Synthesize, Diagnose, and Optimize: Towards Fine-Grained Vision-Language Understanding

Wujian Peng[1,2] Sicheng Xie[1,2] Zuyao You[1,2] Shiyi Lan[3] Zuxuan Wu[1,2†]

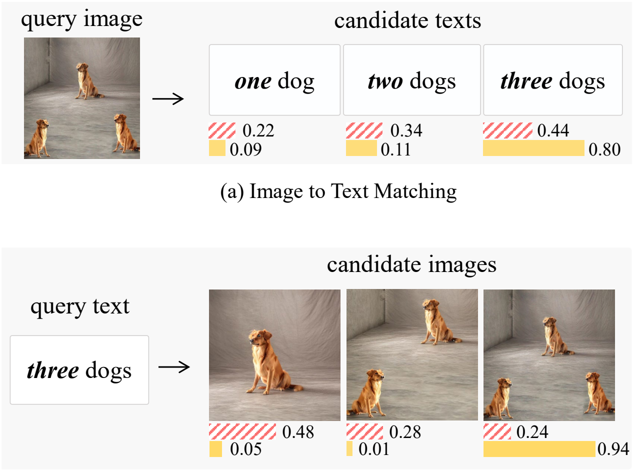
[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
[3]NVIDIA

Abstract

Vision language models (VLM) have demonstrated remarkable performance across various downstream tasks. However, understanding fine-grained visual-linguistic concepts, such as attributes and inter-object relationships, remains a significant challenge. While several benchmarks aim to evaluate VLMs in finer granularity, their primary focus remains on the linguistic aspect, neglecting the visual dimension. Here, we highlight the importance of evaluating VLMs from both a textual and visual perspective. We introduce a progressive pipeline to synthesize images that vary in a specific attribute while ensuring consistency in all other aspects. Utilizing this data engine, we carefully design a benchmark, SPEC, to diagnose the comprehension of object size, position, existence, and count. Subsequently, we conduct a thorough evaluation of four leading VLMs on SPEC. Surprisingly, their performance is close to random guess, revealing significant limitations. With this in mind, we propose a simple yet effective approach to optimize VLMs in fine-grained understanding, achieving significant improvements on SPEC without compromising the zero-shot performance. Results on two additional fine-grained benchmarks also show consistent improvements, further validating the transferability of our approach. Code and data are available at https://github.com/wjpoom/SPEC.

1. Introduction

Vision and Language Foundation Models (VLMs) pretrained on large-scale image-text data [13, 17, 23, 29, 41] have consistently demonstrated impressive performance across a wide range of well-established evaluating tasks, *i.e*. image classification [5], image captioning [1, 18], visual question answering [3] and cross-modal image-text retrieval [18, 38]. Their remarkable performance is gradually convincing the community that these currently available VLMs are almost robust and powerful enough to be transferred to a broad spectrum of downstream tasks, either through finetuning or even in a zero-shot manner.



(a) Image to Text Matching

(b) Text to Image Matching

Figure 1. **We conduct a symmetrical assessment of VLMs in fine-grained comprehension**, considering both linguistic and visual perspectives. The bars in ▨ and 🟨 represent the image-text matching scores for CLIP [23] and our method, respectively. It is evident that CLIP struggles with tasks related to quantity comprehension, whereas our method significantly enhances the model in understanding fine-grained details.

However, recent research has shattered this captivating illusion, revealing that even state-of-the-art VLMs [17, 23, 29, 41] exhibit significant limitations in understanding visual-linguistic concepts that require fine-grained compositional reasoning, especially in tasks involving object attributes or inter-object relationships [21, 31, 33, 40, 43]. This raises a crucial question: *to what extent and in what aspects are VLMs excelling or struggling?* To answer this, previous effort evaluates fine-grained capabilities of VLMs through the image-to-text matching task, as shown in Fig. 1(a). This involves providing a query image and retrieving the matching text from a set of confusing candi-

[†] Corresponding author.

dates, differing subtly in texts. For example, when assessing the counting ability, it is crucial to ensure that quantity is the unique variable and other clues are kept the same. Therefore, a straightforward way to do so is modifying quantity and adjective words used in the texts. While manipulating texts to construct confusing candidate sets has been well-studied due to the sparsity of the text space and advancements in Large Language Models (LLMs) [19, 21, 40, 43], the visual side remains relatively under explored, largely due to the complexity of visual signals and the absence of powerful tools. We posit that exploring the visual dimension in a fine-grained manner is also essential for a comprehensive understanding of VLMs.

