VisualGPTScore: Visio-Linguistic Reasoning with Multimodal Generative Pre-Training Scores

Zhiqiu Lin*[1] **Xinyue Chen***[1] **Deepak Pathak**[1] **Pengchuan Zhang**[2] **Deva Ramanan**[1]
[1]CMU [2]Meta

Open-source code in webpage

Abstract

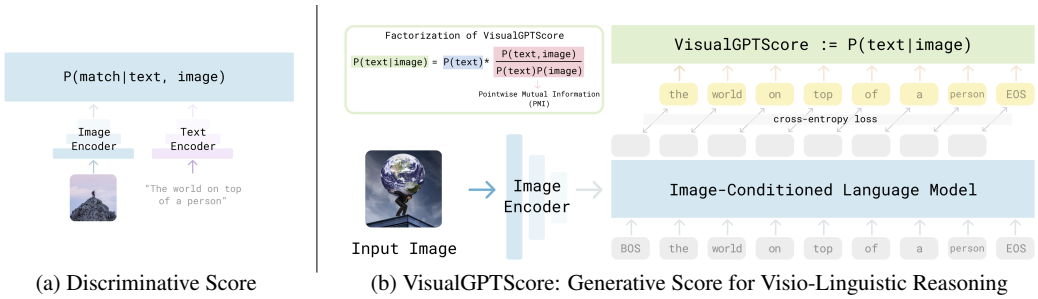
Vision-language models (VLMs) discriminatively pre-trained with contrastive image-text matching losses such as $P(\text{match}|\text{text}, \text{image})$ have been criticized for lacking compositional understanding. This means they might output similar scores even if the original caption is rearranged into a different semantic statement. To address this, we propose to use the **V**isual **G**enerative **P**re-**T**raining Score (**VisualGPTScore**) of $P(\text{text}|\text{image})$, a *multimodal generative* score that captures the likelihood of a text caption conditioned on an image using an image-conditioned language model. Contrary to the belief that VLMs are mere bag-of-words models, our off-the-shelf VisualGPTScore demonstrates top-tier performance on recently proposed image-text retrieval benchmarks like ARO and Crepe that assess compositional reasoning. Furthermore, we factorize VisualGPTScore into a product of the *marginal* P(text) and the *Pointwise Mutual Information* (PMI). This helps to (a) diagnose datasets with strong language bias, and (b) debias results on other benchmarks like Winoground using an information-theoretic framework. VisualGPTScore provides valuable insights and serves as a strong baseline for future evaluation of visio-linguistic compositionality.

1 Introduction

Latest large language models (LLMs) like ChatGPT [9] and GPT-4 [10] have reached human-level performance on tasks requiring complex compositional reasoning [11, 12, 13, 14, 15]. Although pre-trained on massive-scale web data, contemporary vision-language models (VLMs) such as CLIP [1] fail to encode compositional relationships and struggle with recently proposed image-text retrieval benchmarks [3, 8, 16, 17, 18] that humans find relatively trivial. For instance, ARO [3] reveals that state-of-the-art VLMs trained with image-text contrastive (ITC) or image-text matching (ITM) objectives exhibit bag-of-words behaviors and cannot distinguish between nuanced phrases with shuffled words such as "horse eating grass" and "grass eating horse".

Visual Generative Pre-Training Score (our approach). To challenge the prevailing belief that VLMs are bag-of-words models, we draw inspiration from the exceptional compositional reasoning capabilities of *generative pre-trained* LLMs, like GPT [10, 19, 20], which models the entire sequence likelihood via next-token prediction [19, 21, 22]. We leverage the popular *image-conditioned* language model BLIP [6, 7], pre-trained with both discriminative (ITC and ITM) and generative (next-token prediction) objectives. We show that **Visual G**enerative **P**re-**T**raining Score, ie., **the conditional likelihood of the text given an image** $P(\text{text}|\text{image})$, significantly surpasses prior art on a suite of challenging compositionality benchmarks, such as ARO [3], Crepe [4], VL-CheckList [16], where discriminative approaches like ITCScore and ITMScore have failed.

Preprint. Under review.



(a) Discriminative Score (b) VisualGPTScore: Generative Score for Visio-Linguistic Reasoning

Figure 1: ***Generative* score for enhanced evaluation of multimodal compositionality.** Figure (a) illustrates mainstream VLMs evaluated with discriminative scores [1, 2] that model $P(\text{match}|\text{text}, \text{image})$ with contrastive or matching losses. Such approaches exhibit bag-of-words behaviors, failing to generalize to compositional reasoning benchmarks [3, 4, 5]. Figure (b) illustrates **Visual G**enerative **P**re-**T**raining **Score** (VisualGPTScore) based on VLMs [6, 7] trained with generative language modelling loss. Contrary to the prevailing belief that VLMs are mere bag-of-words models [3], VisualGPTScore demonstrates top-tier performance on recently proposed image-text retrieval benchmarks that assesses compositional reasoning (section 3). Moreover, by factorizing VisualGPTScore as a product of marginal probability (P(text)) and Pointwise Mutual Information (PMI) (section 4), we (a) diagnose datasets with strong language bias (section 5), and (b) improve on other retrieval benchmarks like Winoground [8] (section 6). VisualGPTScore offers insights for future evaluation of visio-linguistic compositionality.

Have we closed the compositionality gap? Although VisualGPTScore sometimes achieves *near-perfect* performance on certain benchmarks, we do not claim that it bridges the compositionality gap. To analyze recent benchmarks, we propose to factorize VisualGPTScore into a product of the unimodal marginal $P(\text{text})$ and Pointwise Mutual Information (PMI), which is widely adopted in information retrieval tasks [23, 24, 25, 26]. Through an approximation of $P(\text{text})$ using Monte Carlo sampling, our experiments expose that recent benchmarks can be *partially* addressed with only the language model $P(\text{text})$, indicating a significant design flaw. For instance, many image-to-text retrieval datasets [3, 4] create challenging negative captions by drastically modifying the original ones, without taking into account whether these altered captions still make sense. This bias in language allows the negatives to be easily dismissed without even looking at the image; we show that text-only solutions, such as a "blind" version of VisualGPTScore that conditions on random gaussian noise images, can exceed previous best results achieved by discriminative approaches using VLMs.

A systematic approach for diagnosing language bias and improving on retrieval tasks. We present a method to adjust the individual components of VisualGPTScore ($P(\text{text})$ and PMI) by tuning a scalar value $\alpha \in [0, 1]$. This framework repurposes VisualGPTScore as a diagnostic tool to examine the extent of language bias in different datasets. Additionally, it mirrors the classic PMI^k method [27], which controls the strength of debiasing. We show that it enhances performance on both compositionality benchmarks like Winoground [8] and classic retrieval tasks like COCO [28].

Summary. We present an image-text retrieval method grounded on a *multimodal generative* score:

- Our VisualGPTScore significantly outperforms existing solutions on recent visio-linguistic compositionality benchmarks and challenges the prevailing notion that VLMs are bag-of-words models [3, 4, 8]. Notably, our approach does not entail additional training or computational costs and can rival the performance of discriminative methods [6, 7, 29] on most datasets.
- We present an information-theoretic factorization of VisualGPTScore to address bias in retrieval tasks and to diagnose datasets that exhibit strong language bias. To this end, we leverage a *content-free* technique to reduce the sampling cost of Monte Carlo method for estimating marginal probabilities. We hope our framework provides insights for future evaluations of visio-linguistic compositionality and generative scores in retrieval.

2 Related Works

Vision-language modelling. State-of-the-art VLMs like CLIP [1], ALBEF [2], and FLIP [30] are pre-trained on web-scale image-text datasets [31, 32] using discriminative objectives such as

image-text contrastive (ITC) [1, 30, 33, 34] and image-text matching (ITM) [2, 6, 7, 35, 36, 37, 38] loss. These pre-trained models exhibit robust zero-shot [39, 40, 41] and few-shot [42, 43, 44] performance on traditional discriminative tasks [28, 45, 46], often on par with fully-supervised models. Some recent models [40, 47], like Flamingo [48] and BLIP [6, 7], incorporate generative language modelling objectives (next-token prediction [21, 22]) primarily for downstream tasks such as image captioning [28, 49] and visual-question answering [50, 51, 52].

