Fruitcakes and Cupcakes Emerging from Noise: The ComposiGen Dataset of Compounds and their Compositionality

Jule Godbersen[1,2], **Sinan Cem Kurtyigit**[1,3], **Emma Raimundo Schulz**[1], **Tonmoy Rakshit**[1], **Diego Frassinelli**[4], **Sabine Schulte im Walde**[1], **Carina Silberer**[1]
[1]Institute for Natural Language Processing, University of Stuttgart
[2]Department of Language Science and Technology, Saarland University
[3]School of Computation, Information and Technology, Technical University of Munich
[4]MaiNLP, Center for Information and Language Processing, LMU Munich
jgodbersen@lst.uni-saarland.de, carina.silberer@ims.uni-stuttgart.de

Abstract

Compounds are a complex linguistic phenomenon, as variation in their degree of compositionality often makes their interpretation non-straightforward. We consider the task of visual-linguistic compositionality prediction for English noun-noun compounds, i.e., predicting the degrees to which a compound's meaning is predictable from its constituents. We introduce a new dataset, *ComposiGen*, which provides constituent-specific human-elicited compositionality ratings for compounds of different concreteness categories, and includes generated visual representations for both compounds and their constituents. To enable controlled comparisons, we structure *ComposiGen* such that head constituents are shared across multiple compounds (e.g., *wedding cake*, *cup cake*). We suggest a novel parameter-based approach leveraging constituent-to-compound image transformations to predict different degrees of visual constituent contributions to compound meaning. While our novel approach requires further exploration for validation, our overall results show that the generated images, in particular in combination with text, provide valuable information, and that simple late fusion outperforms multimodal transformers. Taken together, our findings highlight a promising avenue for future research on more efficient multimodal models for compositionality prediction. Our novel dataset offers a rich resource for future in-depth research, including the exploration of visual, constituent-based compound formation.

Keywords: noun compounds, compositionality ratings, image generation, vision models, multimodality, transformers

1. Introduction

Compounds and other multiword expressions are an important part of language because they help us express complex ideas efficiently. However, their meanings are not always easy to infer as compounds might vary in their degree of compositionality, i.e., the extent to which the meaning of the whole can be derived from the meanings of its parts. For example, *wedding cake* is highly compositional, as its meaning is transparently related to both its constituents *wedding* and *cake*, whereas the meaning of *novelty cake* is not a simple composition of the meanings of *novelty* and *cake*. Moreover, the transparency of the constituents may differ; e.g., in *cupcake*, the head constituent *cake* is semantically more transparent regarding the compound's meaning than the modifier *cup*.

We consider the task of **visual-linguistic compositionality prediction** for English noun-noun compounds, defined as predicting for each constituent of a compound a real-valued score expressing the degree to which the compound's meaning can be derived from this constituent. We include visual and textual representations of compounds and their constituents into our experiments as representations across multiple modalities can offer complementary insights that go beyond traditional textual feature-based approaches. For instance, while the word *cake* typically evokes a sweet, baked dessert, an image of a *potato cake* highlights a different concept (cf. Figure 1).

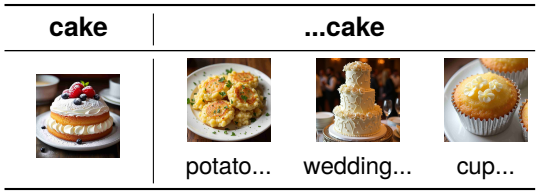

Figure 1: Example compounds sharing "cake" as head constituent, along with generated images.

Furthermore, we suggest that image transformations, specifically the extent of change required to transform an image of a constituent (e.g., *cake*) into an image of a compound (e.g., *wedding cake*) could transparently reveal aspects of compositionality that remain opaque with text-based representations. Finally, word concreteness is yet another factor that modulates the textual-visual divide: concrete concepts are typically more readily depicted, whereas abstract concepts present greater representational challenges (cf. Khaliq et al., 2024; Tater et al., 2024). Despite the complementary insights that the visual and textual modalities can contribute to the understanding of compounds (cf. Günther et al., 2020), most previous work on compositionality prediction focuses on unimodal methods (e.g., Reddy et al., 2011; Pezzelle et al., 2016; Cordeiro et al., 2019; Miletic and Schulte im Walde, 2023), with a

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9321–9338
11–16 May 2026.

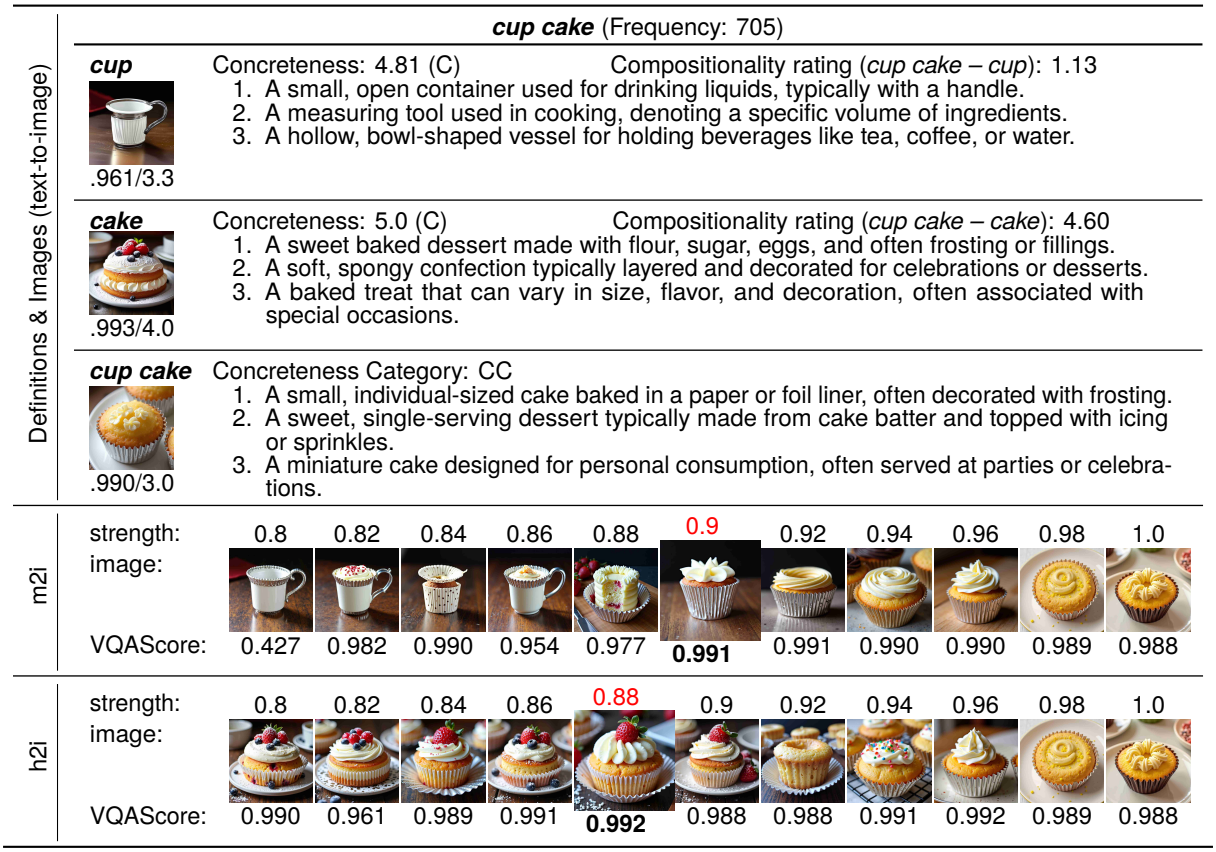

Figure 2: Example item *cup cake* in *ComposiGen*: We include the compound frequency from the ENCOW16AX corpus and the concreteness category (Sect. 3.1), the human-elicited compositionality ratings (Sect. 3.2), generated noun definitions (Sect. 3.3), generated images for the compound and its constituents using text-to-image (t2i; Sect. 3.4) and image-to-image approaches (i.e., modifier-to-image (m2i) or head-to-image (h2i); Sect. 4.2). Numbers below images denote automatic/human image evaluation scores (i.e., VQAScores/human ratings for images and their first definition/token, respectively; Sect. 3.4).

