Removing Distributional Discrepancies in Captions Improves Image-Text Alignment

Yuheng Li[1], Haotian Liu[1], Mu Cai[1], Yijun Li[2], Eli Shechtman[2], Zhe Lin[2], Yong Jae Lee[1], and Krishna Kumar Singh[2]

[1] University of Wisconsin-Madison
[2] Adobe Research

Abstract. In this paper, we introduce a model designed to improve the prediction of image-text alignment, targeting the challenge of compositional understanding in current visual-language models. Our approach focuses on generating high-quality training datasets for the alignment task by producing mixed-type negative captions derived from positive ones. Critically, we address the distribution imbalance between positive and negative captions to ensure that the alignment model does not depend solely on textual information but also considers the associated images for predicting alignment accurately. By creating this enhanced training data, we fine-tune an existing leading visual-language model to boost its capability in understanding alignment. Our model significantly outperforms current top-performing methods across various datasets. We also demonstrate the applicability of our model by ranking the images generated by text-to-image models based on text alignment. Project page: `https://yuheng-li.github.io/LLaVA-score/`

1 Introduction

Recent years have seen rapid advances in multimodal research, encompassing both visual generation [2, 29, 31] and visual understanding [1, 25, 27]. Training capable multimodal models generally requires extensive datasets of image-text pairs, e.g., LAION [32], CC12M [4], and others [11, 33], which are collected on a web-scale and thus tend to be noisy. This noise in the data contributes to certain challenges. For instance, vision-language models often struggle with hallucination [20] and face difficulties in mastering compositional reasoning [36, 40]. Similarly, text-to-image models frequently fail to generate accurate images when processing complex sentence prompts [10].

Given these challenges, the ability to automatically assess whether an image and a caption are semantically aligned plays a crucial role. This capability is essential not only for cleaning the data used to pre-train these models but also for evaluating and enhancing the performance of both text-to-image and image-to-text generation models. The most frequently employed metric for this task is the CLIP score [13], which calculates the cosine distance between the CLIP [28] embeddings of the paired text and image. However, it has been observed that CLIP, along with other methods such as BLIP [19] and Flava [34], tends to

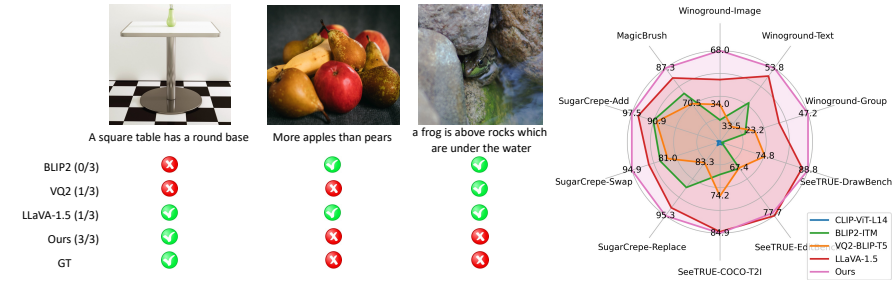

Fig. 1: Left: Qualitative examples for image-text alignment prediction, where our approach can distinguish fundamental concepts such as positioning, counting, and attributes. **Right:** Our approach shows superior performance in image-text alignment.

operate on a bag-of-words basis. They sometimes even cannot tell the difference between simple cases, such as *horse eating grass* versus *grass eating horse* [14, 23, 40].

The conventional approach to training image-text alignment models involves generating negative captions that are misaligned with the images, which are paired with the original positive captions as the training data. For example, prior works randomly shuffle words to create negative captions [40], or employ language models to generate coherent negative sentences [39]. We follow the latter and leverage an LLM to generate a mix of different types of negative captions during training. Specifically, this includes the *replacing* type, where one linguistic element is substituted with an arbitrary counterpart (*e.g.* '*a knife is on the table*' → '*a spoon is on the table*'), and the *swapping* type, where words within the same sentence are rearranged (*e.g.* '*an apple is to the left of a banana*' → '*a banana is to the left of an apple*'). The former type can aid in enhancing an image-text alignment model's perceptual skills as it needs to distinguish between the original and replaced elements, while the latter type can help the model's reasoning capabilities as elements in the caption remain the same but the relationship between them changes.

However, critically, we find that this standard approach of generating negative captions, and ensuring that they have coherence (e.g., proper grammar) on a *per-instance* level, is insufficient. In particular, this approach cannot ensure consistency at the *distribution* level between positive and negative captions. Distributional biases may originate from the initial dataset or from the rules or models employed to generate negative captions. For instance, the COCO dataset [17] contains significantly more captions including the word *giraffe* compared to *elephant*. Yet, when using GPT to generate negative captions, we observe a tendency for GPT to substitute *giraffe* with *elephant*, resulting in a surplus of *giraffe* mentions in positive captions but more *elephant* in negative captions. Unfortunately, this means that an image-text alignment model trained on such data can be biased to predict sentences containing *elephant* to be a positive caption and those that contain *giraffe* to be a negative caption, independent of the paired image. To address this imbalance, our approach consists of fine-tuning a classifier on text captions only (without paired image inputs) to remove biased data that

the text classifier can correctly predict with high confidence. It's important to note that this bias is not unique to the GPT model or how one prompts it, but can arise in other rule-based methods or pre-trained models as well.

