Revisiting the Role of Language Priors in Vision-Language Models

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

Abstract

Vision-language models (VLMs) are impactful in part because they can be applied to a variety of visual understanding tasks in a zero-shot fashion, without any fine-tuning. We study *generative VLMs* that are trained for next-word generation given an image. We explore their zero-shot performance on the illustrative task of image-text retrieval across nine popular vision-language benchmarks. Our first observation is that they can be repurposed for discriminative tasks (such as image-text retrieval) by simply computing the match score of generating a particular text string given an image. We call this probabilistic score the *Visual Generative Pre-Training Score* (VisualGPTScore). While the VisualGPTScore produces near-perfect accuracy on some retrieval benchmarks, it yields poor accuracy on others. We analyze this behavior through a probabilistic lens, pointing out that some benchmarks inadvertently capture unnatural language distributions by creating adversarial but unlikely text captions. In fact, we demonstrate that even a "blind" language model that ignores any image evidence can sometimes outperform all prior art, reminiscent of similar challenges faced by the visual-question answering (VQA) community many years ago. We derive a probabilistic post-processing scheme that controls for the amount of linguistic bias in generative VLMs at test time without having to retrain or fine-tune the model. We show that the VisualGPTScore, when appropriately debiased, is a strong zero-shot baseline for vision-language understanding, oftentimes producing state-of-the-art accuracy.

[*]Equal contribution [1]CMU [2]Meta. Correspondence to: Zhiqiu Lin <zhiqiul@andrew.cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Vision-language models (VLMs) trained on web-scale datasets will likely serve as the foundation for next-generation visual understanding systems. One reason for their widespread adoption is their ability to be used in an "off-the-shelf" (OTS) or zero-shot manner without fine-tuning for specific target applications. In this study, we explore their OTS use on the task of image-text retrieval (e.g., given an image, predict the correct caption out of K options) across a suite of nine popular benchmarks.

Challenges. While the performance of foundational VLMs is impressive, many open challenges remain. Recent analyses (Kamath et al., 2023; Yuksekgonul et al., 2022) point out that leading VLMs such as CLIP (Radford et al., 2021) may often degrade to "bag-of-words" that confuse captions such as "`the horse is eating the grass`" and "`the grass is eating the horse`". This makes it difficult to use VLMs to capture *compositions* of objects, attributes, and their relations. But somewhat interestingly, large-scale language models (LLMs) trained for autoregressive next-token prediction (Brown et al., 2020) seem to be able to discern such distinctions, which we investigate below. A related but under-appreciated difficulty is that of *benchmarking* the performance of visio-linguistic reasoning. Perhaps the most well-known example in the community is that of the influential VQA benchmarks (Antol et al., 2015), which could be largely solved by exploiting linguistic biases in the dataset – concretely, questions about images could often be answered by "blind" language-only models that did not look at the image (Goyal et al., 2017). Notably, we find that such blind algorithms still excel on many contemporary image-text retrieval benchmarks where VLMs may struggle.

Generative models for discriminative tasks. We tackle the above challenges by revisiting the role of language priors through a probabilistic lens. To allow for a probabilistic treatment, we focus on generative VLMs that take an image as input and stochastically generate text via next-token prediction (Li et al., 2022; 2023). We first demonstrate that such models can be easily repurposed for discriminative tasks (such as retrieval) by setting the match score for an image-text pair to be the probability that the VLM would generate that text from the given image, or $P(\text{text}|\text{image})$. We call this probability score the Visual Generative Pre-Training

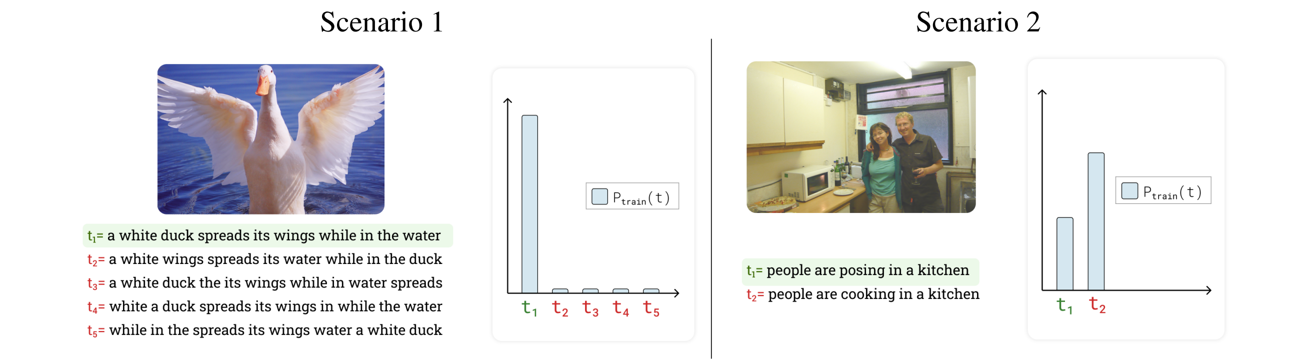


Figure 1. **Two train-test shifts encountered in image-to-text retrieval tasks.** Scenario 1 (**left**) constructs negative captions by shuffling words in the true caption (as in ARO-Flickr (Yuksekgonul et al., 2022)), but this produces implausible text such as "`white a duck spreads its wings in while the water`". Here, exploiting the language bias of the training set will help since it will downweight the match score for such implausible negative captions. In fact, we show that a blind language-only model can easily identify the correct caption. Scenario 2 (**right**) constructs negative captions that are curated to be plausible (as in SugarCrepe (Hsieh et al., 2023)). Here, the language bias of the training set may hurt, since it will prefer to match common captions that score well under the language prior; i.e., the incorrect caption of "`people are cooking in a kitchen`" is slightly more likely than the true caption of "`people are posing in a kitchen`" under the language prior, and so removing the language bias improves performance. We present simple training-free approaches for removing such language biases, and show this significantly improves performance on challenging benchmarks that fall into Scenario 2.

Score, or VisualGPTScore. Computing the VisualGPTScore is even more efficient than next-token generation since given an image, all tokens from a candidate text string can be evaluated in parallel. Though conceptually straightforward, such an approach is not a common baseline. In fact, the generative VLMs (Li et al., 2022) that we analyze train *separate* discriminative heads for matching/classifying image-text pairs, but we find that their language generation head itself produces better scores for matching (since it appears to better capture compositions). Indeed, the OTS VisualGPTScore performs surprisingly well on many benchmarks, even producing near-perfect accuracy on ARO (Yuksekgonul et al., 2022). But it still struggles on other benchmarks such as Winoground (Thrush et al., 2022). We analyze this below.

The role of language priors. We analyze the discrepancy in performance across benchmarks from a probabilistic perspective. Our key insight is that many benchmark biases can be formalized as mismatching distributions over text between foundational pre-training data and benchmark test data – $P_{train}(\text{text})$ versus $P_{test}(\text{text})$. We use a first-principles analysis to account for distribution shift by simply reweighting the VisualGPTScore with the Bayes factor $P_{test}(\text{text})/P_{train}(\text{text})$, a process we call *debiasing*. To compute the Bayes reweighting factor, we need access to both the train and test language prior. We compute $P_{train}(\text{text})$ from an OTS VLM by drawing Monte-Carlo samples of $P_{train}(\text{text}|\text{image})$ from the trainset or Gaussian noise images. Because $P_{test}(\text{text})$ may require access to the test set, we explore practical variants that assume P_{test} is (a) identical to $P_{train}(\text{text})$, (b) uninformative/uniform, or (c) learnable from a small held-out valset. Our analysis helps explain the strong performance of the VisualGPTScore on certain benchmarks and its poor performance on others. Moreover, our analysis offers simple strategies to improve performance through debiasing without requiring any retraining. We conclude by showing a theoretical connection between debiasing and mutual information, which can be seen as a method for removing the effect of marginal priors when computing joint probability scores.

Empirical analysis. We conduct a thorough empirical evaluation of the OTS VisualGPTScore (and its debiased variants) for open-sourced image-conditioned language models (Li et al., 2022; 2023; Liu et al., 2023) across nine popular vision-language benchmarks. We first point out that the VisualGPTScore by itself produces SOTA accuracy on certain benchmarks like ARO (Yuksekgonul et al., 2022) where their inherent language biases help remove incorrect captions that are also unnatural (such as "`a white duck the its wings while in water`" as shown in Fig. 1). In fact, we show that blind baselines also do quite well on these benchmarks, since language-only models can easily identify such implausible captions. However, such language biases do not work well on benchmarks where incorrect captions are carefully constructed to be realistic. Here, VisualGPTScore should be debiased so as not to naively prefer more common captions that score well under its language prior. Debiasing consistently improves performance on benchmarks such as Flickr30K (Young et al., 2014) and Winoground (Thrush et al., 2022). Interestingly, we find that debiasing can also improve accuracy on the *train* set used to learn the generative VLMs, indicating that such models learn biased estimates of the true conditional distribution $P_{train}(text|image)$. We describe this further in our Appendix A. Finally, our approach sets a new state-of-the-art on image-text alignment (Thrush et al., 2022; Wang et al., 2023), showing potential to replace the widely-used CLIPScore (Hessel et al., 2021) in text-to-image evaluation. In fact, our latest work (Lin et al., 2024; Li et al., 2024) extends VisualGPTScore to more powerful vision-language models trained on visual-question-answering (VQA) data, achieving further improvements.