few exceptions (e.g., Köper and Schulte im Walde, 2017; Kurtyigit et al., 2025).

To support further research on multimodal compositionality, we introduce *ComposiGen* — a novel, multimodal dataset of 200 English noun-noun compounds paired with human-elicited compositionality ratings. *ComposiGen* is structured into 36 sets of compounds that share the same head, thus allowing a comparative study of visual transformations from constituents to compounds (such as in *potato cake* vs. *novelty cake*). We also include abstract and concrete constituent pairings to ensure they reflect the full range of concreteness. Within *ComposiGen*, each target (compound or constituent) is represented by three automatically generated noun definitions and one image, using generative models; in addition, we enrich the compound representation with two image sequences resulting from the corresponding two constituent-to-compound image transformations (example given in Figure 2).

Using *ComposiGen*, we evaluate text-based, vision-based, and multimodal (i.e. visual-linguistic) approaches, and also test whether transformations resulting from text-guided constituent-to-compound image generation capture signals of compositionality. We summarize the contributions of our work as follows:

- We present a novel dataset of English noun-noun compounds ***ComposiGen***[1], annotated with constituent-specific human-elicited compositionality ratings and automatically augmented with noun definitions and images using generative models.
- We compare traditional feature-based and state-of-the-art models on the task of compositionality prediction, and present a novel approach using constituent-to-compound image transformations.
- Our findings indicate that a multimodal setup can outperform unimodal approaches, while

[1]The data and code are available at https://github.com/jule-go/ComposiGen.

also demonstrating that the characteristics of the dataset reveal differences on compositionality predictions across methods.

2. Related Work

Generative Data Augmentation. With the rise of large language models (LLMs), generative data augmentation has become a common complement to manual dataset collection (Eigenschink et al., 2023). For instance, Piedboeuf and Langlais (2023) evaluate generation models such as ChatGPT for extending text-based datasets. Beyond text, researchers have applied image generation to build visual resources like metaphor datasets (cf. Chakrabarty et al., 2023; Khaliq et al., 2024). Inspired by Kurtyigit et al. (2025), we use generative models to obtain noun definitions of compounds and their constituents and, using these as prompts, generate corresponding images. Pickard et al. (2025) also use diffusion models in their construction of a multimodal dataset of compounds, but they prompt them with human-written scene descriptions to generate the images. Furthermore, their AdMIRe task is on ranking a set of images based on how well they depict the meaning of a compound in a specific context, while we address a numeric approach to compositionality prediction. To this end, we augment the instances in our dataset with human-elicited compositionality ratings.

Compositionality Prediction. Traditionally, computational approaches to automatically predicting compositionality have relied on feature-based models that compare the textual representations of compounds with their constituents (e.g., Reddy et al., 2011; Cordeiro et al., 2019; Shwartz and Dagan, 2019; Miletic and Schulte im Walde, 2023). Some studies also explore visual information (e.g., Pezzelle et al., 2016; Günther et al., 2020) or adopt basic multimodal setups that merge textual and visual features (e.g., Roller and Schulte im Walde, 2013; Köper and Schulte im Walde, 2017). Our work takes this further by investigating multimodal transformer models and comparing them with standard fusion strategies. While Kurtyigit et al. (2025) use text-to-image models to obtain accurate images of constituents and compounds in their feature-based approach, we present a first attempt to additionally treat the text-guided transformation process of a constituent image into a compound image as a potential signal of compositionality.

3. *ComposiGen* Dataset Creation

We release our novel multimodal *ComposiGen* dataset containing text and image data on English noun-noun compounds along with associated human compositionality ratings. Below we detail the selection of compounds, our elicitation of ratings, and how we use generative models to create descriptions and images of compounds and their constituents. An example instance is given in Figure 2.

Cat.	Examples			
AA	 fantasy romance (1.59 / 2.19)	 health service (2.28 / 2.21)	 incentive scheme (2.26 / 2.41)	 plea agreement (2.39 / 2.22)
AC	 ego trip (1.74 / 3.71)	 farewell party (2.21 / 3.89)	 safety helmet (2.37 / 4.92)	 service area (2.21 / 3.72)
CA	 child protection (4.78 / 2.50)	 data quality (3.93 / 2.18)	 hair loss (4.97 / 2.19)	 word processing (3.56 / 2.03)
CC	 ballet shoe (4.04 / 4.97)	 corn dog (4.96 / 4.85)	 hockey game (4.31 / 4.50)	 street food (4.75 / 4.80)

Figure 3: Example compounds (with generated images and constituent concreteness scores) for our concreteness categories AA, AC, CA, and CC.

3.1. Set of Noun-Noun Compounds

Starting from the English web corpus ENCOW16AX (Schäfer, 2015), we extracted all noun-noun compounds based on the automated part-of-speech tags provided in the corpus: we define noun-noun compounds as noun-noun bigrams whose immediate left and right neighbors are not nouns, in line with prior work (e.g., Rassem et al., 2024; Maurer et al., 2023). We restricted the set to compounds that occurred at least 100 times in the corpus.

The concreteness of the constituents of a compound may influence its semantic transparency and, consequently, its perceived degree of compositionality (cf. Schulte im Walde, 2024). To facilitate further research on this relationship, we require *ComposiGen* to contain compounds with varied pairings of abstract and concrete constituents. We consider this especially important in light of our use of visual representations, as concrete concepts are typically easier to depict, while abstract ones pose greater challenges. Specifically, based on the concreteness of a compound's constituents as determined using Brysbaert et al. (2014)'s ratings, compounds are required to fall into one of the following four categories: AA, AC, CA, or CC, where 'A' stands for abstract (ratings 1.0–2.5) and 'C' stands for concrete (ratings 3.5–5.0). In the abbreviations, the first letter refers to the concreteness category of the modifier constituent, and the second letter refers to the category of the head constituent. Note that we do not disambiguate constituent senses

but rely on word-level concreteness ratings which may lead to unintuitive classifications for some compounds.[2] We furthermore only retained those compounds whose head constituent occurs with both a concrete and an abstract modifier at least once (e.g., *novelty cake* (AC) and *fruit cake* (CC)). Finally, from this set of about $26,800$ compounds, we manually selected a set of 200 noun-noun compounds, composed of 172 different constituents, while enforcing overlaps in head constituents across compounds, diversity in modifier variation, and a distribution of concreteness. Prior work on German noun-noun compounds addresses overlaps in constituents (Schulte im Walde et al., 2016), and we extend this consideration to English noun-noun compounds with a special focus on overlapping head constituents in our *ComposiGen*. About 80% of all distinct heads pair at least with 5 different modifiers into compounds. For instance, the set of compounds with head *cake* contains *novelty cake* (AC), *cup cake* (CC), *fruit cake* (CC), *potato cake* (CC), *rice cake* (CC) and *wedding cake* (CC). Out of the 200 compounds, 16 fall into the AA category, 49 into AC, 28 into CA, and 107 into CC. Figure 3 shows example compounds from each category.