While previous research [14] noticed distribution differences due to implausible and non-fluent negative captions, our work is the first to eliminate the dataset-level distribution differences using a language model to implicitly encompass all linguistic aspects, without the need for explicit identification. This approach allows us to incorporate considerations like word frequency which is not identified beforehand, thereby providing a more holistic optimization over the dataset's distribution.

For our image-text alignment model, we leverage a state-of-the-art vision-language model, LLaVA [24, 25], and finetune it with our curated data. Our model achieves significantly better results compared with other image-text alignment models, demonstrated in Figure 1. We also demonstrate that our curated data can improve other models like BLIP2 [18]. Finally, we further show that our image-text alignment model can help with other vision-language tasks like ranking generated images from T2I models [15]. In summary, we have four main contributions:

- We identify new dataset-level distribution differences between positive and negative captions that lead to biased image-text alignment models.
- To address this, we propose a method to maintain consistency between positive and negative caption distributions, which is critical to ensure that an image-text alignment model relies on both image and text to measure alignment instead of only the biases present in the text.
- We use our curated training data to finetune existing visual-language models like LLaVA, and obtain state-of-the-art results for image-text alignment.
- In addition, we demonstrate the application of our image-text alignment model in ranking the image generations produced by generative models.

2 Related Work

Challenges in Compositional Understanding. Image-text pairs form a crucial interface between visual and linguistic modalities, thus evaluating if a given image-text pair is aligned is important for both data curation and model performance evaluation. The pioneering work of CLIP [28] demonstrated this by leveraging an extensive corpus of such pairs for image-text contrastive training. More recent works such as BLIP [18,19] and LLaVA [24,25] further utilize large language models (LLMs) to achieve image-text alignment via the text generation objective. Such models usually inherit the frozen CLIP visual encoder to produce a set of visual tokens, and then feed such tokens and the language instructions into the LLM.

It has been observed that vision-language models like CLIP have limited capability in understanding compositionality [26,30,36,40,42]. Specifically, they find it challenging to recognize the permutation of words within sentences [36].

Moreover, these models often struggle to identify the binding of attributes to multiple objects within a single sentence or to discern the relationships between objects [30, 40, 42]. While models built on top of large language models (LLMs), such as BLIP [19] and LLaVA [25], exhibit improved understanding capabilities, they still face difficulties with complex compositional understanding largely due to their lack of training on sufficiently challenging data.

Enhancing Vision-Language Compositionality. Several studies focus on improving models' understanding of language compositionality [8,9,12,17,23,39, 40]. Some strategies [15,39] involve using language models to decompose text into multiple succinct assertion phrases, which are then evaluated by VQA models on an individual basis. [23] also finds that assessing the conditional probability of predicting text based on the input image offers more accurate outcomes than traditional discriminative approaches like contrastive or classification scores used in BLIP [19]. Still, the most widely used method remains the explicit fine-tuning of models to differentiate between hard negatives and correct captions [9, 39, 40]. [40] randomly shuffles words to generate negative captions, but subsequent analysis by [14] points out the approach's flaw: models could simply rely on textual cues (e.g., grammar correctness) for predictions. [39] attempts to employ LLMs to create negative captions, ensuring the negative captions are both fluent and meaningful. Nonetheless, we observed that current fine-tuning methods do not generate a diverse range of negative prompts. Moreover, merely evaluating the grammar or logical coherence of negative captions is insufficient to eliminate data bias in the distribution of negative captions.

3 Approach

In our approach, we assume access to images accompanied by accurately labeled positive captions, similar to those found in the MSCOCO [21] dataset. Following [14], contrary to adopting rule-based techniques that often generate illogical sentences, our method utilizes large language models (LLMs) like GPT4 to transform positive captions into negative ones. We first outline our strategy for generating various types of negative captions and then present a straightforward technique to mitigate biases inherent in the distributions of positive and negative captions. We show the entire pipeline in Figure 2.

3.1 Constructing Diverse Negative Captions

Suppose we have a dataset consisting of image-text pairs $\{I, T_p\}$, where I represents an image and T_p is its associated positive caption. Previous research [39,40] showed that merely randomly shuffling image-text pairs to generate negative samples is insufficient to learn capable vision-language models that properly understand the structure of a caption and its relationship with the image. These studies have underscored the value of constructing hard negative captions to significantly improve the model's language compositional abilities. Following this

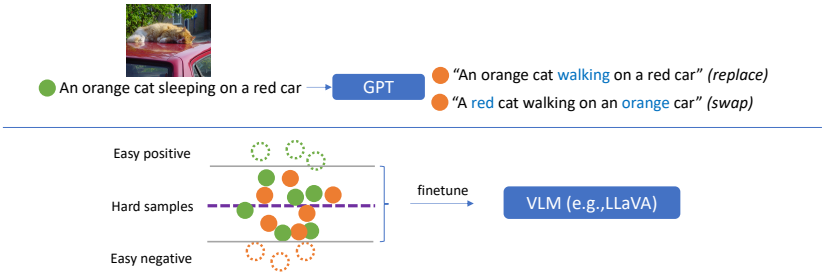

Fig. 2: Top: We feed the positive caption (green dots) into GPT to create two types of negative captions (red dots): substituting one linguistic element with any plausible alternative or swapping the positions of two components. The blue part in negative captions highlights the modifications. **Bottom:** we remove easy negative samples using only text data and utilize the remaining samples to fine-tune vision-language models.

insight [39], we utilize large language models (LLMs) to generate such hard negative captions. Specifically, we create two types of hard negative captions.

