 5: Spearman's ρ for `Combined` predictions across α values.

E ChatGPT Predictions of Compositionality

We query ChatGPT to predict compound-constituent compositionality ratings on a scale from 0 to 1 for the 88 compounds of interest, and we correlate them with the gold ratings. ChatGPT achieves strong correlations of .736 for modifiers and .738 for heads. This performance surpasses both Scenario and SkipGram and approaches the state-of-the-art results reported in the literature (Cordeiro et al., 2019; Miletić and Schulte im Walde, 2023).

F Compounds by Concreteness

Table 3 reports the human-generated concreteness scores of 60 compounds. We will make the full set of ratings publicly available upon paper acceptance.

G Rank Differences

Table 4 reports the rank differences (RDs) between Scenario predictions and the gold ratings for modifiers and heads.

Compound	Concreteness	Compound	Concreteness
car park	5.0	crash course	2.5
human being	4.9	couch potato	2.5
swimming pool	4.9	snake oil	2.5
credit card	4.7	climate change	2.4
parking lot	4.7	night owl	2.4
polo shirt	4.7	sitting duck	2.4
ground floor	4.6	sacred cow	2.4
call centre	4.6	game plan	2.4
brick wall	4.6	eye candy	2.3
cocktail dress	4.6	rock bottom	2.3
application form	4.4	monkey business	2.3
zebra crossing	4.4	face value	2.2
health insurance	4.4	role model	2.2
video game	4.3	melting pot	2.2
law firm	4.3	agony aunt	2.2
bank account	4.2	graveyard shift	2.2
engine room	4.1	cash cow	2.2
radio station	4.1	guilt trip	2.1
grandfather clock	4.1	memory lane	2.1
balance sheet	4.1	shrinking violet	2.1
head teacher	4.1	gravy train	2.1
speed limit	4.0	kangaroo court	2.0
gold mine	3.9	lip service	2.0
graduate student	3.9	ivory tower	2.0
brass ring	3.9	blame game	2.0
lotus position	3.9	rat run	2.0
panda car	3.8	swan song	2.0
search engine	3.7	rat race	1.9
china clay	3.6	crocodile tear	1.9
research project	3.6	cloud nine	1.9

Table 3: Top 30 (left) and bottom 30 (right) compounds ranked by (mean) concreteness, based on human-judgements. Scale: 0 (abstract) to 5 (concrete).

	Scenario		Skip-gram			Scenario		Skip-gram	
Compound	Mod	Head	Mod	Head	**Compound**	Mod	Head	Mod	Head
couch potato	1.0	0.0	2.0	13.0	mailing list	3.5	29.0	8.5	18.0
parking lot	3.0	0.5	5.0	60.5	memory lane	20.5	13.0	32.0	7.5
guilt trip	4.0	0.0	9.0	16.0	cocktail dress	26.0	8.5	25.0	1.5
graveyard shift	4.0	1.0	34.5	10.5	snail mail	11.5	26.0	7.0	25.0
rat run	4.0	3.0	37.0	12.5	swimming pool	27.5	10.0	1.0	5.0
grandfather clock	3.0	4.5	37.0	17.5	blame game	16.0	23.0	16.0	2.0
case study	7.0	4.0	12.0	4.0	diamond wedding	6.0	34.0	35.0	30.0
graduate student	12.0	1.5	10.0	5.5	end user	34.0	6.0	51.5	6.0
think tank	10.0	4.0	50.0	8.0	web site	16.0	26.0	40.0	26.0
rush hour	9.5	6.0	12.0	14.0	brass ring	35.0	8.0	10.0	1.0
crash course	5.0	11.0	7.0	9.0	sitting duck	27.0	16.5	10.5	17.0
research project	7.0	9.0	1.0	20.0	fine line	33.0	14.0	29.0	4.0
front runner	7.0	9.0	43.5	18.0	silver spoon	9.0	38.5	22.0	37.0
zebra crossing	14.0	2.0	29.0	10.0	video game	23.0	24.5	2.0	11.5
balance sheet	4.0	12.5	22.0	43.5	cash cow	13.0	35.0	8.0	21.0
rock bottom	14.0	3.0	4.0	9.0	agony aunt	14.5	36.5	11.0	30.0
nest egg	12.0	5.5	8.0	3.5	call centre	21.0	31.0	42.0	23.5
human being	4.5	13.0	2.5	24.0	bank account	45.0	7.0	9.0	6.0
spelling bee	9.0	9.0	24.0	11.0	public service	44.5	8.5	9.5	4.5
game plan	7.0	11.5	28.0	20.5	face value	31.0	23.0	25.5	14.0
melting pot	6.0	15.0	2.0	16.0	silver bullet	15.0	40.0	8.0	26.0
gravy train	3.0	18.0	24.0	26.0	chain reaction	15.0	41.5	32.0	12.0
radio station	11.5	9.5	19.5	4.0	fashion plate	22.0	37.0	6.0	20.0
eye candy	13.0	9.5	32.5	21.0	ground floor	47.5	15.0	45.0	15.5
polo shirt	13.0	10.5	34.0	2.5	rat race	59.0	4.0	26.0	18.0
credit card	2.5	21.5	4.5	13.5	brick wall	34.0	32.0	34.0	41.0
search engine	18.0	7.0	11.0	17.0	kangaroo court	53.0	14.0	37.0	3.0
cheat sheet	10.0	15.0	5.5	6.0	gold mine	7.0	60.0	25.0	56.0
interest rate	23.0	2.5	19.0	8.0	lotus position	16.0	53.0	46.0	60.0
flea market	13.5	12.0	11.5	49.0	car park	38.0	32.0	32.5	28.0
ivory tower	1.5	24.0	6.5	0.5	smoking jacket	20.0	50.5	13.0	9.5
head teacher	4.0	21.5	33.0	17.5	monkey business	47.0	24.0	54.0	24.0
spinning jenny	23.0	3.5	2.5	41.5	application form	19.0	52.5	14.0	56.5
climate change	13.5	13.0	0.5	41.0	lip service	33.0	39.0	37.0	22.0
health insurance	1.0	26.0	6.0	7.5	shrinking violet	29.0	45.5	31.5	1.5
snake oil	22.0	5.0	20.0	5.5	cloud nine	41.0	34.5	31.0	19.5
role model	26.0	1.0	9.0	37.0	rocket science	70.0	7.0	15.0	2.0
firing line	10.0	19.0	14.0	0.5	speed limit	47.0	42.5	16.0	34.5
china clay	9.0	21.0	2.5	7.0	acid test	50.5	39.5	14.5	5.5
cutting edge	10.0	20.0	21.0	0.0	engine room	16.5	75.5	23.5	45.5
silver screen	21.0	9.0	17.5	16.0	night owl	38.0	54.5	7.0	23.5
smoking gun	1.5	29.0	9.0	15.0	sacred cow	36.0	61.0	6.0	27.0
law firm	1.0	30.0	29.0	34.0	panda car	62.0	52.0	1.0	1.0
swan song	7.5	25.0	15.0	31.0	crocodile tear	86.0	39.0	16.0	18.0

Table 4: Modifier and head RDs between `Scenario` predictions and the gold ratings, sorted by increasing average `Scenario` RD. As a textual point of comparison, we add RDs for `Skip-gram` predictions.