Hit or Be Hit: Tests of (Pre)Compositional Abilities in Vision and Language Models

Mădălina Zgreabăn, Albert Gatt, Pablo Mosteiro
Utrecht University, The Netherlands
b.m.zgreaban@uu.nl, a.gatt@uu.nl, p.j.mosteiroromero@uu.nl

Abstract

Although many benchmarks have been recently proposed to test the compositional abilities of vision and language models, there has been little systematic review or comparison between them. Our study classifies such datasets, their benefits, and shortcomings. Additionally, we expand the VALSE benchmark with two large datasets of the active/passive and ditransitive/dative constructions, on which CLIP, Flava, and LiT were tested. We show models achieve roughly random-chance performance in differentiating captions from foils. We provide evidence suggesting that the linguistic constructions or the order of thematic arguments in captions or foils do not change the models' performance. However, their grammaticality, simplicity, and origin dataset seem to do so.

1 Introduction

Compositionality is constructing the meaning of a structure by its elements and their combination (Partee, 2004). General compositional abilities are important for models to generalize well on new tasks or data, especially in retrieval, text-to-image generation (Ma et al., 2023), and general artificial intelligence (Xu et al. 2022).

This definition of compositionality presupposes that models have robust representations of basic linguistic phenomena, enabling them to decode their combinations. However, Vision and Language Models (VLMs) seem to lack such robustness, as they struggle with syntax (Nikolaus et al., 2022), store less information about grammar than language models (Milewski et al., 2022), and poorly identify the correct word order of sentences (Yuksekgonul et al., 2023). They also struggle to identify primitive concepts, such as color or shape (Yun et al., 2023), attributes of objects, and the spatial relationships between them (Clark and Jaini, 2023; Lewis et al., 2023). Such results generated doubts not only about their compositional abilities (Doveh et al., 2023a; Sinha et al., 2021; Ettinger, 2020), but the possibility of developing them (Trager et al., 2024). However, other studies suggest models encode separately minimally different captions (Diwan et al., 2022), and that entity or relationship recognition is done by special attention heads, hinting at the presence of compositional abilities (Li et al., 2020).

Amid such debates, several image-to-text benchmarks testing compositional abilities of VLMs have been made available (Parcalabescu et al., 2022; Thrush et al., 2022; Yuksekgonul et al., 2023; Ma et al., 2023; Hsieh et al., 2023; Burapacheep et al., 2024; Nikolaus et al., 2022). They evaluate models on whether they can differentiate a *foil* — a caption minimally modified to no longer match the image — from the correct original caption, and, sometimes, a *hard positive* — a minimally modified caption that still correctly describes the image (Kamath et al., 2024). This assumes that compositional abilities help models differentiate sentences only similar in form (i.e., using the same words) by their meanings. These benchmarks (i) have diverse stimuli; (ii) show models score around random chance, despite their good performance in more traditional downstream vision-and-language tasks (Parcalabescu et al., 2022; Thrush et al., 2022, among others); and (iii) investigate the interaction between model size, training data, and simplicity of the captions and models' performance.

Despite the strengths mentioned above, previous benchmarks (i) have not been reviewed on the alignment of their definitions of compositionality, making it unclear if they test basic linguistic phenomena, or combinations thereof; and (ii) have not tested if grammaticality of captions or the order of their thematic arguments interact with models' performance, despite prior evidence of models leveraging grammatical cues for improved performance (Hsieh et al., 2023), and the debate about their (in)sensitivity to word order (Sinha et al., 2021;

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 306–317, 2025

Pham et al., 2021; Ettinger, 2020; Yuksekgonul et al., 2023).

To bridge these gaps, we review previous compositional benchmarks and extend the VALSE benchmark (Parcalabescu et al., 2022) with the active/passive and the dative/ditransitive constructions. In contrast to the BLA benchmark (Chen et al., 2023), our active/passive dataset is larger, and we introduce a new ditransitive/dative dataset. Both datasets also contain simpler and complex versions of the same captions and foils. Besides these datasets, we analyzed how various factors, such as the linguistic structure, argument order, origin dataset, and grammaticality of sentences, interact with models' choices. Thus, the research question of this paper is the following:

Does linguistic structure, argument order, or origin dataset influence VLMs' choice of the correct caption?

Our main contributions and results are:

1. A categorization of VLMs benchmarks testing (pre)compositional abilities, their results, trends, and shortcomings.
2. Two datasets summing up to 9145 pairs of captions-foils.
3. Evidence showing that the linguistic structure and argument order do not influence models' results. Grammaticality, complexity, and noisiness of the origin dataset seem to do so, however.

Section 2 delves into previous literature on compositional abilities, by reviewing benchmarks and their limitations. Section 3 outlines the evaluation procedure, covering the selection of linguistic constructions and models, dataset construction, and statistical tests. Following this, Section 4 presents the results, which are then discussed in Section 5. The paper concludes with Section 6, with limitations addressed in Section 7.