Visio-linguistic compositionality. While VLMs have been successfully applied to diverse tasks requiring visio-linguistic reasoning, such as text-to-image generation [53, 54, 55], benchmarks like Winoground [8], ARO [3], and Crepe [4] cast doubt on their compositional reasoning capabilities. These benchmarks, however, only evaluate discriminative scores (ITCScore and ITMScore) for image-text retrieval. Similarly, we find that all concurrent works focus on discriminative scores of VLMs for compositionality assessment. For example, SyViC [56] discriminatively finetunes VLMs on million-scale synthetic images to enhance spatial, attributive, and relation understanding. Other approaches include (a) training on standard image-text datasets [28, 46] with curated negative captions [57, 58, 59], such as NegCLIP [3], and (b) using carefully-designed regularizers or architectures [5, 17, 60, 61]. In contrast, we demonstrate that an off-the-shelf multimodal generative score can rival previous discriminative approaches on the same benchmarks without additional finetuning.

Generative pre-training and scoring. A plethora of studies suggest that vision models trained with *discriminative* objectives lack incentives to learn structure information [3, 62, 63]. Similarly, early LLMs pre-trained and evaluated with *discriminative* approaches, such as BERT [64] and RoBERTa [65], have also been criticized as bag-of-words models insensitive to word order [66, 67, 68, 69]. Conversely, generative pre-trained LLMs [10, 19, 20] demonstrate exceptional compositional understanding while pre-trained solely with a next-token prediction [19, 21, 22] loss. Furthermore, generative scores of language models [10, 70, 71] have flexible usage in downstream tasks, eg., text evaluation [72] with GPTScore [73] and reranking [74] with pointwise mutual information (PMI [24, 25, 26]). In this work, we also factorize VisualGPTScore as a product of marginal probability and PMI, in order to examine the contribution of each part in different retrieval benchmarks. Specifically, we perform Monte Carlo sampling of our multimodal generative scores to estimate unimodal marginal probabilities ($P(\text{text})$). Our analysis uncovers that a surprisingly large number of recent visio-linguistic datasets [3, 4, 16] can be partially addressed with solutions that focus solely on $P(\text{text})$, completely disregarding the visual modality.

Leveraging language prior in visio-linguistic benchmarks. Visio-linguistic benchmarks, such as image-captioning [75, 76] and visual-question-answering [51, 52, 77, 78, 79, 80, 81], can be shortcutted by exploiting imbalanced unimodal (image or language) prior, often without the need to consider the other modality. Recent compositionality benchmarks are not exempt from this issue as their datasets' language bias closely reflects real-world texts; for instance, sentences with grammatical errors are less likely to be positive. To prevent degenerate unimodal solutions, we advocate for benchmarks like Winoground [8, 17], whose dataset creation and evaluation protocol ensures uniform marginal probabilities among samples.

3 Visual Generative Pre-Training Score

In this section, we formalize our approach of **Visual G**enerative **P**re-**T**raining Score and evaluate it on recently proposed visio-linguistic benchmarks of image-to-text and text-to-image retrieval tasks.

Prior discriminative approaches (Figure 1-a). The majority of mainstream VLMs are pre-trained and evaluated with $P(\text{match}|\text{image}, \text{caption})$, typically modelled via ITCScore and ITMScore. ITC-Score [1, 33] employs a dual-encoder architecture that separately encodes images and texts, which is then followed by a contrastive objective between positive and negative image-text pairs. ITM-Score [2, 6], on the other hand, jointly encodes an image-text pair using a fusion encoder, followed by a binary classification objective indicating whether the pair matches. However, without additional fine-tuning [3, 5, 59], such approaches often fail to generalize to compositional reasoning tasks.

Preliminaries. For the scope of this paper, we assume an image-conditioned language model trained with next-token prediction loss [20, 21, 22]. This encompasses VLMs that have been pre-trained with language generation (captioning) objectives, such as BLIP [6, 7], CoCa [40], GIT [47], and Flamingo [48]. We adopt the open-sourced BLIP models pre-trained on public image-text corpus with both discriminative (ITC and ITM) and generative (next token prediction) objectives for ablation.

VisualGPTScore (Figure 1-b). The likelihood of a text $\mathbf{t} = \{t_1, t_2, \cdots, t_m\}$ conditioned on an image $\mathbf{i}$ can be naturally factorized as a product of conditional probabilities:

$$P(\mathbf{t}|\mathbf{i}) = \prod_{k=1}^{m} P(t_k|t_{<k}, \mathbf{i}) \tag{1}$$

In our implementation, we calculate a weighted sum of the log-likelihoods of t_k at each token position k and apply an exponent to cancel the log:

$$\textbf{VisualGPTScore}(\mathbf{t}, \mathbf{i}) := e^{\sum_{k=1}^{m} w_k \log(P(t_k|t_{<k}, \mathbf{i}))} \tag{2}$$

We set $w_k = \frac{1}{m}$ following prior works [72, 73]. To condition on an input image, BLIP uses a multimodal casual self-attention mask [6, 7] in its image-grounded text decoder, ie., each text token can attend to all its preceding vision and text tokens. While text *generation* requires *sequential* token-by-token prediction, we emphasize that Eq. 2 can be computed in *parallel* at all positions through cross-entropy losses between output logits and ground truth text tokens. As such, our *generative* VisualGPTScore incurs the same computational cost as the state-of-the-art ITMScore [2, 6, 7, 29], which uses a bi-directional attention masking transformer to encode an image-text pair.

Benchmarks and evaluation protocols. We strictly adhere to prior works when reporting results on each benchmark. Notably, ARO [3], Crepe [4], and VL-CheckList [16] focus on image-to-text (I-to-T) retrieval, reporting Recall@1 (R@1) as each image has a single positive caption and multiple negative captions. For ARO [3], we report on four datasets: VG-Relation, VG-Attribution, COCO-Order, and Flickr30k-Order. For Crepe [4], we use the entire productivity set (with complexities ranging from 4 to 12) and report on three datasets: AtomFoils, Negate, and Swap. For VL-CheckList [16], we report on the Relation dataset by averaging the performance of action and spatial tag. For Winoground [8] and EqBen [17], where each sample contains two pairs of image and text, we follow the original protocol to report (a) text score for image-to-text (I-to-T) retrieval, and (b) image score for text-to-image (T-to-I) retrieval. For example, the text score gains 1 point if, for both images, the matching caption score is higher than the non-matching caption score. The calculation for the image score is similar. For Winoground, we report on both the entire testset (400 samples) and the No-Tag subset (171 samples) [18], which tests compositional reasoning rather than other capabilities, e.g., detecting small and out-of-focus objects. We report on the public valset of EqBen [17] because it contains ground-truth labels for computing text and image score respectively. We refer readers to Appendix section 9 for more details and visualization of random samples. Our open-sourced code can be found at link[1]. We include comprehensive reports, such as group scores, fine-grained performance on subtags for Winoground [18], and BLIP-2 results in Appendix.

Performance on Compositionality Benchmarks. In Table 1, we show that VisualGPTScore, based on the BLIP model (pre-trained on LAION-114M [31] with ViT-L image encoder), achieves state-of-the-art results on the majority of recent visio-linguistic benchmarks without any finetuning. Notably, VisualGPTScore outperforms the best discriminative approaches (including BLIP's ITMScore) on all I-to-T retrieval tasks of ARO, Crepe, and VL-CheckList. It also rivals on T-to-I retrieval tasks of Winoground and EqBen, although it falls short on their corresponding I-to-T retrieval tasks, a point which we will address in the following section.

4 Information-Theoretic Factorization of VisualGPTScore

To comprehend the performance discrepancy across benchmarks, we observe that Winoground [8] and EqBen [17] follow distinct creation procedures compared to other compositionality datasets. ARO [3], Crepe [4], VL-CheckList [16] construct hard negative captions for each image by modifying the ground truth caption. As a result, these negative captions *do not have any matching images* in the dataset. On the other hand, both Winoground and EqBen include two image-text pairs in a test case, which means each caption (as well as image) has a equal chance of being positive. We propose that Winoground and EqBen pose a greater challenge due to their curation and evaluation protocol which enforces a **balanced marginal distribution** [52]. We now formally analyze this discrepancy through an information-theoretic factorization of VisualGPTScore.

