A Visual Tour Of Current Challenges In Multimodal Language Models

Shashank Sonkar, Naiming Liu, Richard G. Baraniuk
Rice University
{ss164, nl35, richb}@rice.edu

Abstract

Transformer models trained on massive text corpora have become the de facto models for a wide range of natural language processing tasks. However, learning effective word representations for *function words* remains challenging. Multimodal learning, which visually grounds transformer models in imagery, can overcome the challenges to some extent; however, there is still much work to be done. In this study, we explore the extent to which visual grounding facilitates the acquisition of function words using stable diffusion models that employ multimodal models for text-to-image generation. Out of seven categories of function words, along with numerous subcategories, we find that stable diffusion models effectively model only a small fraction of function words – a few pronoun subcategories and relatives. We hope that our findings will stimulate the development of new datasets and approaches that enable multimodal models to learn better representations of function words.

1 Introduction

Transformer models [1, 2] are currently state-of-the-art across many natural language processing (NLP) tasks such as question answering [3, 4], information retrieval [5], inference [6, 7], and machine translation [8, 9]. Transformers are masked language models which use the self-attention mechanism [8] to output contextualized word embeddings. However, not all words can be modeled effectively using context information [10–14]. Function words like conjunctions, pronouns, prepositions etc are difficult to learn using the masked language modeling loss [15, 16].

An alternative method to learn the representations of function words is to use multimodal learning to ground the language models visually in natural images. These multimodal language models (MLMs) [17–19] learn an aligned representation of images and text. Recently, stable diffusion models (SDMs) [20] have gained popularity for text-to-image generation. SDMs take a natural language prompt as input, encode the prompt using a MLM, and then generate an image capturing the semantics of the prompt.

The key point of this short paper is that SDMs can be used to gain new and useful insights into the workings of MLMs. In our study, we use carefully crafted prompts with seven different categories of function words and their sub-categories [21] to probe SDMs. Next, we visually inspect whether the images capture the semantics of the function words. Despite the MLM in SDMs being visually grounded, we discover that, for the majority of function words, the images generated do not accurately convey the meaning entailed by the prompts.

Our findings can inspire innovative research in the construction of datasets to improve MLMs' understanding of function words, which are fundamental building blocks of English grammar. We also provide the code on github for readers to replicate our findings and explore further.

2 Background

2.1 Stable Diffusion Model

Stable diffusion model (SDM) is a trending, open source text-to-image generation model which utilizes latent diffusion model conditioned on the text embeddings. As shown in figure 1, SDM is composed of three main components: 1) a

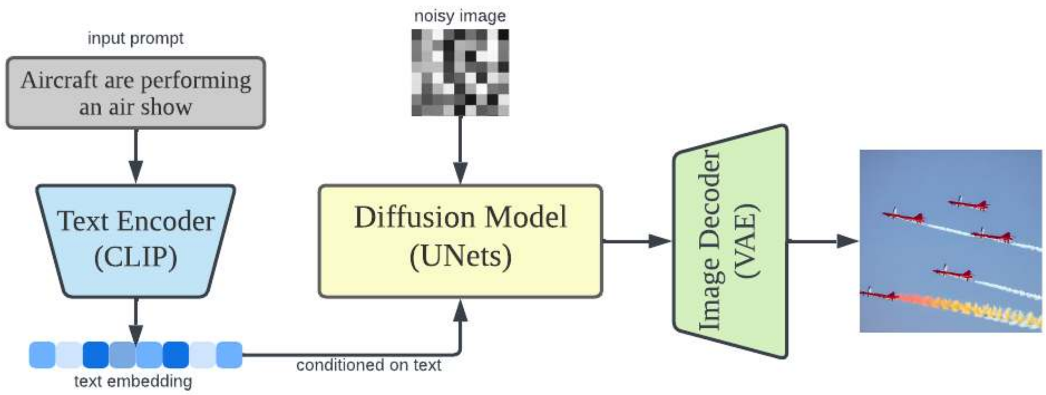


Figure 1: A Stable Diffusion Model (SDM) architecture [20] has three main components: a frozen CLIP ViT-L/14 as text encoder, UNets as diffusion model, and variational autoencoders (VAE) as image decoder. The text encoder embeds the input prompt into a high-dimensional representation, which is fed to the diffusion model together with a noisy sample. Then, the image decoder converts the diffusion model's latent space image representation to a real image that captures the semantics of the input prompt.