2 Background

2.1 (Pre)Compositional Benchmarks

Current Benchmarks: Text-to-image benchmarks testing compositionality use compositional aspects (*compositional* benchmarks) or representations of linguistics phenomena required for them (*precompositional* benchmarks): VALSE (Parcalabescu et al., 2022), Winoground (Thrush et al., 2022), ARO (Yuksekgonul et al., 2023), CREPE (Ma et al., 2023), SugarCrepe (Hsieh et al., 2023), Colorswap (Burapacheep et al., 2024), Nikolaus et al. (2022), Kamath et al. (2024), VL-CheckList (Zhao et al., 2023), and BLA (Chen et al., 2023). We will now describe several aspects by which these benchmarks can be categorized.

Minimally Modfied Captions: Foils or *hard positives* can be created by (i) *swapping* words, e.g. going from 'a white chair and black blackboard' to 'a black chair and white blackboard' by changing attributes of objects in ColorSwap; (ii) *replacing* words with ones from other word classes; (iii) *adding* elements to the original captions, i.e. negation (see SugarCrepe); (iv) by *mixing* and combining multiple strategies systematically, e.g. using both *swap* and *replace*, such as VALSE. Their effect on whether they do or do not match the image depends on if the benchmark is intended to contain only foils, *hard positives*, or both. The column 'Minimal Differences' from Table 1 offers a complete overview of strategies for creating these captions.

Targeted phenomena: The benchmarks target visual or linguistic phenomena, where the latter are generally textual aspects reflected visually, while the former are more image-dependent, i.e., symbolic images of captions (Winoground), or relationships traced back to the image (CREPE). Concerning their coverage, linguistic phenomena can be a) specific, such as existence, plurality (VALSE), or negation (CREPE); or b) general, such as verbs, adverbs, or prepositions, named 'relations' in Winoground and 'attributes' in ColorSwap. See columns 'Type Properties' and 'Phenomena Covered' of Table 1 for an inventory of them.

Compositional property tested: The benchmarks can test (i) a *prerequisite* of compositional abilities, such as identifying the correct word order (Winoground); (ii) a compositional *ability*, such as systematicity; or (iii) an *effect* of compositionality, such as differentiating between two minimally different sentences. Note that some benchmarks might test several such aspects, see column 'Compositionality' for a full breakdown of the benchmarks' tests.

Fine-tuning: Models' performance can be negatively influenced by their training data, such as noisy, partial, or simplistic captions (Yuksekgonul et al., 2023; Doveh et al., 2023a); their training method, i.e. textual alignment, (Ossowski et al., 2024; Doveh et al., 2023a); or their textual encoder (Yuksekgonul et al., 2023; Li et al., 2023; Kamath et al., 2023; Clark and Jaini, 2023). Given compositional benchmarks have been systematically and carefully modified, they have been used for fine-tuning models, resulting in greatly improved performance, i.e. ARO, ColorSwap, or BLA (Doveh et al., 2023a). However, artifacts can influence such results, as models show smaller improvements — under 10%— on carefully controlled benchmarks such as SugarCrepe. See the column 'Tests' for an overview of benchmarks used previously for fine-tuning.

2.2 Previous results of the benchmarks

General results: Generally, results on VALSE, Winoground, SugarCrepe, ColorSwap, VL-CheckList, BLA, and ARO suggest poor compositional abilities. Specifically, models perform around or below the random chance threshold of 50%, with some exceptions, a good example being Winogrund, where model scores on text-to-image matching are as much as 70% lower than those of humans. Even though fine-tuning on foils can improve models' scores up until 70%, they still drop by 20% when confronted with another type of nominally different captions, the *hard positives* in Kamath et al. (2024).

Complexity of the caption, training datasets & model size: Models generally perform worse on longer, more complex or previously unseen captions (Winoground, CREPE). No interaction was found between the training dataset (Hsieh et al., 2023; Ma et al., 2023; Nikolaus et al., 2022) or the model size (Ma et al., 2023) and models' performance.

2.2.1 Shortcomings of evaluation benchmarks

Size: Yuksekgonul et al. (2023) argue that benchmarks that are composed of around 400 data points may not render significant statistical results. Many benchmarks, such as Winground, Colorswap, and each linguistic category from VALSE are not much bigger, with no more than 1000 data points.

Uncontrolled Artifacts: Grammaticality or plausibility artifacts (Hsieh et al., 2023) are not always controlled for, such as in ARO or CREPE, which models could exploit if fine-tuned on them, see Hsieh et al. (2023).

Inconsistent Definitions: Definitions of linguistic, visual properties, and minimal swapped units vary across studies. A phenomenon like 'pragmatics' is labeled as linguistic in Winoground and as visual in ARO. The minimal sentence elements swapped can be a singular visual concept (i.e. atom Ma et al., 2023), or a part of speech (Thrush et al., 2022). This can lead to a difficult comparison of results across studies.

Abilities tested: Current compositionality tests evaluate more effects than aspects of compositional abilities, such as systematicity (CREPE). This can lead to minimalization of the aspects tested and over/underestimation of models' abilities.

Targeted Phenomena: Visual and linguistic phenomena are arbitrarily chosen. For example, Parcalabescu et al. (2022) chose the linguistic phenomena of their benchmark based on their pervasiveness, but provided no empirical evidence to support this claim. This can lead to an over/underestimation of models' abilities, as the chosen phenomena can prove less prevalent than assumed.