3.2. Gold Standard Compositionality Ratings

We used the Prolific platform[3] to elicit two compositionality ratings for each of the 200 compounds, one for their modifier and one for their head. Specifically, we asked the annotators to evaluate the extent to which the overall meaning of a compound can be related to the meaning of each of its constituents on a scale from 0 (= not compositional) to 5 (= compositional). The annotation guideline, along with an example, is provided in Appendix A.1.

We randomly split the set of compounds into questionnaires of 25 items and included additional control cases to eliminate spammers. Each compound-constituent pair was annotated by 15 annotators, and the final rating was computed as the average of their judgments.

Only English native speakers with an approval rate above 90% in earlier tasks were eligible to participate in the annotation. They were compensated for each submitted questionnaire, yielding an average hourly wage of £9. In total, ratings from 100 different annotators were approved.

As shown in Figure 4, the majority of compound-constituent pairs received a high compositionality rating (x-axis), indicating that for most compounds

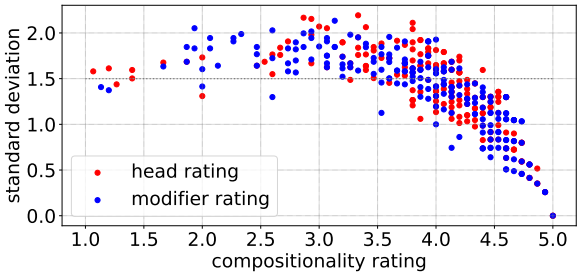

Figure 4: Mean compositionality ratings and standard deviations for modifier and head judgments.

the constituents provide a strong contribution to the overall meaning. The corresponding standard deviations (y-axis) are very low for strongly compositional compounds, but range between 1.3 and 2.3 for less compositional compounds. Standard deviation values in the range of 1.5 (cf. dotted horizontal line in Figure 4) suggest considerable annotator disagreement, which in turn may signal that our compound-constituent pairs are particularly difficult to judge (see Pollock (2018); Knupleš et al. (2023); Paisios et al. (2023) for in-depth analyses of disagreements on ratings on a scale), and thus may also pose challenges for automatic prediction approaches.

3.3. Textual Description Generation

We prompt an LLM to generate textual descriptions in the form of noun definitions for the 200 compounds as well as their 172 different constituents. The prompt is provided in Appendix B. This generation approach ensures consistent definitions with a predefined structure for all items. We used OpenAI (2025)'s ChatGPT–4o model[4] to generate three noun definitions for each target. Our work does not consider multiple senses of constituents or compounds, relying on a single sense per target. For follow-up experiments, we therefore use only the first definition generated by ChatGPT.

3.4. Image Generation

Analogously to the textual modality, we obtain visual representations for both the compounds and constituents of *ComposiGen*. Specifically, we generate one image for each target by prompting a text-to-image diffusion model with the corresponding noun definition (cf. Section 3.3) as the textual guidance signal. Compared to other acquisition methods such as image retrieval, this approach requires less effort, offers a controlled setup and ensures that images with a consistent style are available for all concepts. We run models in inference mode on a GPU, with the maximum number of inference steps set to 30.

[2] *Tea service*, for instance, is automatically classified as CA compound, even though *service* does not denote an abstract concept like assistance or duty, but instead refers to a tangible set of items used in serving tea.

[3] Available at https://www.prolific.com.

[4] Via the free web version on 13/14 January 2025.

For each compound, we furthermore generate image sequences as part of a parameter-based compositionality prediction approach using text-guided image-to-image generation, which we will detail in Section 4.2.

Using the Diffusers library on Hugging Face (von Platen et al., 2022), we apply PixArtSigma (Chen et al., 2025)[5] for constituent images, and FLUX (Black Forest Labs, 2024)[6] for compound images. FLUX also supports image-to-image generation, our choice is thus consistent with the text-guided image-to-image approach for compound image generation (cf. Section 4.2).

Note that our image generation process does not incorporate an explicit mechanism to differentiate between concrete and abstract targets. Instead, both are generated in the same manner.

Image Quality. We evaluated the faithfulness of the text-to-image generated images, i.e., if they accurately show the intended concepts, using both human judgments and an automatic metric. To this end, we collected image-word alignment ratings for the 372 instances in *ComposiGen*, each consisting of a target word (consituent or compound) and its generated image (cf. Section 3.4). For the human evaluation, three annotators assessed each pair by rating how well the image aligns with the word on a scale from 0 (= not at all) to 4 (= perfectly). Details of the annotation procedure are given in Appendix A.2. As an automatic metric, we rely on common sense understanding of visual-question-answering (VQA) models and employ VQAScore (Li et al., 2024).[7] For each image-target pair, we prompt a pretrained VQA model with the question:

"[target definition] Does this figure show [target]?"

The model's probability for the answer "yes" yields the VQAScore. Target definitions are those described in Section 3.3.

All results are shown in Table 3 (Appendix A.2); here we focus on the main findings. Overall, human ratings and VQAScore exhibit similar patterns. As expected, and as illustrated by the examples in Figure 3, abstract concepts are more difficult to depict than concrete ones. Specifically, the generated images for fully concrete compounds (CC) and concrete constituents (C) are most accurate — humans and VQAScore rated these image-word pairs the highest (human $2.99/2.67$, VQAScore $0.90/0.89$, respectively) — while images of fully abstract compounds and abstract constituents receive the lowest ratings ($1.69/1.49$ and $0.76/0.84$, respectively). Interestingly, the human judges rated the images of compounds as more accurate (2.61) compared to constituents, where modifiers are found as more accurately depicted than heads (2.36 vs. 1.98).

To quantify the agreement between the two evaluation methods, we computed Spearman's rank-order correlation coefficient between the VQAScores and human ratings. The results exhibit a moderate mean correlation of $\rho = 0.66$, with the lowest correlation for compounds where both constituents are abstract ($\rho = 0.42$), and with the highest correlation for compounds consisting of a concrete modifier and an abstract head ($\rho = 0.80$). Importantly, the moderate correlation between the measures suggests that VQAScore is a valid automatic proxy for compound image accuracy, which could enable scaling of the dataset (with the caveat that its reliability varies depending on the concreteness of the compounds and constituents). In the present work, we will use VQAScore for the denoising approach put forward in Section 4.2.

4. Compositionality Prediction

To model compositionality, we explore traditional feature-based methods using prediction-based and transformer-based models, as well as a novel parameter-based approach that leverages denoising strength values derived from image transformations. Across all approaches, predictions are intrinsically evaluated by computing the correlation between predicted compositionality scores and human-elicited compositionality ratings.