Motivated by the great progress achieved in generative models, we present an effective framework for generating high-quality image candidates that are suitable for evaluating the performance of VLMs. This framework ensures that images within the same candidate set only differ in the specified property of interest, while all other properties remain consistent. We break down this task into several simple and manageable steps. As illustrated in Fig. 2, we start by utilizing a text-to-image model [22, 24] to generate images featuring a single object. Then, a segmenter [14, 15] is employed to separate the objects from their backgrounds, yielding a library of foreground instance spanning various categories. From there, we select instances and arrange them on a blank canvas (manipulating attributes such as size, position, existence, and quantity of a specific object at this stage is straightforward). Finally, we use an inpainting model [22, 24] to fill the missing background portions, producing an photo-realistic image. It is worth noting that, during the inpainting process, we design a progressive background filling strategy, effectively ensuring consistency in the background across all images in the same candidate set.

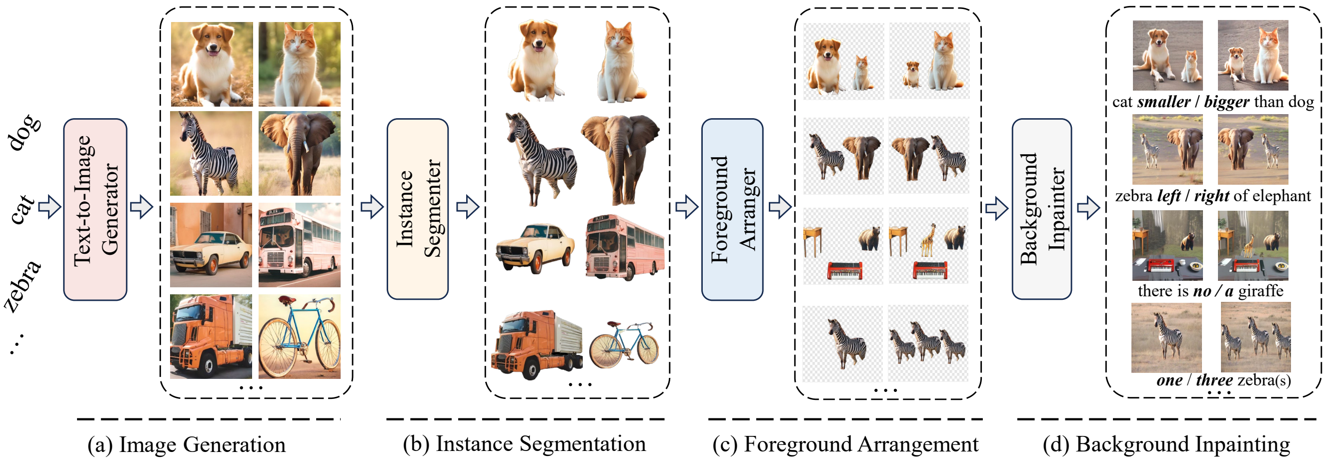
Empowered by this data construction pipeline, we carefully develop a new benchmark, named as **SPEC**, to evaluate the proficiency of VLMs in comprehending fine-grained concepts including **S**ize, **P**osition, **E**xistence and **C**ounting. We systematical test four VLMs [17, 23, 29, 41] on this newly created test bed. Surprisingly, even state-of-the-art models perform at chance-level, exposing significant performance deficiencies. Following this, we implement a straightforward approach to remedy this by incorporating hard negative examples (*i.e.*, confusing images or texts within the same candidate set) into the same training batch. This encourages the model to discern subtle differences among candidate examples, leading to a significant improvement in performance on SEPC while preserving the original zero-shot capability. Furthermore, to demonstrate the model's generalization ability, we conduct additional tests on two existing datasets [33, 40], which also focus on compositional reasoning. The consistent improvement further validates that our method effectively guide the model to acquire essential and transferable comprehending abilities at a finer granularity. Our main contributions are:

1. **A progressive data constructing pipeline**. We present a progressive data construction pipeline designed for creating a candidate image set. Within each candidate set, images vary exclusively in a specified attribute while ensuring consistency across other aspects. Such data are valuable for conducting text-to-image matching tasks, as depicted in Fig. 1(b), offering a visual perspective for evaluating VLMs.
2. **A carefully curated benchmark: SPEC**. We meticulously craft a novel benchmark, SPEC, with a specific focus on evaluating VLMs' understanding of fine-grained visual-linguistic concepts, encompassing object size, position, existence, and count. The introduction of SPEC enables a symmetrical evaluation of VLMs from both image and text perspectives, addressing the previous lack of image-centric testing data.
3. **A simple and effective remedy**. We evaluate four VLMs on SPEC, revealing significant limitations. In response, we propose a method to enhance the understanding of fine-grained visual-linguistic concepts. Experimental results indicate notable improvement not only on SPEC but also consistent results on two additional datasets, while preserving zero-shot capability.