Notation. Given an image $\mathbf{i}$ and a text $\mathbf{t}$, $P(\mathbf{i}, \mathbf{t})$ represents the *joint probability* of a matching image-text pair $(\mathbf{i}, \mathbf{t})$ being sampled from the distribution. Likewise, $P(\mathbf{i})$ and $P(\mathbf{t})$ models the

[1] `https://github.com/linzhiqiu/visual_gpt_score/`

Mode	Metric	Benchmark	Dataset	Random	ITMScore	VisualGPTScore	SOTA Method	
							Name	Score
I-to-T	Recall@1	ARO [3]	VG-relation	50.0	58.7	**89.1** (+30.4)	NegCLIP [3]	81.0
			VG-Attribution	50.0	90.3	**95.3** (+ 5.0)		71.0
			COCO-Order	20.0	45.1	**99.4** (+54.3)		91.0
			Flickr30K-Order	20.0	51.3	**99.5** (+48.2)		86.0
		Crepe [4]	Atom-Foils	16.7	29.5	**73.2** (+43.7)	ALBEF [2]	28.5
			Negate	16.7	25.5	**79.6** (+54.1)		48.7
			Swap	16.7	20.7	**78.1** (+57.4)		20.4
		VL-CheckList [16]	Relation	50.0	67.7	**90.8** (+23.1)	Syn-BLIP [56]	70.2
	Text Score	Winoground [8]	All	25.0	**35.8**	27.0 (- 8.8)	CLIP [1]	28.5
			No-Tag [18]	25.0	**41.9**	34.9 (- 7.0)		26.2
		EqBen [17]	Val-Set	25.0	**26.0**	9.6 (-16.4)		26.3
T-to-I	Image Score	Winoground [8]	All	25.0	15.8	**21.5** (+ 5.7)	CLIP [1]	10.5
			No-Tag [18]	25.0	21.5	**28.5** (+ 7.0)		8.72
		EqBen [17]	Val-Set	25.0	20.3	**26.1** (+ 5.8)		19.6

Table 1: **Performance on diverse visio-linguistic benchmarks.** We compare SOTA discriminative approaches with our proposed VisualGPTScore (Eq. 2), implemented on the open-sourced BLIP [6] model pre-trained on LAION-114M [31]. VisualGPTScore significantly outperforms the best discriminative approaches (such as ITMScore) across a variety of I-to-T retrieval benchmarks including ARO [3], Crepe [4], and VL-CheckList [16]. On COCO-Order and Flickr30K-Order, VisualGPTScore achieves *near-perfect* performance. It also shows superior T-to-I retrieval performance on Winoground [8] and EqBen [17]. We improve the respective I-to-T performance via debiasing in section 6. We report the performance of the largest CLIP [1] variant (ViT-L/14) on Winoground and EqBen since we do not have access to SOTA pre-trained models [17]. For Winoground, we also report performance on the No-Tag [18] subset that focuses solely on compositional reasoning.

(unimodal) *marginal probability* that a randomly sampled pair contains the image $\mathbf{i}$ or the text $\mathbf{t}$. Due to the non-negligible distribution shift between the train and test datasets, we denote these probabilities with P_{train} and P_{test}, respectively.

VisualGPTScore as a product of $P(\mathbf{t})$ and PMI. As VisualGPTScore models the image-conditioned likelihood of text, we propose to decompose it as a product of *marginal probability* (of text) and *Point-wise Mutual Information (PMI [23, 82])*:

$$\textbf{VisualGPTScore}(\mathbf{t}, \mathbf{i}) := P_{train}(\mathbf{t}|\mathbf{i}) \tag{3}$$

$$=: P_{train}(\mathbf{t}) * \text{pmi}_{P_{train}}(\mathbf{t}, \mathbf{i}) \tag{4}$$

where

$$\text{pmi}_P(\mathbf{t}, \mathbf{i}) = \frac{P(\mathbf{t}, \mathbf{i})}{P(\mathbf{t})P(\mathbf{i})} = \frac{P(\mathbf{t}|\mathbf{i})}{P(\mathbf{t})} = \frac{P(\mathbf{i}|\mathbf{t})}{P(\mathbf{i})} \tag{5}$$

PMI is an information-theoretic measure that quantifies the *association* between an image and a text [83, 84, 85]. It measures how much more (or less) likely the image-text pair co-occurs than if the two were independent. Eq. 5 has found applications in diverse sequence-to-sequence modelling tasks [24, 25, 26] as a retrieval (reranking) objective. Compared to the conditional likelihood $P(\mathbf{t}|\mathbf{i})$, PMI reduces the learned bias for preferring "common" texts with high marginal probabilities $P(\mathbf{t})$ [24, 25, 26]. We now try to estimate $P(\mathbf{t})$ and show how this factorization (Eq. 4) can expose the contributions of each part across benchmarks.

Estimating marginal probabilities using Monte Carlo sampling (oracle approach). We can estimate $P_{train}(\mathbf{t})$ from an image-conditioned language model ($P_{train}(\mathbf{t}|\mathbf{i})$) via Monte Carlo sampling [86], by drawing n images from the train distribution:

$$P_{train}(\mathbf{t}) \approx \frac{1}{n}\sum_{k=1}^{n} P_{train}(\mathbf{t}|\mathbf{i}_k) = \frac{1}{n}\sum_{k=1}^{n} \textbf{VisualGPTScore}(\mathbf{t}, \mathbf{i}_k) \tag{6}$$

Reducing sampling cost with content-free (gaussian noise) images (our approach). The Monte Carlo method outlined above, while straightforward, can be computationally expensive to achieve robust estimates. To address this, we draw inspiration from [87], which uses a *content-free* "null" text *prompt* (such as "N/A") to calculate the probability of a text from LLMs, ie., $P(\mathbf{t}) \approx P(\mathbf{t}|\text{"N/A"})$.

Benchmark	Dataset	SOTA	LLMs ($P_{LLM}(\mathbf{t})$)			VisualGPTScore	
			BART	FLAN-T5	OPT	$P(\mathbf{t}\|\mathbf{null})$	$P(\mathbf{t}\|\mathbf{i})$
ARO	VG-Relation	81.0	**81.1**	**84.4**	**84.7**	**87.6**	**89.1**
	VG-Attribution	71.0	**73.6**	**76.5**	**79.8**	**80.7**	**95.3**
	COCO-Order	91.0	**95.0**	**98.0**	**97.9**	**98.6**	**99.4**
	Flickr30K-Order	86.0	**95.2**	**98.2**	**98.6**	**99.1**	**99.5**
Crepe	Atom-Foils	28.5	**38.8**	**43.0**	**53.3**	**55.4**	**73.2**
	Negate	48.7	44.4	13.6	5.0	**60.8**	**79.6**
	Swap	20.4	**53.3**	**69.5**	**72.7**	**69.7**	**78.1**
VL-CheckList	Relation	70.2	45.1	49.3	51.0	**75.9**	**90.8**

(a) R@1 of $P(\mathbf{t})$ on ARO/Crepe/VL-CheckList

Dataset	VisualGPTScore	PMI
	$P(\mathbf{t}\|\mathbf{i})$	$\frac{P(\mathbf{t}\|\mathbf{i})}{P(\mathbf{t})}$
Winoground	27.0	**33.0** (+6.0)
EqBen	9.6	**19.8** (+10.2)

(b) Text Score of PMI on Winoground/EqBen

Table 2: **Examining the contribution of $P(\mathbf{t})$ and PMI on different I-to-T benchmarks.** Table (a) shows the performance of pure LLMs ($P_{LLM}(\mathbf{t})$) and VisualGPTScore with 3 null (gaussian noise) images ($P_{train}(\mathbf{t}) \approx P_{train}(\mathbf{t}|\mathbf{null})$) with mean of 0.4 and std of 0.25 on ARO/Crepe/VL-CheckList. We **bold** all results that are better than SOTA discriminative approaches (third column in gray color). Surprisingly, "blind" approaches that ignore all visual evidence can outperform SOTA, though still lower than VisualGPTScore (in green color). Note that such approaches can only achieve a 0 text score on Winoground and EqBen, because it must match the correct caption for both images in a test case. Table (b) shows that on such balanced benchmarks, replacing the VisualGPTScore with PMI (by multiplying $\frac{1}{P_{train}(\mathbf{t})}$) significantly improves performance, correcting the bias of VisualGPTScore towards more "common" texts regardless of the image.

Our approach requires much fewer gaussian images (as few as 3) as "null" *images* to compute Eq. 6. We find this method to be less computationally demanding and just as effective as sampling thousands of images from trainset (LAION [31]). Sampling details are in Appendix section 8.