4.1. Feature-Based Approaches

Classic feature-based compositionality prediction involves two steps. First, feature vectors are extracted for compounds and their constituents using well-established representation models. Second, a compound's degree of compositionality is estimated by the cosine similarities between its vector and that of each of its constituents (yielding one score for the head and one for the modifier, respectively). A higher similarity between a constituent's representation and that of its compound indicates a higher predicted degree of compositionality.

Unimodal Features. We extract textual and visual features for the compounds and constituents in the *ComposiGen* dataset. We follow Kurtyigit et al. (2025) in training a `Skip-gram` model (Mikolov et al., 2013) on the ENCOW16AX corpus, where all occurrences of particular compounds are replaced with unique tokens such that the model learns dedicated embeddings. We use a pretrained vision transformer (`ViT`; Dosovitskiy et al., 2021) to extract a visual feature vector for each target's image in *ComposiGen*. For that, we use the embedding from the final hidden layer, preceding the classification layer.

[5] `PixArt-alpha/PixArt-Sigma-XL-2-1024-MS`

[6] `black-forest-labs/FLUX.1-dev`

[7] We use `clip-flant5-xxl` as underlying model.

Setup	Generated images and scores											
m2i	strength σ:	[initial]	0.80	0.82	0.84	0.86	0.88	0.90	0.92	0.94	0.96	0.98	1.00
	image:												
	VQAScore:	—	0.194	0.129	0.986	0.147	0.991	0.989	0.989	0.989	**0.992**	0.989	0.984
h2i	strength σ:	[initial]	0.80	0.82	0.84	0.86	0.88	0.90	0.92	0.94	0.96	0.98	1.00
	image:												
	VQAScore:	—	0.990	0.961	0.989	0.991	**0.992**	0.988	0.988	0.991	0.992	0.989	0.988

Figure 5: Example images generated with different denoising strength values for "wedding cake", along with their VQAScores. The highest evaluation score for each generation process is shown in bold. The number in red is taken as compositionality score as the according image got the highest score.

Multimodal Models. In an `early fusion` approach, we use the extracted text embeddings f_t from `Skip-gram` and image embeddings f_v from `ViT`, and compute multimodal representations $f_{multi} = \alpha f_t + (1-\alpha) f_v$ with a weighting factor $\alpha = 0.8$.[8] Beyond this, we explore more complex fusion strategies, specifically testing state-of-the-art transformer-based vision-language models `CLIP` (Radford et al., 2021) and `FLAVA` (Singh et al., 2022), which learned aligned embeddings for text and images via contrastive and multimodal objectives from large multimodal datasets. We feed the target definition to the textual encoder of these models, while the visual encoder is fed the target image. For `CLIP`, multimodal embeddings result from mean-pooling the encoded text and image; for `FLAVA`, from mean-pooling the hidden embeddings of the first multimodal encoder layer.[9] Finally, in a `late fusion` approach, we combine the predictions s_t and s_v from `Skip-gram` and `ViT`, respectively, into a multimodal prediction score $\hat{s}_{multi} = \beta s_t + (1-\beta) s_v$ with weight β set to 0.7.[8]

4.2. Parameter-Based Approach

Our approach to modeling compositionality draws inspiration from text-guided image-to-image generation with denoising diffusion models. In this setup, an initial image gets noisified to some extent and is then used as starting point for the denoising process, in which an image is to be generated that matches the text prompt. The denoising strength parameter σ determines how much noise is added to the initial image, and thus influences how much the newly generated image may deviate from the initial one. The compound's definition t serves as textual guidance during generation.

In the context of transforming constituent images into compound images, we hypothesize that highly compositional compounds permit more preservation of the constituent images, since their relation is also visually transparent and require only little changes — and therefore less noise needs to be added to the constituent image — whereas minimally compositional compounds require greater initial noise to sufficiently disrupt the constituent image and permit greater deviation during the generation process.

For example, when transforming an image of a *cake* into an image of a *wedding cake*, much of the information from the initial constituent image can be preserved, so only a small amount of noise needs to be added (= low denoising strength value). In contrast, when transforming a *dog* into a *corn dog*, the initial image provides little useful information, and a larger amount of noise (= high denoising strength value) must be added to allow for substantial changes during generation.

For the set of generated images we then ask: *The generated image v of which noise level does depict the compound?* We use VQAScore to assess for each generated image individually how faithful it depicts the compound. The image with the highest VQAScore, while requiring only minimal noise, is selected, and the corresponding denoising strength value is taken as predicted compositionality score $\hat{s}$:

$$\hat{s}_{denoise} = \arg\min_{\sigma \in D} \max \text{VQAScore}(v(\sigma), t)$$

The generation of compound images is performed individually for each of the two compound-constituent pairs. Figure 5 illustrates an example, showing images generated across different denoising strengths starting from the modifier constituent (m2i) and from the head constituent (h2i) for the compound *wedding cake*. In the h2i setup, the generated *wedding cake* images progressively depict a multi-tiered cake decorated with flowers, while con-

[8] α and β were empirically selected based on performance across different tested values (cf. Appendix C.1).

[9] We describe the best-performing setups; variations in input, layer choice and pooling led to weaker results.

sistently keeping the *cake* itself in focus. In contrast, in the m2i setup, with a lower level of noise, the model struggles to position the cake within the *wedding* scene (e.g., at strength $\sigma_{0.84}$ the cake replaces the bride). Compared to h2i, m2i requires a higher noise level before the *wedding cake* becomes the central focus. We observe a similar tendency in other examples (see also Figure 2).

5. Experiments

We conduct experiments on *ComposiGen* to assess how well different modalities capture compositionality. Model predictions are compared to human ratings and analyzed in relation to *ComposiGen*'s properties.

5.1. Setup

Data and Metrics. We compute Spearman's rank-order correlation ρ between model predictions and *ComposiGen*'s human compositionality ratings to assess the degree of alignment between predicted scores and human judgments.[10] Ideally, the resulting correlation values are close to 1, indicating that the predictions and human ratings produce similar rankings of compound-constituent pairs.

Models. As text-only model we train a `Skip-gram` model on ENCOW16AX.[11] As image-only model we load a pretrained `ViT`.[12] Both, `early` and `late fusion` are based on `ViT` and `Skip-gram`. We furthermore report results for the multimodal models `CLIP` and `FLAVA`.[13]

For the parameter-based approach `denoising`, we rely on FLUX[6] to generate compound images starting from the respective constituent images, and vary the denoising strength parameter σ over the range $D = [0.8, 1.0]$ in increments of 0.02. At the maximum value ($\sigma_{1.0}$), the initial constituent image is effectively replaced by pure noise. The choice of range D is empirically motivated. As the effect of σ depends on the underlying scheduler and no standard interval is specified in prior work, we conducted exploratory experiments over a broader set of values. We observed that values below 0.8 produced no perceptible changes in the generated compound images, whereas larger values increasingly degraded the constituent image content. To automatically evaluate the generated images we use VQAScore as reported in Section 3.4.

As a reference point, we report results obtained with ChatGPT-4o[14] in a few-shot setting. The model was given instructions similar to the human annotation guidelines, along with three examples to illustrate the expected output (cf. Appendix C.2). In the following, we refer to this model as `ChatGPT`.

