An Examination of the Compositionality of Large Generative Vision-Language Models

Anonymous ACL submission

Abstract

With the success of Large Language Models (LLMs), many Generative Vision-Language Models (GVLMs) have been constructed via multimodal instruction tuning. However, the performance of GVLMs in multimodal compositional reasoning remains largely unexplored, as existing evaluation metrics and benchmarks focus predominantly on assessing contrastive models like CLIP. In this paper, we examine both the evaluation metric (VisualGPTScore) and current benchmarks for evaluating the compositionality of GVLMs. We find that the VisualGPTScore is sensitive to sentence syntax rather than visual contents, so the curation methods of current benchmarks lead to severe morphological bias when evaluating with VisualGPTScore. To combat this, we define a MorphoBias Score to quantify the morphological bias and propose a novel LLM-based strategy to calibrate the benchmarks. Moreover, a novel and challenging task is added to evaluate the robustness of GVLMs against inherent inclination toward syntactic correctness. We include the calibrated dataset and the task into a new benchmark, namely MOrphologicall De-biased Benchmark (MODE). Our study provides the first unbiased benchmark for the compositionality of GVLMs, facilitating future research in this direction. We will release our code and datasets.

1 Introduction

A surge of research on vision-language models (VLMs) has demonstrated success in a wide range of tasks, including zero-shot visual recognition (Radford et al., 2021; Gao et al., 2021; Zhou et al., 2022), visual question answering (Alayrac et al., 2022; Chen et al., 2022), and image-to-text retrieval (Alayrac et al., 2022; Gong et al., 2023a). Previous VLMs are mainly constructed based on the contrastive learning framework (CVLMs), where visual and linguistic embeddings are projected into a joint space to match (Radford et al., 2021; Jia et al., 2021; Li et al., 2022). However, with the advent of large language models (LLMs) like ChatGPT, GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023a), recent studies have extended the LLM architecture to multimodal settings, which is named as Generative VLMs (GVLMs) (Liu et al., 2023a; Zhu et al., 2023; Li et al., 2023a; Gao et al., 2023). The GVLMs deviate from the CVLMs in projecting visual features into the latent lexical space of LLMs, and leveraging the generation capacity of LLMs to solve vision-language tasks. In the training process, most work follows the recipe of freezing the main body of visual encoders and LLMs, only updating the negligible parameters of projecting layers, which is also called “bridge architecture" (Rajesh et al., 2023).

Despite the emergence of research on GVLMs, the understanding of compositionality in GVLMs has remained an enigmatic black box, with no thorough investigations conducted thus far. Previous research studies (Thrush et al., 2022; Zhao et al., 2022; Yuksekgonul et al., 2022a; Ma et al., 2023; Ray et al., 2023a) in multimodal compositionality focus on establishing retrieval-based benchmarks for evaluating CVLMs on object relations and attribute understanding, order sensitiveness of sentence elements, and atom-level understanding. The CVLMs have demonstrated abilities to discriminate position captions from negative ones based on the image-text similarity, where the disparities between the positive and negative captions are relatively subtle, such as “an old person kisses a young person" and “a young person kisses an old person" (Thrush et al., 2022).

Recently, log-likelihood based scores are leveraged to evaluate the generative models (Fu et al., 2023; Liu et al., 2023b; Lin et al., 2023; Li et al., 2023b). Given an image-text pair consisting of an image with positive and negative references, the GVLMs estimate the log-likelihood probabili-

ties of generating positive and negative references conditioned on the image, respectively. It quantitatively reflects the GVLMs' understanding of alignment between image and positive/negative references. Following Lin et al. (2023), we alias the log-likelihood score as VisualGPTScore.

In this paper, we examine both the VisualGPTScore and current benchmarks of vision-language compositionality. We found that:

- Using VisualGPTScore to evaluate GVLMs is not sensitive to bags-of-words problems that broadly exist in the evaluation of CLIP with similarity scores. The bags-of-words phenomenon during evaluation is related to metrics, and metrics based on similarity suffer from it.

- VisualGPTScore sometimes prefers syntactical correctness rather than content-related correctness. It scores negative references with reasonable syntax, but unrelated contents higher than positive references. In contrast, CLIP pays more attention to the correlation of visual contents but is not sensitive to the order of atoms in references.

- There exists a morphological bias in current multimodal compositional reasoning benchmarks. Current compositional reasoning benchmarks are curated for evaluating contrastive VLMs, and the method of constructing negative references may not be suitable for generative VLM evaluation.

Based on these observations, our tactics comprise:

- We quantitatively assess the morphological bias by leveraging large language models, and name it as the MorphoBias Score for statistical analysis.

- Based on the MorphoBias Score, we propose a MOrphologicall De-biased Benchmark (MODE) based on current benchmarks for more robust multimodal compositionality evaluation. We adopt different strategies to calibrate existing benchmarks. We add a new challenging assessment in MODE to evaluate the content understanding across visual and language modalities. The performance of several GVLMs is reported on MODE, and we show the robustness and faithfulness to human judgments of our new benchmark.