3 Methodology

3.1 Datasets

VALSE: We extended VALSE due to its focus on linguistic constructions, control for unimodal collapse (Parcalabescu et al., 2022), grammaticality, and plausibility of actant swapping.

Actant Swap Category: We chose the actant swap category, which in Parcalabescu et al. (2022) was obtained by generating image captions from the SWiG dataset, and by swapping their thematic arguments to create foils. For illustration, considering the caption in (1a): its corresponding foil, shown in (2a), is obtained by swapping the subject and object, thus by *actant swapping*. Swapping might be a more difficult test than replacing/adding words (see Hsieh et al., 2023), as small syntactical changes might go unnoticed in models with less syntactic information (Milewski et al., 2022).

Added constructions: We extended the actant swap category by firstly creating passive and ditransitive versions of active captions, thereby introducing the active/passive construction, e.g. (1a) and (1b), and the dative/ditransitive construction, e.g. (1c) and (1d). Afterwards, we generate corresponding foils for them, illustrated with examples in (2), which were generated

Benchmark	Compositionality	Minimal Differences	Type Properties	Phenomena Covered	Size	Tests
VALSE (Parcalabescu et al., 2022)	Effect (Differentiation)	Mixed (*foils*)	Linguistic	Existence, Plurality, Counting, Spatial Relations, Actions, Coreference	8782	No
Winoground (Yuksekgonul et al., 2023)	Prerequisite (Word order); Effect (Differentiation)	Swap (*foils*)	Mixed	Object (l), Relation (l), Symbolic (v), Pragmatics (v), Series (v)	897	No
ColorSwap (Burapacheep et al., 2024)	Effect (Differentiation)	Swap (*foils*)	Linguistic	Color Attributes	1000	Yes
CREPE (Ma et al., 2023)	Ability (Systematicity, Productivity); Effect (Differentiation)	Mixed (*foils/hard negatives*)	Linguistic	Negatives, Atomic Swaps	100200	No
SugarCrepe (Hsieh et al., 2023)	Effect (Differentiation)	Mixed (*foils/hard negatives*)	Linguistic	Objects, Attributes, Relations	7512	Yes
ARO (Yuksekgonul et al., 2023)	Prerequisite (Word order); Effect (Differentiation)	Mixed (*foils*)	Linguistic	Relations, Attributes	52685	Yes
VL-CheckList (Zhao et al., 2023)	Effect (Differentiation)	Replace (*foils*)	Linguistic	Object, Attribute, Verb Replacements	410000	No
BLA (Chen et al., 2023)	Effect (Differentiation)	Swap (*foils*)	Linguistic	Actants, Predicates and Clauses Swaps	1939	Yes
Kamath et al. (2024)	Effect (Differentiation)	Mixed (*hard positives*)	Linguistic	Attributes, Relations and Word Order	55191	Yes

Table 1: **Overview of (Pre)Compositional Benchmarks.** *Compositionality*: (pre)compositional aspects tested, i.e. *prerequisite*, *ability* or *effect*, with phenomena specified in parenthesis; *Minimal differences*: foils/*hard positives* formed by mixed methods (‘add’, ‘swap’, ‘replace’), or singular methods (‘add’/‘swap’/‘replace’); *Type properties*: the type of properties measured, i.e. linguistic or mixed (linguistic and visual); *Phenomena Covered*: the specific phenomena covered, with types specified in parenthesis for mixed properties (*visual*, v; *linguistic*, l); *Tests*: fine-tuned models on the datasets presented in the original papers.

by swapping nouns in (1). We chose these two constructions because their variants allow isolating the effect of the construction itself from the order of thematic arguments. For example, when compared, (2a) and (1b) still have thematic arguments in similar positions in the sentence, having only the construction as the only difference between the foil and the caption. Depending if a sentence is a foil or caption it will be referred to as active caption (1a) or foil (2a), dative caption (1c) or foil (2c), and so on. See Appendix A for example images for the captions.

1. Captions
   a) A player hits a ball.
   b) A ball is hit by a player.
   c) A woman gives a book to the girl.
   d) A woman gives the girl a book.
2. Foils
   a) A ball hits a player.
   b) A player is hit by a ball.
   c) A woman gives the girl to a book.
   d) A woman gives a book the girl.

Dataset construction: We obtained each type of construction and its corresponding foils, by adapting the code of Raam (2022) to use spaCy `en_core_web_trf` parsers, chosen for its higher accuracy in part-of-speech (POS) tagging. Passive captions were obtained by changing the active verb inflections of sentences that had a subject and an object, while dative sentences were created by alternating sentences that had a subject, object, and beneficiary. Foils were obtained by swapping the object and subject in active/passive sentences, or the beneficiary and the object in ditransitive/dative ones. For each construction, only sentences that could have an alternation and an existing image were selected. Thus, each image has 2 captions and 2 foils, regardless of the linguistic construction.

We parsed (i) 1 million sentences from Conceptual Captions (CC, Sharma et al., 2018) and the actant swap dataset from VALSE for the active/passive construction; (ii) 2 million CC sentences for the ditransitive construction, double the amount of active sentences parsed given the rarity of the construction. All the code and datasets obtained will be made available upon request, similarly to Koplenig et al. (2017).