	Approach	Mode	Mod	Head
Feature-based	Skip-gram	T	0.46	0.39
	ViT	V	0.30	0.24
	early fusion	T+V	0.34	0.28
	late fusion	T+V	**0.50**	**0.41**
	CLIP	T+V	0.41	0.27
	FLAVA	T+V	0.22	0.25
Parameter-based (denoising)		V	0.11	0.03
ChatGPT		T	0.44	0.41

Table 1: Spearman's ρ for predicted compositionality scores and human compositionality ratings. With the exception of the parameter-based approach, these correlation values are statistically significant ($p < 0.005$).

	human	Skip-gram	ViT	late fusion
Mod	1.13	0.47	0.16	0.38
Head	4.60	0.85	0.58	0.77

Figure 6: Human ratings and model predictions for the example item *cup cake* detailed in Figure 2.

5.2. Results

Table 1 reports Spearman's ρ correlation scores across approaches. Overall, the correlations range from weak to moderate, indicating that *ComposiGen* is a challenging dataset. Except for `FLAVA`, all models perform better on modifier ratings (column Mod) than on head ratings (column Head).

Among the unimodal feature-based models, `Skip-gram` performs best.[15] Almost all multimodal feature-fusion models and image-only `ViT` perform worse than text-only `Skip-gram`. `Late fusion`, in contrast, which simply combines the unimodal predictions of `ViT` and `Skip-gram`, achieves the best overall performance on both modifier and head, with correlations of 0.5 and 0.41, respectively. This shows that, first, *ComposiGen*'s

[10] For `denoising`, we use $1 - \hat{\sigma}_{denoise}$ as predictions.

[11] Trained using the gensim library (Řehůřek and Sojka, 2010), with a window size of 20, minimum count of 5, and 300 dimensions.

[12] From TorchVision (Marcel and Rodriguez, 2010): `vit_h_14`.

[13] From Hugging Face (Wolf et al., 2020): `openai/clip-vit-base-patch16`; `facebook/flava-full`.

[14] Via the free web version on 27 August 2025.

[15] We tested also other unimodal models, BERT (Devlin et al., 2019), FastText (Bojanowski et al., 2017), ResNet (He et al., 2016), but their results were worse.

Modifier	Generated images and scores for "[modifier] cake"											
fruit	strength σ:	[initial]	0.8	0.82	0.84	0.86	0.88	0.9	0.92	0.94	0.96	0.98	1.0
	image:												
	VQAScore:	—	0.673	0.661	0.855	0.427	0.815	0.875	0.832	0.876	**0.961**	0.938	0.938
rice	strength σ:	[initial]	0.8	0.82	0.84	0.86	0.88	0.9	0.92	0.94	0.96	0.98	1.0
	image:												
	VQAScore:	—	0.661	0.798	0.919	0.638	0.807	0.927	0.869	0.759	0.882	0.909	**0.951**

Figure 7: Example images generated in h2i-setup with different denoising strength values for compounds with "cake" as head constituent.

Set	Skip-gram		ViT		late fusion	
	Mod	Head	Mod	Head	Mod	Head
All	0.46	0.39	0.30	0.24	0.50	0.41
AA	0.25	−0.16	0.00	0.01	0.11	−0.15
AC	0.32	0.48	0.04	0.42	0.33	0.54
CA	0.48	−0.24	0.20	−0.13	0.50	−0.25
CC	0.50	0.48	0.41	0.20	0.54	0.48

Table 2: Spearman's ρ for selected feature-based approaches and considering concreteness-driven subsets of target compounds. Cells in green show improvements over the All setting.

images contain valuable information for compositionality prediction that visual encoders can at least partially capture; and second, multimodal information can enhance performance over unimodal features. The latter is consistent with Kurtyigit et al.'s (2025) late fusion results on compositionality prediction. Our results, however, underscore that effective integration of the two modalities is essential to fully exploit their complementary strengths.

Text-only `ChatGPT` performs on par with `late fusion` for head ratings, but, notably, worse than `Skip-gram` on modifiers. Kurtyigit et al. (2025) report a higher performance of `ChatGPT` on Reddy et al.'s (2011) dataset (0.74 for modifier and 0.74 for head ratings). A reason for this difference could be that *ComposiGen* is more challenging, or that Reddy et al.'s (2011) data may have been part of `ChatGPT`'s training data, facilitating compositionality prediction on their target words.

Finally, our `denoising` approach performs poorly, with correlations of 0.11 for modifiers and 0.03 for heads. Contrary to our hypothesis that denoising strength can serve as a direct predictor of compositionality, our results suggest that this assumption does not hold in practice, at least not in the way implemented here. We explored several potential reasons for this weak performance: First, the approach depends on the quality and variability of the generated compound images. The choice of range and step size for denoising strength influences the resulting perturbations, and restricting the values to the empirically viable range of 0.8 to 1.0 may not have been sufficient to generate images that elicit meaningful compositional differences. However, replicating the experiments with SDXL (Podell et al., 2024) yielded comparable results, suggesting that the issue is not tied to a single image-to-image generation model. Second, the reliance on VQAScore may be insufficient for assessing compositional structure across different denoising strength values. Additional experiments using CLIPScore (Hessel et al., 2021) as alternative metric for selecting the best compound image from a sequence of generated images showed similarly weak effects. Determining the exact cause would require including human supervision into the compound image selection process. More fundamentally, the poor results may reflect that our use of denoising strength does not express compositionality in a sufficient manner to function as a reliable predictor.

5.3. Analysis

Our novel *ComposiGen* dataset offers an in-depth analysis of the compositionality of noun-noun compounds in terms of varying **concreteness levels** of the constituents, and how, given an individual head constituent, **varying modifiers** drive the visual transformation process from head to compounds. We conduct our analyses below with the unimodal and multimodal models that performed best in our experiments (Section 5.2), i.e., `Skip-gram`, `ViT`, and `late fusion`. Figure 6 shows their compositionality predictions for the example in Figure 2.

Concreteness Categories. Recall that we grouped the compounds in *ComposiGen* into the four concreteness categories AA, AC, CA, and CC (Section 3.3). Table 2 reports the correlation

values on these compound subsets (rows AA–CC). For comparison, row All repeats the performance on the entire set of compounds from Table 1. Note that the number of compounds within the subsets varies, each being considerably smaller than the entire set. Cells highlighted in green indicate correlation values exceeding the All setting. These improvements are statistically significant ($p < 0.01$), whereas correlations that do not exceed the All setting are generally not statistically significant. The comparison of the results between the AA and CC subsets to All strongly indicates that compositionality of *fully* concrete compounds is easier to predict than that of *fully* abstract ones; this applies to both the textual and the visual modalities, except `ViT` on Head. Image-only `ViT` fails to account for human ratings on both, abstract head and modifier constituents (the best score is 0.04 on Mod), suggesting that abstract concepts cannot be sufficiently captured by a single image (cf. Section 3.4). Interestingly, however, on both subsets of *abstract head* constituents (rows AA and CA) we observe a negative correlation across nearly all models for the head ratings, including text-only `Skip-gram`. This suggests that abstract heads systematically pose a challenge for models in encoding compositionality, and that the difficulty of accurately depicting abstract concepts does not fully explain this. Due to the overall weak performance of the parameter-based approach, we did not perform a subset analysis in this setting. It therefore remains unclear whether the higher correlations observed for compounds including concrete constituents would generalize to parameter-based compositionality predictions. One possibility is that concrete constituent images allow for more targeted changes into compound images, whereas abstract constituent images might require fewer content modifications in general. This question remains for future investigation.