Category	Sub-category	Examples
Pronouns	Subject	he, she, they
	Object	him, her, them
	Possessive	his, her, our
	Indefinite	nobody, everyone
	Reflexive	himself, herself
Conjunctions		and, but, yet, or
Interrogatives		who, which, where

Category	Sub-category	Examples
Determiners	Article	a, an, the
	Numeral	one, two, ten
	Quantifier	little, many, few
Qualifiers		not, always, very
Prepositions	Place	in, on, under
	Movement	up, down, towards
	Particle	on, off, with

Table 1: Categories of function words with examples. For each category, we carefully design prompts to probe stable diffusion models to explore their understanding of function words.

text encoder (a frozen CLIP ViT-L/14 [18]), 2) a diffusion model (U-Nets [22]) and 3) an image decoder (variational autoencoder [23]). The text encoder takes a natural language prompt as input and transforms it into a high-dimensional embedding using the self-attention mechanism [8]. Then, using this embedding and noise sample as input, the diffusion model and the image decoder output the target image.

2.2 Multimodal Language Models

We discuss a few specifics of multimodal language models (MLMs) in this section. We focus on CLIP, a type of MLM, since it is used by the SDM that we use for our experiments. CLIP aligns images and their textual descriptions to embed text and image in the same vector space. CLIP contains a text encoder and an image encoder. Both are optimized simultaneously using the principle of contrastive learning [24] by maximizing the cosine similarity between paired text and image embeddings while minimizing the cosine similarity between unpaired ones. CLIP representations have many applications including visual question-answering [25], automatic image captioning [26] and object navigation [27].

2.3 Linguistics Review: Function Words

Words can be broadly classified into two categories — function words like determiners, prepositions, pronouns, etc and content words like nouns, verbs, adjectives, etc. It is hard to capture the semantics of content words as compared to function words using the contextual information [10–14]. Since language models using context to predict the missing/next word in the pre-training objectives, they have been shown to perform poorly on function words [15, 16]. We believe that grounding text encoder in natural images can alleviate these shortcomings, but the question remains to what extent. As mentioned before, we primarily focus on the CLIP text encoder and explore how visually-grounded language model performs as text encoder of SDM when it comes to modeling functions words.

(a) *He* is dancing in the rain

(b) *She* is dancing in the rain

(c) *We* are dancing in the rain

(d) *Nobody* in the group is wearing a hat

(e) *Some* in the group are wearing a hat

(f) *Everyone* in the group is wearing a hat

(g) I shook hands with *myself*

(h) The boy punched *himself* in the face

(i) She patted *herself* for a job well done

Figure 2: Sample images depicting SDM's success (green border) and failure (red border) in capturing the semantics of different subcategories of **pronouns**. (a)–(c) show that the information about gender and count implicit in subject pronouns like he, she, we is accurately depicted. But, for indefinite pronouns, SDMs fail to capture the notion of negatives ((d) nobody), existenial ((e) some), and universals ((f) everyone). Likewise SDMs fail to capture the meaning of reflexive pronouns like (g) myself, (h) himself, (i) herself.

3 Experiments

In this section, we divide function words into seven categories (listed in table 1) and visually inspect each category to check if it can be modeled through SDMs. We list out language prompts used to probe SDM[1] for each category and present a figure that contains multiple images that spans all its subcategories. Green/red border around the images are used to identify if SDM successfully/unsuccessfully outputs an image that captures the semantics of the input prompt.

Note that we provide only a sample of images in the experiment section. Please refer to the appendix for more samples.