Contributions:

- We introduce VisualGPTScore to repurpose generative VLMs for discriminative (image-text retrieval) tasks.
- Our analysis shows that language priors play a key role in addressing train-test distribution shifts, leading to a zero-shot debiasing technique that significantly improves performance on challenging benchmarks.
- We find that many recent benchmarks for foundational VLMs like ARO can be largely solved by blind solutions (e.g., P(text)) that ignore images. This underscores the need to reevaluate language priors in vision-language benchmarks.

2. Related works

Vision-language models. State-of-the-art VLMs like CLIP (Radford et al., 2021) are pre-trained on web-scale image-text datasets (Schuhmann et al., 2022) using discriminative objectives like image-text contrastive (ITC) (Radford et al., 2021) and image-text matching (ITM) (Li et al., 2021) loss, typically formulated as $P(match|image, text)$. These pre-trained models exhibit robust zero-shot and few-shot (Lin et al., 2023; Wortsman et al., 2022) performance on traditional discriminative tasks (Deng et al., 2009; Lin et al., 2014), often on par with fully-supervised models. More recently, image-conditioned language models like Flamingo (Alayrac et al., 2022) and BLIP (Li et al., 2022; 2023) incorporate generative objectives primarily for downstream tasks such as captioning (Agrawal et al., 2019) and VQA (Goyal et al., 2017).

Visio-linguistic compositionality. Benchmarks like ARO (Yuksekgonul et al., 2022), Crepe (Ma et al., 2022), Winoground (Thrush et al., 2022), EqBen (Wang et al., 2023), VL-CheckList (Zhao et al., 2022), and SugarCrepe (Hsieh et al., 2023) show that discriminative scores of VLMs, such as ITCScore and ITMScore, fail on their image-text retrieval tasks that assess compositional reasoning. Concurrently, advances on these tasks often involve fine-tuning discriminative VLMs with more data. One of the most popular approaches, NegCLIP (Yuksekgonul et al., 2022), augments CLIP using programmatically generated negatives from original texts. Extending this, subsequent studies propose more expensive and heavily-engineered solutions. SyViC (Cascante-Bonilla et al., 2023) fine-tunes VLMs on million-scale synthetic images to augment spatial, attributive, and relation understanding. SGVL (Herzig et al., 2023) and Structure-CLIP (Huang et al., 2023) sample negatives using costly scene graph annotations. MosaiCLIP (Singh et al., 2023) and SVLC (Doveh et al., 2022) use linguistic tools such as scene graph parsers and LLMs to design better negative captions. The most recent DAC (Doveh et al., 2023) leverages a combination of foundation models including BLIP2, ChatGPT, and SAM to rewrite and augment image captions. In contrast, we demonstrate that OTS generative scores can outperform these costly approaches on compositionality benchmarks.

Generative pre-training and scoring. Vision models trained with *discriminative* objectives often lack incentives to learn structure information (Brendel & Bethge, 2019; Tejankar et al., 2021). Similarly, early LLMs trained with *discriminative* approaches, such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), have also been criticized as bag-of-words models insensitive to word order (Bertolini et al., 2022; Hessel & Schofield, 2021; Papadimitriou et al., 2022; Sinha et al., 2021). Conversely, generative pre-trained LLMs (Radford et al., 2019) demonstrate exceptional compositional understanding while pre-trained solely with a next-token prediction (Bengio et al., 2003) loss. Furthermore, generative scores of LLMs (OpenAI, 2023; Chung et al., 2022; Zhang et al., 2022) have flexible usage in downstream tasks, such as text evaluation (Yuan et al., 2021; Fu et al., 2023) and reranking (Keskar et al., 2019). While generative scores from VLMs have been previously used for discriminative tasks (Tschannen et al., 2023; Miech et al., 2021), our work uniquely investigates the critical role of language priors and introduces the first debiasing solution that improves retrieval without the need for retraining.

3. The role of language priors

In this section, we present a simple probabilistic treatment for analyzing the role of language priors in image-conditioned language models (or generative VLMs). Motivated by their strong but inconsistent performance across a variety of image-text retrieval benchmarks, we analyze their behavior when there exists a mismatch between training and test distributions, deriving simple schemes for addressing the mismatch with reweighting. We emphasize that the training data that we refer to is the foundational pre-training dataset, while the test data is always a given benchmark dataset; in fact, most benchmarks we analyze do not even provide a trainset. We conclude by exposing a connection to related work on mutual information.

Computing $P(t|i)$. To begin our probabilistic treatment, we first show that image-conditioned language models (that

(a) $P_{train}(t|i)$ through generative VLMs

(b) $P_{train}(t)$ via Monte Carlo sampling

Figure 2. Estimating $P_{train}(t|i)$ and $P_{train}(t)$ from generative VLMs. Figure (a) shows how image-conditioned language models such as Li et al. (2022) that generate text based on an image can be repurposed for computing $P_{train}(t|i)$, which is factorized as a product of $\prod_{k=1}^{m} P(t_k|t_{<k}; i)$ for a sequence of m tokens. These terms can be efficiently computed in parallel, unlike sequential token-by-token prediction for text generation. Figure (b) shows two approaches for Monte Carlo sampling of $P_{train}(t)$. While the straightforward approach is to sample trainset images, we find that using "null" (Gaussian noise) images can also achieve robust estimates.

probabilistically generate text based on an image) can be repurposed for computing a score between a given image i and text caption t. The likelihood of a text sequence $t = \{t_1, t_2, \ldots, t_m\}$ conditioned on image i is naturally factorized as an autoregressive product (Bengio et al., 2003):

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

Image-conditioned language models return back m softmax distributions corresponding to the m terms in the above expression. Text generation requires sequential token-by-token prediction, since token t_k must be generated before it can be used as an input to generate the softmax distribution over token t_{k+1}. Interestingly, given an image i and a text sequence t, the above probability can be computed in parallel because the entire sequence of tokens $\{t_k\}$ is already available as input. Figure 2-a shows a visual illustration.

Train-test shifts. Given the image-conditioned model of $P(t|i)$ above, we now analyze its behavior when applied to test data distributions that differ from the trainset, denoted as P_{test} versus P_{train}. Recall that any joint distribution over images and text can be factored into a product over a language prior and an image likelihood $P(t, i) = P(t)P(i|t)$. Our analysis makes the strong assumption that the image likelihood $P(i|t)$ is identical across the train and test data, but the language prior $P(t)$ may differ. Intuitively, this assumes that the visual appearance of entities (such as a "white duck") remains consistent across the training and test data, but the frequency of those entities (as manifested in the set of captions $P(t)$) may vary. We can now derive $P_{test}(t|i)$ via Bayes rule:

$$P_{test}(t|i) \propto P(i|t)P_{test}(t) \tag{2}$$

$$= P(i|t)\frac{P_{train}(t)}{P_{train}(t)}P_{test}(t) \tag{3}$$

$$\propto P_{train}(t|i)\frac{P_{test}(t)}{P_{train}(t)} \tag{4}$$

The above shows that the generative pre-training score $P_{train}(t|i)$ need simply be weighted by the ratio of the language priors in the testset versus trainset. Intuitively, if a particular text caption appears more often in the testset than the trainset, one should increase the score reported by the generative model. However, one often does not have access to the text distribution on the testset. For example, real-world deployments and benchmark protocols may not reveal this. In such cases, one can make two practical assumptions; either the language distribution on test is identical to train, or it is uninformative/uniform (see Figure 1):

Scenario 1:

$$P_{test}(t) = P_{train}(t) \Rightarrow \text{Optimal score is } P_{train}(t|i) \tag{5}$$

Scenario 2:

$$P_{test}(t) \text{ is uniform.} \Rightarrow \text{Optimal score is } \frac{P_{train}(t|i)}{P_{train}(t)} \tag{6}$$

Tunable . In reality, a testset might be a mix of both scenarios. To model this, we consider a soft combination where the language prior on the testset is assumed to be a flattened version of the language prior on the trainset, for some temperature parameter $\alpha \in [0, 1]$:

$$P_{test}(t) \propto P_{train}(t)^{1-\alpha} \Rightarrow \text{Optimal score is } \frac{P_{train}(t|i)}{P_{train}(t)^{\alpha}} \tag{7}$$

By setting α to 0 or 1, one can obtain the two scenarios described above. Some deployments (or benchmarks) may benefit from tuning α on a held-out valset, if available.