We refer to the first method for creating negative captions as the *replacing* strategy, which identifies key components in a language and uses a language model to replace it with other plausible substitutes. The replaced component can be any linguistic part such as a noun, adjective, preposition, etc. For example, *"a photo of a broken down stop sign"* could be replaced with *"a photo of a brand new stop sign"*; *"a cute cat looking at a bird"* could be changed to *"a cute dog looking at a bird"*. Conceptually, this type of negative caption aims to enhance the model's recognition capabilities. For executing this task, we engage GPT by providing it with specific instructions followed by some contextual examples, as illustrated in Figure 3.

We refer to the second method for creating negative captions as the *swapping* strategy. This approach involves generating a new sentence by utilizing the original language components, typically resulting in the swapping of two or several identical linguistic elements. Specifically, we first ask GPT to break down the original positive sentence into its key components, and then request the model to construct a new sentence with these elements. For example, if the input caption is *"an airplane is flying in the blue sky"*, GPT is tasked with identifying the key components including *"airplane"*, *"flying"*, *"blue"*, and *"sky"*. The newly crafted sentence could be *"a blue airplane is flying in the sky"*. Please note that for some short captions, there may not be enough language elements to form a reasonably different sentence in meaning. Thus, we also need to judge if the new sentence makes sense or not. Please refer to the Figure 3 for our prompt used in GPT.

3.2 Addressing Distribution Discrepancies

Previous studies [14] have demonstrated that artifact issues arising from incorrect construction in negative captions can lead to distribution disparities between positive and negative captions. To mitigate this, commonsense and grammar

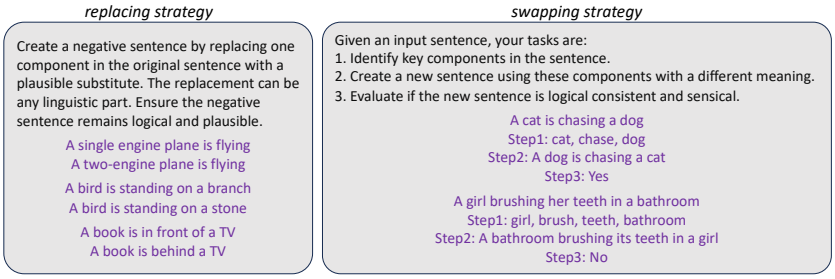

Fig. 3: Prompts used in GPT for generating two types of negative captions, with in-context examples shown in purple.

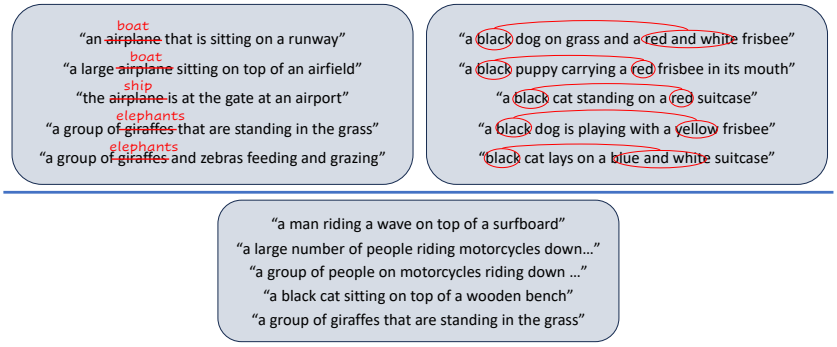

Fig. 4: Top prediction based on a text-only binary classifier. **Top left:** Negative captions generated through replacement strategy. **Top right:** Negative captions generated through swapping strategy. **Bottom:** Positive captions from the COCO dataset.

models are employed. Nonetheless, our findings indicate that *merely assessing the sensibility of sentences* is insufficient for aligning the distributions of positive and negative sentences because of two following two reasons.

First, each image-pair dataset inherently establishes its own unique distribution. For instance, the COCO dataset [21], despite its widespread use and diverse content, predominantly encompasses everyday scenes, objects, and activities. Common subjects such as people and motorcycles, along with frequent activities like skiing and surfing, characterize the distinct distribution of COCO. Second, when employing rule-based methods or prompting language models to generate negative captions, any strategy adopted will inevitably reflect biases from human preconceptions or pre-trained models. This leads to negative captions originating from a different distribution.

To illustrate this concept, we develop a blind text-only binary classifier that processes positive and negative captions without seeing any images. Surprisingly, we find that despite the coherence and logical structure of negative captions, the classifier is adept at distinguishing between negative and positive captions based solely on the text. Figure 4 shows the most confidently correct predictions for both our *replace* and *swap* data categories constructed from COCO.