3.1 Pronouns

Pronouns are used in English grammar as a substitute for nouns. They are divided into five categories: 1) subject pronouns e.g., he, she, we 2) object pronouns e.g., him, her, them 3) possessive adjectives e.g., his, her, our 4) indefinite pronouns e.g., few, many, nobody, everyone and 5) reflexive pronouns e.g., himself, herself, ourselves.

Subject pronouns, object pronouns, and possessive pronouns reflect the gender and count of the entity they refer to. Through language prompts like "*He* is dancing in the rain', '*She* is dancing in the rain', and '*We* are dancing in the rain', We can test the diffusion model's ability to produce visuals that appropriately depict the gender and count of entities that each language prompt embodies. Our experiments reveal that, for the most part, the images in figure 2a–2c did accurately represent the gender and count.

However, that is not the case for indefinite pronouns and reflexive pronouns. Indefinite pronouns are divided into negatives (none, no one, nobody), assertive existential (some, someone, somebody), and universals (everyone, everybody). We probed the semantics of indefinite pronouns using prompts like '*No one* in the group is wearing a hat', '*Some* in the group are wearing a hat', and '*Everyone* in the group is wearing a hat'. We found that the diffusion model is not able to differentiate amongst the three subcategories of indefinite pronouns as can be seen in the figure 2d–2f.

[1]We use "*Stable Diffusion v1-4*" model released at `https://huggingface.co/CompVis/stable-diffusion-v1-4`

(a) The ball is *in* the box (b) The ball is *on* the box (c) The ball is *next to* the box (d) The ball is *behind* the box

(e) He is walking *up* the stairs (f) He is walking *down* the stairs (g) The girl ran *from* the dog (h) The girl ran *towards* the dog

(i) The lights are *on* (j) The lights are *off* (k) He eats oats *with* milk (l) He eats oats *without* milk

Figure 3: Sample images depicting SDM's success (green border) and failure (red border) in capturing the semantics of different subcategories of **prepositions**. Even for prepositions of place which can be learned easily through visual grounding, images for (a) in, (b) on, (c) next, and (d) behind show that SDMs do not understand this subcategory of preposition. Same analysis holds for prepositions of movement like (e) up, (f) down, (g) from, and (h) towards. Unsurprisingly, SDMs also fail for the hardest abstract category of particles, which include (i) on, (j) off, (k) with, and (l) without.

Reflexive pronouns make reference to the formerly mentioned noun and include words that end with which $-self$ and $-selves$. We investigated these pronouns using prompts like 'The boy punched *himself* in the face', 'She patted *herself* for a job well done.' and 'I shook hands with *myself*'. We found that images in figure 2g–2i did not reflect the reflexive nature of these pronouns.

3.2 Prepositions and Particles

Prepositions are an interesting category of words that are simple to illustrate with visuals since they convey spatial and temporal relationships. Despite their simplicity we noted that modeling prepositions is rather challenging for current diffusion models. Spatial relations are covered by prepositions of place e.g. in, on, under, etc. For testing preposition of place, we used prompts like 'The ball is *in* the box', 'The ball is *on* the box', 'The ball is *under* the box', etc.

We observed that the diffusion models were not able to differentiate amongst prepositions of place as can be seen in the figure 3a–3d. Even though the model successfully outputs the image in figure 3a for the prompt 'The ball is *in* the box', it outputs similar images for other prepositions of place 3b–3d, thereby raising the question — did SDM really

(a) *A* dice rolled on the table
(b) *An* aircraft performing an air show
(c) The dog is guiding *the* sheep
(d) There is *one* orange in the photo
(e) There are *two* oranges in the photo
(f) There are *ten* oranges in the photo
(g) *Few* oranges in the basket
(h) *Many* oranges in the basket
(i) *Little* milk in the bottle
(j) *Lot* of milk in the bottle
(k) The plate is a *little* dirty
(l) The plate is *very* dirty
(m) The sky is *not* orange
(n) The sky is *always* orange