Simplified Datasets and Grammaticality: Each obtained dataset was further simplified to test potential artifacts of sentence length, by re-parsing all sentences to keep only the head of noun phrases for the thematic arguments involved in the swap. For example, we omitted extra descriptions of objects or datives containing embedded sentences (i.e. from ‘the girl *whom I met yesterday*’ to ‘the girl’). To test the efficiency of our method of simplification, we randomly sampled 100 simplified active/passive sentences — 400 captions and foils — and observed that 12% of them were not completely simplified. See Table 2 for a complete breakdown of the numbers of each dataset.

To also statistically test if models exploit the grammaticality of sentences for their answers, we used GRUEN (Zhu and Bhat, 2020) to obtain a grammaticality score for all our sentences. Unlike Parcalabescu et al. (2022), we do not select sentences with a certain grammaticality threshold, to observe overall if more ungrammaticality has a bigger effect on models’ choices.

Dataset	SWiG	CC	Total
Active/Passive	683	5450	6133
Ditransitive	0	3012	3012
Simplified Datasets			
Active/Passive	683	4737	5420
Ditransitive	0	2998	2998

Table 2: **Datasets**. **First table**: captions after the first round of parsing for both types of constructions; **Second table**: remaining simplified data points. Note the SWiG dataset only contained 7 ditransitive sentences, which were excluded, given they were not sufficient for the statistical analysis.

3.2 Models

For evaluation, we chose three contrastive-learning models, namely CLIP-ViT 32, Flava, and LiT-ViT B16B (Zhai et al., 2022), which have a stronger alignment between their textual and image embeddings, given their training to align the two modalities.

CLIP-ViT 32: CLIP is a dual-stream model with a vision (Dosovitskiy et al., 2021) and text (Vaswani et al., 2023) transformer as backbones, trained on WIT, a 400 million image-text pairs dataset scraped from multiple online sources. We chose CLIP given it has the best results on the actant swap category from VALSE and we wanted to investigate if this generalizes on our dataset, where actant swapping is one of the alternations made.

Flava: Flava is a dual-stream model with the same ViT image encoder and a ViT-b/16 textual encoder. It was trained on PMD, a public dataset of 70 million pairs containing, among others, CC sentences. Flava is also trained to perform well in unimodal language tasks, scoring 20% more than CLIP (Singh et al., 2022). Thus, Flava was selected for its unimodal performance, and for improved performance in multimodal (Singh et al., 2022), and even compositional tasks, such as ColorSwap, when compared to CLIP.

LiT-ViT B16B: LiT uses a ViT image encoder and a BERT textual encoder, using contrast-tuning during training. Previous studies argued for the importance of the textual modality in VLMs (Wu et al., 2023). Instead of fine-tuning both image and text encoders, LiT only fine-tunes the textual encoder, obtaining better zero-shot image classification performance than CLIP (Zhai et al., 2022). LiT is trained on a dataset containing 4 billion image-text pairs, some originating from CC.

3.3 Statistical tests

Two Linear Mixed Models (lmer, Bates et al., 2015) were conducted in RStudio (RStudio Team, 2020)[1].

The First Lmer We tested if the following factors had any significant effects on models’ choices: (i) the higher grammaticality of either the caption or foil; (ii) the simplification of the caption or foil; (iii) the linguistic construction, i.e. passive or ditransitive, as opposed to the active; (iv) the origin dataset; (v) the model; (vi) and model size. Note that we ran two tests with respect to the linguistic constructions (active vs. passive or active vs. ditransitive), for all models directly, and for each model separately.

The second lmer We tested if captions or foils are chosen more by models if the order of thematic arguments in the compared sentences is different (active caption vs. active foil) or similar (passive caption vs. active foil). Grammaticality, simplification, the origin dataset, model, and model sizes were also investigated alongside word order. **Contrasts and random effects.** All tests had random effects for images, and binary or ternary orthogonal contrasts for all factors, corresponding to their levels. Note that models were introduced as fixed factors, due to the small sample of VLMs tested.

[1] For a review of linear mixed models see Gałecki and Burzykowski (2012).

4 Results

Accuracy ratios For overall visualization, Table 3 presents the accuracy ratios of models for SWiG, and Table 4 presents the accuracy ratio for models, tested on both constructions from CC.

First lmer Our results did not show any significant effect of the linguistic construction. Regardless of the model, the chances of choosing captions go down by almost half if foils are equally grammatical to captions, and slightly down if foils are more grammatical than captions. The same grammaticality value for captions and foils, and bigger foil grammaticality have a lower effect on ditransitive sentences, refer to Figure 1 for an illustration. The chance of choosing a caption is slightly increased by simplified sentences, and by their origin from SWiG, a cleaner dataset. For FLAVA or LiT, foils become more preferred, as also seen in the ratios. **Models Subsamples.** For all models captions were chosen more when comparing simplified sentences.

Second lmer No significant effect of word order was found. The trends of the first lmer test in grammaticality, models, and simplification are similar. Equally grammatical foils and captions have a smaller effect on sentences with different argument structures, while higher foil grammaticality has a bigger effect on them. For an illustration, see Figure 2. See Appendix C for detailed significance values.