2 Background

2.1 Generative vision-language models

In this paper, we define GVLMs as models that combine visual encoders with large language models (LLMs) trained on large text corpora. The prevailing approach in recent research connects a frozen visual encoder with an LLM by training mapping layers on images-text pairs, followed by fine-tuning using multi-modal instructional data to facilitate multi-turn conversations (Liu et al., 2023a; Gao et al., 2023; Zhu et al., 2023; Dai et al., 2023; Su et al., 2023; Gong et al., 2023b; Sun et al., 2023). This approach is anchored in the idea of treating visual tokens the same as linguistic ones. The visual tokens are mapped into a lexical embedding space and harnessed to generate textual content in an autoregressive manner. Formally, given an image I and the visual encoding $g(I)$ from encoders like Vision Transformer (Dosovitskiy et al., 2020), the mapping process can be formulated as:

$$\boldsymbol{z} = \mathbf{M}(g(I)), \boldsymbol{z} = \{z_1, z_2, ..., z_N\}, \quad (1)$$

where N is the number of visual tokens and $\mathbf{M}$ is the mapping layers. Different from CVLMs that utilize contrastive loss, the training objective of multimodal autoregressive training is to maximize the log-likelihood of the next true token. Denote the tokenized instructions as $\boldsymbol{p}$ and the output words as $t_i, (1 \leq i \leq K)$, the GVLM training objective is defined as:

$$\max_{\theta_M, \theta_\sigma} \sum_{i=1}^{K} \log P(t_i | \boldsymbol{p}, \boldsymbol{z}, t_1, t_2, ..., t_{i-1}; \theta_M, \theta_\sigma) \quad (2)$$

where θ_M refers to the learnable parameters of mapping layers $\mathbf{M}$ and θ_σ refers to other tunable parameters like adapter layers in LLaMA-Adapter V2 (Gao et al., 2023), or visual abstractor and LoRA in mPLUG-Owl (Ye et al., 2023).

In comparison, the training objective of CVLMs (CLIP-based (Radford et al., 2021)) is based on the contrastive loss between vision and language. Please refer to Appendix A.1 for CVLMs formulation.

2.2 Vision-language compositionality

Recent works on vision-language compositionality focus on introducing benchmarks to evaluate the CVLMs, mainly on CLIP (Radford et al., 2021). Winoground (Thrush et al., 2022) is one of the pioneers in building benchmarks for multimodal

compositionality, curating 400 test items to evaluate the pragmatics, symbolic and series factors of VLMs. Afterwards, several benchmarks have been proposed to challenge the objects, relations and attributes understanding of VLMs, including VL-CheckList (Zhao et al., 2022), ARO (Yuksekgonul et al., 2022a), CREPE (Ma et al., 2023) and Cola (Ray et al., 2023b) *etc.* These benchmarks are in the form of image-text retrieval, requiring the model to differentiate positive references from negative references based on the visual contents of the images. See Fig 6 in the Appendix for the details of the image-text retrieval format. These benchmarks are curated for evaluating contrastive VL models like CLIP, where similarity scores between images and references serve as the criteria for selecting references. Then, the accuracy of selecting positive samples across all data samples will be reported to assess the model's compositional understanding capability.

2.3 Evaluation metrics for multimodal retrieval

Since previous benchmarks have been carefully curated for evaluating CLIP, contrastive scores naturally emerge as the metric for assessing the compositional similarity between images and references. For generative models, an intuitive way is reference-based, measuring the quality of generated captions with metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015). Among the reference-based metrics, BERTScore (Zhang et al., 2019) tackles superficial matching between captions and references in lexical expression, delving deeper into the semantic similarity matching. GPTScore (Fu et al., 2023) proposes to leverage emergent abilities of generative models to score generated texts. Inspired by GPTScore, recent works (Lin et al., 2023; Li et al., 2023b; Liu et al., 2023b) measure the GVLMs using the log-likelihood of directly generating reference sentences conditioned on the image. We follow the Lin et al. (2023) to abbreviate the kind of method as VisualGPTScore, which can be formulated as:

$$\text{VisualGPTScore}(\boldsymbol{r}|\mathcal{I}) = \sum_{t=1}^{m} w_t \log P(r_t|\boldsymbol{r}_{<t}, \boldsymbol{p}, \mathcal{I}; \theta_{GVLM}) \quad (3)$$

where $\mathcal{I}$, $\boldsymbol{r}$, $\boldsymbol{p}$ represents the image, reference sentence and instructions. θ_{GVLM} refers to parameters of GVLMs and $w_t = \frac{1}{m}$. The VisualGPTScore is directly estimated conditioned on images and thus reference-free. In this work, we examine the VisualGPTScore and discuss the potential influence of using it in current benchmarks for vision-language compositionality.

3 Evaluation Metric Examination

VisualGPTScore measures the probability of generating specific references conditioned on the given images, as defined in Eqn. 3. The generative evaluation method is based on the inherent attribute of GVLMs and used in image-text retrieval (Lin et al., 2023; Li et al., 2023b; Liu et al., 2023b). Since current benchmarks on VL compositions consists of image-text pairs, we follow Lin et al. (2023) to utilize VisualGPTScore for evaluating VL compositionality of GVLMs. However, our primary focus in this section is to examine the robustness of VisualGPTScore.

3.1 Experimental setup

Model choices: In this section we leverage two state-of-the-art GVLMs, namely LLaVA (Liu et al., 2023a) and MiniGPT-4 (Zhu et al., 2023), to conduct experiments. LLaVA is the first method to project visual features into LLaMA (Touvron et al., 2023a) latent space via multimodal instruction tuning. A linear projection layer and the parameters of the LLM are tuned on conversations, detailed descriptions, and complex reasoning datasets. MiniGPT-4 (Zhu et al., 2023) maps visual embeddings obtained from ViT and Q-Former (Li et al., 2022) into Vicuna (Chiang et al., 2023) via a linear projection layer. We adopt the model version of "LLaVA-7B-v0" and "Minigpt4-aligned-with-Vicuna7B" to evaluate. However, we found that when using VisualGPTScore to evaluate compositionality, both models exhibited similar patterns. Therefore, for the sake of brevity, we only present the results for LLaVA.