Figure 4: Sample images depicting SDM's success (green border) and failure (red border) in capturing the semantics of different subcategories of **determiners** and **qualifiers**. Unlike the case of subject pronouns, images (a)–(c) show that SDMs cannot capture the notion of singularity implicit in articles like a, an, and the. They also exhibit weak understanding of cardinal numerals like (d) one, (e) two, and (f) ten. Concept of *less* and *more* suggested by quantifiers like (g) few, and (h) many for countable nouns and quantifiers like (i) little, and (j) lot for uncountable nouns is also not modeled by SDMs. (k)–(n) Qualifiers, which likewise cover the concept of *less* and *more* for adjectives and adverbs, too fail to be modeled by SDMs.

understand the meaning of 'in' in figure 3a? In this paper, we assume that SDM does not in fact understand the meaning of *in* in figure 3a since it outputs similar images irrespective of the input.

Likewise for prepositions of movements which model temporal relations, creative prompts like 'He is walking *up* the stairs', He is walking *down* the stairs', 'The girl ran *towards* the dog', 'The girl ran *away* from the dog' were used. In this experiment as well, we concluded from figure 3e–3h that the semantic properties of prepositions in questions were not accounted for.

We also consider a special type of prepositions: particles which reflect the state of an entity. Examples include on, off, with, without, etc. Using inventive prompts like 'The lights are *on*', 'The lights are *off*', 'He eats oats *with* milk.', 'He eats oats *without* milk.', we notice that diffusion models are ineffective at modeling particles either as seen in figure 3i–3l.

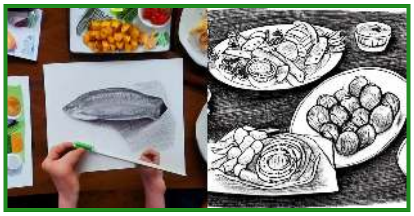
3.3 Determiners and Qualifiers

Determiners include articles, cardinal numerals, and quantifiers. Articles like, a, an, and the modify the nouns by placing a restriction on them to show how particular or generic they are. Articles generally indicate a single *unit* of noun that is being modified. Cardinal numerals and quantifiers also modify the nouns by indicating their quantity.

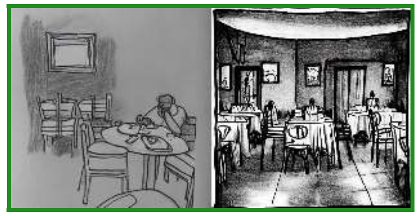
(a) *What* are we eating for dinner

(b) *Where* did we eat the dinner

(c) *Who* are we eating with for dinner



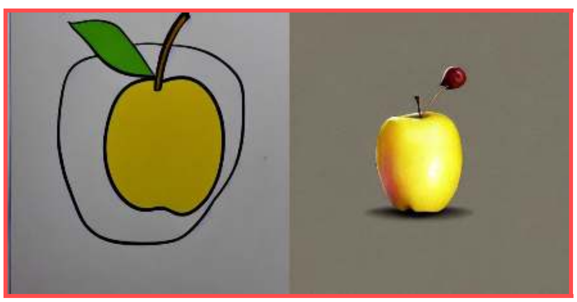
(d) Draw a picture of *what* we had for the dinner



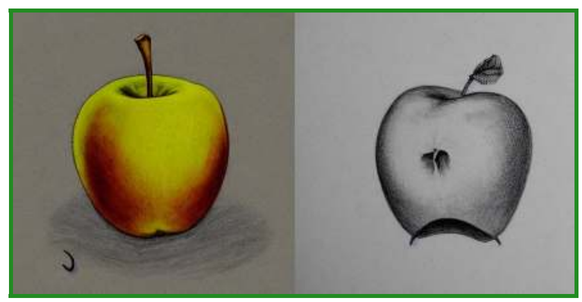
(e) Draw a picture of *where* we had the dinner

(f) Draw a picture of *who* we had the dinner with

Figure 5: Sample images depicting SDM's success (green border) and failure (red border) in capturing the semantics of **interrogatives** and **relatives**. SDMs are unable to understand that the answers to questions like (a) what, (b) where, and (c) who are respectively, things, locations, and persons. However, when (d) what, (e) where, and (f) who are used as relatives, SDMs show that MLMs can capture their essence. This category of function words is the simplest of all function words since it comprises words that co-occur with different contexts in texts, making it understandable even without multimodal learning.