Implications for retrieval benchmarks. We speculate some benchmarks like ARO-Flickr (Yuksekgonul et al., 2022) are close to Scenario 1 because they include negative captions that are *implausible*, such as "*a white duck the its wings while in water spreads*". Such captions will have a low score under the language prior $P_{train}(t)$ and so reporting the raw generative score $P_{train}(t|i)$ (that keeps its language prior or bias) will improve accuracy. In fact, we show that applying a *blind* language model (that ignores all image evidence) can itself often identify the correct caption. On the other hand, for test datasets with more *realistic* negative captions (Scenario 2), it may be useful to remove the language bias of the trainset, since that will prefer to match to common captions (even if they do not necessarily agree with the input image). This appears to be the case for SugarCrepe (Hsieh et al., 2023), which uses LLMs like ChatGPT to ensure that the negative captions are realistic.

An information-theoretic derivation of α-debiasing. Our approach to debiasing is reminiscent of mutual information, which can also be seen as a method for removing the effect of marginal priors when computing joint probability scores (Daille, 1994). In fact, α-debiasing (Eq. 7) is equivalent to a form of pointwise mutual information (PMI) known as PMI^k (Role & Nadif, 2011). PMI is a classic information-theoretic measure that quantifies the association between two variables (Yao et al., 2010; Henning & Ewerth, 2017; Shrivastava et al., 2021). In the context of image-text retrieval, PMI measures how much more or less likely the image-text pair co-occurs than if the two were independent:

$$pmi_P(t, i) = \frac{P(t, i)}{P(t)P(i)} = \frac{P(i|t)}{P(i)} = \frac{P(t|i)}{P(t)} \tag{8}$$

However, directly applying PMI (Eq. 8) for retrieval tends to overly inflate scores for rarer texts (Role & Nadif, 2011). Consequently, the PMI^k approach was introduced to control the strength of debiasing. Below, we rewrite the Eq. 7 using the language of PMI^k:

$$\frac{P_{train}(t|i)}{P_{train}(t)^{\alpha}} = \frac{P_{train}(t, i)}{P_{train}(i)P_{train}(t)^{\alpha}} \tag{9}$$

$$\propto \frac{P_{train}(t, i)^{\frac{1}{\alpha}}}{P_{train}(i)P_{train}(t)}, \text{ as } P_{train}(i) \text{ is constant in I-to-T} \tag{10}$$

$$= pmi^{k}_{P_{train}}(t, i), \text{ where } k = \frac{1}{\alpha} \geq 1 \tag{11}$$

Eq. 11 shows that our α-debiasing is equivalent to PMI^k for $k = \frac{1}{\alpha}$. PMI^k is widely adopted in information retrieval tasks (Li et al., 2016; Li & Jurafsky, 2016; Wang et al., 2020). This alternative derivation could explain why α-debiasing remains effective across various testing benchmarks (as we show next), even when our previous probabilistic assumptions may not hold.

4. Experimental results on I-to-T retrieval

In this section, we verify our hypothesis on I-to-T retrieval benchmarks using state-of-the-art multimodal generative VLMs. In particular, we adopt image-conditioned language models such as BLIP (Li et al., 2022) as the learned estimator of $P_{train}(t|i)$. Then, we discuss how we perform Monte Carlo estimation of $P_{train}(t)$, including a novel efficient sampling method based on "content-free" Gaussian noise images. Finally, we show the state-of-the-art results of our generative approach on recent I-to-T retrieval benchmarks.

Preliminaries. We leverage OTS image-conditioned language models to estimate $P_{train}(t)$. Most of our diagnostic experiments focus on the open-sourced BLIP (Li et al., 2022; 2023) model, trained on public image-text corpora using discriminative (ITC and ITM) and generative (captioning) objectives. Discriminative objectives typically model $P(\text{match}|t, i)$. For example, ITCScore calculates cosine similarity scores between image and text features using a dual-encoder; ITMScore jointly embeds image-text pairs via a fusion-encoder and returns softmax scores from a binary classifier. We term the generative score as **Visual Generative Pre-Training Score** (**VisualGPTScore**). While BLIP is pre-trained using all three objectives, this generative score has not been applied to discriminative tasks before our work. Lastly, our approach can be extended to other generative VLMs. We also present some additional results using LLaVA-1.5 (Liu et al., 2023), a recent state-of-the-art VLM (Liu et al., 2023) that produces SOTA accuracy on several challenging benchmarks.

Implementing VisualGPTScore. Our method calculates an average of the log-likelihoods of t_k at each token position k and applies an exponent to cancel the log:

$$\text{VisualGPTScore}(t, i) := e^{\frac{1}{m}\sum_{k=1}^{m} \log(P(t_k|t_{<k}, i))} \tag{12}$$

To condition on an input image, BLIP uses a multimodal casual self-attention mask (Li et al., 2022) in its image-grounded text decoder, i.e., each text token attends to all its preceding vision and text tokens. We emphasize that VisualGPTScore has the same computational cost as ITMScore, which uses the same underlying transformer but with a bi-directional self-attention mask to encode an image-text pair. We address potential biases of this estimator in Appendix A.

Estimating $P_{train}(t)$ using Monte Carlo sampling (oracle approach). Given $P_{train}(t|i)$, we can estimate $P_{train}(t)$ via classic Monte Carlo sampling (Shapiro, 2003), by drawing n images from the train distribution, such as LAION114M (Schuhmann et al., 2021) for BLIP:

$$P_{train}(t) \approx \frac{1}{n}\sum_{k=1}^{n} P_{train}(t|i_k) \tag{13}$$

Reducing sampling cost with Gaussian noise images (our approach). The above Equation 13 requires many trainset samples to achieve robust estimates. To address this, we draw inspiration from (Zhao et al., 2021), which uses a content-free text prompt "N/A" to calibrate the probability of a text from LLMs, i.e., $P(t|\text{"N/A"})$. To apply this to our generative VLMs, we choose to sample "null" inputs as Gaussian noise images. It turns out Eq. 13 can be estimated using as few as 1-3 Gaussian noise images (with a mean and standard deviation calculated from trainset distribution). We provide a visual illustration of this method in Figure 2-b. We find this method to be less computationally demanding and just as effective as sampling thousands of images from trainset. We ablate sampling procedures in Appendix B and show that our method generalizes across BLIP and BLIP-2 architectures in Appendix C.

Benchmarks and evaluation protocols. We comprehensively report on four recent I-to-T retrieval benchmarks that assess compositionality, including ARO (Yuksekgonul et al., 2022), Crepe (Ma et al., 2022), SugarCrepe (Hsieh et al., 2023), and VL-CheckList (Zhao et al., 2022). In these datasets, each image has a single positive caption and multiple negative captions. ARO (Yuksekgonul et al., 2022) has four datasets: VG-Relation, VG-Attribution, COCO-Order, and Flickr30k-Order. SugarCrepe (Hsieh et al., 2023) has three datasets: Replace, Swap, and Add. For Crepe (Ma et al., 2022), we use the entire productivity set and report on three datasets: Atom, Negate, and Swap. VL-CheckList (Zhao et al., 2022) has three datasets: Object, Attribute, and Relation. Appendix E visualizes these datasets.

SOTA performance on all four benchmarks. In Table 1, we show that our OTS generative approaches, based on the BLIP model pre-trained on LAION-114M with ViT-L image encoder, achieves state-of-the-art results on all benchmarks. We outperform the best discriminative VLMs, including LAION5B-CLIP, and consistently surpass other heavily-engineered solutions, including NegCLIP, SyViC, MosaiCLIP, DAC, SVLC, SGVL, Structure-CLIP, all of which fine-tune CLIP on much more data. Details on how we report the baseline results can be found in Appendix D. For reference, we also include results of text-only Vera and Grammar from Hsieh et al. (2023). To show that even the most recent SugarCrepe is not exempt from language biases, we run two more text-only methods:

1. $P_{LLM}(t)$: passing captions into a pure LLM, such as BART-base (Yuan et al., 2021), FLAN-T5-XL (Chung et al., 2022), and OPT-2.7B (Zhang et al., 2022), to compute a text-only GPTScore (Fu et al., 2023).
2. $P_{train}(t)$: passing both captions and Gaussian noise images to BLIP as shown in Figure 2.