Modifier Variation. Another key property of the *ComposiGen* dataset is that each head combines with multiple modifiers to form various compounds (e.g., the head *cake* is combined with *rice* or *cup* as modifiers). This property provides insights when analyzing the image sequences in *ComposiGen* (cf. Section 4.2), as it allows us to trace how an initial constituent image evolves into a compound image, revealing how the head constituent guides compound formation. Specifically, we can directly compare image sequences sharing the same head to examine how the generated images deviate from this head in dependence of different modifiers.

Figure 7 illustrates this with an example of *fruit cake* and *rice cake*. Across both compounds, we observe that with denoising strengths up to 0.88 the generated images largely resemble the constituent image, without yet depicting the compound. For *fruit cake* this resemblance persists even up to $\sigma_{0.9}$ and no unexpected changes occur across the different denoising strength values. Contrary to our expectation that the fruits from the initial image would remain in the compound image, they disappear as well when using $\sigma_{0.92}$ and higher values. In the case of *rice cake*, changes are visible already at early steps (e.g. $\sigma_{0.84}$), where rice grains become apparent. A particularly large shift occurs between $\sigma_{0.96}$ and $\sigma_{0.98}$, though the reason for this abrupt change is unclear given the preceding sequence.

6. Conclusion

We introduced *ComposiGen*, a novel, structured multimodal dataset with constituent-specific human-elicited compositionality ratings of compounds paired with images of the compounds and their constituents. Our experiments showed that images, in particular in combination with text, provide valuable information to predict the degree of compositionality of compounds. However, the choice of feature representations and the way of their combination is crucial — simple but specifically trained `Skip-gram` as well as simple `late fusion` outperformed state-of-the-art multimodal transformer models.

A novel method that uses denoising strength in text-guided constituent-to-compound image generation as direct compositionality predictor resulted in very weak correlations. This suggests that denoising strength is not a reliable indicator of compositionality, or that VQAScore may not have been effective for strength value selection. To determine the exact cause for that as well as for the underperformance of transformer-based models, further research towards more effective methods is needed.

The new *ComposiGen* dataset provides a rich resource for that purpose. Its coverage of varying concreteness levels furthermore allows for systematic analysis of factors influencing compositionality, and it supports exploration of how constituent images transform into compound images.

Limitations

Our approach is constrained by the choice of models, reliance on single representations, and simplification to one sense per constituent. Additionally, we retain all data for our experiments without outlier removal and our approach has not yet been applied to sets of compounds from other datasets. Addressing these aspects could further validate and extend our findings.

Ethics Statement

We report no ethical concerns related to this work. All human participation was voluntary, with fair compensation (£9/hour). No personally identifiable information was collected. All modeling experiments were conducted using properly cited open-source libraries. Materials for reproducibility, including data, prompts, and code, are available at https://github.com/jule-go/ComposiGen.

Acknowledgments

This research was supported by the DFG Research Grant SCHU 2580/4-1 *Multimodal Dimensions and Computational Applications of Abstractness*. We also thank the annotators for their contributions to the data annotation process, and the reviewers for useful feedback and suggestions.

7. Bibliographical References

Black Forest Labs. 2024. FLUX. Image generation model.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Marc Brysbaert, Amy Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I Spy a Metaphor: Large Language Models and Diffusion Models Co-Create Visual Metaphors. In *Findings of the Association for Computational Linguistics: ACL*, pages 7370–7388, Toronto, Canada.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2025. PIXART-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *The 18th European Conference on Computer Vision 2024*, pages 74–91, Milano, Italy.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 45(1):1–57.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, pages 4171–4186, Minneapolis, MN, USA.

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. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Peter Eigenschink, Thomas Reutterer, Stefan Vamosi, Ralf Vamosi, Chang Sun, and Klaudius Kalcher. 2023. Deep generative models for synthetic data: A survey. *IEEE Access*, 11:47304–47320.

Fritz Günther, Marco Alessandro Petilli, 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*, 112:104104.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, Las Vegas, NV, USA.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic.

Mohammed Khaliq, Diego Frassinelli, and Sabine Schulte Im Walde. 2024. Comparison of Image Generation Models for Abstract and Concrete Event Descriptions. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 15–21, Mexico City, Mexico (Hybrid).

Urban Knupleš, Diego Frassinelli, and Sabine Schulte im Walde. 2023. Investigating the Nature of Disagreements on Mid-Scale Ratings: A Case Study on the Abstractness-Concreteness Continuum. In *Proceedings of the SiGNLL Conference on Computational Natural Language Learning*, pages 70–86, Singapore.

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.

Sinan Kurtyigit, Diego Frassinelli, Carina Silberer, and Sabine Schulte im Walde. 2025. A Couch Potato is not a Potato on a Couch: Prompting Strategies, Image Generation, and Compositionality Prediction for Noun Compounds. In *Findings of the Association for Computational Linguistics*, pages 10766–10776, Vienna, Austria.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. 2024. Evaluating and Improving Compositional Text-to-Visual Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5290–5301, Seattle, WA, USA.

Sébastien Marcel and Yann Rodriguez. 2010. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1485–1488, New York, NY, USA.

Maximilian Maurer, Chris Jenkins, Filip Miletić, and Sabine Schulte im Walde. 2023. Classifying Noun Compounds for Present-Day Compositionality: Contributions of Diachronic Frequency and Productivity Patterns. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 40–51, Ingolstadt, Germany. Association for Computational Lingustics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space.

Filip Miletic 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.

OpenAI. 2025. GPT-4o. Large language model.

Dmitri Paisios, Nathalie Huet, and Elodie Labeye. 2023. Addressing the Elephant in the Middle: Implications of the Midscale Disagreement Problem Through the Lens of Body-Object Interaction Ratings. *Collabra: Psychology*, 9(1).

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.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. SemEval-2025 task 1: AdMIRe - advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, pages 2597–2609, Vienna, Austria.

Frédéric Piedboeuf and Philippe Langlais. 2023. Is ChatGPT the ultimate Data Augmentation Algorithm? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15606–15615, Singapore.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria.

Lewis Pollock. 2018. Statistical and Methodological Problems with Concreteness and other Semantic Variables: A List Memory Experiment Case Study. *Behavior Research Methods*, 50:1198–1216.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763.

Malak Rassem, Myrto Tsigkouli, Chris W Jenkins, Filip Miletić, and Sabine Schulte im Walde. 2024. Visualising changes in semantic neighbourhoods of english noun compounds over time. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 240–246.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta.

Stephen Roller and Sabine Schulte im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the*

2013 Conference on Empirical Methods in Natural Language Processing, pages 1146–1157.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28 – 34, Lancaster.

Sabine Schulte im Walde. 2024. Collecting and investigating features of compositionality ratings. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword expressions in lexical resources*, number 6 in Phraseology and Multiword Expressions. Language Science Press, Berlin.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016. GhoSt-NN: A Representative Gold Standard of German Noun-Noun Compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2285–2292, Portorož, Slovenia.

Vered Shwartz and Ido Dagan. 2019. Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A Foundational Language And Vision Alignment Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15617–15629, New Orleans, LA, USA.