(a) Draw an apple *and* a banana



(b) Draw *either* an apple *or* a banana

Figure 6: Sample images depicting SDM's success (green border) and failure (red border) in capturing the semantics of **conjunctions**. This category offers an intriguing scenario because it explores the area of logic and reasoning. Conjunctions *either/ or* and *and* express a decision choice, however the images (a) and (b) show that SDM is unable to comprehend the choice implied by the conjunctions.

Using nouns that have the same single and plural form, we examine whether the diffusion model captures the singularity implied in the articles. Prompts include '*A* dice rolled on the table', '*An* aircraft performing an air show', 'The dog is guiding *the* sheep'. Images in figure 4a–4c reveal that the diffusion models do not capture the singularity.

Cardinal numerals also offer a similar class of words to test the diffusion models. Simply using number words like one, two, and ten to fill the *mask* and looking at images generated for prompts 'There is/are *mask* orange/oranges in the photo', we observed that diffusion models do not understand the concept of numbers as can be seen in figure 4d–4f.

Next interesting category of determiners is quantifiers which includes words like little, few, lot, many etc. Using sentences like 'There is *little* milk in the bottle', 'There is a *lot of* milk in the bottle', '*Few* oranges in the basket', and *Many* bananas in the basket', we find that the diffusion models lack the comprehension for quantifiers as well as can be seen in figure 4g– 4j.

Qualifiers are similar to quantifiers and numbers which also limit or enhance another word's meaning, but are associated with the adjectives and adverbs rather than the nouns. Examples include not, never, always, a little, very, etc. We designed few prompts like 'The sky is *not* orange', 'The sky is *never* orange', 'The sky is *always* orange', 'The plate is a *little* dirty', 'The plate is *very* dirty'. Unsurprisingly as with the case with quantifiers and numbers, figure 4k–4n shows that the diffusion models fails to recognize the negation as well as cannot model the intensity of the qualifier .

3.4 Interrogatives and Relatives

What, Where, and Who are examples of interrogative words used to pose a query. The reason we use these words to probe the diffusion model lies in the answer to these queries. The answer for what, where, and who are objects, places, or person respectively. 'Wh-' words can also act as relatives in descriptive sentences where they do not pose a question.

Thus, to probe the diffusion model with 'wh-' words we use both questions as well as descriptive prompts like '*What* are we eating for dinner', '*Where* did we eat the dinner?', '*Who* are we eating dinner with?', 'Draw a picture of *what* we had for the dinner' 'Draw a picture of *where* we had the dinner.', and 'Draw a picture of *who* we had the dinner with'. We conclude from the images that the diffusion model does not understand the semantics of the answers to the interrogative (figure 5a–5c). But surprisingly when used in descriptive sentence, it was able to comprehend the answers (figure 5d–5f).

3.5 Conjunctions

Conjunctions play an important role in english grammar by connecting two sentences. Examples include and, but, yet, either, or etc. We use *or* and *and* since they provide a choice between two alternatives. With language prompts like 'Draw an apple and a banana' and 'Draw either an apple or a banana', we can probe the diffusion model to understand if it can understand the notion of a choice implicit in *or* vs *and*. We find that the model generates just an apple both the cases (figure 6a and 6b).

4 Conclusions

We have explored the limitations of learned representations of function words in multimodal language models by a visual tour of images generated by stable diffusion models. Our results indicate that the semantics of function words are poorly understood by these language models. In particular, stable diffusion models only work for select pronoun subcategories and the category of relatives out of the seven categories of function words and their multiple subcategories. Future work by the research community should focus on methods to remedy these shortcomings, such as the construction of function word datasets.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[3] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: deep contextualized entity representations with entity-aware self-attention. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6442–6454. Association for Computational Linguistics, 2020.

[4] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[5] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*, 2021.