Discussion on α-debiasing. Table 2 shows that debiasing affects benchmarks differently depending on their construction; benchmarks with unrealistic negative captions (such as ARO-Flickr) benefit from a language prior that can identify such negative examples. Here, debiasing with large α hurts performance. On the other hand, benchmarks with realistic negative captions (such as SugarCrepe) tend to benefit from debiasing because it reduces the influence of the language prior. Our findings are reminiscent of the lessons from the VQA benchmark (Goyal et al., 2017), known to be solvable by "blind" algorithms that do not look at the image, e.g., questions such as "Is there a clock" have an answer of "Yes" 98% of the time. However, we also find that some recent benchmarks such as Winoground (Thrush et al., 2022) and EqBen (Wang et al., 2023) introduce strict evaluation protocols that aggressively penalize such blind algorithms. We discuss these challenging Scenario 2 benchmarks (with far lower SOTA accuracy) in the next section.

5. Additional Challenging Benchmarks

In this section, we apply our OTS generative approaches to five more Scenario 2 benchmarks: (a) Winoground (Thrush et al., 2022) and EqBen (Wang et al., 2023) for image-text alignment; (b) COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014) for large-scale retrieval; (c) ImageNet (Deng et al., 2009) for zero-shot image classification. While naively applying OTS VisualGPTScore leads to inferior performance on these benchmarks, our training-free α-debiasing consistently improves its performance even with a fixed α=1, without accessing the held-out valset (Table 3-a). We also derive the optimal text-to-image (T-to-I) retrieval objective and show that OTS generative scores can achieve robust T-to-I performance (Table 3-b). Lastly, we apply VisualGPTScore and its α-debiased version to a state-of-the-art VLM, LLaVA-1.5 (Liu et al., 2023), and outperform widely-used methods such as CLIPScore (Hessel et al., 2021) on the challenging Winoground and EqBen benchmarks. This suggests that VisualGPTScore is a supe-

Table 1. OTS generative VLMs are SOTA on image-to-text retrieval benchmarks. We begin by evaluating blind language models (in red). Surprisingly, this already produces SOTA accuracy on certain benchmarks such as ARO-Flickr, compared to the best discriminative approaches (in gray). We also nd that blind inference of generative VLMs, $P_{train}(t)$ via sampling Gaussian noise images (in blue) often performs better and achieve above-chance performance even on the most recent SugarCrepe. Next, we show that simply repurposing a generative VLM's language generation head for computing image-text scores (VisualGPTScore in yellow) which corresponds to = 0, consistently produces SOTA accuracy across all benchmarks. Finally, debiasing this score by tuning on valset (in green) further improves performance, establishing the new SOTA.

Score	Method	ARO			
		Rel	Attr	COCO	Flickr
Random	-	50.0	50.0	20.0	20.0
Text-Only	Vera	61.7	82.6	59.8	63.5
	Grammar	59.6	58.4	74.3	76.3
$P_{LLM}(t)$	BART	81.1	73.6	95.0	95.2
	Flan-T5	84.4	76.5	98.0	98.2
	OPT	84.7	79.8	97.9	98.6
$P_{train}(t)$	BLIP	87.6	80.7	98.6	99.1
$P(\text{match}|t;i)$	CLIP	59.0	62.0	46.0	60.0
	LAION2B-CLIP	51.6	61.9	25.2	30.2
	LAION5B-CLIP	46.1	57.8	26.1	31.0
	NegCLIP	81.0	71.0	86.0	91.0
	Structure-CLIP	83.5	85.1	-	-
	SyViC	80.8	72.4	92.4	87.2
	SGVL	-	-	87.2	91.0
	MosaiCLIP	82.6	78.0	87.9	86.3
	DAC-LLM	81.3	73.9	94.5	95.7
	DAC-SAM	77.2	70.5	91.2	93.9
	BLIP-ITC	63.1	81.6	34.3	41.7
	BLIP-ITM	58.7	90.3	45.1	51.3
$\frac{P_{train}(t|i)}{P_{train}(t)}$	Ours (= 0)	89.1	95.3	99.4	99.5
	Ours (= 1)	68.1	87.9	32.4	44.5
	Ours (=)	89.1	95.4	99.4	99.5

(a) Accuracy on ARO

Score	Method	VL-CheckList		
		Object	Attribute	Relation
Random	-	50.0	50.0	50.0
Text-Only	Vera	82.5	74.0	85.7
	Grammar	58.0	52.4	68.5
$P_{LLM}(t)$	BART	52.0	51.0	45.1
	Flan-T5	60.3	55.0	49.3
	OPT	59.3	48.8	60.0
$P_{train}(t)$	BLIP	68.2	58.7	75.9
$P(\text{match}|t;i)$	CLIP	81.6	67.6	63.1
	LAION2B-CLIP	84.7	67.8	66.5
	LAION5B-CLIP	87.9	70.3	63.9
	NegCLIP	81.4	72.2	63.5
	SyViC	-	70.4	69.4
	SGVL	85.2	78.2	80.4
	SLVC	85.0	72.0	69.0
	DAC-LLM	87.3	77.3	86.4
	DAC-SAM	88.5	75.8	89.8
	BLIP-ITC	90.6	80.3	73.5
	BLIP-ITM	89.9	80.7	67.7
$\frac{P_{train}(t|i)}{P_{train}(t)}$	Ours (= 0)	92.6	78.7	90.8
	Ours (= 1)	90.4	77.6	77.8
	Ours (=)	94.4	82.1	92.8

(b) Accuracy on VL-CheckList

Score	Method	SugarCrepe		
		Replace	Swap	Add
Random	-	50.0	50.0	50.0
Text-Only	Vera	49.5	49.3	49.5
	Grammar	50.0	50.0	50.0
$P_{LLM}(t)$	BART	48.4	51.9	61.2
	Flan-T5	51.4	57.6	40.9
	OPT	58.5	66.6	45.8
$P_{train}(t)$	BLIP	75.9	77.1	70.9
$P(\text{match}|t;i)$	CLIP	80.8	63.3	75.1
	LAION2B-CLIP	86.5	68.6	88.4
	LAION5B-CLIP	85.0	68.0	89.6
	NegCLIP	88.3	76.2	90.2
	BLIP-ITC	85.8	73.8	85.7
	BLIP-ITM	88.7	81.3	87.6
$\frac{P_{train}(t|i)}{P_{train}(t)}$	Ours (= 0)	93.3	91.0	91.0
	Ours (= 1)	83.2	85.5	85.9
	Ours (=)	95.1	92.4	97.4

(c) Accuracy on SugarCrepe

Score	Method	Crepe		
		Atom	Swap	Negate
Random	-	16.7	16.7	16.7
Text-Only	Vera	43.7	70.8	66.2
	Grammar	18.2	50.9	9.8
$P_{LLM}(t)$	BART	38.8	53.3	44.4
	Flan-T5	43.0	69.5	13.6
	OPT	53.3	72.7	5.0
$P_{train}(t)$	BLIP	55.4	69.7	60.8
$P(\text{match}|t;i)$	CLIP	22.3	26.6	28.8
	LAION2B-CLIP	23.6	24.8	18.0
	LAION5B-CLIP	24.2	23.9	20.1
	BLIP-ITC	24.8	17.7	26.5
	BLIP-ITM	29.5	20.7	25.5
$\frac{P_{train}(t|i)}{P_{train}(t)}$	Ours (= 0)	73.2	78.1	79.6
	Ours (= 1)	20.6	28.3	35.6
	Ours (=)	73.3	78.1	79.6

(d) Accuracy on Crepe

rior choice for measuring image-text alignment.

Balanced evaluation protocols for retrieval. Winoground and EqBen evaluate image-text alignment through retrieval tasks, and we nd their evaluation protocols discourage blind solutions. We refer the reader to the benchmarks for more details, but in summary, both benchmarks operate on pairs of image-text pairs $\{(i_0;t_0);(i_1;t_1)\}$ and construct two I-to-T retrieval (text score) tasks with a single image and two candidate captions. The text score is awarded 1 point only if *both* retrieval tasks are correct. Consider the

Table 2. -debiasing on I-to-T benchmarks and $P_{train}(t)$ frequency charts of both positive and negative captions. Increasing from 0 to 1 hurts performance on benchmarks with non-sensical negative captions like ARO-Flickr. ARO's negative captions are easier to identify because of their low score under the language prior $P_{train}(t)$, implying such benchmarks may even be solved with blind algorithms that avoid looking at images. On the other hand, for benchmarks like SugarCrepe with more balanced $P_{train}(t)$ between positive and negative captions, tuning leads to performance gain. Appendix D shows analysis on all datasets.

Alpha-Tuning	Prior Frequency	Alpha-Tuning	Prior Frequency
ARO-Flickr		SugarCrepe-Add	

Table 3. Additional results on Winoground/EqBen/COCO/Flickr30K/ImageNet1K. Table (a) shows the importance of -debiasing on these compositionality and large-scale retrieval benchmarks. While OTS generative scores do not work well, debiasing with a larger close to 1 can consistently and often signi cantly improve I-to-T performance. To highlight the improvement, we mark results without debiasing ($= 0$) (in yellow), debiasing with a xed $= 1$ (in pink), and cross-validation using held-out valsets ($_{val}$) (in green). Table (b) shows that OTS generative scores can obtain favorable results on all T-to-I retrieval tasks, competitive with the ITMScore.