2. Related Work

Vision and Language Models (VLMs). Models such as CLIP [23], ALIGN [13], CyCLIP [10] CoCa [39], OmniVL [32] and Open-VCLIP [34, 35] have demonstrated impressive performance across a wide range of downstream tasks. These models include two separate unimodal encoders, each designed to extract representations for visual and textual input. To achieve alignment between the two modalities, they typically employ a huge number of visual-textual pairs for contrastive learning. By pretraining on 400M noisy data, CLIP [23] achieves a top-1 accuracy on ImageNet-1K [5] comparable to that of ResNet-50 [11], even though it is never specifically trained on ImageNet and is evaluated in a zero-shot manner. However, as highlighted in recent work [31, 40], these advancements are primarily attributed to the simplicity of evaluation tasks which requires no reasoning or compositional capabilities. The performance of these models are limited on tasks that require fine-grained understanding [21, 40].

Benchmarking VLMs in Finer Granularity. To assess the model's understanding of nuanced visual-linguistic concepts, several new benchmarks have been proposed. Winoground [31] is curated by experts with a focus on compositional understanding. VALSE [21] and VL-Checklist [43] investigate several linguistic phenomena by transforming real captions into confusing alternatives.



(a) Image Generation (b) Instance Segmentation (c) Foreground Arrangement (d) Background Inpainting

Figure 2. **The overall illustration of our data progressive construction pipeline.** We initiate the process by generating a batch of images containing a single object. Subsequently, we extract the object from the background in the images. Following that, we arrange the background-free images on a blank canvas according to specifications (with control over attributes). Finally, we meticulously fill in the missing background, ensuring consistency across candidates.

ARO [40] diagnoses VLMs in attribution, relation and ordering. Eqben [33] assesses whether the model is sensitive to visual semantic changes. Most existing benchmarks solely focus on subtle textual changes [12, 19, 21, 40, 43], as crafting confusing text candidates is straightforward that can be achieved through either LLMs or simple rules. Winoground [40] and Eqben [33] are most relevant to us, since we focus on minimal semantic changes in both images and texts, enabling a more comprehensive evaluation across modalities. However, the scale of Winoground is restricted by its costly curation, and the image diversity of Eqben is limited by virtual engines. In contrast, our data construction pipeline is scalable and can produce diverse images.

Enhancing VLMs for Fine-grained Understanding. To mitigate challenges in fine-grained recognition, various approaches have been explored. Syn-CLIP [2] utilize data synthesized by 3D simulation engines to enhance the model's understanding of concepts beyond nouns. EQSIM [33] incorporates an additional regularization loss to generalize VLMs to nuanced multimodal compositions. TSVLC [7] and ViLEM [4] introduce negative texts generated by LLMs [6, 26] to inject fine-grained knowledge. Construct-VL [30] addresses these challenges from a continual learning perspective. These methods compel the model to focus on subtle differences by introducing confusing texts as hard negatives. However, we argue that the absence of visual hard negatives limits its performance. Thus, we introduce hard negatives for both modalities, simultaneously enhancing the visual and textual encoders.

3. Synthesize: Data Construction Pipeline

Our goal is to build a set of perplexing image, wherein each image differs solely in a specified attribute while ensuring consistency in all other aspects. We first emphasize the importance of preserving consistency among candidate images for effective evaluation (Sec. 3.1). To address this, we break down this problem and introduce a progressive data construction pipeline (Sec. 3.2). Then, we carefully devise a benchmark that centers on evaluating VLMs' grasp of fine-grained visual-linguistic concepts (Sec. 3.3).

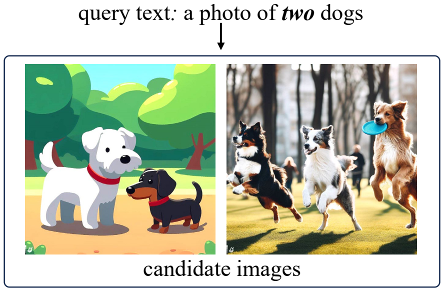


Figure 3. **Ensuring consistency among candidates is crucial to avoid ambiguity.** The images above not only differ in quantity but also show a significant variation in the appearance of the objects. Consequently, attributing the model's correctness or errors solely to the understanding of quantity is not convincing.