[6] Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. Entailment as few-shot learner. *ArXiv*, abs/2104.14690, 2021.

[7] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[9] Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3159–3166, 2019.

[10] Nicholas Asher, Tim van de Cruys, et al. Content vs. function words: The view from distributional semantics. In *Proceedings of Sinn und Bedeutung*, volume 22, pages 1–21, 2018.

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

[12] Raffaella Bernardi, Georgiana Dinu, Marco Marelli, and Marco Baroni. A relatedness benchmark to test the role of determiners in compositional distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 53–57, 2013.

[13] Karl Moritz Hermann, Edward Grefenstette, and Phil Blunsom. " not not bad" is not" bad": A distributional account of negation. *arXiv preprint arXiv:1306.2158*, 2013.

[14] Tal Linzen, Emmanuel Dupoux, and Benjamin Spector. Quantificational features in distributional word representations. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 1–11, 2016.

[15] Najoung Kim, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, et al. Probing what different nlp tasks teach machines about function word comprehension. *arXiv preprint arXiv:1904.11544*, 2019.

[16] Rui P Chaves and Stephanie N Richter. Look at that! bert can be easily distracted from paying attention to morphosyntax. *Proceedings of the Society for Computation in Linguistics*, 4(1):28–38, 2021.

[17] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018.

[18] 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. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

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

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[21] Thomas P Klammer. *Analyzing English Grammar, 6/e*. Pearson Education India, 2007.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[23] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[24] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[25] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

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

[27] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838, 2022.

A Appendix

In the appendix, we layout creative prompts that could demonstrate the characteristics of each category of the function words. For each prompt, we provide four additional images generated by the stable diffusion model.

A.1 Pronouns

Pronouns contains five categories: subject pronouns (SP), object pronouns (OP), possessive adjectives (PA), indefinite pronouns (IP), reflexive pronouns (RP). Table 2 and Table 3 shows additional prompts for each category and their corresponding generated images.

Prompts	Images
He is dancing in the rain. **(SP)**	
She is dancing in the rain. **(SP)**	
We are dancing in the rain. **(SP)**	
The teacher hugs **him**. **(OP)**	
The teacher hugs **her**. **(OP)**	
The boy is holding **his** hand. **(PA)**	
The boy is holding **her** hand. **(PA)**	

Table 2: Images generated by Stable Diffusion Model for prompts with pronouns. This table covers first three subcategories of pronouns: subject pronouns (SP), object pronouns (OP), possessive adjectives (PA). The pronoun words are colored in red.

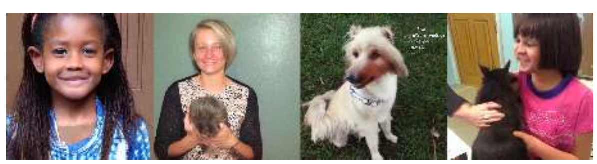
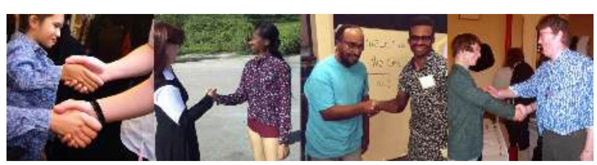
Prompts	Images
Nobody in the group is wearing a hat. **(IP)**	
No one in the group is wearing a hat. **(IP)**	
Some in the group are wearing hats. **(IP)**	
Someone in the group is wearing a hat. **(IP)**	
Everyone in the group is wearing a hat. **(IP)**	
Everybody in the group is wearing a hat. **(IP)**	
The boy punched **himself** in the face. **(RP)**	
She patted **herself** for a job well done. **(RP)**	
I shook hands with **myself**. **(RP)**	

Table 3: Images generated by Stable Diffusion Model for prompts with pronouns. This table covers last two subcategories of pronouns: indefinite pronouns (IP), reflexive pronouns (RP). The pronoun words are colored in red.

A.2 Prepositions and Particles

Prepositions include two categories, prepositions of place (PoP), which models entity positions and prepositions of movement (PoM) which signify temporal relations. We also include a special type of prepositions – particle (Par), which shows the state of an object. Table 4 and Table 5 shows prompts and generated images for propositions.