$P(\mathbf{t})$ plays a key role in addressing ARO/Crepe/VL-CheckList (Table 2-a). We posit that some I-to-T benchmarks can be *partially* addressed simply by considering the marginal probabilities of text. Especially, ARO [3], Crepe [4], and VL-CheckList [16] construct hard negative captions by drastically altering the original captions, often resulting in sentences that lack semantic coherence or violate grammatical rules. For example, COCO-Order [3] randomly shuffles all words in a caption, transforming *"two dogs sharing a frisby in their mouth in the snow"* into *"in dogs the in frisby sharing two mouth their a snow"*. Such adversarially constructed negative captions will inherently have low marginal probabilities in any image-text distributions, or even low probability in real-world text distributions. As Table 2 shows, one can achieve impressive performance on these benchmarks by "blindly" modeling $P(\mathbf{t})$ (without considering any visual evidence) through two simple approaches:

1. $P_{LLM}(\mathbf{t})$: passing captions into a pure LLM (such as BART-base [72], FLAN-T5-XL [70], and OPT-2.7B [71]) to compute a text-only GPTScore [73].
2. $P_{train}(\mathbf{t}|\mathbf{null})$: passing both captions and "null" (gaussian noise) images to BLIP to compute a "blind" version of VisualGPTScore.

Replacing VisualGPTScore with PMI boosts performance on Winoground/EqBen(Table 2-b). We demonstrate that PMI, the "debiased" version of VisualGPTScore via multiplying $\frac{1}{P_{train}(\mathbf{t})}$, can significantly boost its performance on balanced benchmarks such as Winoground and EqBen. Intuitively, this debiasing procedure mitigates the tendency of VisualGPTScore to always assign higher scores (regardless of the image) to more "common" texts (like "the person on top of the world") compared to less "common" texts (like "the world on top of the person"), since both texts have the same chance of being positive in Winoground testset.

5 Diagnosis of Visio-Linguistic Benchmarks

In this section, we investigate the major discrepancy among various I-to-T retrieval benchmarks: the **shift in $P(\mathbf{t})$ from train to test data**. As Table 2 suggests, different benchmarks may rely more on certain parts of VisualGPTScore ($P_{train}(\mathbf{t})$ or PMI). Therefore, we repurpose VisualGPTScore as a diagnostic tool by introducing a tunable alpha α that weighs the contribution of each component, allowing us to systematically analyze recent visio-linguistic benchmarks.

Optimal I-to-T retrieval objective. To simplify our analysis, we make the assumption that the conditional $P(\mathbf{i}|\mathbf{t})$ stays the same across training and test distributions. In other words, we assume

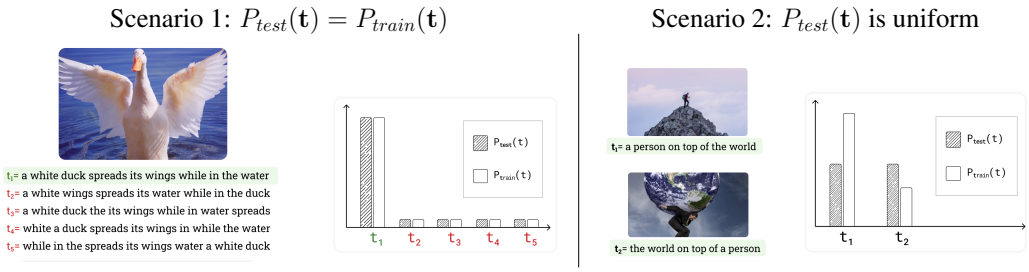


Figure 2: **Two hypothetical scenarios of $P_{test}(\mathbf{t})$ across I-to-T retrieval benchmarks.** Scenario 1 (Flickr30K-Order [3]) happens when $P_{test}(\mathbf{t})$ follows $P_{train}(\mathbf{t})$; for instance, the negative captions are, by construction, less likely to be a real-world caption due to obvious grammatical mistakes. Scenario 2 (Winoground [8]) stands for datasets constructed in a balanced manner; for instance, each caption in Winoground has a matching caption, and thus $P_{test}(\mathbf{t})$ is uniformly distributed. For visualization, we renormalize P_{train} and P_{test} on captions of a single test sample.

that the visual appearance $\mathbf{i}$ of a $\mathbf{t}$ = `"white duck"` will be consistent across the train and test datasets, but the marginal frequency $P(\mathbf{t})$ of the `"white duck"` textual string may change. Given an image $\mathbf{i}$, we can rewrite the optimal retrieval objective $P_{test}(\mathbf{t}|\mathbf{i})$ via Bayes rule:

$$P_{test}(\mathbf{t}|\mathbf{i}) \propto P(\mathbf{i}|\mathbf{t})P_{test}(\mathbf{t}) \tag{7}$$

$$= P(\mathbf{i}|\mathbf{t})\frac{P_{train}(\mathbf{t})}{P_{train}(\mathbf{t})}P_{test}(\mathbf{t}) \tag{8}$$

$$\propto \textbf{VisualGPTScore}(\mathbf{t},\mathbf{i})\frac{P_{test}(\mathbf{t})}{P_{train}(\mathbf{t})} \tag{9}$$

The above derivation suggests that for optimal retrieval, we need access to both $P_{train}(\mathbf{t})$ and $P_{test}(\mathbf{t})$. Because we can approximate the former via Monte Carlo sampling (or "null" version of VisualGPTScore), we now consider two scenarios of $P_{test}(\mathbf{t})$ (Figure 2).

Scenario 1: $P_{test}(\mathbf{t})$ is equal to $P_{train}(\mathbf{t})$. We assume for all candidate captions of an image $\mathbf{i}$:

$$P_{test}(\mathbf{t}) = P_{train}(\mathbf{t}) \quad \Rightarrow \quad \text{Optimal score is } \textbf{VisualGPTScore}(\mathbf{t},\mathbf{i}). \tag{10}$$

Benchmarks such as ARO [3], Crepe [4], VL-CheckList [16] likely fall into this category (Eq. 10), as their negative captions are usually adversarially constructed and have much lower marginal probabilities. In other words, such adversarial negatives rarely occur in natural real-world image-text pairs, which is also why NegCLIP [3] and similar approaches [5, 57, 59, 88] train on hard negative captions generated from original captions or ground-truth scene graphs.

Scenario 2: $P_{test}(\mathbf{t})$ is uniform. In contrast, Winoground-like benchmarks [8, 17] ensures that the marginal probabilities of all captions are uniform across the test-set, since each caption appears twice, once as a positive and once as a negative (Figure 2):

$$P_{test}(\mathbf{t}) \text{ is uniform.} \quad \Rightarrow \quad \text{Optimal score is } \frac{\textbf{VisualGPTScore}(\mathbf{t},\mathbf{i})}{P_{train}(\mathbf{t})} = \text{pmi}_{P_{train}}(\mathbf{t},\mathbf{i}). \tag{11}$$

Tunable α. We introduce a tunable temperature parameter $\alpha \in [0,1]$ to interpolate between the two scenarios, where $\alpha = 0$ implies the first and $\alpha = 1$ implies the second:

$$P_{test}(\mathbf{t}) \propto P_{train}(\mathbf{t})^{1-\alpha} \quad \Rightarrow \text{Optimal score is } \textbf{VisualGPTScore}_{\alpha}(\mathbf{t},\mathbf{i}) = \frac{\textbf{VisualGPTScore}(\mathbf{t},\mathbf{i})}{P_{train}(\mathbf{t})^{\alpha}} \tag{12}$$