Metric	Benchmark	ITMScore	$\frac{P_{train}(t\|i)}{P_{train}(t)}$ =0	=1	= $_{val}$	$_{val}$
Text Score	Winoground EqBen	$35.5_{(2:4)}$ $26.1_{(0:3)}$	$27.5_{(2:3)}$ $9.6_{(0:2)}$	$33.7_{(2:4)}$ $19.8_{(0:3)}$	$36.6_{(2:6)}$ $19.8_{(0:3)}$	$0.855_{(0:023)}$ $0.992_{(0:007)}$
R@1 / R@5	COCO Flickr30k	71.9 / 90.6 88.8 / 98.2	19.7 / 40.6 34.6 / 59.0	46.2 / 73.1 58.7 / 88.0	48.0 / 74.2 63.6 / 89.2	0.819 0.719
Accuracy	ImageNet1K	37.4	18.6	36.2	40.0	0.670

(a) -debiasing on valsets for I-to-T retrieval

Metric	Benchmark	ITMScore	$P_{train}(t\|i)$
Image Score	Winoground EqBen	15.8 20.3	21.5 26.1
R@1 / R@5	COCO Flickr30k	54.8 / 79.0 77.8 / 93.9	55.6 / 79.2 76.8 / 93.4

(b) T-to-I retrieval

common case where one caption is more likely under a language prior; here the common caption will be correctly retrieved for one of the tasks but will be incorrectly retrieved for the other, implying no points will be awarded. Similarly stringent metrics are used for T-to-I retrieval (image score). The nal group score is awarded 1 point only if all 4 retrieval tasks are correct.

-debiasing consistently improves I-to-T retrieval. Table 3-a shows that simply debiasing VisualGPTScore with a xed $= 1$ signi cantly improves performance on challenging I-to-T benchmarks. One can also do slightly better by using a held-out valset to tune for the optimal $\in [0, 1]$. For Winoground and EqBen, we sample half of the data as a valset and perform a grid search for $_{val}$ (using a step size of 0.001), reporting the performance on the other half. We repeat this process 10 times and report the mean and standard deviation. For COCO and Flickr30K, we perform -debiasing using Recall@1 (R@1) on the of cial valset. We report the zero-shot classi cation accuracy on ImageNet1K, which can be viewed as an I-to-T retrieval task that retrieves the best textual label (out of 1000) for each image. We simply use one-shot samples from Lin et al. (2023) to cross validate on ImageNet, which incurs negligible costs. Appendix B details the debiasing procedure for each dataset. Lastly, we observe that generative approaches still lag behind the ITMScore of BLIP for the two large-scale retrieval benchmarks. This motivates us to study biases of generative models from the statistical perspective of biased estimators, brie y examined in Appendix A.

Extending to T-to-I retrieval. Though not the focus of our work, we show that image-conditioned language models can be applied to T-to-I retrieval. Given a text caption t, we can rewrite the Bayes optimal T-to-I retrieval objective as:

$$P_{test}(i|t) \propto P_{train}(t|i) \cdot P_{train}(i) \tag{14}$$

Equation 14 is hard to implement because we do not have access to $P_{train}(i)$. However, when $P_{train}(i)$ is approximately uniform, one can directly apply $P_{train}(t|i)$ for optimal performance. We report T-to-I performance in Table 3-b, where our generative approach obtains competitive results compared against ITMScore, likely because T-to-I retrieval is less affected by language biases.

Table 4. Superior performance of VisualGPTScore on challenging image-text alignment benchmarks. We compare VisualGPTScore (and its α=1 version) against popular image-text scoring methods such as CLIPScore and those that combine VLMs with additional LLMs like ChatGPT. On Winoground and EqBen, our VisualGPTScore (α=0) outperforms all methods using only a state-of-the-art VLM (LLaVA-1.5). Moreover, debiasing with α=1 (using a single Gaussian noise image) consistently improves I-to-T retrieval, thereby increasing the text and group score. To ensure a fair comparison, we use the publicly available model checkpoints and corresponding code of prior works. Method descriptions and implementation details can be found in Appendix D.

Method	LLMs used	Winoground			EqBen		
		Text	Image	Group	Text	Image	Group
Random Chance	–	25.0	25.0	16.7	25.0	25.0	16.7
Official implementation							
CLIPScore	–	31.3	11.0	8.8	35.0	33.6	21.4
VPEval	ChatGPT	12.8	11.0	6.3	34.3	25.7	21.4
LLMScore	ChatGPT	21.3	17.8	12.5	32.9	27.9	22.9
Our results based on LLaVA-1.5							
TIFA	Llama-2	22.8	18.5	15.5	30.0	30.0	21.4
VQ2	FlanT5	14.0	27.3	10.0	22.9	40.7	20.0
Davidsonian	ChatGPT	21.0	16.8	15.5	26.4	20.0	20.0
VisualGPTScore (α=0)	–	36.3	37.0	24.8	25.7	42.1	21.4
VisualGPTScore (α=1)	–	44.3	37.0	27.5	42.9	42.1	29.3

State-of-the-art image-text alignment. Text-to-image generative models such as DALL-E 3 (Betker et al., 2023) are often evaluated with models that score the agreement (or alignment) between the generated image and the input caption, such as the CLIPScore (Hessel et al., 2021). However, as CLIP struggles with compositional texts (Kamath et al., 2023), recent studies such as VPEval (Cho et al., 2023b) and LLMScore (Lu et al., 2023) combine VLMs with LLMs like ChatGPT to more accurately score image-text alignment. Most recently, TIFA (Hu et al., 2023), VQ2 (Yarom et al., 2023), and Davidsonian (Cho et al., 2023a) use LLMs to generate a set of Q&A from input captions, then score the image based on the accuracy of a VQA model. Appendix D describes these methods in details. Table 4 shows that VisualGPTScore (and its debiased α=1 version) outperforms such complex approaches for image-text alignment, needing only an OTS state-of-the-art VLM, LLaVA-1.5 (Liu et al., 2023). This suggests that image-conditioned language models can already serve as robust alignment metrics. We also encourage readers to explore our latest research on VQAScore (Lin et al., 2024; Li et al., 2024), which adapts VisualGPTScore to more advanced generative models trained with visual-question-answering (VQA) datasets.

6. Discussion and Limitations

Summary. Our study shows the efficacy of *generative* pre-training scores in solving *discriminative* tasks. We present a first-principles analysis to account for mismatching distributions over text between train and test data. Our analysis motivates a training-free (zero-shot) solution to effectively debias language priors in generative scores. We hope our analysis can encourage future work to revisit the issue of language biases in vision-language benchmarks.

Limitations and future work. VisualGPTScore depends on VLMs pre-trained on noisy and imbalanced web data, which may result in biases (Mehrabi et al., 2021; Parashar et al., 2024). We make several simplified assumptions in the main paper to offer an intuitive explanation of VisualGPTScore. For instance, the image-conditioned language model might not accurately represent $P_{train}(t|i)$ and assigns higher scores towards more common texts. We examine this phenomenon in Appendix A. Future work may attempt other sampling methods like coreset selection (Guo et al., 2022; Wu et al., 2023) to estimate $P_{train}(t)$ with improved efficiency. As VisualGPTScore shows competitive performance, distilling it into discriminative CLIPScore (Miech et al., 2021) can reduce its inference cost. Finally, VQAScore (Lin et al., 2024; Li et al., 2024) applies VisualGPTScore to the latest vision-language models trained on visual-question-answering (VQA) datasets to achieve the state-of-the-art performance. This demonstrates that generative scoring is a more reliable alternative to CLIPScore (Hessel et al., 2021) for automated evaluation of text-to-image models.

Impact Statement

VisualGPTScore is developed with the important goal of advancing the field of vision-language models. It has many positive societal impacts, such as improving the scientific evaluation of generative models (Lin et al., 2024; Li et al., 2024). Nonetheless, we encourage future work to study its biases, especially since the underlying models are trained on noisy and imbalanced data (Parashar et al., 2024; Mehrabi et al., 2021).

References

Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., and Anderson, P. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8948–8957, 2019.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

Bertolini, L., Weeds, J., and Weir, D. Testing large language models on compositionality and inference with phrase-level adjective-noun entailment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4084–4100, 2022.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2023.

Brendel, W. and Bethge, M. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cascante-Bonilla, P., Shehada, K., Smith, J. S., Doveh, S., Kim, D., Panda, R., Varol, G., Oliva, A., Ordonez, V., Feris, R., et al. Going beyond nouns with vision & language models using synthetic data. *arXiv preprint arXiv:2303.17590*, 2023.