In the dataset with replaced negative data, it is observed that GPT frequently substitutes the terms *"boat"* and *"elephant"* for the original COCO captions' *"airplane"* and *"giraffe"*, respectively. This pattern makes it straightforward for a text-only model to identify these captions as negative. In our *swap* data, the text-only classifier is able to identify captions with colorful animals interacting with a black object as negative. Note that these negative prompts are logically coherent and grammatically correct, thus cannot be easily detected by a grammar model. Conversely, the top accurately predicted positive captions often depict common activities like surfing and motorcycle riding, which align closely with the typical content style of COCO. It's important to recognize that data bias is not solely a characteristic of our method for generating data nor the manner in which we prompt GPT. In the experimental section, we demonstrate that such distribution discrepancy is also present in other approaches [14] that employ GPT for creating negative prompts.

The presence of bias in the data can obstruct a vision-language model's ability to truly understand image structures and learn language compositions, as the model might rely solely on textual cues for making predictions. To address this issue, we propose a straightforward solution aimed at reducing data distribution differences by selectively removing data that is predicted with high confidence by a text-alone model.

Figure 2 (bottom) depicts our conceptual goal of data filtering: to eliminate straightforward or biased positive and negative samples. This is done to ensure that the remaining text data cannot be distinguished by any text classifier, based solely on the text information. Essentially, our aim is to maximize the entropy of the text information within our dataset. In practice, we organize our dataset into N partitions. For each iteration, one partition is designated as the test set while the remaining $N-1$ partitions serve as the training set. We employ a pretrained Bert model [7] as our text-only classifier, and train it on the training set. Subsequently, this trained classifier is applied to the designated test set. We then rank the correctly predicted samples by the classifier's confidence level, removing the top $k\%$ of these samples for both positive and negative class predictions. The rest of the data is retained as our refined dataset. This procedure is repeated for each partition, ensuring a comprehensive reduction of bias across the dataset.

Note that while our observation and data filtering approach is applied to address the image-text alignment issue, it could be a more general problem for any multimodal data scenario where bias in one modality might negatively impact model training.

3.3 Finetuning VLMs for Image-Text Alignment Scoring

Once we have the unbiased data, we opt to fine-tune a vision-language model (VLM) to enhance its language compositional understanding capabilities with respect to images. In our study, we mainly select LLaVA-1.5 [24] due to its superior performance in image and text understanding. Since LLaVA-1.5 is designed to generate text, to adapt it for use as a image-text score calculator, we employ

the following prompt formatting: *"Does this image I match the following caption T. Answer Yes or No directly."* Given that LLaVA relies on a language model to produce subsequent words, we manually extract the logits associated with the responses *Yes* and *No* for the next word. We then define the matching score as:

$$\frac{e^{\mathbf{P}(Yes|prompt)}}{e^{\mathbf{P}(Yes|prompt)} + e^{\mathbf{P}(No|prompt)}} \tag{1}$$

We discover that this straightforward approach is quite effective and can already outperform many existing state-of-the-art baselines. However, LLaVA-1.5 was not specifically trained for this type of matching problem. Therefore, to enhance its performance, we finetune it with the same prompt formatting introduced above using our curated data. We assign the labels *Yes* and *No* to positive and negative pairs, respectively.

Our curated dataset is not restricted to LLaVA-1.5. In our experiments, we also finetune a Q-former equipped with the Image-Text Matching (ITM) head in BLIP2 [18]. The ITM head is essentially a binary classification layer, identical to our dataset's structure. We thus fine-tune the model using the standard cross-entropy loss.

4 Results

We evaluate our fine-tuned LLaVA-1.5 model against various baselines across different datasets. Additionally, we conduct ablation studies to evaluate the impact and importance of our dual-strategy approach for generating diverse negative captions and our method for addressing data distribution discrepancies. Since our main model is built upon LLaVA-1.5, we name our score as LLaVA-score.

4.1 Baselines

We evaluate our model against a range of leading multimodal models:
CLIP-ViT-L/14, the largest variant of OpenAI CLIP models [28].
BLIP-2 [18], which incorporates both image-text matching and image-text contrastive learning approaches.
NegCLIP, as introduced in [40], fine-tuned on challenging negative captions generated by randomly shuffling words within sentences.
VisualGPTScore [22] proposes utilizing the probability of generating specific text given an image (i.e., $P(T|I)$) as an effective metric for calculating image-text alignment scores. For this, we utilize LLaVA-1.5 [24], the state-of-the-art vision-language model, to calculate their proposed VisualGPTScore, employing a reweighting technique as suggested by [22].
Image-Reward [38]: A reward model, trained on human preferences of image-text pairs produced by Text-to-Image (T2I) models in DiffusionDB [37].
VQ2 [39] first extracts a set of candidate QA pairs from the text, then uses a VQA model to score each pair. For this baseline, we report results using both the PaLI model [6]—directly citing their paper as the model is not publicly

available—and BLIP-T5 [18] which is accessible via their official GitHub repository [3]. Note that while TIFA [15] similarly utilizes a QA pairs approach, we exclude it from our results due to VQ2 demonstrating superior performance, aiming for simplicity in our comparison.
PaLI fine-tuned on the SeeTrue dataset [39]: A version of PaLI specifically fine-tuned for the alignment task, using a curated dataset. Performance metrics are cited directly from the original paper [39].
LLaVA-1.5 [24], Originally a text generation model, we employ a specific prompt as introduced in Sec. 3.3. By using Equation 1, we convert it into a scoring function. Through empirical testing, this prompt is found to be the most effective, and it is utilized to finetune our model.