Datasets: We use Winoground (Thrush et al., 2022), VL-Checklist (Zhao et al., 2022), ARO (Yuksekgonul et al., 2022a) and CREPE (Ma et al., 2023) in the evaluation analysis, totaling 52,189 images and 129,558 reference sentences. All benchmarks necessitate the model's selection of positive reference sentences from negative ones. For Winoground, we report text score, image score

and group score as the paper (Thrush et al., 2022). For other datasets, Recall@1 accuracy is reported.

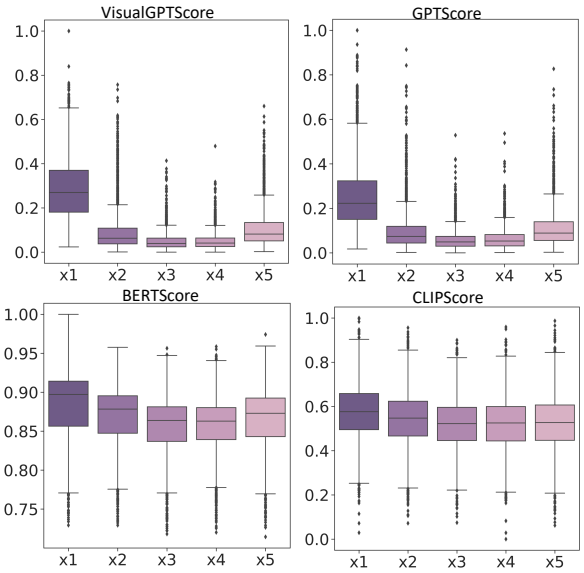


Figure 1: Box plots of scaled score distributions for original (`x1`) and perturbed captions (`x2-x5`, `x2: shuffle nouns & adj`, `x3: shuffle all but nouns & adj`, `x4: shuffle within trigrams`, `x5: shuffle trigrams`). The distribution gap between the original captions and the shuffled captions is evident for the generative scores, while the contrastive score (BERTScore) is significantly less affected by the order of words. The CLIPScore sub-figure illustrates the distribution of similarity scores generated by the CLIP model, which is compared with the first three sub-figures of LLaVA-7B.

3.2 Sensitivity to bags-of-words

Previous research works have pointed out that CVLMs suffer from the bags-of-words phenomenon when doing compositional reasoning due to the pre-training recipe using visual and text data in instances-level (Yuksekgonul et al., 2022b; Diwan et al., 2022). However, we observe that the bags-of-words problem is not only related to the models, but also highly correlated to the evaluation metrics, and VisualGPTScore is not sensitive to the bags-of-words phenomenon.

We explore the influence of different metrics in sensitivity to the order of tokens in sentences for GVLMs. Following CREPE (Ma et al., 2023), we randomly sample 2.5K image-text pairs from the COCO dataset (Lin et al., 2014) and adopt the following strategies to shuffle the elements of captions: *Shuffle only nouns & adjectives*, *Shuffle all but nouns & adjectives*, *Shuffle within trigrams*, *Shuffle trigrams*. Then, we calculate the VisualGPTScore, GPTScore (Fu et al., 2023) and BERTScore (Zhang et al., 2019) based on LLaVA-7B. The distribution of normalized scores are shown in Fig. 1, where `x1` represents positive references and `x2-x5` represents shuffled references, respectively.

It can be observed that to the same model, LLaVA-7B, VisualGPTScore is similar to GPTScore, more sensitive to the order and structure of reference sentences compared with contrastive metric BERTScore. We also report the score distribution of the CLIP model using contrastive similarity (CLIPScore in Fig. 1), which is similar to the distribution of BERTScore results on LLaVA-7B. It implies the bags-of-words problem may be attributed to the evaluation metrics based on similarity score, but generative scores mitigate the problem to some extent.

3.3 Sensitivity to syntax and contents

Based on the observation that ViusalGPTScore mitigates the bags-of-words problem to some extent, we are curious about whether they lean more towards evaluating syntactic correctness than content relevance when assessing the compositionality of GVLMs. To examine it, we design an experiment using the test set of Flickr30K dataset (Young et al., 2014). Specifically, we sample 507 image-text pairs and construct three types of evaluation cases as shown in Fig. 2. Given an image, the task is to retrieve the positive reference from the cases below. The final scores are averaged over 507 test samples. In `Case 1`, each positive reference sentence is accompanied by two hard negatives with shuffled nouns, adjectives and trigrams. In `Case 2`, the provided negatives are fluent and syntactically correct captions sampled from COCO, which are unrelated to the visual contents. In `Case 3`, we keep only adjectives and nouns in the positive reference sentences by removing all the adverbs, pronouns and modifiers.

We present Recall@1 of VisualGPTScore for the GVLM (LLaVA-7B), and vision-language similarity for the CVLM (CLIP) in three evaluation cases. As shown in Fig. 3, the LLaVA model can easily discriminate the right reference sentences from the shuffled ones, reaching 98.62% with the help of VisualGPTScore. However, if the negatives are random reference sentences in `Case 2`, the performance degradation is up to 31.56%. In `Case 3`, where the sentences are syntactically incorrect, the performance drops to 27.02%. In contrast, CLIP excels at excluding negative sentences that are con-

Case 1	VisualGPTScore
Right caption: an elderly asian woman wearing a straw-like hat sits outside near a bicycle while a gray car is about to pass by.	0.405
Shuffled caption: an like gray hat wearing a bicycle - asian woman sits outside near a straw while a about car is elderly to pass by.	0.051
Shuffled caption: elderly an asian wearing a woman hat sits straw-like a near outside bicycle a while is gray car pass to about by	0.077
Case 2	
Right caption: an elderly asian woman wearing a straw-like hat sits outside near a bicycle while a gray car is about to pass by	0.405
Random caption: the two cats are laying on the chair together	0.231
Random caption: two giraffes in an outdoor setting eating grass	0.432
Case 3	
Content caption: elderly asian woman, straw-like hat, bicycle , gray car	0.322
Random caption: the two cats are laying on the chair together	0.231
Random caption: two giraffes in an outdoor setting eating grass	0.432