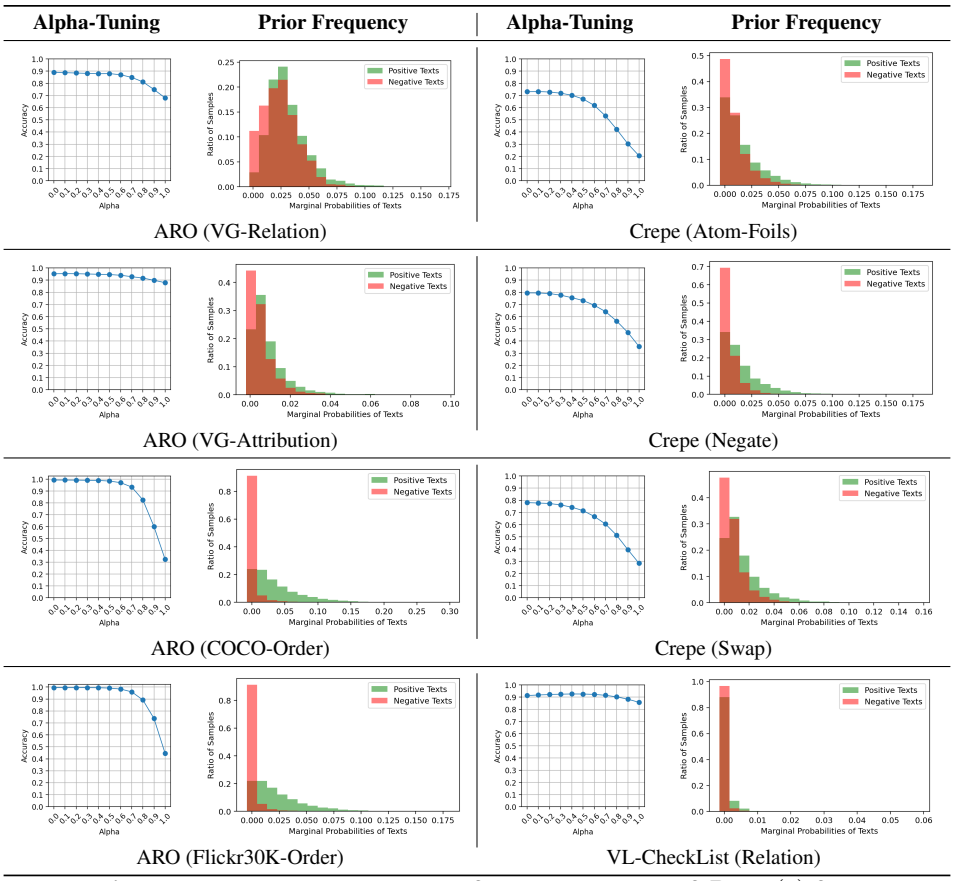


Table 3: **α-tuning on I-to-T benchmarks and frequency charts of $P_{train}(\mathbf{t})$ for both positive and negative captions.** Increasing α from 0 to 1 hurts performance on benchmarks whose marginal probabilities of positive and negative captions differ. Notably, this usually occurs on datasets whose negative captions are adverserially constructed, ie., by shuffling words in the positive caption.

Interestingly, the above can be rewritten using the language of PMI^k [27, 89], a well-known variant of PMI that controls the amount of debiasing [24, 25, 26].

$$\textbf{VisualGPTScore}_{\alpha}(\mathbf{t},\mathbf{i}) = \frac{P_{train}(\mathbf{t},\mathbf{i})}{P_{train}(\mathbf{i})P_{train}(\mathbf{t})^{\alpha}} \tag{13}$$

$$\propto \frac{P_{train}(\mathbf{t},\mathbf{i})^{\frac{1}{\alpha}}}{P_{train}(\mathbf{i})P_{train}(\mathbf{t})} \quad \text{, as } P_{train}(\mathbf{i}) \text{ is constant in I-to-T} \tag{14}$$

$$= \text{pmi}^{k}_{P_{train}}(\mathbf{t},\mathbf{i}), \text{ where } k = \frac{1}{\alpha} \geq 1 \tag{15}$$

Results of different α's on I-to-T retrieval benchmarks. We plot the results of α-tuning in Table 3. We show side-by-side frequency charts of $P_{train}(\mathbf{t})$ for positive and negative captions. As α increases from 0 to 1, we observe that performance decreases the most for datasets like COCO-Order and Flickr-Order, which are constructed with adversarial negative captions whose $P(\mathbf{t})$ are close to 0 and can satisfy the first scenario in a trivial fashion.

Implication for multimodal compositionality evaluation. Despite our approach showing encouraging results on ARO/Crepe/VL-CheckList, Eq. 4 reveals that vanilla VisualGPTScore is biased towards "common" captions. Moreover, as Table 2 demonstrates, solutions that ignore images can still outperform SOTA algorithms trained on carefully-tuned negative samples [5, 59]. This makes it hard to interpret the progress these methods have made in bridging the visio-linguistic compositionality gap. Notably, “blind" solutions that ignore visual evidence achieve a score of 0 on Winoground and EqBen. These two datasets enforce that all captions have the same chance of being positive, reinforcing the lesson learned from previous visio-linguistic benchmarks like balanced VQA-2 [52].

Therefore, we urge future work to construct and evaluate on such balanced benchmarks when testing visio-linguistic compositionality.

6 Additional Experimental Results

We now present results on Winoground, EqBen, and classic retrieval benchmarks (COCO [28] and Flickr30k [46]). We also show that α-tuning using a held-out validation set, as per Equation 12, can regulate the intensity of debiasing, consistently improving performance on these tasks.

Tuning α through cross validation. We show improved results on Winoground and EqBen using a validation set to tune for optimal $\alpha \in [0, 1]$. We sample half of the data as validation set to search for α^*_{val} (using a step size of 0.001) and report the performance on the other half. We repeat this process 10 times to compute mean and std in Table 4-a. We also perform alpha-tuning on classic I-to-T retrieval tasks of COCO and Flickr30k using the provided validation split. Instead of sampling gaussian noise images, we directly approximate $P_{train}(\mathbf{t})$ by averaging the scores of testset images, without incurring computational cost.

T-to-I retrieval on COCO/Flickr30k. For completeness, we also report T-to-I performance on these benchmarks in Table 4-b, where VisualGPTScore again achieves competitive results, presumably because T-to-I retrieval is less affected by learned language bias.

Metric	Benchmark	ITMScore	VisualGPTScore$^{\alpha}_{\text{I2T}}$			
			α=0	α=1	α=α^*_{val}	α^*_{val}
Text Score	Winoground [8]	$\underline{35.5}_{(2.4)}$	$27.5_{(2.3)}$	$33.7_{(2.4)}$	$\mathbf{36.6}_{(2.6)}$	$0.855_{(0.023)}$
	EqBen [17]	$\mathbf{26.1}_{(0.3)}$	$9.6_{(0.2)}$	$19.8_{(0.3)}$	$\underline{19.8}_{(0.3)}$	$0.992_{(0.007)}$
R@1 / R@5	COCO [28]	**71.9 / 90.6**	19.7 / 40.6	46.2 / 73.1	<u>48.0 / 74.2</u>	0.819
	Flickr30k [46]	**88.8 / 98.2**	34.6 / 59.0	58.7 / 88.0	<u>63.6 / 89.2</u>	0.719

(a) α-tuning on val sets for I-to-T retrieval

Metric	Benchmark	ITMScore	VisualGPTScore
R@1 / R@5	COCO [28]	54.8 / 79.0	55.6 / 79.2
	Flickr30k [46]	77.8 / 93.9	76.8 / 93.4

(b) T-to-I retrieval on COCO/Flickr30k

Table 4: **Additional results on retrieval benchmarks.** Table (a) shows that grid searching for optimal alpha on validation sets can significantly improve I-to-T results on test sets of both compositionality and classic benchmarks. Table (b) shows that VisualGPTScore also obtains favorable results on classic T-to-I retrieval tasks, potentially because it does not require debiasing for language bias. In each row, we **bold** the best result and <u>underline</u> the second best result.

Other ablation studies. We summarize the conclusions of selected ablation studies in the Appendix. We include retrieval results using BLIP-2 [7] in Appendix section 11. Interestingly, our findings suggest that simply appending the output tokens of VLMs to *frozen* LLMs (as done in BLIP-2 FLAN-T5 model) does not always enhance its visio-linguistic reasoning capabilities, while incurring more computational costs. In fact, it sometimes reduces the performance, possibly due to the strong language bias introduced by the LLMs. We also compare different Monte Carlo sampling methods in Appendix section 8, showing that our sampling approaches can achieve strong performance with less computation overhead compared to sampling LAION (trainset) images.

7 Discussion and Limitations

Summary. Our study shows the efficacy of a *generative* pre-training score in solving *discriminative* tasks that require *multimodal* compositional reasoning. With the rise of generative pre-training in recent models like GPT-4 [10], we see VisualGPTScore as a reliable starting point for future tasks. We also propose an information-theoretic factorization of VisualGPTScore to highlight language bias in recent visio-linguistic benchmarks and offer a systematic way to debias in common retrieval tasks.