4.2 Datasets and Metrics

We evaluate on the following datasets and metrics:
Winoground [36]: This dataset uniquely comprises quartets, each including two images and two texts, necessitating a nuanced interpretation of both linguistic and visual elements for accurate matching. The metrics reported for this dataset encompass an image score, a text score, and a group score.
SeeTRUE [39]: A benchmark designed for assessing vision-language models, featuring a test set combining multiple sources. Current sources include Drawbench, EditBench, and COCO-t2i, pairing real texts with synthetic images. Following [39], we utilize ROC-AUC as the evaluation metric.
SugarCrepe [14]: A recently introduced benchmark focuses on generating creative negative captions using a language model and employs grammar and commonsense models for data cleaning, marking a novel approach in benchmark design.
MagicBrush [41]: This benchmark facilitates human-driven image editing, constructed using Dall-E 2, and includes captions for both the original and edited images. Featuring quartets similar to Winoground [36], we adopt the same methodology for calculating the group score with modification due to nature of data. Refer supp for details.

4.3 Implementation Details

For our dataset curation, we select COCO, which provides positive image-text pairs. To diversify our image dataset and enhance our model's robustness to synthetic images, we incorporate a subset of the training images from SeeTRUE [39], specifically the *coco_train_t2i* set. To construct negative data, we use GPT to generate two types of negative captions (*swap* and *replace*), and we perform random sampling to maintain an equal amount of positive and negative data. In the process of curating our dataset, we ensure that neither images nor captions included in the training dataset are present within the test dataset.

For our model development, we choose to refine LLaVA-1.5 [24] using our curated data. We train the model with a batch size of 64 on 8 NVIDIA A100 GPUs for a single epoch, setting the learning rate at 2e-6. For our data filtering, we use k to be 30% and N to be 5. Find more details in the supp.

Table 1: We evaluate our model alongside baselines across multiple datasets. The best results are shown in bold. For three subsets of the SeeTRUE dataset, we present the ROC AUC scores following the original paper [39]. For the SugarCrepe dataset, accuracy is employed as the performance metric. For the MagicBrush [41] dataset, we report the group score. We use LLaVA-1.5 to calculate the VisualGPT score [23].

	Winoground			SeeTRUE			SugarCrepe			MagicBrush
	image	text	group	DrawBench	EditBench	COCO-T2I	replace	swap	add	
Chance Performance	25.00	25.00	16.67	50.0	50.0	50.0	50.0	50.0	50.0	33.33
CLIP-ViT-L-14 [28]	10.50	28.50	7.75	61.4	62.1	59.2	79.4	61.4	74.8	52.89
NegCLIP [40]	11.75	30.75	8.25	63.2	66.0	62.8	85.3	75.3	87.2	61.12
BLIP2-ITM [18]	24.25	41.75	19.00	60.8	67.5	68.0	88.9	83.9	91.8	75.32
BLIP2-ITC [18]	12.00	28.50	8.50	64.9	67.9	69.9	86.7	66.9	92.3	67.85
Image-Reward [38]	15.25	43.00	12.75	70.4	70.2	77.0	88.2	81.0	95.2	70.28
VisualGPT [23]	37.00	44.25	27.50	77.0	74.2	69.1	88.2	87.1	95.5	78.31
VQ2 (PaLI) [39]	42.25	47.00	30.50	82.6	73.6	83.4	-	-	-	-
VQ2 (BLIP-T5) [39]	34.00	33.50	23.25	74.8	67.4	74.2	83.3	81.0	90.9	70.46
PaLI (ft on SeeTRUE) [39]	38.00	46.50	28.75	86.8	77.2	83.2	-	-	-	-
LLaVA-1.5 [24]	49.75	51.00	34.25	86.9	**78.3**	84.5	93.5	88.3	95.8	82.61
LLaVA-score (Ours)	**68.00**	**53.75**	**47.25**	**88.8**	77.7	**84.9**	**95.3**	**94.9**	**97.5**	**87.28**

4.4 Main results

Table 1 presents a comparison between our models and various strong baselines across different datasets. It is evident that our model outperforms others in nearly all datasets. Interestingly, utilizing Eq. 1, LLaVA-1.5 zero-shot performance is the second-best method overall. Particularly for Winoground, a benchmark well-known for its challenges in visual and linguistic reasoning, some models such as CLIP, BLIP2, and Image-Reward perform at or below chance level. Our fine-tuned LLaVA-1.5 substantially enhancing its reasoning capabilities showing the importance of our strategy for training data curation. For VQ2, the VQA models (BLIP or PaLI) underperform our model due to their lack of training on challenging negative examples. The finetuned PaLI model also falls short, attributed to its training dataset's lack of diversity and absence of data filtering. Our model also shows impressive results on synthetic image benchmarks like SeeTRUE and MagicBrush.

4.5 Performance on Attribute, Counting, Spatial Reasoning

In the preceding section, we demonstrated that our model surpasses baseline models across various datasets. This section offers an alternative perspective on our model's performance, particularly in areas where both vision-language models and recent text-to-image (T2I) models encounter significant challenges. These challenges include multiple attributes binding, counting objects, and understanding spatial relationships between objects. Addressing these issues is crucial for both visual understanding and evaluating the capabilities of generative models.