Figure 2: **An example of three `Cases` of captions we construct to validate the preference of syntax and contents.** `Right caption`: the original caption of the image, `Shuffled caption`: caption that the sentence elements are shuffled, `Random caption`: fluent and syntactically correct captions from other datasets (COCO), `Content caption`: caption that keeps only adjectives and nouns to keep the contents like objects and attributes. We present the normalized VisualGPTScore of every reference sentences in this example. The scores of the `Right caption` and `Content caption` may be lower compared to the `Random caption` (0.405, 0.322 vs. 0.432). This indicates that in this example, generative VLMs tend to prioritize syntactically correct sentences over ones that are more relevant to the content.

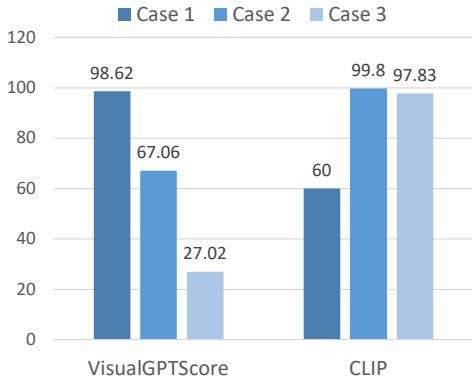


Figure 3: We report the accuracy of VisualGPTScore based on LLaVA-7B and contrastive score based on CLIP in the sampled 507 image-text pairs, each pair is consisted of three cases like the example in Fig. 2.

textually unrelated to the image, but suffers from insensitive to syntax and sentence order.

The potential reason for the above results is the difference in the pre-training paradigm. Specifically, the generative model pre-training is to maximize the likelihood of the next token prediction in an auto-regressive manner. In contrast, the training objective of CVLM is to maximize the alignment between positive image-text pairs and minimize that between negative ones. Previous research (Yuksekgonul et al., 2022b) shows that CLIP takes the short-cut strategy of not encoding the order information, but only object features for retrieval/captioning tasks, which conforms to our finding. We also conclude that *the generative VLMs may take the short-cut strategy of not fully mapping the visual and linguistic features, but leveraging the emerging capacity to generate based on limited visual cues.*

4 Benchmarks Examination

From above, we know current benchmarks are curated for evaluating CLIP based on similarity score originally. Hence, we also examine the impact of using these datasets for evaluating GVLMs with VisualGPTScore.

4.1 Morphological bias in current benchmarks

According to the observation made in Section 3, it is evident that auto-regressive vision-language models exhibit sensitivity toward the syntax and order of phrases. Hence, existing benchmarks that generate hard negatives by swapping, shuffling, or replacing specific entities promote a morphological bias, which refers to a preference for models to rely on the morphological structure of words. Consequently, this bias can be exploited by GVLMs to effortlessly differentiate between positive and negative samples.

To show that the bias exists in current compositional reasoning benchmarks, we conduct the ablation of utilizing both GVLMs and CVLMs to reason nonsensical images with normal reference sentences. Specifically, we construct the image-text pairs by replacing the original images with images composed of *random noises*. We observe the performance drop in both the GVLMs and CVLMs. As shown in Fig. 4, the performance degradation of CLIP (ViT-B/32) is large, approaching the `Recall@1` accuracy of randomly choosing. However, as for the LLaVA-7B, the trend of performance dropping is not obvious, indicating the GVLMs make the right choices **solely** based on

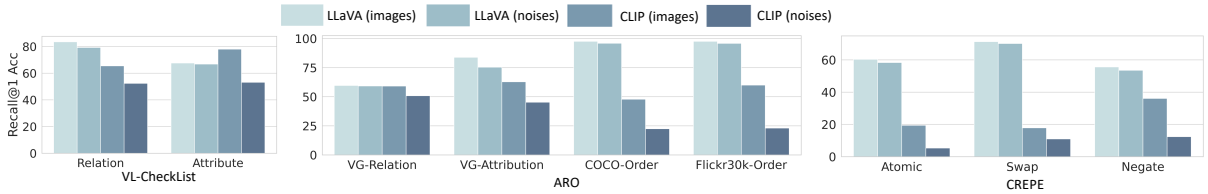


Figure 4: The drop in performance of the LLaVA model when performing compositional reasoning on **nonsensical noisy images** is minimal in existing benchmarks, whereas the CLIP model exhibits a significant decrease. This indicates current benchmarks are exploited by the LLM part of GVLMs, not effective in measuring the multimodal compositionality.

the linguistic reference sentences without visual features. Therefore, the benchmarks lean towards evaluating the compositionality of the LLM part in GVLMs, rather than that of visio-linguistic representations.