Cho, J., Hu, Y., Garg, R., Anderson, P., Krishna, R., Baldridge, J., Bansal, M., Pont-Tuset, J., and Wang, S. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023a.

Cho, J., Zala, A., and Bansal, M. Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328*, 2023b.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Valter, D., Narang, S., Mishra, G., Yu, A. W., Zhao, V., Huang, Y., Dai, A. M., Yu, H., Petrov, S., hsin Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction- netuned language models. *ArXiv*, abs/2210.11416, 2022.

Daille, B. *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et ltres linguistiques*. PhD thesis, Ph. D. thesis, Université Paris 7, 1994.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Diwan, A., Berry, L., Choi, E., Harwath, D., and Mahowald, K. Why is winoground hard? investigating failures in visuolinguistic compositionality. *arXiv preprint arXiv:2211.00768*, 2022.

Doveh, S., Arbelle, A., Harary, S., Panda, R., Herzig, R., Schwartz, E., Kim, D., Giryes, R., Feris, R., Ullman, S., et al. Teaching structured vision&language concepts to vision&language models. *arXiv preprint arXiv:2211.11733*, 2022.

Doveh, S., Arbelle, A., Harary, S., Alfassy, A., Herzig, R., Kim, D., Giryes, R., Feris, R., Panda, R., Ullman, S., et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. *arXiv preprint arXiv:2305.19595*, 2023.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

Fu, J., Ng, S.-K., Jiang, Z., and Liu, P. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. 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*, pp. 6904–6913, 2017.

Guo, C., Zhao, B., and Bai, Y. 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, pp. 181–195. Springer, 2022.

Henning, C. A. and Ewerth, R. Estimating the information gap between textual and visual representations. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 14–22, 2017.

Herzig, R., Mendelson, A., Karlinsky, L., Arbelle, A., Feris, R., Darrell, T., and Globerson, A. Incorporating structured representations into pretrained vision & language models using scene graphs. arXiv preprint arXiv:2305.06343, 2023.

Hessel, J. and Schofield, A. 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), pp. 204–211, 2021.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718, 2021.

Hsieh, C.-Y., Zhang, J., Ma, Z., Kembhavi, A., and Krishna, R. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. arXiv preprint arXiv:2306.14610, 2023.

Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., and Smith, N. A. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. arXiv preprint arXiv:2303.11897, 2023.

Huang, Y., Tang, J., Chen, Z., Zhang, R., Zhang, X., Chen, W., Zhao, Z., Lv, T., Hu, Z., and Zhang, W. Structure-clip: Enhance multi-modal language representations with structure knowledge. arXiv preprint arXiv:2305.06152, 2023.

Kamath, A., Hessel, J., and Chang, K.-W. Text encoders are performance bottlenecks in contrastive vision-language models. arXiv preprint arXiv:2305.14897, 2023.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217, 2019.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. Ctrl: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858, 2019.

Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., Xia, X., Zhang, P., Neubig, G., and Ramanan, D. Evaluating and improving compositional text-to-visual generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024.

Li, J. and Jurafsky, D. Mutual information and diverse decoding improve neural machine translation, 2016.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. 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, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34: 9694–9705, 2021.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning, pp. 12888–12900. PMLR, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740–755. Springer, 2014.

Lin, Z., Yu, S., Kuang, Z., Pathak, D., and Ramana, D. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. arXiv preprint arXiv:2301.06267, 2023.

Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., and Ramanan, D. Evaluating text-to-visual generation with image-to-text generation. arXiv preprint arXiv:2404.01291, 2024.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. arXiv preprint arXiv:2304.08485, 2023.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

Lu, Y., Yang, X., Li, X., Wang, X. E., and Wang, W. Y. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. arXiv preprint arXiv:2305.11116, 2023.

Ma, Z., Hong, J., Gul, M. O., Gandhi, M., Gao, I., and Krishna, R. Crepe: Can vision-language foundation models reason compositionally? arXiv preprint arXiv:2212.07796, 2022.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6):1–35, 2021.

Miech, A., Alayrac, J.-B., Laptev, I., Sivic, J., and Zisserman, A. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9826–9836, 2021.

OpenAI. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

Papadimitriou, I., Futrell, R., and Mahowald, K. When classifying grammatical role, bert doesn't care about word order... except when it matters. arXiv preprint arXiv:2203.06204, 2022.

Parashar, S., Lin, Z., Liu, T., Dong, X., Li, Y., Ramanan, D., Caverlee, J., and Kong, S. The neglected tails of vision-language models. arXiv preprint arXiv:2401.12425, 2024.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763. PMLR, 2021.

Role, F. and Nadif, M. 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, pp. 218–223, 2011.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402, 2022.

Shapiro, A. Monte carlo sampling methods. Handbooks in operations research and management science, 10:353–425, 2003.

Shrivastava, A., Selvaraju, R. R., Naik, N., and Ordonez, V. Clip-lite: information efficient visual representation learning from textual annotations. arXiv preprint arXiv:2112.07133, 2021.

Singh, H., Zhang, P., Wang, Q., Wang, M., Xiong, W., Du, J., and Chen, Y. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. arXiv preprint arXiv:2305.13812, 2023.

Sinha, K., Jia, R., Hupkes, D., Pineau, J., Williams, A., and Kiela, D. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. arXiv preprint arXiv:2104.06644, 2021.

Tejankar, A., Sanjabi, M., Wu, B., Xie, S., Khabsa, M., Pirsiavash, H., and Firooz, H. A fistful of words: Learning transferable visual models from bag-of-words supervision. arXiv preprint arXiv:2112.13884, 2021.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. Winoground: Probing vision and language models for visio-linguistic compositionality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5238–5248, 2022.

Tschannen, M., Kumar, M., Steiner, A., Zhai, X., Houlsby, N., and Beyer, L. Image captioners are scalable vision learners too. arXiv preprint arXiv:2306.07915, 2023.

Wang, T., Lin, K., Li, L., Lin, C.-C., Yang, Z., Zhang, H., Liu, Z., and Wang, L. Equivariant similarity for vision-language foundation models. arXiv preprint arXiv:2303.14465, 2023.

Wang, Z., Feng, B., Narasimhan, K., and Russakovsky, O. Towards unique and informative captioning of images. In European Conference on Computer Vision (ECCV), 2020.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al. Robust fine-tuning of zero-shot models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7959–7971, 2022.

Wu, X., Deng, Z., and Russakovsky, O. Multimodal dataset distillation for image-text retrieval. arXiv preprint arXiv:2308.07545, 2023.

Yao, T., Mei, T., and Ngo, C.-W. Co-reranking by mutual reinforcement for image search. In *Proceedings of the ACM international conference on image and video retrieval*, pp. 34–41, 2010.

Yarom, M., Bitton, Y., Changpinyo, S., Aharoni, R., Herzig, J., Lang, O., Ofek, E., and Szpektor, I. What you see is what you read? improving text-image alignment evaluation. *arXiv preprint arXiv:2305.10400*, 2023.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. 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.

Yuan, W., Neubig, G., and Liu, P. Bartscore: Evaluating generated text as text generation. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27263–27277. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhao, T., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, 2021.

Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., and Yin, J. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

A. Is VisualGPTScore a Biased Estimator of $P_{\text{train}}(t|i)$?

Retrieval performance on trainset (LAION). This paper is built on the assumption that VisualGPTScore is a reliable estimator of $P_{\text{train}}(t|i)$. However, this simplifying assumption does not completely hold for the BLIP model we examine. We speculate that such OTS generative scores are biased towards more common texts. We witness this same phenomenon in Table 5, where we perform image-text retrieval on random subsets from training distribution LAION-114M (Li et al., 2022).

Table 5. Retrieval performance on randomly sampled training (LAION114M) subsets with varied sizes. Table (a) shows that while OTS generative scores are robust for T-to-I retrieval, its performance degrades on I-to-T retrieval tasks when the number of candidate texts increases. This implies that OTS generative scores suffer from language biases towards certain texts even in the training set. Nonetheless, we show that our debiasing solution using either $= 1$ or optimal $\in [0;1]$ with a step size of 0.001, can consistently boost the performance. Figure (b) visualizes debiasing results on LAION subsets, where each curve represents a different sample size.

Dataset Size	I-to-T Retrieval					T-to-I Retrieval	
	ITM	$\frac{P_{\text{train}}(t\|i)}{P_{\text{train}}(t)}$				ITM	$P_{\text{train}}(t\|i)$
		$=0$	$=1$	$=$			
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

Modelling the language bias in VisualGPTScore. As evidenced in Table 5, 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:

$$\text{VisualGPTScore}(t;i) := \hat{P}_{\text{train}}(t|i) = P_{\text{train}}(t|i) \cdot P_{\text{train}}(t) ; \tag{15}$$

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.