To quantify the improvements our model achieves in addressing these common challenges, we construct three specialized datasets. Specifically, we prompt GPT to generate scenarios involving attribute binding, object counting, and spatial relationships, providing in-context examples to guide the generation process.

Table 2: Comparison of our model against baselines across three prevalent challenges, highlighting the highest average accuracy in bold.

	Attribute			Counting			Spatial		
	avg	pos	neg	avg	pos	neg	avg	pos	neg
Chance Performance	50	50	50	50	50	50	50	50	50
Threshold-Independent Models (with oracle)									
CLIP-ViT-L-14 [28]	63	52	74	58	68	48	53	18	88
NegCLIP [40]	65	78	52	59	66	52	57	48	66
BLIP2-ITC [18]	66	72	60	57	36	78	57	90	24
VisualGPT [5]	73	90	56	65	52	78	62	56	68
Inherent Decision Models									
BLIP2-ITM [18]	58	100	16	53	96	10	51	100	2
Image-Reward [38]	70	100	40	61	100	22	57	98	16
VQ2 (BLIP-T5) [39]	66	94	38	56	82	30	54	94	14
LLaVA-1.5 [24]	71	98	44	62	96	28	57	98	16
LLaVA-score (Ours)	**81**	90	66	**71**	86	56	**81**	76	86

Subsequently, the T2I model [35] is employed to generate 50 images for each prompt. From these, we manually select one positive and one negative image based on image-text alignment. We opt for synthetic images as it offers greater flexibility, as illustrated in Figure 5, it can incorporate diverse styles (e.g., painting styles for birds), create unconventional images (e.g., the examples involving soccer and cats), and has attribute/object merging capabilities of T2I models to enrich our dataset with varied negative cases (e.g., the negative example of mistakenly combining clock and apple).

Table 2 presents our results. We report classification accuracy as the metric, alongside showing both positive and negative results in gray color. We categorize the baseline models into two distinct groups for comparison purposes. 1) Upper section of the table, consists of models that generate scores without an established classification threshold. Consequently, we calculate the accuracy by using an oracle threshold, which involves optimizing over all predicted scores to find the best cutoff (cheating on the test set). 2) Lower part of the table, operates with an inherent decision boundary for classification.

Our model demonstrates superior performance across all three challenging cases, beating competitors that perform only marginally better than random guessing, particularly on tasks requiring spatial understanding. It's noteworthy that many models exhibit a bias, often predicting the same outcome across the test dataset. This tendency is underscored by their high accuracy on positive samples contrasted with poor performance on negative ones. Given that our images are synthesized using text-to-image (T2I) models, essential linguistic elements are likely to be present even in images labeled as negative (refer to Figure 5). This observation suggests that baseline models operate similarly to a bag-of-visual-words approach, predicting an image as matching the textual description as long as it contains certain key visual concepts.

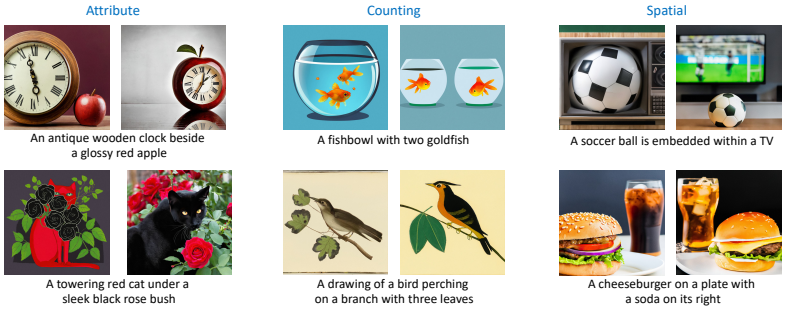

Fig. 5: Our curated test datasets feature captions paired with one positive image and one negative image each. All the positive images are displayed on the left side.

4.6 Ablation Studies

Importance of different negative data. Table 3 begins with our baseline model, LLaVA-1.5 [24], outlining its performance. The following two rows illustrate the effects of finetuning the LLaVA-1.5 model with only *replace* and *swap* negative data, with filtering processes applied. As shown in the subcategories of the SugarCrepe dataset, *replace* and *swap* data enhance performance in their respective categories but do not significantly impact other cases. Nevertheless, our final model excels across almost all tests, indicating that combining *replace* and *swap* negative data has a complementary effect, particularly noticeable in the *add* category of the SugarCrepe dataset and other evaluated datasets.
Importance of data filtering. The subsequent row presents outcomes from merging two types of negative data without any filtering. This combination has shown to be beneficial for specific datasets, like Winoground and SugarCrepe, although the results don't match those of the final model with filtering. However, biases in the text data cause the fine-tuned model to show no improvement on datasets like MagicBrush and SeeTRUE. Since data filtering reduces the amount of data, the fifth row displays results from randomly subsampling a training set to match the quantity of filtered data. This comparison highlights the effectiveness of our filtering technique as without it, the model may get biased to just rely on language distribution and ignore the image when making predictions.
Effect of data filtering. Fig. 6 shows the impact of removing the top k% of biased data on the performance across four test datasets. Here, k varies from 0 to 90, representing the progression from no data removal to the exclusion of 90% of the data. The observed trend indicates that as the biased data is progressively removed, performance improves, peaking at approximately 30% to 40%. Beyond this point, performance declines due to the diminishing volume of training data.