4.2 MorphoBias score

To alleviate the morphological bias in current benchmarks, we first quantify the bias for analysis. In an ideal scenario, in the absence of visual intervention, the quantified scores generated by GVLMs for positive and negative reference sentences should be equivalent. Therefore, we define the *MorphoBias Score* to measure the morphological discrepancy between positive and negative reference sentences. Formally, the MorphoBias Score is calculated using the generative scores of positive and negative text produced by auto-regressive language models:

$$
\begin{aligned}
&Score_{MorphoBias} \\
&= \Delta(\sum_{i=1}^{m} w_i \log P(p_i|\boldsymbol{p}_{<i};\theta) \\
&- \sum_{j=1}^{n} \hat{w}_j \log P(n_j|\boldsymbol{n}_{<j};\theta)),
\end{aligned}
\tag{4}
$$

where $\Delta, \mathbf{p}, \mathbf{n}, \theta$ represent normalization, positive tokens, negative tokens, and parameters of LLMs respectively. we leverage a strong LLM, Vicuna-13B-v1.5 (Chiang et al., 2023), to compute the MorphoBias Score, which are normalized between -1 and 1. We present the visualization of MorphoBias Score distributions over different benchmarks in Fig. 5. We find that most of the mainstream benchmarks except Winoground are biased towards positive captions with distribution centers located to the right, *which makes the generative scores of GVLMs on these benchmarks overvalued.*

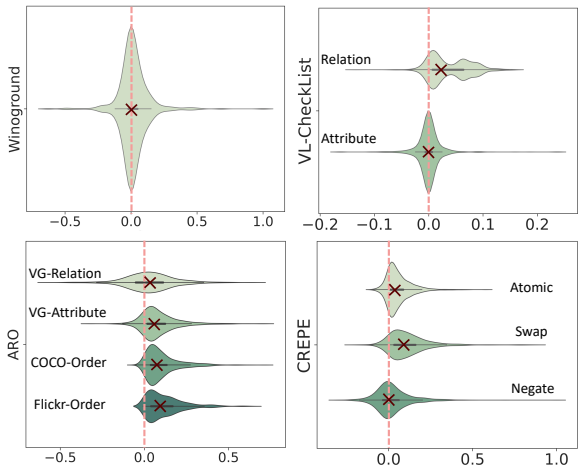


Figure 5: We visualize the distribution of *MorphoBias Score* in current benchmarks. The *MorphoBias Score* is defined as the difference between the LLM-based generative scores of positive and negative references. For ARO, VL-CheckList and CREPE, the distribution of the MorphoBias scores is situated towards the positive end (to the right of the red line), implying that these benchmarks are biased to positive captions morphologically.

5 Calibrate the Bias in Benchmarks

In this section, we propose a strategy to modify the benchmarks and calibrate the morphological bias in order to provide a better evaluation of GVLMs. Specifically, we filter current datasets leveraging LLMs and add a novel challenge to evaluate visual content understanding. We name the new benchmark as **MO**rphologicall **De**-biased Benchmark, abbreviated as **MODE**. In the following, we describe the filtering details of each dataset and the new challenge. Then we show human evaluation to show the effectiveness of MODE.

5.1 Winoground.

The Winoground (Thrush et al., 2022) dataset comprises 400 image-text pairs, with each pair consisting of two images and two captions. The two captions exhibit identical sets of morphemes, albeit in

	Comprehensive	Relation		Attribute		Atomic	Negate	Content	
	Winoground	VL-CheckList	VG(ARO)	VL-CheckList	VG(ARO)	VG(CREPE)	VG(CREPE)	COCO	Flickr30K
num of images	800	5,193	2,328	5,858	5,193	1,954	1,930	2500	500
num of references	800	10,386	4,656	11,716	10,386	11,724	11,580	7,500	1,500
metrics	Group Score	Recall@1							
random results	16.7%	50.0%	50.0%	50.0%	50.0%	16.7%	16.7%	33.3%	33.3%
Human Evaluation (closer to 0 is better)									
origin ref.	-	3.18	1.73	0.95	3.29	1.67	2.11	-	-
MODE ref.	-	1.40	0.62	0.35	1.01	0.94	1.63	-	-

Table 1: Taxonomy of **MODE** benchmark and human evaluation results on rating bias. Each branch undergoes human evaluation based on 50 reference sentences from the original dataset and 50 from MODE.

different orders. Different from other benchmarks that construct hard negatives by simply altering the positive texts, both positive and negative texts in Winoground are fluent, meaningful, and can match related images. Thus, we include all samples in Winoground into the MODE benchmark without further calibrating, aiming to evaluate the comprehensive multimodal compositional understanding of GVLMs, especially on the *pragmatics, symbolic and series* factors as introduced in (Thrush et al., 2022).

5.2 Relations and attributes.

Real-world natural scenes are inherently intricate, encompassing a multitude of specific attributes such as colors, materials, and object relationships. Models that can tackle compositional reasoning require a nuanced understanding that goes beyond mere object-level analysis. Hence, we collect relation and attribute branches from ARO (Yuksekgonul et al., 2022a) and VL-CheckList (Zhao et al., 2022). To calibrate the morphological bias, we compute the MorphoBias score of the samples as described in Eqn. 4 and filter out ones that have a higher score than the threshold. The idea is to ensure that samples with strong morphological bias are excluded for better vision-language compositional evaluation.

We choose the filtering thresholds of the MorphoBias score to be close to zero (specifically, by ensuring the $p - value$ of the MorphoBias score is statistically below $1e - 5$). The filtered data includes 5,193 items from VL-CheckList and 2,328 items from Visual Genome (Krishna et al., 2017) to measure relation reasoning, and 5,858 items from VL-CheckList as well as 5,193 items from Visual Genome to evaluate attribute reasoning. Specifically, for VL-CheckList, the *Relation* branch contains two subclasses, *i.e. action* and *spatial*, and the *Attribute* branch includes *action*, *color*, *material*, *size* and *state*. The number of items in each subclass is elaborated in Table 1.