Limitations and future work. Our approach depends on a model pre-trained on noisy web datasets, which may result in inherited biases [90]. We do not explore fine-tuning techniques due to computational constraints, but it is possible to enhance I-to-T retrieval performance using hard negative samples during training, such as with controllable generation [74, 91]. Furthermore, our analysis is based on simplified assumptions. For instance, the model might not accurately represent $P_{train}(\mathbf{t}|\mathbf{i})$, a phenomenon we examine in Appendix section 10. Estimating $P_{train}(\mathbf{t})$ by sampling gaussian noise images is potentially imprecise. Future VLMs could directly model $P_{train}(\mathbf{t})$, or use techniques like coreset selection [92] or dataset distillation [93, 94] to sample more representative images. Our debiasing method may also apply to other generative models for tackling discriminative tasks, such as text-to-image models [53, 95, 96].

Appendix

8 Ablation Studies on α-Tuning

Estimating $P(\mathbf{t})$ via null (gaussian noise) images is more cost-effective. We use Winoground testset to show that sampling gaussian noise images in order to calculate $P(\mathbf{t}|\text{null})$ can be more efficient than Monte Carlo sampling of LAION (trainset) images. As demonstrated in Table 5, a limited number of Gaussian noise images (e.g., 3 or 10) can surpass the results obtained with 1000 LAION images. Moreover, using null images produces less variance in the results.

Sample Size	$P(\mathbf{t}\|\text{null})$		$P(\mathbf{t})$	
	α=α^*_{test}	α^*_{test}	α=α^*_{test}	α^*_{test}
3	$35.95_{(0.5)}$	$0.821_{(0.012)}$	$32.20_{(1.6)}$	$0.706_{(0.150)}$
10	$36.25_{(0.4)}$	$0.827_{(0.016)}$	$33.60_{(0.9)}$	$0.910_{(0.104)}$
100	$36.35_{(0.1)}$	$0.840_{(0.010)}$	$34.70_{(0.6)}$	$0.910_{(0.039)}$
1000	$36.25_{(0.0)}$	$0.850_{(0.000)}$	$35.15_{(0.3)}$	$0.960_{(0.033)}$

Table 5: **Comparing sampling of null images ($P(\mathbf{t}|\mathbf{null})$) and trainset images ($P(\mathbf{t})$).** We show the text score results of α-tuning on Winoground I-to-T retrieval task. We ablate 3/10/100/1000 gaussian noise and LAION samples and report both mean and std using 5 sampling seeds. The optimal $\alpha^* \in [0, 1]$ is grid searched on testset via a step size of 0.001. The gaussian noise images are sampled with a mean calculated from the LAION subset and a fixed std of 0.25.

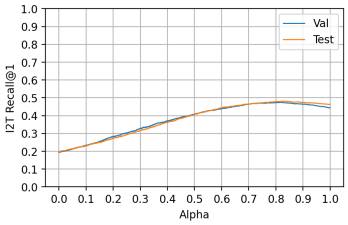
Details of gaussian noise samples. Unless otherwise specified, the gaussian noise images are sampled with a mean of 1.0 and a standard deviation of 0.25. By default, we use 100 images for Winoground, 30 images for EqBen, and 10 images for the rest of the benchmarks. We also fix the sampling seed in our code to ensure reproducibility. We leave more advanced techniques of generating null images to future works.

Alternative approach on COCO/Flickr30k: estimating $P(\mathbf{t})$ using testset images. For classic I-to-I retrieval benchmarks like COCO [28] and Flickr30k [46], we can directly average scores of all candidate images (in the order of thousands) to efficiently approximate $P(\mathbf{t})$ without the need to sample additional gaussian noise images. This approach incurs no computation overhead as we have already pre-computed scores between each candidate image and text. We show in Table 6 that using testset images indeed results in better performance than sampling 3 null gaussian images.

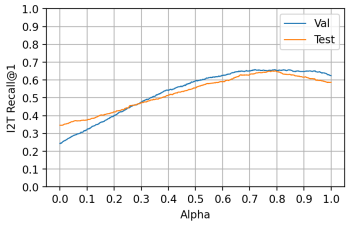
Metric	Benchmark	VisualGPTScore	Sampling Method	$\text{VisualGPTScore}^{\alpha}_{\text{I2T}}$		
				α=1	α=α^*_{val}	α^*_{val}
R@1 / R@5	COCO [28]	19.7 / 40.6	Testset Images	46.2 / 73.1	48.0 / 74.2	0.819
			Null Images	24.4 / 52.6	40.4 / 66.6	0.600
	Flickr30k [46]	34.6 / 59.0	Testset Images	58.7 / 88.0	63.6 / 89.2	0.719
			Null Images	27.8 / 62.2	48.5 / 79.0	0.427

Table 6: **I-to-T retrieval on COCO/Flickr30k using different sampling methods.** Estimating $P(\mathbf{t})$ by averaging the scores of test set images demonstrates superior performance compared to sampling additional gaussian noise images. Although this approach doesn't impose any additional computational overhead, it assumes access to all candidate images of the benchmark.

Tuning α with a validation set. In Table 7, similar performance trends are observed across validation and test splits of COCO and Flickr30k I-to-T retrieval benchmarks using the same $\alpha \in [0, 1]$. Furthermore, α^*_{test} and α^*_{val} are empirically close. As such, our method can function as a reliable training-free debiasing method. Future studies may explore fine-tuning methods to further improve the debiasing performance.



(b) Alpha-tuning on COCO Retrieval



(c) Alpha-tuning on Flickr Retrieval

Table 7: **α-tuning results on both val set and test set for COCO/Flickr30k I-to-T retrieval.** We observe that validation and test performance are strongly correlated while we interpolate $\alpha \in [0, 1]$.

9 Benchmark Details

We include random samples from each benchmark in Table 8.

Dataset	Image	Positive Caption	Negative Caption(s)
VG-Relation		the bus is to the right of the trees	the trees is to the right of the bus
VG-Attribution		the striped zebra and the large tree	the large zebra and the striped tree
COCO-Order		two dogs sharing a frisby in their mouth in the snow	two frisby sharing a mouth in their snow in the dogs in dogs the in frisby sharing two mouth their a snow two dogs sharing in a frisby their mouth in snow the a frisby in the snow two dogs sharing their mouth in
Flickr30K-Order		a white duck spreads its wings while in the water	a white wings spreads its water while in the duck a white duck the its wings while in water spreads white a duck spreads its wings in while the water while in the spreads its wings water a white duck
Crepe-AtomFoils		microwave in a kitchen, and sink in a kitchen.	microwave in a cupboard, and sink in a kitchen microwave in a bar, and sink in a kitchen line in a kitchen, and sink in a kitchen microwave in a kitchen, and shower in a kitchen microwave in a kitchen, and tap in a kitchen
Crepe-Negate		a chair next to a table, with the back of the chair visible.	A chair is not next to a table, with the back of the chair visible A chair next to a table, with the back not of the chair visible A chair next to a table, with the back of the chair visible A chair next to a table, with something of the chair visible. There is no back. There is no chair next to a table, with the back of the chair visible
Crepe-Swap		a car driving on a road with a line next to a tree.	a car driving on a bright green leaves with a line next to a tree a bright green leaves driving on a road with a line next to a tree a car driving on a tree with a line next to a road a car driving on a road with a line next to a white car a car driving on a road with a line next to a street
VL-CheckList Relation (spatial)		person read book	person carry book
VL-CheckList Relation (action)		sign near boy	sign far from book
Winoground		a person on top of the world	the world on top of a person
		the world on top of a person	a person on top of the world
EqBen		The person is touching the dish which is in front of him/her.	The person is holding the dish which is in front of him/her.
		The person is holding the dish which is in front of him/her.	The person is touching the dish which is in front of him/her.

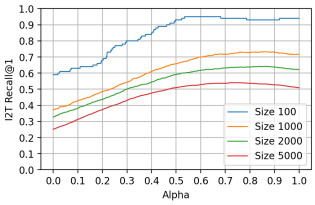
Table 8: **Visualization of benchmarks.** ARO (VG-Relation/VG-Attribution/COCO-Order/Flickr30K-Order), Crepe (AtomFoils/Negate/Swap), and VL-CheckList-Relation (spatial/action) are constructed by generating hard negative captions for an image-text pair. On the other hand, each sample of Winoground and EqBen contains two image-text pairs. This ensures that unimodal solutions, which only look at the text (or the image), fail to perform better than random chance.

10 Is VisualGPTScore a Biased Estimator of $P(\mathbf{t}|\mathbf{i})$?

Retrieval performance on trainset (LAION). This paper is built on the assumption that VisualGPTScore is a reliable estimator of $P_{train}(\mathbf{t}|\mathbf{i})$. However, this simplifying assumption does not completely hold for the BLIP model we examine. As highlighted in the main paper, VisualGPTScore tends to assign higher scores to more common texts. We witness this same phenomenon in Table 9, where we perform image-text retrieval on random subsets of training set (LAION-114M [6, 31]).