We assess the quality of our filtered data by training a text-only model on 80% of the filtered data and testing it on the remaining 20%. Note that this evaluation is performed on filtered data, and a model should reach an accuracy close to 50% in ideal data indicating it cannot distinguish between positive and negative caption just using text. We vary the parameter k from 0% to 90%, in increments of 10%, resulting in corresponding filtered data quality percentages

Table 3: Upper: Ablating our training data based on LLaVA-1.5 [24]. **Lower:** Fine-tuning a BLIP2 model with our data.

	Winoground			SeeTRUE			SugarCrepe			MagicBrush
	image	text	group	DrawBench	EditBench	COCO-T2I	replace	swap	add	
LLaVA-1.5 [24]	49.75	51.00	34.25	86.9	78.3	84.5	93.5	88.3	95.8	82.61
Only replace w/ filter	51.25	**54.25**	38.50	87.4	77.6	83.7	95.2	88.8	95.3	85.70
Only swap w/ filter	63.50	49.25	42.25	81.0	74.0	79.2	92.6	95.5	91.3	76.63
w/o filter	65.75	51.50	46.75	88.4	76.4	81.1	94.5	91.7	95.0	82.99
Random subsample	64.25	50.00	43.25	88.2	76.8	81.8	94.5	92.0	93.6	83.12
LLaVA-score (Ours)	**68.00**	53.75	**47.25**	**88.8**	**77.7**	**84.9**	**95.3**	**94.9**	**97.5**	**87.28**
BLIP2-ITM	24.25	41.75	19.00	60.8	67.5	68.0	88.9	83.9	91.8	75.32
BLIP2-ITM (ft on our data)	**39.5**	**42.75**	**28**	**87.5**	**76.2**	**82.8**	**94.3**	**91.4**	**96.0**	**79.62**

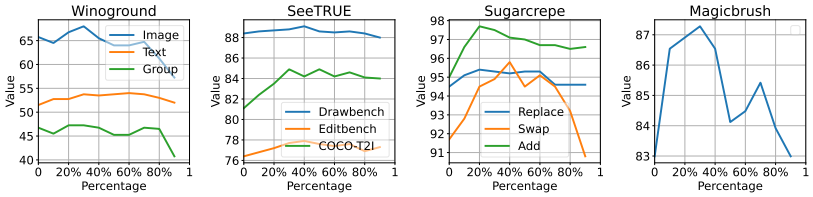

Fig. 6: Ablation on filter percentage of data. For most of the data, performance peaks around 30% and reduces after that due to decrease in training data size.

of 75.9%, 68.2%, 60.7%, 56.4%, 53.8%, 51.3%, 50.6%, 51.0%, 49.8%, and 47.9%. Empirically, we find that data quality below 60% is satisfactory for our purposes. Consequently, we decide to discard 30% of the data in our experiment as it's a good trade-off between quality of filtered data and amount of training data.

To demonstrate that the distribution gap we identified also exists in datasets created by others, we apply the same evaluation method to SugarCrepe [14], a test benchmark also generated by GPT, that was refined using grammar and common sense models. Nevertheless, it still exhibits an accuracy of 69.0%, indicating the presence of the bias.

Generalization on other model. The bottom section of Table 3 shows that when BLIP2 is fine-tuned with the ITM head (a binary classification head) using our curated data, there is an observable improvement in performance. It's important to highlight that we only finetune and use the Q-former in BLIP2 without including any LLMs, and this model surpasses nearly all baselines in benchmarks that do not utilize LLMs (refer table 1). As expected, the fine-tuned BLIP2 model does not outperform the model fine-tuned using LLaVA, particularly on the Winoground dataset, which is recognized for its challenging reasoning tasks. This suggests the importance of utilizing a more advanced vision-language model as the foundation model for achieving better results. Nonetheless, given that BLIP2's model size (180M) is significantly smaller compared to LLaVA-1.5 (13B), it offers a more lightweight option for simpler applications.

4.7 Application: Text Alignment Ranking for Image Generation

In this subsection, we showcase an application of our model for image generation. A major challenge with T2I models is their inconsistent adherence to complex

Fig. 7: We show T2I generation results ranked by our model according to image-text alignment. The most aligned images are ranked higher.

Table 4: Performance evaluation of ranking correlation on the TIFA [15] dataset.

	CLIP-ViT-L-14 [28]	Neg CLIP [40]	BLIP2-ITM [18]	BLIP2-ITC [18]	Image Reward [38]	Visual GPT [23]	VQ2 (T5) [39]	TIFA [15]	LLaVA-1.5 [24]	Ours
Spearman ρ	28.5	40.1	43.3	51.5	62.8	36.9	55.1	59.7	61.5	**64.5**
Kendall τ	19.5	28.7	29.9	36.9	46.3	28.7	41.5	47.2	45.8	**49.8**

compositional prompts [10]. Inspired by [16], one can generate a series of images from a T2I model and then employ our model to rerank them based on relevance, ultimately presenting the highest-ranked images to the user.