Normal Captions			
Comparison	CLIP	Flava	LiT
1	69.38	46.25	64.01
2	45.17	40.11	43.12
3	64.32	56.51	62.42
4	45.96	57.64	49.08
5	52.48	43.77	45.17
Simplified Captions			
Comparison	CLIP	Flava	LiT
1	69.38	46.25	64.01
2	45.09	40.17	43.04
3	64.27	56.74	62.37
4	45.96	57.80	49.08
5	52.41	43.84	45.09

Table 3: **Ratios on SWiG**. *Comparison 1*: Active caption vs. foil; *Comparison 2*: Passive caption vs. foil; *Comparison 3*: Active vs. Passive captions; *Comparison 4*: Active vs. Passive foils; *Comparison 5*: Passive caption vs. Active foil. For each comparison, the value represents the preference of the the first term over the second.

Normal Captions			
Comparison	CLIP	Flava	LiT
Act/pass 1	64.20	24.93	52.33
Act/pass 2	54.46	42.49	44.50
Act/pass 3	64.81	29.00	59.59
Act/pass 4	44.01	56.54	46.20
Act/pass 5	48.14	53.26	35.85
Dit/Dat 1	63.44	27.22	49.71
Dit/Dat 2	55.84	35.50	41.92
Dit/Dat 3	64.87	36.35	47.12
Dit/Dat 4	45.24	52.91	46.37
Dit/Dat 5	50.31	37.04	42.12
Simplified Captions			
Comparison	CLIP	Flava	LiT
Act/pass 1	65.78	36.96	62.17
Act/pass 2	53.32	45.55	45.92
Act/pass 3	52.56	41.05	55.75
Act/pass 4	42.28	53.52	41.15
Act/pass 5	64.41	48.00	53.75
Dit/Dat 1	61.44	36.72	53.65
Dit/Dat 2	62.05	37.76	51.60
Dit/Dat 3	60.20	46.19	46.47
Dit/Dat 4	51.38	50.25	48.88
Dit/Dat 5	54.60	40.12	49.63

Table 4: **Ratios for CC**. Each 'Comparison' number indicates a different pairing of captions or foils between the active/passive construction (Act/pass) and the ditransitive one (Dit/Dat). Note that comparisons are only performed within the same construction. *Comparison 1*: Active (or Ditransitive) caption vs. foil; *Comparison 2*: Passive (or Dative) caption vs. foil; *Comparison 3*: Active (or Ditransitive) vs. Passive (or Dative) captions; *Comparison 4*: Active (or Ditransitive) vs. Passive (or Dative) foils; *Comparison 5*: Passive (or Dative) caption vs. Active (or Dative) foil. For each comparison, the value represents how much the first term of comparison is preferred over the second.

5 Discussion

General results Our overall random-chance performance indicates poor compositional abilities, in line with results obtained on VALSE, SugarCrepe, or Winoground. Our general results also provide evidence that models might be influenced by many factors, such as the simplicity, noisiness, and grammaticality of the compared sentences, where by noisiness we mean how clean the original dataset of sentences is. All aspects point to a lack of compositional ability, as such factors should not influence models' choices.

Linguistic Constructions and Argument Structure Our results do not offer evidence that specific linguistic constructions or word order of semantic arguments influence models' choices.

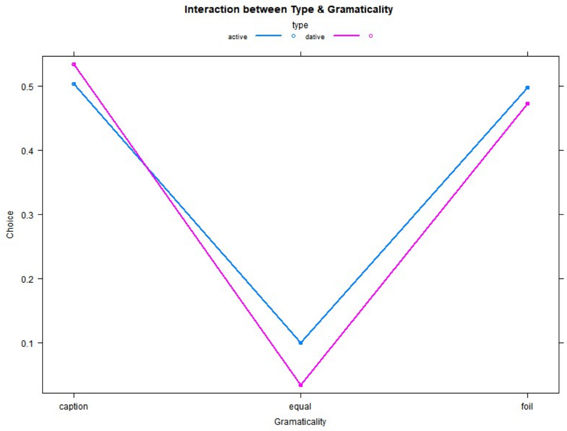

Figure 1: Plotted interaction between *Type*, with levels ‘active’ and ‘ditransitive’, and *Grammaticality*, with levels ‘caption’, ‘foil’, and ‘equal’. Note that a bigger grammatical value between a *caption* or *foil* is assigned to the one with a bigger GRUEN grammaticality score. The pink line is the ‘active’, while the blue one is the ‘ditransitive’ construction.