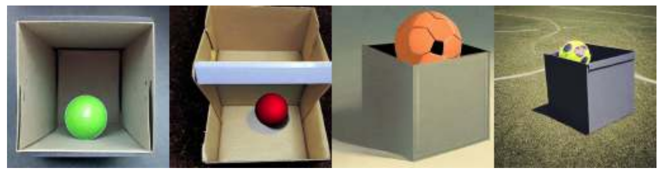
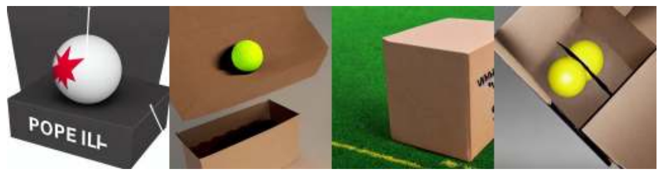
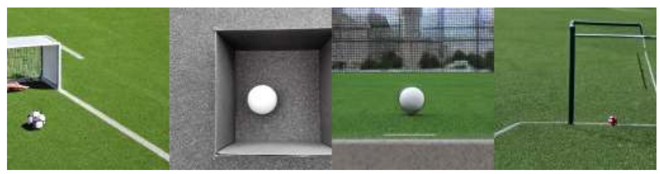
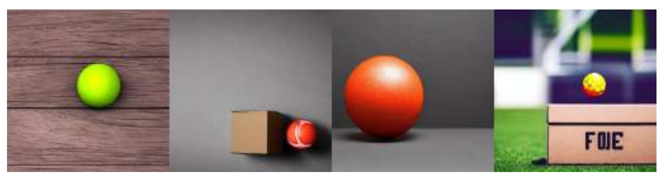
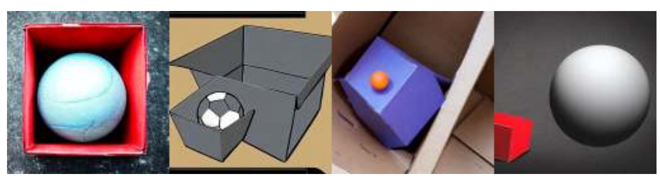
Prompts	Images
The ball is **in** the box. **(PoP)**	
The ball is **on** the box. **(PoP)**	
The ball is **under** the box. **(PoP)**	
The ball is **next to** the box. **(PoP)**	
The ball is **behind** the box. **(PoP)**	
The ball is **in front of** the box. **(PoP)**	
The ball is **in between** the boxes. **(PoP)**	

Table 4: Images generated by Stable Diffusion Model for prompts with prepositions. This table covers the first subcategory of prepositions: preposition of place (PoP). The preposition words are colored in red.

Prompts	Images
He is walking **up** the stairs. **(PoE)**	
He is walking **down** the stairs. **(PoE)**	
The girl ran **towards** the dog. **(PoE)**	
The girl ran **from** the dog. **(PoE)**	
The light is **on**. **(Par)**	
The light is **off**. **(Par)**	
Oats **with** milk. **(Par)**	
Oats **without** milk. **(Par)**	
Coffee **with** creamer. **(Par)**	
Coffee **without** creamer. **(Par)**	

Table 5: Images generated by Stable Diffusion Model for prompts with prepositions. This table covers the last two subcategory of prepositions: preposition of movement (PoE) and Particles (Par). The prepositions are colored in red.

A.3 Determiners and Qualifiers

Determiners is composed of three types: articles (AR), cardinal numerals (CN), and quantifiers (QUAN). Table 6 and Table 7 demonstrates some examples in this category. Another interesting type of function words is qualifiers (QUAL), where some creative prompts and images are presented in Table 7.