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

$$\hat{P}_{\text{train}}(t) := \frac{1}{n}\sum_{k=1}^{n} \hat{P}_{\text{train}}(t|i_k) = P_{\text{train}}(t)^{1+} : \tag{16}$$

Rewriting optimal I-to-T objective with $\hat{P}$. We can rewrite Equation 4 as:

$$P_{\text{test}}(t|i) \propto P_{\text{train}}(t|i)\frac{P_{\text{test}}(t)}{P_{\text{train}}(t)} \tag{17}$$

$$= \hat{P}_{\text{train}}(t|i)\frac{P_{\text{test}}(t)}{P_{\text{train}}(t)^{1+}} \tag{18}$$

$$= \hat{P}_{\text{train}}(t|i)\frac{P_{\text{test}}(t)}{\hat{P}_{\text{train}}(t)} \tag{19}$$

-debiasing with $\hat{P}$. Using Equation 19, we can reformulate debiasing (Equation 7) as follows:

$$P_{\text{test}}(t) \propto P_{\text{train}}(t)^{1-\hat{}} \quad) \quad \text{Optimal score is } \frac{\hat{P}_{\text{train}}(t|i)}{\hat{P}_{\text{train}}(t)} \tag{20}$$

where $= \frac{\hat{}+}{1+}$. Notably, the above equation has the same structure as before (Equation 7). This implies that even if $P_{\text{train}}(t) = P_{\text{test}}(t)$, we still anticipate $= \frac{}{1+} \neq 0$. This accounts for why the optimal is not 0 when we perform I-to-T retrieval on trainset in Table 5.

Implication for vision-language modelling. Our analysis indicates that similar to generative LLMs (Li et al., 2016; Li & Jurafsky, 2016), contemporary image-conditioned language models also experience issues related to imbalanced learning (Kang et al., 2019). Potential solutions could be: (a) re ned sampling techniques for Monte Carlo estimation of $P(t)$ such as through dataset distillation (Wu et al., 2023), and (b) less biased modelling of $P(t|i)$ such as through controllable generation (Keskar et al., 2019).

B. Ablation Studies on -Debiasing

Details of Gaussian noise samples. BLIP and BLIP-2 experiments sample Gaussian noise images 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, 1 image for ImageNet, and 3 images for the rest of the benchmarks.

Estimating $P_{train}(t)$ via Gaussian noise images is more sample-ef cient. We use Winoground to show that sampling Gaussian noise images to calculate $P_{train}(t)$ can be more ef cient than sampling trainset images. As demonstrated in Table 6, 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.

Table 6. Comparing sampling of Gaussian noise images and trainset images for estimating $P_{train}(t)$. We report text scores of -debiasing 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 $2 [0;1]$ is 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 xed std of 0.25.

Sample Size	Guassian Noise Images		Trainset Images	
	$= _{test}$	$_{test}$	$= _{test}$	$_{test}$
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	$36:25_{(0:0)}$	$0.850_{(0:000)}$	$35:15_{(0:3)}$	$0.960_{(0:033)}$

Alternative approach on COCO/Flickr30k: estimating $P_{train}(t)$ using testset images. For large-scale retrieval benchmarks like COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014), we can directly average scores of all candidate images (in the order of thousands) to ef ciently approximate $P_{train}(t)$ without the need to sample any Gaussian noise images. This approach incurs zero computation cost as we have already pre-computed scores between each candidate image and text. We show in Table 7 that using testset images indeed results in better performance than sampling 3 Gaussian noise images.

Table 7. I-to-T retrieval on COCO/Flickr30k using different sampling methods. Estimating $P_{train}(t)$ by averaging the scores of testset images (with zero computational cost) demonstrates superior performance compared to sampling additional Gaussian noise images.

Metric	Benchmark	$P_{train}(t|i)$	Sampling Method	$\frac{P_{train}(t|i)}{P_{train}(t)}$		
				$=1$	$= _{val}$	$_{val}$
R@1 / R@5	COCO	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	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

Tuning with a valset. In Table 8, similar performance trends are observed across validation and test splits of COCO and Flickr30k I-to-T retrieval benchmarks using the same $2 [0;1]$. Furthermore, $_{test}$ and $_{val}$ are empirically close. As such, our method can function as a reliable training-free debiasing method.

C. Experiments with BLIP-2

We provide BLIP-2 results for completeness.

BLIP-2 (Li et al., 2023) overview. BLIP-2 leverages frozen pre-trained image encoders (Fang et al., 2022) and large language models (Chung et al., 2022; Zhang et al., 2022) to bootstrap vision-language pre-training. It proposes a lightweight

Table 8. α-debiasing results on both the valset and testset for COCO/Flickr30k I-to-T retrieval. We observe that validation and test performance are strongly correlated while we interpolate $\alpha \in [0, 1]$.

(b) Alpha-tuning on COCO Retrieval

(c) Alpha-tuning on Flickr Retrieval

Querying Transformer (Q-Former) that is trained in two stages. Similar to BLIP (Li et al., 2022), 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 (Chung et al., 2022) or OPT (Chung et al., 2022), 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 (Table 9 and Table 10). 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 (Chung et al., 2022) as the frozen LLM. Our α-debiasing solutions generalize to all variants of BLIP-2.

Table 9. BLIP-2 on ARO/Crepe/VL-CheckList/SugarCrepe.

Benchmark	Dataset	Random	w. Q-Former			w. Flan-T5
			ITC	ITM	$P_{train}(t\|i)$	$P_{train}(t\|i)$
ARO	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	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	Object	50.0	89.7	89.2	90.1	84.1
VL-CheckList	Attribute	50.0	76.6	79.3	73.9	70.6
VL-CheckList	Relation	50.0	70.5	72.3	89.9	56.7
SugarCrepe	Replace	50.0	86.7	88.5	93.0	82.4
SugarCrepe	Swap	50.0	69.8	80.9	91.2	80.8
SugarCrepe	Add	50.0	86.5	88.0	92.7	76.2

Table 10. BLIP-2 on Winoground/EqBen.

Benchmark	Model	I-To-T (Text Score)						T-To-I (Image Score)		
		ITC	ITM	$\frac{P_{train}(t\|i)}{P_{train}(t)^{\alpha}}$				ITC	ITM	$P_{train}(t\|i)$
				$\alpha=0$	$\alpha=1$	$\alpha=\alpha^*$	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

D. Additional Reports

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

Details of Table 1. For CLIP (Radford et al., 2021), LAION2B-CLIP, and LAION5B-CLIP (Schuhmann et al., 2022), we report the results from Hsieh et al. (2023) using the ViT-B-32, ViT-bigG-14, and xlm-roberta-large-ViT-H-14 models respectively. The results of NegCLIP (Yuksekgonul et al., 2022), Structure-CLIP (Huang et al., 2023), SVLC (Doveh et al., 2022), SGVL (Herzig et al., 2023), DAC-LLM, and DAC-SAM (Doveh et al., 2023) are directly copied from their original papers. We run BLIP-ITC and BLIP-ITM using our own codebase, which will be released to the public.

Method descriptions for Table 4. CLIPScore (Hessel et al., 2021) measures the cosine similarity (dot product) score between an image and text, each embedded using the CLIP image and text encoder, respectively. VPEval (Cho et al., 2023b) utilizes GPT-3.5 to translate the text prompt into a Python-like program that invokes vision foundation models such as CLIP, BLIP, and GroundingDINO, to examine fine-grained image details. LLMScore (Lu et al., 2023) uses BLIP-2 to first caption the image, then uses ChatGPT to score the difference between the BLIP-generated caption and the text prompt. TIFA (Hu et al., 2023) and Davidsonian (Cho et al., 2023a) first use LLMs such as a finetuned Llama-2 or GPT-3.5 to generate a set of Q&A given the text prompt, then return the accuracy score of the VQA model. VQ2 (Yarom et al., 2023) uses a finetuned FlanT5 to generate the Q&A, then averages the log likelihoods of the generated answers.

Implementation details of Table 4. We report the performance on Winoground (Thrush et al., 2022) and EqBen-Mini, which is an official subset of EqBen (Wang et al., 2023) for benchmarking large foundational VLMs. We follow the official implementation of CLIPScore (Hessel et al., 2021) to report the performance of CLIP-ViT-B-32 (Radford et al., 2021). For VPEval (Cho et al., 2023b) and LLMScore (Lu et al., 2023), we strictly follow the official codebase to benchmark their performance. For TIFA (Hu et al., 2023), VQ2 (Yarom et al., 2023), Davidsonian (Cho et al., 2023a), we strictly follow their released code and adopt their QA-generation language models (or in-context Q&A samples for ChatGPT). However, as we do not have access to the private VQA models they adopted, e.g., PaLI-17B, we implement these approaches using LLaVA-1.5-13B (Liu et al., 2023) as the VQA model. We stick to the default system message to prompt LLaVA-1.5, which can be found on their official GitHub repo. For fair comparison, our VisualGPTScore is also implemented using LLaVA-1.5-13B. We only use the system message without appending any questions when computing $P(\text{text}|\text{image})$. For α-debiasing, we sample a single Gaussian image with a mean of 0 and standard deviation of 0.25 (derived from the statistics of training images used to train LLaVA).