5.3 Atomic and negate.

In CREPE benchmark (Ma et al., 2023), the authors propose to assess the VLMs on captions that atoms are replaced or negated. The atom replacing is like *a **bus** with a side, light, and window v.s. a **train** with a side, light, and window*, whereas the atom or sentence negating is as *Another bowl on a cloth with an orange in it. The another bowl has a reflection and **casts a shadow** v.s. Another bowl on a cloth with an orange in it. The another bowl has a reflection and casts something. **There is no shadow***. There is a considerable proportion of reconstructed captions in CREPE that are fluent and coherent, thereby we also leverage the same method to filter the samples as we do for relations and attributes.

5.4 Replace syntactic perturbation with a content-only understanding challenge.

A plethora of benchmarks perturbs the order information in the reference sentences to measure the word order sensitivity of CVLMs, which tend to treat the captions as *bags of words* as we present in Fig. 1. The hard negative construction methods include swapping atoms, shuffling nouns, adjectives, trigrams, and all words *etc*. However, due to the intrinsic morphological awareness of LLMs, the challenge of order perturbation is not effective in assessing the visio-linguistic compositionality of GVLMs. Hence, we abandon the order challenge and propose a content-only understanding challenge.

Specifically, we modify the positive reference sentences from COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014), keeping only the object- and attribute-related atoms/words. Then, we randomly select fluent, coherent and meaningful reference sentences from other datasets to serve

	Comprehensive	Relation	Attribute	Atomic	Negate	Content
LLaVA-7B (Liu et al., 2023a)	13.00	65.52	70.55	35.01	59.01	42.02
LLaVA-13B (Liu et al., 2023a)	17.00	62.75	72.70	38.33	7.56	49.80
MiniGPT-7B (Zhu et al., 2023)	9.50	66.18	78.48	35.62	24.15	19.92
mPLUG-Owl (Ye et al., 2023)	11.00	65.91	69.04	34.90	54.61	35.73
InstructBLIP (Dai et al., 2023)	**26.00**	**73.87**	79.39	44.37	66.84	**57.83**
LLaMA Adapter V2 (Gao et al., 2023)	7.75	58.67	65.07	31.32	20.26	10.48
Emu (Sun et al., 2023)	4.00	68.54	**85.84**	**51.38**	**87.20**	2.79

Table 2: Evaluation results of GVLMs on MODE benchmark. All the models are instruction-tuned. We present the average performance of two sub-branches within the categories of *Relation*, *Attribute* and *Content*.

as hard negatives, which are unrelated to the visual content. Examples of this challenging task can be found in Fig. 8 in the Appendix. *The task poses a challenge and exemplifies the robustness of GVLMs against their inherent inclination towards syntactically correct reference sentences.*

5.5 Human evaluation of MODE.

In order to illustrate that our proposed MODE alleviates the morphological bias, we ask two annotators to rate the disparity between positive and negative reference sentences. The rating score ranges from -5 to 5, where the higher the score, the more reasonable text is for positive reference sentences. Conversely, the lower the score, the more reasonable the text is for negative ones. The definition of *reasonable* comprises fluency, syntax, and the meaning of sentences. Note the reference sentences from the original dataset or MODE are agnostic to the annotators and we average the ratings of them. Table 1 clearly demonstrates that the reference sentences in our MODE benchmark substantially mitigate bias, as indicated by the score of human judgments approaching zero. The drop implies that the morphological disparity between positive and negative reference sentences is drastically narrowed.

5.6 Results of GVLMs on MODE

Based on the MODE benchmark, we report the performance of more concurrent GVLMs based on the VisualGPTScore metric in Table 2. It can be observed that InstructBLIP (Dai et al., 2023) and Emu (Sun et al., 2023) hold the top-2 positions in almost all dimensions of our benchmark. However, the abysmal performance on *Comprehensive* and *Content* implies the vulnerability of Emu when negative reference sentences are hard and challenging. In contrast, InstructBLIP and LLaVA-13B (Liu et al., 2023a) are more robust to the *Content* challenge and achieve high performance on hard negatives. This provides the first de-biased and comprehensive evaluation of recent GVLMs in terms of visual compositionality. **Note that we do not claim that MODE can better measure the performance of GVLMs in all aspects.** However, it can better measure their compositionality with less morphological bias, which is supported by the reduction of MorphoBias Score and the human evaluation in Table 1. We believe this benchmark can facilitate a unified and fair comparison for future GVLM research.

6 Conclusion

In this work, we evaluate the compositionality of "bridge-architecture" generative VLMs via generative multimodal score, VisualGPTScore. We examine both the VisualGPTScore and current benchmarks for evaluating the multimodal compositional understanding of GVLMs. Based on the examinations, we identify the morphological bias that exists in current datasets for GVLMs, and define the bias with MorphoBias Score quantitatively. We then propose a MODE benchmark that calibrates the morphological bias and provides a better content understanding evaluation for GVLMs. We report the results of multiple GVLMs on our proposed MODE benchmark and uncover new findings of the GVLMs' capabilities.

7 Limitations

We discuss the potential limitations of this paper from two aspects. First, our proposed novel benchmark cannot be proved to better measure the performance of generative VLMs in all aspects, including emergent capability, vision understanding and complex reasoning. Our benchmark just evaluates the GVLMs in terms of VL compositionality more fairly by removing the morphological bias in pre-

vious benchmarks. Second, our new benchmark is based on calibrating the previous ones, and sampling from them to lower the MorphoBias Score. Thus, the scale of the whole dataset is relatively small, limiting the generalization of the benchmark to some extent.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023a. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023b. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

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

Zejun Li, Ye Wang, Mengfei Du, Qingwen Liu, Binhao Wu, Jiwen Zhang, Chengxing Zhou, Zhihao Fan, Jie Fu, Jingjing Chen, et al. 2023b. Reform-eval: Evaluating large vision language models via unified re-formulation of task-oriented benchmarks. *arXiv preprint arXiv:2310.02569*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2023. Visualgptscore: Visio-linguistic reasoning with multimodal generative pre-training scores. *arXiv preprint arXiv:2306.01879*.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Kousik Rajesh, Mrigank Raman, Mohammed Asad Karim, and Pranit Chawla. 2023. Bridging the gap: Exploring the capabilities of bridge-architectures for complex visual reasoning tasks. *arXiv preprint arXiv:2307.16395*.