Dataset Size	I-to-T Retrieval					T-to-I Retrieval	
	ITMScore	VisualGPTScore$^{\alpha}_{\text{I2T}}$				ITMScore	VisualGPTScore
		α=0	α=1	α=α^*	α^*		
100	**96.0**	59.0	94.0	**95.0**	0.535	95.0	**97.0**
1000	**90.9**	37.1	71.7	85.7	0.733	92.0	**93.1**
2000	**87.2**	32.8	62.3	64.3	0.840	87.8	**89.8**
5000	**79.8**	25.1	50.9	54.1	0.727	81.9	**84.4**

(a) Performance on LAION trainset retrieval



(b) Alpha-tuning on LAION

Table 9: **Retrieval performance on randomly sampled LAION subsets with varied sizes.** Table (a) provides a performance comparison between ITMScore and VisualGPTScore on both I-to-T and T-to-I retrieval tasks on randomly sampled LAION subsets. VisualGPTScore performs competitively on T-to-I retrieval. However, when it comes to I-to-T retrieval (with $\alpha = 0$), performance deteriorates when the number of candidate texts increases. Nevertheless, debiasing VisualGPTScore with $\alpha = 1$, or better still, grid searching optimal $\alpha^* \in [0, 1]$ with a step size of 0.001, can consistently boost the results. This experiment suggests that VisualGPTScore shows a bias towards more common texts even in the (imbalanced) training set. Table (b) presents the performance curves of VisualGPTScore$^{\alpha}$ on LAION subsets with different sample sizes.

Modelling the language bias in VisualGPTScore. As evidenced in Table 9, we believe VisualGPTScore is biased towards more common texts due to modelling error. To consider this error in our analysis, we rewrite the VisualGPTScore as:

$$\textbf{VisualGPTScore}(\mathbf{t}, \mathbf{i}) := \hat{P}_{train}(\mathbf{t}|\mathbf{i}) = P_{train}(\mathbf{t}|\mathbf{i}) \cdot P_{train}(t)^{\beta}, \tag{16}$$

where $\hat{P}$ represents the (biased) model estimate and P represents the true distribution. The model bias towards common texts is encoded by an unknown parameter β.

Monte Carlo estimation using $\hat{P}$. Because our Monte Carlo sampling method relies on $\hat{P}_{train}(\mathbf{t}|\mathbf{i})$, it is also a biased estimator of $P_{train}(\mathbf{t}|\mathbf{i})$:

$$\hat{P}_{train}(\mathbf{t}) := \frac{1}{n}\sum_{k=1}^{n} \hat{P}_{train}(\mathbf{t}|\mathbf{i}_k) = P_{train}(\mathbf{t})^{1+\beta}. \tag{17}$$

Rewriting optimal I-to-T objective with $\hat{P}$. We can rewrite Equation 9 as:

$$P_{test}(\mathbf{t}|\mathbf{i}) \propto P_{train}(\mathbf{t}|\mathbf{i}) \frac{P_{test}(\mathbf{t})}{P_{train}(\mathbf{t})} \tag{18}$$

$$= \hat{P}_{train}(\mathbf{t}|\mathbf{i}) \frac{P_{test}(\mathbf{t})}{P_{train}(\mathbf{t})^{1+\beta}} \tag{19}$$

$$= \textbf{VisualGPTScore}(\mathbf{t}, \mathbf{i}) \frac{P_{test}(\mathbf{t})}{\hat{P}_{train}(\mathbf{t})} \tag{20}$$

α-tuning with $\hat{P}$. Using Equation 20, we can reformulate α-tuning (Equation 12) as follows:

$$P_{test}(\mathbf{t}) \propto P_{train}(\mathbf{t})^{1-\hat{\alpha}} \quad \Rightarrow \text{Optimal score is } \textbf{VisualGPTScore}_{\alpha}(\mathbf{t}, \mathbf{i}) = \frac{\textbf{VisualGPTScore}(\mathbf{t}, \mathbf{i})}{\hat{P}_{train}(\mathbf{t})^{\alpha}} \tag{21}$$

where $\alpha = \frac{\hat{\alpha}+\beta}{1+\beta}$. This implies that even if $P_{train}(\mathbf{t}) = P_{test}(\mathbf{t})$, we still anticipate $\alpha = \frac{\beta}{1+\beta} \neq 0$. This accounts for why the optimal α is not 0 in Table 9. It also provides an explanation for the slight deviation from 0 often observed in the best alpha on ARO/Crepe/VL-CheckList in Table 3.

Implication for vision-language modelling. Our analysis indicates that similar to generative LLMs [24, 25], contemporary image-conditioned language models also experience issues related to imbalanced learning [97]. Potential solutions could be: (a) refined sampling techniques for Monte Carlo estimation of $P(\mathbf{t})$ such as through coreset selection or dataset distillation, and (b) less biased modelling of $P(\mathbf{t}|\mathbf{i})$ such as through hard negatives mining and controllable generation.

11 Experiments with BLIP-2

We provide a brief overview of BLIP-2 [7] and compare its results with BLIP in this section.

BLIP-2 [7] overview. BLIP-2 leverages frozen pre-trained image encoders [98] and large language models [70, 71] to bootstrap vision-language pre-training. It proposes a lightweight Querying Transformer (Q-Former) that is trained in two stages. Similar to BLIP [6], Q-Former is a mixture-of-expert model that can calculate ITC, ITM, and captioning loss given an image-text pair. Additionally, it introduces a set of trainable query tokens, whose outputs serve as *visual soft prompts* prepended as inputs to LLMs. In its first training stage, Q-Former is fine-tuned on the same LAION dataset using the same objectives (ITC+ITM+captioning) as BLIP. In the second stage, the output query tokens from Q-Former are fed into a frozen language model (such as FLAN-T5 [70] or OPT [70]) after a linear projection trained only with captioning loss. BLIP-2 achieves state-of-the-art performance on various vision-language tasks with significantly fewer trainable parameters.

BLIP-2 results. We present retrieval performance of the BLIP-2 model that uses ViT-L as the frozen image encoder. We report results for both the first-stage model (denoted as Q-Former) and the second-stage model which employs FLAN-T5 [70] as the frozen LLM. Table 10 reveals that neither using a powerful visual backbone nor coupling Q-Former with a frozen LLM can improve results on ARO/Crepe/VL-CheckList. We conjecture that (a) the frozen visual backbone is unable to leverage compositional reasoning capabilities of LLMs, and (b) the frozen LLM contribute a strong language bias. Similarly, Table 11 shows that while a stronger visual backbone moderately enhances the ITMScore on Winoground/EqBen, it does not improve VisualGPTScore. As our findings suggest that frozen unimodal models does not advance visio-linguistic capabilities of VLMs, future work should investigate better modelling and training techniques to effectively incorporate the compositional reasoning capabilities of LLMs into VLMs.

Benchmark	Dataset	Random	w. Q-Former			w. Flan-T5
			ITC	ITM	VisualGPT	VisualGPT
ARO [3]	VG-Relation	50.0	46.4	67.2	90.7	89.1
	VG-Attribution	50.0	76.0	88.1	94.3	90.9
	COCO-Order	20.0	28.5	25.2	96.8	99.3
	Flickr30K-Order	20.0	25.3	28.6	97.5	99.7
Crepe [4]	Atom-Foils	16.7	20.8	20.9	74.7	69.7
	Negate	16.7	13.4	14.2	79.1	90.0
	Swap	16.7	13.4	18.0	79.5	79.1
VL-CheckList [16]	Relation	50.0	70.5	72.3	89.9	56.7

Table 10: **BLIP-2 on ARO/Crepe/VL-CheckList.** Using powerful frozen unimodal models (pre-trained image encoders and LLMs) does not lead to improved compositionality on these tasks.