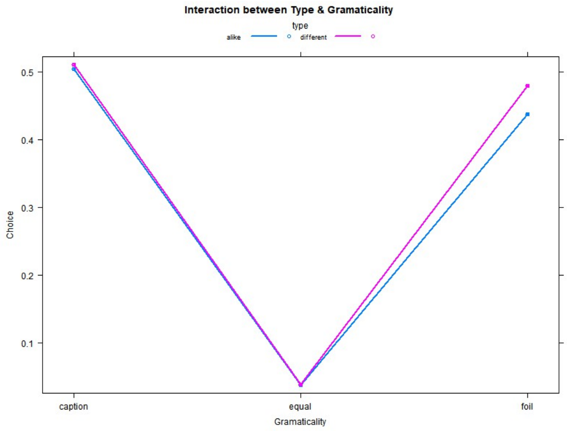

Figure 2: Plotted interaction between *Type*, with levels ‘alike’ and ‘different’, and *Grammaticality*, with levels ‘caption’, ‘foil’, and ‘equal’. The pink line is ‘alike’, the blue one is ‘different’

Grammaticality Our results suggest captions are chosen slightly less when a foil is grammatical but significantly less when captions and foils are equally grammatical. Thus, our results indicate that models favor foils, shown by the intercept of our first test (i.e. $\beta = 0.37$), but also that they might leverage grammatical cues in their decisions, in line with Hsieh et al. (2023). Tests done for each model show grammaticality affects types of constructions unevenly, i.e. passive or alternation more. This might be because our methods might introduce more ungrammaticality, or complexity to one of the structures, aspects previously shown to decrease performance (Thrush et al., 2022; Ma et al., 2023).

Simplification and Noisiness Noisier datasets lower the performance of VLMs (Thrush et al., 2022; Ma et al., 2023), which was also shown by our improved results on SWiG. Our results also show that simplifying the compared caption and foil makes models perform better (Ma et al., 2023). Simplicity and low complexity of the sentences compared might improve performance because the more complex the text is, the more actors might be in it. In such cases, models insensitive to word order will regard many more sentences as likely to match the image, thereby lowering the chance of choosing the actual caption.

Models As shown by our ratios, Flava performs the worst, followed by LiT and CLIP. All models obtain mostly chance-like performance, in line with results on ARO, Winoground, or VALSE. Note that both Flava and LiT were found previously to perform better than CLIP on other tasks. This difference in performance further reinforces the fact that traditional benchmarks do not focus on testing compositional aspects. Our results seem to suggest all models perform better on simpler sentences, unlike what was shown in Wingorund or CREPE where CLIP did better on lengthier captions. Our results could differ from those of previous studies, given that we specifically controlled for certain artifacts, for example, by comparing the same sentences in their original or simplified form, which was not previously done. All models also performed better on SWiG, a less noisy dataset, suggesting that cleaner foils and captions improve performance (Doveh et al., 2023b). This is interesting considering Flava and LiT were trained on datasets containing CC. **Flava.** Flava performs significantly worse than CLIP, unlike in benchmarks such as Winoground, and ColorSwap. This might be because Flava is worse than CLIP at recognizing good word order, as shown by ARO. **LiT.** Compared to Flava, LiT might obtain an improved performance due to its increased training dataset size, which could result in the model seeing more minimal differences in the sentences. LiT might also perform better due to its textual encoder, as Zhai et al. (2022) showed that the BERT textual encoder improves performance. **CLIP.** The performance of CLIP and its relationship to its training dataset is hard to indicate.

6 Conclusions

We reviewed previous studies about the compositional abilities of VLMs, classifying them and their downsides. The current study introduced two new VALSE datasets of 9145 pairs of captions and foils, almost doubling the VALSE benchmark.

Our tests also indicate that many aspects could influence models' performance, such as the grammaticality of the compared sentences, and their simplicity or noisiness, further reinforcing the results of previous studies. We offered evidence that these aspects affect models uniformly, unlike previous studies (Thrush et al., 2022; Ma et al., 2023). We also offered evidence that in terms of performance, CLIP is followed by LiT and then Flava. Our results indicate that Flava performs much worse than the other models.

7 Future research and limitations

One-stream models could be considered for testing in the future, given that we only evaluated dual-stream ones. One of the other shortcomings of the current paper is that it tests only three models. Thus, the methodology for the current paper could be used and easily adapted to test new models fine-tuned for better compositional performance, such as X-VLM (shown to reach the highest accuracy on relations in ARO), SigCLIP, or MosaiCLIP (Singh et al., 2023).

The current datasets could be made more balanced. For example, the SWiG dataset could be better represented, since most parsed sentences were taken from CC. A possible future research direction is to generate active constructions from captions that were originally passive. We only created passive constructions by modifying the active ones, and not the reverse. This created an imbalance, as for the ditransitive/dative construction, we created examples from both types of original constructions, i.e. ditransitive or dative. Along similar lines, one of the limitations of the current study is that we did not investigate if any of the generated foils were incorrectly generated, and whether the sentences that remained non-simplified, after simplification, influenced results. Follow-up studies could test these aspects. Generating images from the foils to see if preferences change, as in ColorSwap, could also make an interesting follow-up to the current study. We also built a parser for the 'with/against' construction that could be used for new datasets, see Appendix B, which will also be available upon request. Future studies might also consider trying to account for inconsistencies in definitions of composition ality, to enlarge the current compositional aspects tested, or to apply our methodology for languages other than English.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models usinglme4. *J. Stat. Softw.*, 67(1).

Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. 2024. Colorswap: A color and word order dataset for multimodal evaluation. *Preprint*, arXiv:2402.04492.

Xinyi Chen, Raquel Fernández, and Sandro Pezzelle. 2023. The BLA benchmark: Investigating basic language abilities of pre-trained multimodal models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5817–5830, Singapore. Association for Computational Linguistics.