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.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

A. Human Annotation Studies

A.1. Human Judgments on Compositionality Ratings

For the 200 compounds in *ComposiGen*, we asked fifteen human annotators to evaluate the extent to which the overall meaning of a compound can be related to the meaning of each of its constituents on a scale from 0 (= not compositional) to 5 (= compositional). Figure 8 provides the guidelines that we used to instruct the annotators, and Figure 9 shows an example item.

Compositionality of English compound nouns

Task: Evaluate the extent to which the overall meaning of a compound noun can be related to the meanings of its parts.
Examples:

- The meaning of *search engine* **is** related to the meanings of *search* and *engine*.
- The meaning of *blackmail* **is not** related to the meanings of *black* or *mail*.
- The meaning of *strawberry* **is not** related to the meaning of *straw*, but it **is** related to the meaning of *berry*.

Compositionality can be perceived differently. Therefore, please rate the degree of compositionality of a compound noun on the given scale.
Please note that there are two ratings because each compound noun consists of two parts. Please **answer both ratings** for every given word.
For each compound noun, there is a section where you can optionally add any personal comments about your rating.
Important information:

- Only for **native English speakers**.
- Please **rate all 28 words** on both of their parts.
- We are interested in your subjective opinion. Please decide **spontaneously** on a value on the scale.
- Please **read all text thoroughly**. Some ratings are designed to verify your attention to the task; please select the specified option when prompted. Failure to do so will result in the **rejection** of your submission.

Figure 8: Guidelines for the human annotators for rating the compositionality of compounds with respect to their constituents.

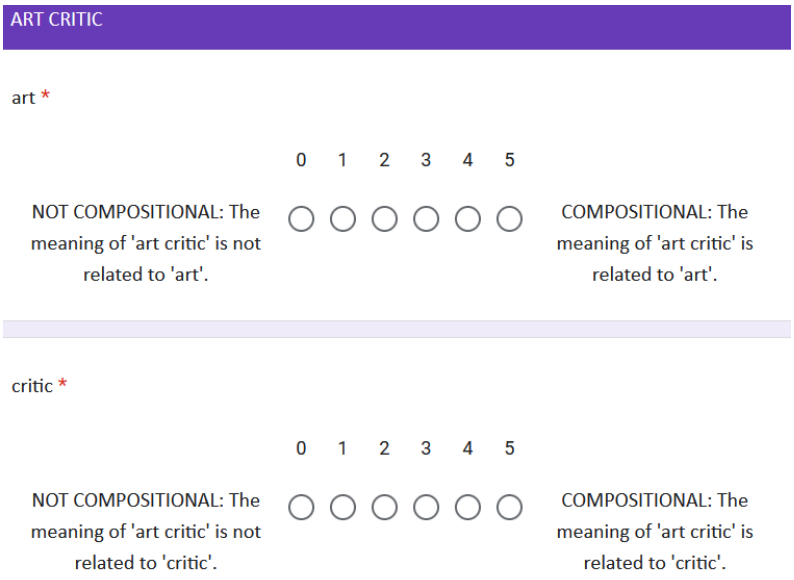

Figure 9: Example question asking for compositionality judgments on *art critic*.

A.2. Human Judgments on Image–Text Alignment

For the 372 instances in *ComposiGen* comprising a target word (consituent or compound) and its generated image (cf. Section 3.4), we asked three human annotators to rate how well the image aligns with the word on a scale from 0 (= not at all) to 4 (= perfectly). Figure 10 provides the guidelines that we used to instruct the annotators, and Figure 11 shows an example item. Each questionnaire contained 35 items, 31 items of *ComposiGen*, and 4 attention checks (i.e., 12 questionnaires in total). The annotators were recruited through the Prolific platform and were compensated with an average award of £11/hour. They were not permitted to complete more than one task.

Word–Image Correspondence

Task Instruction:
You will be shown 35 items, each consisting of a question along with an image. Your task is to judge how well the image depicts the given word mentioned in the question.

Steps:

1. Identify the target word by reading the question.
2. Review the image.
3. Select an option on the scale from 0 (= image depicts word not at all) to 4 (= image depicts word perfectly).

Quality check:
To ensure the quality of the responses, some instances are designed to serve as control checks, with exactly one correct answer. In these instances, please select the option specified in the question. Please note that if you fail to correctly select the specified options, your entire submission will be rejected and you won't receive any payment.

Purpose of the task:
Your responses will be used in a research study.

Important information:

- Only for ***native English speakers***.
- Please ***evaluate all 35 images***. Submitting an incomplete questionnaire will result in the rejection of your submission.
- We are interested in your subjective opinion. Try not to overthink your answer. Please decide ***spontaneously*** on a value on the scale.
- Please ***read all questions thoroughly***. Failure to quality checks will result in the rejection of your submission.
- Feel free to quit at any time without giving a reason (note that you won't be paid in this case).

Figure 10: Guidelines for the human annotators for rating image–text alignment.

Alignment Ratings. The annotation results are represented in Table 3. The average rating across the mean human ratings for each of the 372 items is 2.48 (and a standard deviation of 0.77), so overall the generated images align very well with the constituents/compounds.

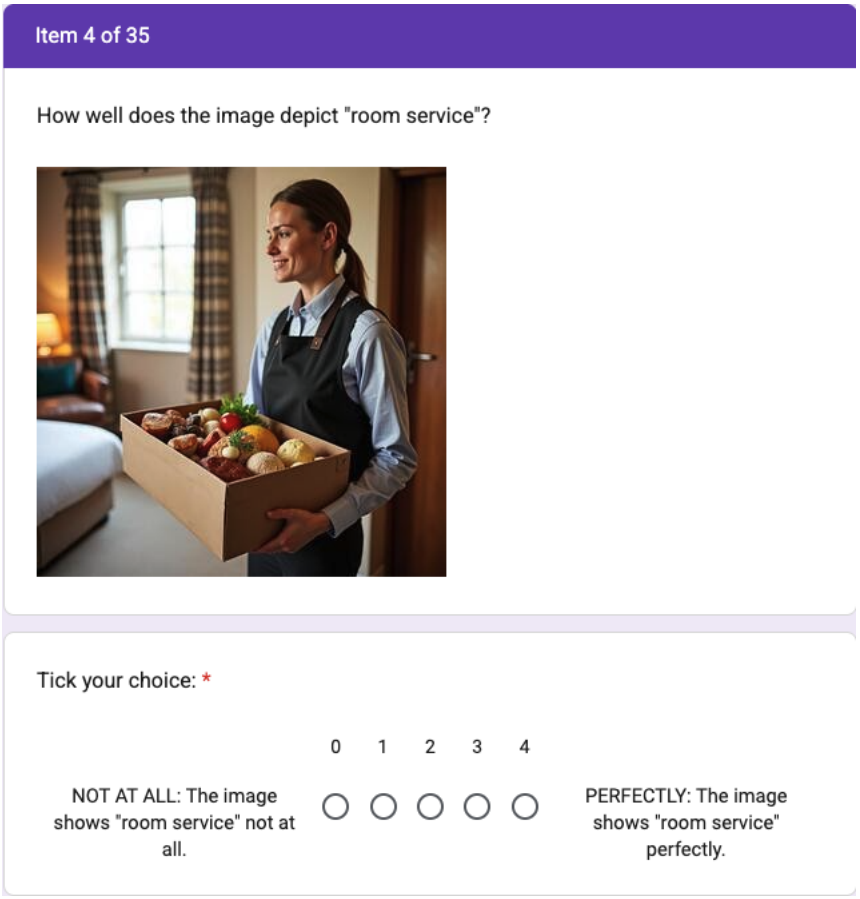

Figure 11: Example question asking for image–text alignment judgments on *room service*.