Benchmark	Model	I-To-T Retrieval (Text Score)						T-To-I Retrieval (Image Score)		
		ITC	ITM	VisualGPT$^{\alpha}_{I2T}$				ITC	ITM	VisualGPT
				α=0	α=1	α=α^*	α^*			
Winoground	BLIP	28.0	35.8	27.0	33.0	36.5	0.836	9.0	15.8	21.5
	BLIP2-QFormer	30.0	42.5	24.3	29.3	33.0	0.882	10.5	19.0	20.0
	BLIP2-FlanT5	-	-	25.3	31.5	34.3	0.764	-	-	19.5
EqBen (Val)	BLIP	20.9	26.0	9.6	19.8	19.8	0.982	20.3	20.3	26.1
	BLIP2-QFormer	32.1	36.2	12.2	21.9	22.2	0.969	23.4	28.4	26.6
	BLIP2-FlanT5	-	-	8.5	22.0	22.0	1.000	-	-	20.9

Table 11: **BLIP-2 on Winoground/EqBen.** While BLIP-2's frozen image encoder moderately boosts the performance of ITMScore, it does not benefit VisualGPTScore even after attaching the visual prompts to a frozen LLM. We posit that better modelling and training techniques are required to leverage the compositional reasoning capabilities of LLMs in VLMs.

12 Additional Reports

Computational resources. All experiments use a single NVIDIA GeForce 3090s GPU.

Group scores on Winoground/EqBen using BLIP (Table 12).

Method	Winoground			EqBen		
	Text Score	Image Score	Group Score	Text Score	Image Score	Group Score
ITCScore	28.0	9.0	6.5	20.9	20.3	10.6
ITMScore	35.8	15.8	13.3	26.0	20.32	12.6
VisualGPTScore$^{\alpha}$	36.5	21.5	16.8	20.4	26.1	11.68

Table 12: Performance comparison of BLIP's ITCScore, ITMScore, and α-tuned VisualGPTScore$^{\alpha^*}$ on Winoground (all) and EqBen (val).

Fine-grained tags on Winoground (Table 13).

Dataset	Size	Method	Text Score	Image Score	Group Score
NoTag	171	ITCScore	32.6	11.6	8.1
		ITMScore	41.9	21.5	19.2
		VisualGPTScore$^{\alpha^*}$	43.0	28.5	23.8
NonCompositional	30	ITCScore	43.3	16.7	16.7
		ITMScore	50.0	23.3	16.7
		VisualGPTScore$^{\alpha^*}$	43.3	33.3	26.7
AmbiguouslyCorrect	46	ITCScore	32.6	8.7	6.5
		ITMScore	28.3	6.5	2.2
		VisualGPTScore$^{\alpha^*}$	26.1	19.6	8.7
VisuallyDifficult	38	ITCScore	29.0	7.9	7.9
		ITMScore	26.3	10.5	7.9
		VisualGPTScore$^{\alpha^*}$	31.6	13.2	7.9
UnusualImage	56	ITCScore	32.5	8.9	8.9
		ITMScore	21.4	10.7	7.1
		VisualGPTScore$^{\alpha^*}$	30.4	10.7	8.9
UnusualText	50	ITCScore	20.0	8.0	6.0
		ITMScore	38.0	12.0	12.0
		VisualGPTScore$^{\alpha^*}$	30.0	18.0	12.0
ComplexReasoning	78	ITCScore	16.7	2.6	1.3
		ITMScore	21.8	5.1	2.6
		VisualGPTScore$^{\alpha^*}$	21.8	10.3	6.4

Table 13: BLIP performance on Winoground subtags [18]. We report the number of test instances for each subtag and their respective text score, image score, group score.

References

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

[2] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[3] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

[4] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? *arXiv preprint arXiv:2212.07796*, 2022.

[5] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs. *arXiv preprint arXiv:2305.06343*, 2023.

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

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[8] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 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, 2022.

[9] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[10] OpenAI. Gpt-4 technical report. 2023.

[11] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[12] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[13] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. *arXiv preprint arXiv:2211.11559*, 2022.

[14] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023.

[15] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

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

[17] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. *arXiv preprint arXiv:2303.14465*, 2023.

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

[19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

[20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[21] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

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

[23] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40, 2009.

[24] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL `https://aclanthology.org/N16-1014`.

[25] Jiwei Li and Dan Jurafsky. Mutual information and diverse decoding improve neural machine translation, 2016.

[26] Zeyu Wang, Berthy Feng, Karthik Narasimhan, and Olga Russakovsky. Towards unique and informative captioning of images. In *European Conference on Computer Vision (ECCV)*, 2020.

[27] Béatrice Daille. *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. PhD thesis, Ph. D. thesis, Université Paris 7, 1994.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 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, 2014.

[29] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[30] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. *arXiv preprint arXiv:2212.00794*, 2022.

[31] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[32] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

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

[34] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35: 6704–6719, 2022.

[35] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[36] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[37] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.

[38] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[39] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

[40] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[41] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

[42] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramana. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. *arXiv preprint arXiv:2301.06267*, 2023.

[43] Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*, 2022.

[44] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

[45] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[46] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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, 2014.

[47] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

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

[49] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

[50] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

[51] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[52] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[53] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

[54] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 423–439. Springer, 2022.

[55] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[56] Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gül Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. Going beyond nouns with vision & language models using synthetic data. *arXiv preprint arXiv:2303.17590*, 2023.

[57] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, et al. Teaching structured vision&language concepts to vision&language models. *arXiv preprint arXiv:2211.11733*, 2022.

[58] Madeline Chantry Schiappa, Michael Cogswell, Ajay Divakaran, and Yogesh Singh Rawat. Probing conceptual understanding of large visual-language models. *arXiv preprint arXiv:2304.03659*, 2023.

[59] Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Tangjie Lv, Zhipeng Hu, and Wen Zhang. Structure-clip: Enhance multi-modal language representations with structure knowledge. *arXiv preprint arXiv:2305.06152*, 2023.

[60] Rohan Pandey, Rulin Shao, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Cross-modal attention congruence regularization for vision-language relation alignment. 2022.

[61] Rohan Pandey, Rulin Shao, Paul Pu Liang, and Louis-Philippe Morency. Does structural attention improve compositional representations in vision-language models?

[62] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

[63] Ajinkya Tejankar, Maziar Sanjabi, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsiavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision. *arXiv preprint arXiv:2112.13884*, 2021.

[64] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[66] Lorenzo Bertolini, Julie Weeds, and David Weir. Testing large language models on compositionality and inference with phrase-level adjective-noun entailment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4084–4100, 2022.

[67] Jack Hessel and Alexandra Schofield. How effective is bert without word ordering? implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, 2021.

[68] Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. When classifying grammatical role, bert doesn't care about word order... except when it matters. *arXiv preprint arXiv:2203.06204*, 2022.

[69] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*, 2021.

[70] Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022.

[71] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[72] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf`.

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

[74] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

[75] Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C Lawrence Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.

[76] Micah Hodosh and Julia Hockenmaier. Focused evaluation for image description with binary forced-choice tasks. In *Proceedings of the 5th Workshop on Vision and Language*, pages 19–28, 2016.

[77] Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Revisiting visual question answering baselines. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 727–739. Springer, 2016.

[78] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5014–5022, 2016.

[79] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020.

[80] Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of gqa. *arXiv preprint arXiv:2103.09591*, 2021.

[81] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1574–1583, 2021.

[82] Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[83] Ting Yao, Tao Mei, and Chong-Wah Ngo. Co-reranking by mutual reinforcement for image search. In *Proceedings of the ACM international conference on image and video retrieval*, pages 34–41, 2010.

[84] Christian Andreas Henning and Ralph Ewerth. Estimating the information gap between textual and visual representations. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 14–22, 2017.

[85] Aman Shrivastava, Ramprasaath R Selvaraju, Nikhil Naik, and Vicente Ordonez. Clip-lite: information efficient visual representation learning from textual annotations. *arXiv preprint arXiv:2112.07133*, 2021.

[86] Alexander Shapiro. Monte carlo sampling methods. *Handbooks in operations research and management science*, 10:353–425, 2003.

[87] Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, 2021.

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

[89] François Role and Mohamed Nadif. Handling the impact of low frequency events on co-occurrence based measures of word similarity. In *Proceedings of the international conference on Knowledge Discovery and Information Retrieval (KDIR-2011). Scitepress*, pages 218–223, 2011.

[90] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[91] Tianshu Wang, Faisal Ladhak, Esin Durmus, and He He. Improving faithfulness by augmenting negative summaries from fake documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11913–11921, 2022.

[92] Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *Database and Expert Systems Applications: 33rd International Conference, DEXA 2022, Vienna, Austria, August 22–24, 2022, Proceedings, Part I*, pages 181–195. Springer, 2022.

[93] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[94] Ilia Sucholutsky and Matthias Schonlau. Soft-label dataset distillation and text dataset distillation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[95] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[96] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[97] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

[98] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.