Kevin Clark and Priyank Jaini. 2023. Text-to-image diffusion models are zero-shot classifiers. *Preprint*, arXiv:2303.15233.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. *Preprint*, arXiv:2211.00768.

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. In *International Conference on Learning Representations*.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. 2023a. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Preprint*, arXiv:2305.19595.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. 2023b. Teaching structured vision & language concepts to vision & language models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2657–2668.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Preprint*, arXiv:1907.13528.

Andrzej Gałecki and Tomasz Burzykowski. 2012. Linear mixed-effects model. In *Linear mixed-effects models using R: a step-by-step approach*, pages 245–273. Springer.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36, pages 31096–31116. Curran Associates, Inc.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. Text encoders bottleneck compositionality in contrastive vision-language models. *Preprint*, arXiv:2305.14897.

Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. 2024. The hard positive truth about vision-language compositionality. *Preprint*, arXiv:2409.17958.

Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Müller-Spitzer. 2017. The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. *PLOS ONE*, 12(3):1–25. Publisher: Public Library of Science.

Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H. Bach, and Ellie Pavlick. 2023. Does clip bind concepts? probing compositionality in large image models. *Preprint*, arXiv:2212.10537.

Jie S. Li, Yow-Ting Shiue, Yong-Siang Shih, and Jonas Geiping. 2023. Augmenters at semeval-2023 task 1: Enhancing clip in handling compositionality and ambiguity for zero-shot visual wsd through prompt augmentation and text-to-image diffusion. *Preprint*, arXiv:2307.05564.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? *Preprint*, arXiv:2212.07796.

Victor Milewski, Miryam de Lhoneux, and Marie-Francine Moens. 2022. Finding structural knowledge in multimodal-bert. *Preprint*, arXiv:2203.09306.

Mitja Nikolaus, Emmanuelle Salin, Stephane Ayache, Abdellah Fourtassi, and Benoit Favre. 2022. Do vision-and-language transformers learn grounded predicate-noun dependencies? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1538–1555, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Timothy Ossowski, Ming Jiang, and Junjie Hu. 2024. Prompting large vision-language models for compositional reasoning. *Preprint*, arXiv:2401.11337.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *Preprint*, arXiv:2112.07566.

Barbara H. Partee. 2004. *Compositionality in formal semantics : selected papers*. Blackwell.

Thang M. Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *Preprint*, arXiv:2012.15180.

Yannick Raam. 2022. Probing models on the active and passive voice: How far do vision and language models look beyond sentence surfaces? Unpublished BA thesis.

RStudio Team. 2020. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629.

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. 2023. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. *Preprint*, arXiv:2305.13812.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. Unnatural language inference. *Preprint*, arXiv:2101.00010.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. *Preprint*, arXiv:2204.03162.

Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. 2024. Linear spaces of meanings: Compositional structures in vision-language models. *Preprint*, arXiv:2302.14383.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Chenwei Wu, Li Erran Li, Stefano Ermon, Patrick Haffner, Rong Ge, and Zaiwei Zhang. 2023. The role of linguistic priors in measuring compositional generalization of vision-language models. *Preprint*, arXiv:2310.02777.

Zhenlin Xu, Marc Niethammer, and Colin Raffel. 2022. Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. *Preprint*, arXiv:2210.00482.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? *Preprint*, arXiv:2210.01936.

Tian Yun, Usha Bhalla, Ellie Pavlick, and Chen Sun. 2023. Do vision-language pretrained models learn composable primitive concepts? *Preprint*, arXiv:2203.17271.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18102–18112.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2023. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *Preprint*, arXiv:2207.00221.

Wanzheng Zhu and Suma Bhat. 2020. Gruen for evaluating linguistic quality of generated text. *Preprint*, arXiv:2010.02498.

A Examples of captions and foils

1. Captions
   - a) A player hits a ball.
   - b) A ball is hit by a player.
   - c) A woman gives a book to the girl.
   - d) A woman gives the girl a book.
2. Foils
   - a) A ball hits a player.
   - b) A player is hit by a ball.
   - c) A woman gives the girl to a book.
   - d) A woman gives a book the girl.

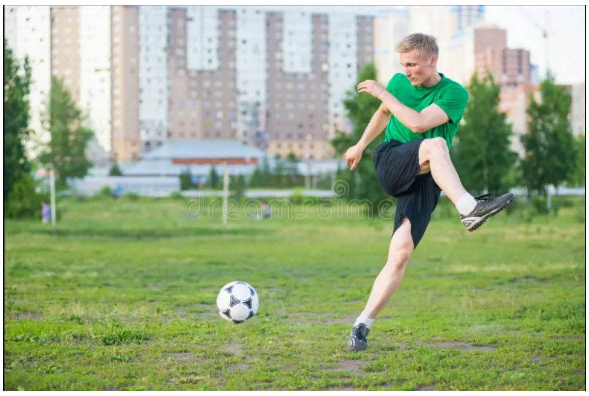

Figure 3: Image for active/passive construction, obtained by Google-searching the caption. **Corresponding captions**: (1a) and (1b). **Corresponding foils**: (2a), (2b)