Category	Human ↑ (0-4)	VQAScore ↑ (0-1)	Correlation ↑ ρ	Count
Compound	2.61 (.78)	0.864 (.18)	0.64	200
Head	1.98 (.58)	0.859 (.19)	0.63	23
Modifier	2.36 (.78)	0.879 (.17)	0.71	136
Head/Modifier	2.51 (.74)	0.878 (.20)	0.85	13
CC	2.99 (.71)	0.904 (.14)	0.52	107
CA	2.35 (1.0)	0.829 (.21)	0.80	28
AC	2.25 (.74)	0.830 (.19)	0.57	49
AA	1.69 (.97)	0.758 (.25)	(0.42)	16
C	2.67 (.74)	0.889 (.17)	0.69	121
A	1.49 (.79)	0.844 (.18)	0.50	51
Mean	2.48 (.77)	0.87 (.18)	0.66	372

Table 3: Image-Text Alignment: Human mean ratings and VQAScores across word categories (compound or constituent; Head/Modifier: constituent tokens that appear as both, head and modifier, e.g., *dog*)) and concreteness categories C(oncrete) and A(bstract) for compounds and constituents. Numbers in parentheses denote standard deviation. ρ: Spearman's rank correlation coefficient between human ratings and VQAScores. All correlations are statistically significant ($p < 0.005$) except for AA. Count: Distribution of image–text pairs across categories in *ComposiGen*.

B. Prompting ChatGPT for Noun Definitions

For each constituent and compound of *ComposiGen* we generated noun definitions using ChatGPT. The instructional prompt we used is listed in Figure 12. Following this instructional prompt, the model was sequentially prompted for compound and constituent noun definitions, within a single session. The generated outputs are available in our repository.

Hi! You are an intelligent machine, generating prompts that are suitable inputs for image generation models. In order to generate good images, it is necessary to have prompts with a fine-grained level of detail that are of high quality. You are the expert for generating such image generation prompts! I will provide you with a list of targets. Those can be unigrams or bigrams. Every target is written in a separate line. !!! Your task is to first understand the targets, and then to list three different noun definitions of this target you consider suitable as image generation prompt !!! The definition prompts can consist of tags or natural language sentences. They should span a maximum of 75 tokens each. Every prompt should be highly informative on itself. Don't include duplicate prompts, the three definition prompts should differ from each other. Please list the definition prompts as if writing to a txt-file. Please refer to the example below for the desired format.

——— Example: my input ———

couch potato
couch
potato
...

——— Example: your output ———

Definitions for "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.'

Definitions for "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.'

Definitions for "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.'

...

Please let me know if we can start or if you have questions that need further clarification!

Figure 12: Prompt for generating noun definitions using ChatGPT.

C. Compositionality Prediction Approaches

C.1. Fusion of Text and Vision

In our `early fusion` approach we experiment with combining text and image embeddings into one multimodal representation using concatenation, mean-pooling, or a weighted aggregation with weighting factor α. Figure 13 shows the influence of the chosen combination method on the compositionality prediction task.

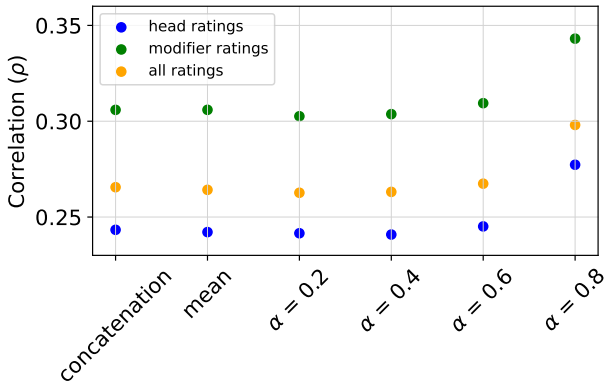

Figure 13: Spearman's ρ for different early fusion combinations of text and vision embeddings.

For our `late fusion` approach we combine text-based and image-based compositionality predictions using a weighting factor β. Figure 14 illustrates the effect of varying β on the compositionality prediction task.

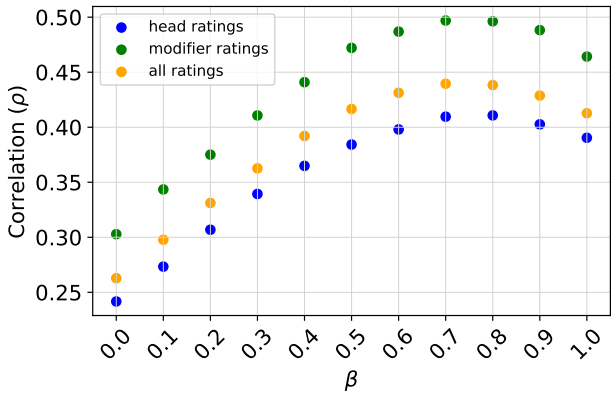

Figure 14: Spearman's ρ for different late fusion combinations of text and vision compositionality predictions.

C.2. Prompting ChatGPT for Compositionality Ratings

The instructional prompt for generating compositionality ratings with ChatGPT is shown in Figure 15. Following this instructional prompt, the model was sequentially prompted for all *ComposiGen*'s compounds to provide compositionality ratings with respect to the head and the modifier, within a single session. The generated outputs are available in our repository.

Hi! You are an expert annotator for linguistic tasks. This time your task involves complex English expressions, in particular noun-noun compounds. Your task is to evaluate the extent to which the overall meaning of a compound can be related to the meaning of its parts. Here are some example compounds with different degrees of compositionality:

- The meaning of search engine is related to the meanings of search and engine.
- The meaning of blackmail is not related to the meanings of black or mail.
- The meaning of strawberry is not related to the meaning of straw, but it is related to the meaning of berry.

I will provide you with a list of noun-noun compounds. Every compound is written in a separate line. Its constituents are separated with whitespace. !!! Your task is to understand the meaning of the compound as well as the meanings of its constituents and then to provide constituent-specific ratings on compositionality. !!! The compositionality ratings should be on a scale between 0 (definitely opaque, i.e. low compositionality) and 5 (definitely transparent, i.e. high compositionality). Feel free to use the whole range. For each compound you are expected to provide two compositionality ratings: one with respect to the first constituent and one with respect to the second constituent. Please provide the ratings in the format "compound,constituent,rating", using a separate line for each compound-constituent combination, as if writing to a txt-file. Please refer to the example below for the desired format.

——— Example: my input ———
flea market
spelling bee
graduate student
...
——— Example: your output ———
flea market,flea,0.379
flea market,market,4.714
spelling bee,spelling,4.815
spelling bee,bee,0.517
graduate student,graduate,4.700
graduate student,student,5.000
...
Please let me know if we can start or if you have questions that need further clarification!

Figure 15: Prompt for collecting compositionality ratings using ChatGPT.