Group scores on Winoground/EqBen using BLIP (Table 11).

Table 11. Performance comparison of BLIP's ITCScore, ITMScore, and α-tuned VisualGPTScore$^{\alpha}$ on Winoground and EqBen.

Method	Winoground (all)			EqBen (val)		
	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	20.3	12.6
VisualGPTScore	36.5	21.5	16.8	20.4	26.1	11.7

Fine-grained tags on Winoground (Table 12).

Performance on SugarCrepe (Table 13).

α-debiasing on ARO/Crepe/SugarCrepe/VL-CheckList (Table 14).

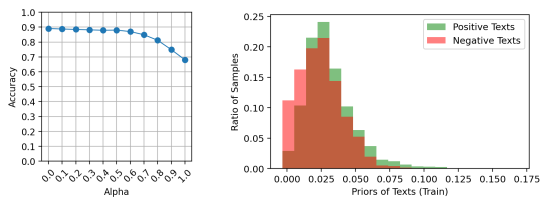
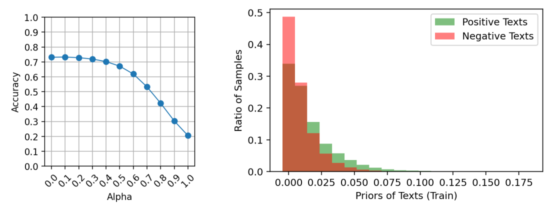
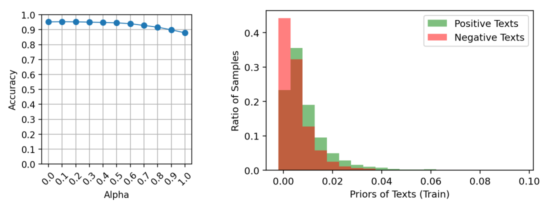
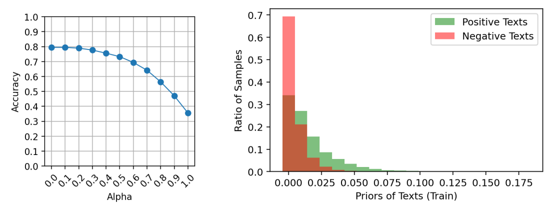
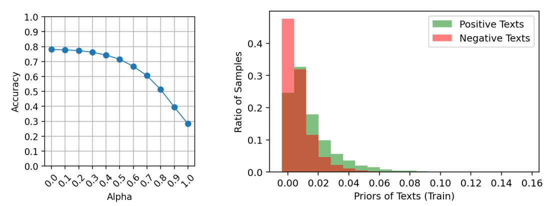
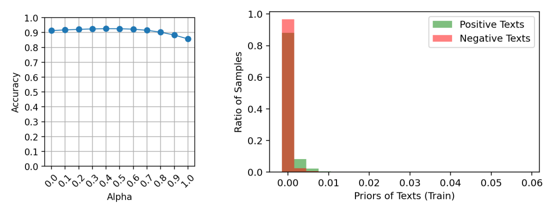
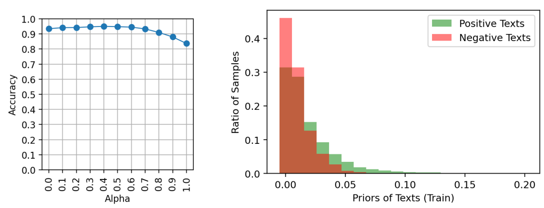
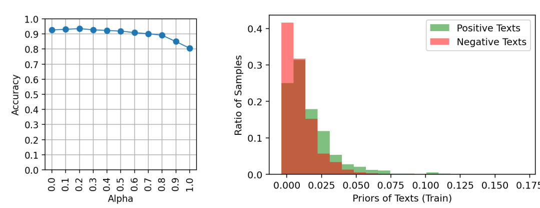
Table 12. BLIP performance on Winoground subtags (Diwan et al., 2022). We report the number of test instances for each subtag and their respective text score, image score, group score.

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	43.0	28.5	23.8
NonCompositional	30	ITCScore	43.3	16.7	16.7
		ITMScore	50.0	23.3	16.7
		VisualGPTScore	43.3	33.3	26.7
AmbiguouslyCorrect	46	ITCScore	32.6	8.7	6.5
		ITMScore	28.3	6.5	2.2
		VisualGPTScore	26.1	19.6	8.7
VisuallyDifficult	38	ITCScore	29.0	7.9	7.9
		ITMScore	26.3	10.5	7.9
		VisualGPTScore	31.6	13.2	7.9
UnusualImage	56	ITCScore	32.5	8.9	8.9
		ITMScore	21.4	10.7	7.1
		VisualGPTScore	30.4	10.7	8.9
UnusualText	50	ITCScore	20.0	8.0	6.0
		ITMScore	38.0	12.0	12.0
		VisualGPTScore	30.0	18.0	12.0
ComplexReasoning	78	ITCScore	16.7	2.6	1.3
		ITMScore	21.8	5.1	2.6
		VisualGPTScore	21.8	10.3	6.4

Table 13. **Performance on SugarCrepe (Hsieh et al., 2023).** SugarCrepe is the most recent visio-linguistic compositionality benchmark which improves upon previous Crepe (Ma et al., 2022) by using state-of-the-art large language models (including ChatGPT), instead of rule-based templates, to generate more natural negative text captions. We show that text-only baselines and LLM-based methods indeed fail to succeed on SugarCrepe. However, our OTS generative approaches still achieve competitive results compared against SOTA discriminative approaches. The results of human performance, text-only baseline, and SOTA CLIP and NegCLIP-SugarCrepe are directly taken from the Hsieh et al. (2023). For other approaches, we evaluate their performance following the same procedure as described in main texts.

Method	Model	SugarCrepe			
		Replace	**Swap**	**Add**	**AVG**
Human Performance	-	98.67	99.50	99.00	99.06
Random Chance	-	50.00	50.00	50.00	50.00
Text-Only Baseline	Vera	49.46	49.30	49.50	49.42
	Grammar	50.00	50.00	50.00	50.00
$P_{LLM}(\mathbf{t})$	Bart	48.41	51.93	61.16	53.83
	Flan-T5	51.41	57.59	40.94	49.98
	OPT	58.53	66.58	45.78	56.96
$P_{train}(\mathbf{t})$	BLIP	75.90	77.14	70.89	74.64
ITCScore	CLIP-LAION2B	86.50	68.56	88.37	81.14
	CLIP-LAION5B	84.98	67.95	89.62	80.85
	BLIP	85.76	73.79	85.66	81.74
	BLIP-2	86.66	69.77	86.50	80.98
	NegCLIP-SugarCrepe	88.27	74.89	90.16	84.44
ITMScore	BLIP	88.68	81.29	87.57	85.85
	BLIP2-Qformer	88.45	80.87	87.96	85.76
$P_{train}(\mathbf{t}\|\mathbf{i})$	BLIP	**93.33**	**91.00**	**90.98**	**91.77**
	BLIP2-Qformer	**93.00**	**91.24**	**92.69**	**92.31**
	BLIP2-FlanT5	82.44	76.57	76.24	78.42
$\frac{P_{train}(\mathbf{t}\|\mathbf{i})}{P_{train}(\mathbf{t})}$	BLIP	**95.09**	**92.39**	**97.36**	**94.95**
	BLIP2-Qformer	**94.62**	**92.27**	**97.58**	**94.82**
	BLIP2-FlanT5	85.69	78.80	91.76	85.42

Table 14. **α-debiasing results on all I-to-T benchmarks and $P_{train}(\texttt{t})$ frequency charts.** Increasing α from 0 to 1 hurts performance on benchmarks with non-sensical negative captions such as ARO and Crepe. These benchmarks can also be largely solved with blind algorithms that avoid looking at images. On the other hand, for benchmarks like SugarCrepe with more balanced $P_{train}(\texttt{t})$ between positives and negatives, tuning α leads to performance gain.

Alpha-Tuning / Prior Frequency	Alpha-Tuning / Prior Frequency
 ARO (VG-Relation)	 Crepe (Atom-Foils)
 ARO (VG-Attribution)	 Crepe (Negate)
ARO (Flickr30K-Order)	 Crepe (Swap)
SugarCrepe (Add)	 VL-CheckList (Relation)
 SugarCrepe (Replace)	 SugarCrepe (Swap)