Prompts	Images
A dice rolled on the table. **(AR)**	
An aircraft performing an air show. **(AR)**	
The sheep is eating grass. **(AR)**	
There is **one** orange in the photo. **(CN)**	
There are **three** oranges in the photo. **(CN)**	
There are **ten** oranges in the photo. **(CN)**	
There is **little** milk in the bottle. **(Quan)**	
There is **a lot of** milk in the bottle. **(Quan)**	

Table 6: Images generated by Stable Diffusion Model for prompts with determiners. This table covers the first three subcategories of determiners, Article (AR), Numeral Cardinals (NC) and Quantifiers (Quan). The determiner words are colored in red.

Prompts	**Images**
There are **few** bananas on the table. **(Quan)**	
There are **many** bananas on the table. **(Quan)**	
Few oranges in the basket. **(Quan)**	
Many oranges in the basket. **(Quan)**	
The sky is **not** green. **(Qual)**	
The sky is **never** green. **(Qual)**	
The sky is **always** green. **(Qual)**	
The plate is **a little** dirty. **(Qual)**	
The plate is **very** dirty. **(Qual)**	

Table 7: Images generated by Stable Diffusion Model for prompts with determiners. This table covers two subcategories of determiners: Quantifiers (Quan) and Qualifiers. (Qual). The determiner words are colored in red.

A.4 Interrogatives and Relatives

Interrogatives and relatives both represent function words that are ‘Wh-’ alike, such as What, Where or Who. Interrogatives (Int) usually raises a question while relatives (Rel) do not. More prompts and generated images are provided in Table 8.

Prompts	Images
Where did we have the coffee? **(Int)**	
What are we eating for dinner? **(Int)**	
Who are we eating with for dinner? **(Int)**	
What is climbing the tree? **(Int)**	
Who is climbing the tree? **(Int)**	
Draw a picture of **what** we had for dinner. **(Rel)**	
Draw a picture of **who** is climbing the tree. **(Rel)**	
Draw a picture of **where** we had our coffee. **(Rel)**	

Table 8: Images generated by Stable Diffusion Model for prompts with interrogatives (Int) and relatives (Rel). The interrogative words are colored in red.

A.5 Conjunctions

In this section, it provides some supplementary prompts and images in Table 9 to show how stable diffusion model could perform with conjunctions.

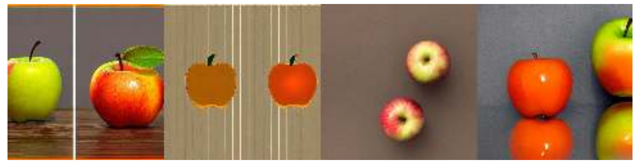
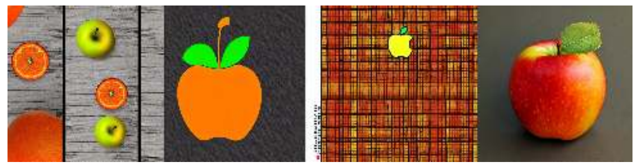
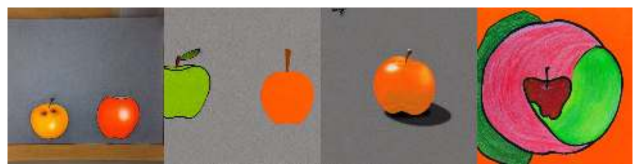
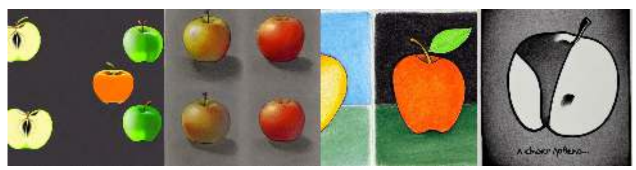
Prompts	Images
Generate an image of **either** an apple **or** an orange.	
Generate an image of apple **and** orange.	
Draw an apple **and** an orange.	
Draw **either** an apple **or** an orange.	
It was sunny **but** now it is raining.	
It was rainy **but** now it is sunny.	
The bottle is taller **than** the cup.	
The cup is taller **than** the bottle.	

Table 9: Images generated by Stable Diffusion Model for prompts with conjunctions. The conjunction words are colored in red.