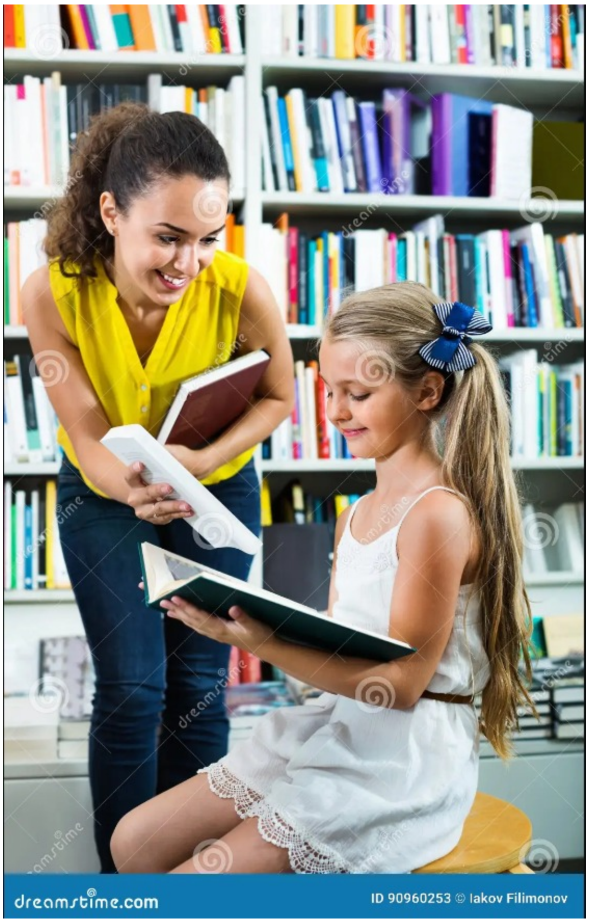

Figure 4: Image for the ditransitive/dative construction, obtained by Google-searching the caption. **Corresponding captions**: (1c) and (1d). **Corresponding foils**: (2c), (2d)

B With/against foil parser examples

1. Captions
   - a) The boy hits the stick against the fence.
   - b) The boy hits the fence with the stick.
2. Foils
   - a) The boy hits the fence against the stick.
   - b) The boy hits the stick with the fence.

C Statistical results

First Lmer Test Captions were coded as 1, and foils as 0, thus any positive effects of a factor on the intercept means improved performance (i.e. captions are chosen over foils). For the first lmer test, the significant values for the fixed effects are *grammaticality equal* (β = -0.43, t-score = -27.63, p-value < 0.05), *grammaticality foil* (β = -0.03, t-score = -10.08, p-value < 0.05), *complexity simplified* (β = 0.05, t-score = 18.14, p-value < 0.05), *dataset SWiG* (β = 0.04, t-score = 7.38, p-value < 0.05), *model FLAVA* (β = -0.18, t-score = -59.3, p-value < 0.05), *model LIT* (β = -0.09, t-score = -25.26, p-value < 0.05). Note that ternary factors (i.e. grammaticality, model) calculate coefficients of one level in comparison to the mean of the other two. Significant interaction values were found for *type ditransitive and grammaticality equal* (β = -0.06, t-score = -2.22, p-value < 0.05), *type ditransitive and grammaticality foil* (β = -0.05, t-score = -8.39, p-value < 0.05).

Second Lmer Test *Grammaticality equal* (β = -0.44, t-score = -19.92, p-value < 0.05), *grammaticality foil* (β = -0.03, t-score = -16.42, p-value < 0.05), *complexity simplified* (β = 0.06, t-score = 25.14, p-value < 0.05), *dataset SWiG* (β = -0.03, t-score = -3.74, p-value < 0.05), *model FLAVA* (β = -0.18, t-score = -60.17, p-value < 0.05), *model LIT* (β = -0.09, t-score = -31.54, p-value < 0.05) had significant values. The significant interaction effects were *type ditransitive and grammaticality equal* (β = -0.02, t-score = -0.49, p-value < 0.05), *type ditransitive and grammaticality foil* (β = 0.03, t-score = 6.45, p-value < 0.05).

Models subsamples The subsamples on models generally follow the trends of the first lmer model, with some exceptions, which are explained for each model. Note that simplifying captions improved performance for all models, i.e. *Flava* (β = 0.07, t-score = 15.16, p-value < 0.05), *CLIP* (β = 0.018, t-score = 3.70, p-value < 0.05), *LiT* (β = 0.06, t-score = 3.44, p-value < 0.05).

Flava Grammaticality of the foils has a significant effect only for the ditransitive construction subsample, slightly making captions more likely (β = 0.03, t-score = 3.44, p-value < 0.05). On non-subsampled data, the effect of a grammatical foil on ditransitive sentences is slightly bigger (β = 0.05, t-score = 5.11, p-value < 0.05) than on active ones.

CLIP On subsamples, only the grammaticality of the foil has a slightly lower effect on passive constructions than active ones in the active/passive subsample (β =-0.05, t-score =-3.89, p-value < 0.05).

LiT On subsampled data, a bigger grammaticality score of the foil affects less passive structures (β=0.03, t-score =-2.80, p-value < 0.05), and more ditransitive captions (β =0.06, t-score =3,36, p-value < 0.05). On both subsamples, no interactions between grammaticality and type are significant.