TIFA [15] introduced a benchmark with images generated using different T2I models from identical prompts with human rankings. Our evaluation on this dataset reveals our model's superior alignment with human rankings, as shown in Table 4. Spearman's ρ and Kendall's τ are statistical measures that evaluate the similarity between two ranking orders, with higher values indicating greater alignment. Figure 7 illustrates qualitative ranking results using our model for the generation results from a T2I model [35].

5 Conclusion

In this paper, we introduced an innovative method for generating high-quality training data for image-text alignment models. By implementing a mixed-type negative caption creation strategy and a novel filtering mechanism, we ensured a balanced distribution between positive and negative captions. This approach not only rectified the biases in existing models but also significantly enhanced the performance of our fine-tuned visual-language models, which achieve state-of-the-art results for image-text alignment. Our findings underscore the importance of high-quality, balanced training data in improving visual-language compositional reasoning and alignment. They also open up new avenues for exploring how these methods can be applied to other tasks or modalities in addition to the visual-language domain. In conclusion, our work offers an effective approach to improving image-text alignment, paving the road for future alignment studies.

This work was supported in part by NSF CAREER IIS2150012, Adobe Data Science award, Microsoft Accelerate Foundation Models Research Program, and Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration) and (No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training).

References

1. 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. NeurIPS **35**, 23716–23736 (2022)
2. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., Ramesh, A.: Improving image generation with better captions (2023), https://cdn.openai.com/papers/dall-e-3.pdf
3. Bitton, Y.: WYSIWYR repository. https://github.com/yonatanbitton/wysiwyr (2023), accessed: 2024-02-24
4. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021)
5. Chen, J., Guo, H., Yi, K., Li, B., Elhoseiny, M.: Visualgpt: Data-efficient adaptation of pretrained language models for image captioning (June 2022)
6. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A.J., Padlewski, P., Salz, D.M., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A.V., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B.K., Riquelme, C., Steiner, A., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: Pali: A jointly-scaled multilingual language-image model. ArXiv **abs/2209.06794** (2022), https://api.semanticscholar.org/CorpusID:252222320
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Doveh, S., Arbelle, A., Harary, S., Herzig, R., Kim, D., Cascante-Bonilla, P., Alfassy, A., Panda, R., Giryes, R., Feris, R., et al.: Dense and aligned captions (dac) promote compositional reasoning in vl models. Advances in Neural Information Processing Systems **36** (2024)
9. Doveh, S., Arbelle, A., Harary, S., Schwartz, E., Herzig, R., Giryes, R., Feris, R., Panda, R., Ullman, S., Karlinsky, L.: Teaching structured vision & language concepts to vision & language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2657–2668 (2023)
10. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A.R., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PUIqjT4rzq7
11. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next

generation of multimodal datasets. Advances in Neural Information Processing Systems **36** (2024)
12. Herzig, R., Mendelson, A., Karlinsky, L., Arbelle, A., Feris, R., Darrell, T., Globerson, A.: Incorporating structured representations into pretrained vision & language models using scene graphs. EMNLP (2023)
13. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7514–7528. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.595, https://aclanthology.org/2021.emnlp-main.595
14. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In: Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
15. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. ICCV (2023)
16. Karthik, S., Roth, K., Mancini, M., Akata, Z.: If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection. arXiv preprint arXiv:2305.13308 (2023)
17. Le, T., Lal, V., Howard, P.: Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. Advances in Neural Information Processing Systems **36** (2024)
18. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (2023), https://api.semanticscholar.org/CorpusID:256390509
19. Li, J., Li, D., Xiong, C., 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)
20. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. EMNLP (2023)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., 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)
22. Lin, Z., Chen, X., Pathak, D., Zhang, P., Ramanan, D.: Revisiting the role of language priors in vision-language models. arXiv preprint arXiv:2306.01879 (2023)
23. Lin, Z., Chen, X., Pathak, D., Zhang, P., Ramanan, D.: Visualgptscore: Visio-linguistic reasoning with multimodal generative pre-training scores. arXiv preprint arXiv:2306.01879 (2023)
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv:2304.08485 (2023)
26. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? arXiv preprint arXiv:2212.07796 (2022)

27. OpenAI: Gpt-4v(ision) system card (2023), https://cdn.openai.com/papers/GPTV_System_Card.pdf
28. 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)
29. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
30. Ray, A., Radenovic, F., Dubey, A., Plummer, B.A., Krishna, R., Saenko, K.: Cola: How to adapt vision-language models to compose objects localized with attributes? (2023)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
32. 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. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
33. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1238, https://aclanthology.org/P18-1238
34. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650 (2022)
35. stabilityai: IF repository. https://github.com/deep-floyd/IF
36. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., 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)
37. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: Large-scale prompt gallery dataset for text-to-image generative models. arXiv:2210.14896 [cs] (2022), https://arxiv.org/abs/2210.14896
38. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024)
39. Yarom, M., Bitton, Y., Changpinyo, S., Aharoni, R., Herzig, J., Lang, O., Ofek, E., Szpektor, I.: What you see is what you read? improving text-image alignment evaluation. arXiv preprint arXiv:2305.10400 (2023)
40. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=KRLUvxh8uaX
41. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. In: Advances in Neural Information Processing Systems (2023)

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