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A Plummer, Ranjay Krishna, and Kate Saenko. 2023a. Cola: How to adapt vision-language models to compose objects localized with attributes? *arXiv preprint arXiv:2305.03689*.

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A Plummer, Ranjay Krishna, and Kate Saenko. 2023b. Cola: How to adapt vision-language models to compose objects localized with attributes? *arXiv preprint arXiv:2305.03689*.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022a. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022b. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting

twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

A.1 Formulations of GVLMs and CVLMs

In accordance with the discussion in the main text, we define GVLMs as models that combine visual encoders with large language models (LLMs) trained on large text corpora. The visual tokens are mapped into a lexical embedding space and harnessed to generate textual content in an autoregressive manner. Formally, given an image I and the visual encoding $g(I)$ from encoders like Vision Transformer (Dosovitskiy et al., 2020), the mapping process can be formulated as:

$$\boldsymbol{z} = \mathbf{M}(g(I)), \boldsymbol{z} = \{z_1, z_2, ..., z_N\}, \quad (5)$$

where N is the number of visual tokens and $\mathbf{M}$ is the mapping layers. Different from CVLMs that utilize contrastive loss, the training objective of multimodal autoregressive training is to maximize the log-likelihood of the next true token. Denote the tokenized instructions as $\boldsymbol{p}$ and the output words as $t_i, (1 \leq i \leq K)$, the GVLM training objective is defined as:

$$\max_{\theta_M, \theta_\sigma} \sum_{i=1}^{K} \log P(t_i | \boldsymbol{p}, \boldsymbol{z}, t_1, t_2, ..., t_{i-1}; \theta_M, \theta_\sigma) \quad (6)$$

where θ_M refers to the learnable parameters of mapping layers $\mathbf{M}$ and θ_σ refers to other tunable parameters like adapter layers in LLaMA-Adapter V2 (Gao et al., 2023), or visual abstractor and LoRA in mPLUG-Owl (Ye et al., 2023).

In comparison, the training objective of CVLMs (CLIP-based (Radford et al., 2021)) is based on the contrastive loss between vision and language. Given an input image I and text T, the encoded visual and linguistic features are denoted as f_v and f_t. Then, two transformation matrices W_v and W_t are employed to project the visual and text features into a joint feature embedding space, which is formulated as:

$$v = \frac{W_v^\top f_v}{||W_v^\top f_v||}, u = \frac{W_t^\top f_t}{||W_t^\top f_t||} \quad (7)$$

In the shared embedding space, contrastive learning narrow the discrepancy of vision and language, aligning the image-text pairs in the same batch. The training objective of this process comprises two components, *i.e.* $\mathcal{L}_{v \to t}$ for text retrieval and $\mathcal{L}_{t \to v}$ for image retrieval. The similarity of matched pairs will be maximized while unmatched ones will be minimized. The formula is:

$$\begin{aligned} \mathcal{L} =& \mathcal{L}_{v \to t} + \mathcal{L}_{t \to v} \\ =& -\frac{1}{|\Omega_v^+|} \sum_{T_j \in \Omega_v^+} \log \frac{\exp(v_i^\top u_j / \tau)}{\sum_{T_k \in \Omega_t} \exp(v_i^\top u_k / \tau)} \\ & -\frac{1}{|\Omega_t^+|} \sum_{I_i \in \Omega_t^+} \log \frac{\exp(u_i^\top v_j / \tau)}{\sum_{I_k \in \Omega_v} \exp(u_i^\top v_k / \tau)} \end{aligned} \quad (8)$$

where Ω_v, Ω_t represent a batch of images and texts while Ω_v^+, Ω_t^+ denote positive subsets matched to image I_i and text T_i.

***POS**: an old person kisses a young person*

***NEG**: a young person kisses an old person*

Figure 6: An data example in current benchmarks. The image, positive and negative references are from Winoground (Thrush et al., 2022).

A.2 Granularity influence of VisualGPTScore.

To explore the influence of granularity of references in the visio-linguistic compositional reasoning, we leverage a language model to enrich the object details and relational phrases for short references in Winoground dataset, where all references are fluent and reasonable. Vicuna-13B-v1.5[1] is adopted as the LLM, which is instruction-following tuned based on LLaMA 2 (Touvron et al., 2023b), one of the strongest open-source LLMs currently. Note

[1]https://huggingface.co/lmsys/vicuna-13b-v1.5/tree/main

Models&References	Text Score	Image Score	Group Score
LLaVA+Original	12.06	12.77	7.45
LLaVA+Fine-grained	8.51(-3.55)	37.23(+24.50)	6.38(-1.07)
MiniGPT-4+Original	18.44	17.02	9.22
MiniGPT-4+Fine-grained	6.03(-12.41)	31.91(+14.89)	4.96(-4.26)

Table 3: Accuracy of LLaVA and MiniGPT-4 on original and fine-grained references of filtered Winoground dataset. The definitions of Text Score, Image Score, and Group Score is specified in Winoground (Thrush et al., 2022).

that we artificially filter out nonsensical and unrelated expanded captions that are not relevant to the image and keep 282 of 400 image-text pairs finally. The expandation of references is shown in Fig. 7.

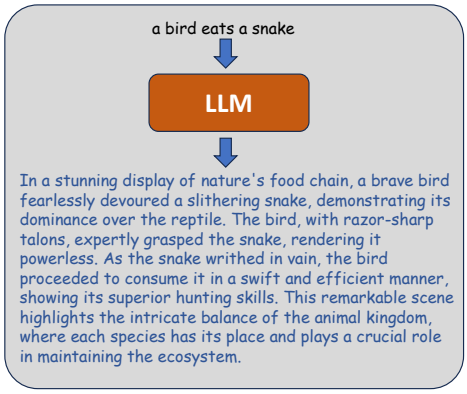


Figure 7: An LLM is leveraged to fine-grain the references.

We present the results in Table 3, and observe that the performance of "Image Score" has been largely improved, indicating the fine-grained references are beneficial for text-to-image retrieval based on the definition of "Image Score" in Winoground.

A.3 Zero-shot answer generation

Unlike contrastive models, GVLMs excel in zero-shot generation when guided by instructions, prompts, or demonstrations. We attempt to prompt and demonstrate the LLaVA and MiniGPT-4 to output the choices of positive or negative reference sentences based on corresponding images. However, we do not consider zero-shot generation of answers in our paper with two reasons. First, zero-shot answer generation cannot reflect the GVLMs' compositional understanding quantitatively, without scores or probabilities to show the confidence of judgements.

Second, when demonstrating the GVLMs to generate the option number of reference sentences directly, it is hard to acquire the direct answer due to the free-form answer format, especially considering the emergent capability is limited in relatively small-scaled GVLMs. In a limited number of instances, we observed successful model outputs where options or inference processes were accurately provided, resembling the blue line in Table 4. However, in the majority of cases, the GVLMs generated fabricated answers that were characterized by a rhetorical tone, similar to the examples shown in Table 4. Also, there are cases that the rationales are correct, but the option number is wrong, conflicting with the reasoning process of GVLMs (shown in orange line in Table 4). Hence, assessing the compositionality of GVLMs solely through direct zero-shot answer generation becomes challenging, particularly when the zero-shot capability is constrained within a relatively small-scale model like the 7B variant. Furthermore, it is not possible to quantitatively analyze the alignment of a single image-text pair using this type of evaluation method.

- Ducks are known to be efficient at swimming, and they spend a significant amount of time in the water. In this image, a duck is swimming on the water while facing the camera. The option B is most similar to the image since it describes a duck walking or swimming in the water.
- Which of the two options provides a better visual representation of the actual image?
- Which character should I output?
- Which one is most similar to the image?
- The girl is standing to the left of the giraffe, and they seem to be in a natural environment. The girl is feeding the giraffe from her hand, and the giraffe has its tongue out, reaching towards her hand. The answer is more likely to be B.

Table 4: Examples of zero-shot answer generation method. Blue: free-form generation, Teal: fabricated answers, Orange: conflicting rationales and answers.

A.4 Examples of content challenge of MODE

We present some examples of items in Content challenge branch in our MODE benchmark in Fig. 8. Each item comprises of one positive reference sentence and two negative ones. The red texts are positive reference sentences that only kept visual contents related phrases, while the black texts are negative reference sentences that extracted randomly from other datasets. The negative reference sentences are fluent, coherent and meaningful, but irrelevant to the contents of images.

The pure content understanding is challenging. Specifically, the intrinsic inclination of GVLMs towards syntactic correctness drives the GVLMs to prefer negative reference sentences. From the perspective of our proposed MorphoBias score, the bias of our Content Challenge is opposite to the current benchmarks, which is biased to the negative reference sentences in morphology. Therefore, GVLMs have to overcome the negative bias in morphology and show the robustness of visual understanding.

baby , bouncy seat , boy , toys

white bathroom with a sink, toilet, garbage can and basket

kitchen with wooden cabinets and granite countertops

woman , peach tank top , mountain bike

a woman skiing down a ski slope in the slope

a group of people are in an inner tube looking boat

girls , tree branch , dog

a female is on the computer playing a car game

there is one snowboarding going down the hill

large green train , wooden crates

two black and one white dog interacting in the grass

a man is standing at edge of a pond, with two dogs and is throwing a branch is water

donuts , paper , coffee cup

a bearded man wearing a denim jacket sits on a bench

a bellhop is pushing luggage around inside a hotel

small crowd , people , doubles match , tennis

young male with glasses, blond-hair and beard, holding a black shovel over a campfire and a barbecue pit, filled with red meat

two people waving their hands in the air and looking up

suit case , large leaf setting , car

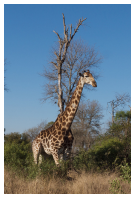
one lone army soldier overlooking an area with binoculars or perhaps a range finder in a sub desert area

black male wearing yellow shirt doing a reading with his equipment

plate full , pizza , corn , cheese

a young man holding a young woman in his arms as they get splashed by water shooting up from a fountain

an old man with a beard is sitting on a milk crate on the street



tall giraffe , tall brush

two people stand at the peak of a mountain

two men wearing martial arts clothing are practicing martial arts

various electronics , floor

a woman with a drink and a woman with a cellphone

a man jumps rope while a crowd of people watch him

living room scene , man , young girl , wii controllers , woman

a woman wearing a pink shirt showing a man with a striped sweater how to do some work with yarn

two teams, one in pink and one in white, play lacrosse on a field

Figure 8: Examples of Content challenge in our MODE benchmark. The red texts denote positive reference sentences that solely capture visual elements while disregarding sentence structure. On the other hand, the black texts represent negative reference sentences that are grammatically sound and meaningful, yet unrelated to the visual contents depicted in the images.