3.1. Importance of Candidate Consistency

As illustrated in Fig. 3, when conducting a matching task with the textual query "a photo of two dogs", the model might mistakenly choose the image on the right (which actually contains three dogs). However, attributing this error solely to counting difficulty is not convincing. The model might select the right-side image due to its more photo-realistic appearance or better alignment with the word "dog", as the left-side image has a cartoon style. Conversely, if the model correctly selects the right-side image for the query "a photo of three dogs", we also cannot assert that the model is proficient in counting. These obscure ambiguities arise because there is no guarantee of the uniqueness of changing factors among candidate images during evaluation. Therefore, ensuring consistency among candidates in all aspects except the one under investigation is crucial. To this end, we propose a progressive data construction method, which will be elaborated in detail as follows.

3.2. Progressive Data Construction

The data construction framework is illustrated in Fig. 2, comprising four progressive steps. Initially, we generate images featuring a single prominent object. Then, we isolate the foreground from the background, resulting in a library of foreground objects spanning various categories. Subsequently, we select objects from this library and arrange them on a blank canvas, adjusting their attributes and relationships. Lastly, we carefully fill in the missing background, ensuring consistency among different candidates.

3.2.1 Generating Images with Single Objects

We initiate the process by utilizing a generation model to obtain a collection of images, each featuring a single and prominent object corresponding to a specific category. Due to the progress in visual generation models [24, 25, 36, 37], these images display a high level of photo-realistic and diverse content. In practice, we use Stable-Diffusion-XL 1.0 [22] as our generator, and prompt it with "a photo of a single and fully visible [class name]". The emphasis on "single and fully visible" is crucial, as the model might otherwise generate images with multiple objects or encounter occlusion issues, as observed in [42]. The [class name] represents a specific category from 80 classes of COCO [18].

3.2.2 Isolating Objects from the Background

For ease in subsequent processes, we need to separate the objects from the backgrounds where they are embedded. To accomplish this, we first utilize an open-set detector, Grounding-DINO [20] to outline the regions containing the objects. Subsequently, we prompt SAM [15] with this bounding box as to obtain the final segmentation results. Thus far, we have established a library containing instances from various categories. These instances are background-free, allowing for composition on a blank canvas, while their attributes and relations can be controlled.

3.2.3 Arranging Objects on the Canvas

Recall that our goal is to manipulate a specific visual-linguistic concept of an image. Currently, this task appears straightforward when we exclude the background from consideration. We can retrieve instances from the library and arrange them on a blank canvas according to our specifications. For example, we can flexibly control the quantity of an object through duplication operations, modify their sizes via resizing, and determine whether an object exists and specify the position of an existing object. This process resembles the concept of copy-paste [9, 42], with a notable distinction: we paste objects onto a blank background and placing emphasis on controlling properties such as the size, position, existence, and quantity of each individual object.

3.2.4 Infilling the Missing Background

We have constructed images with differences in specified attributes, however, they currently lack a suitable background. To fill the missing area, we employ a inpainting model [22] which demonstrates proficiency in filling large holes. It is worth noting that generating backgrounds individually for each image would result in significant differences in the backgrounds, posing a challenge to maintaining consistency among candidate images. To overcome this, we introduce a strategy where images within the same candidate set share a common and consistent background during the inpainting process. As depicted in the upper part of Fig. 4, to present the giraffe in different positions, we start by surrounding it with an initial background. Then, we relocate this initial image on the canvas as required, and fill the remaining blank space. Similarly, in the lower part, we first embed the zebra into a reasonable environment. Following that, we expand the scene horizontally, introduce the elephant, and fill the blank areas through inpainting. In summary, we begin the process by generating an initial background, which is then expanded to the surroundings, effectively ensuring consistency among the candidate images. Additional examples, such as adjusting the size or quantity of an object, can be found in the supplementary.

3.3. SPEC Benchmark

Utilizing the data engine outlined in Sec. 3.2, we carefully devise the SPEC benchmark with the goal of assessing the performance of VLMs in comprehending object size, position, existence and count. An overview of the SPEC benchmark is presented in Fig. 5. SPEC contains six subsets, which will be elaborated as follows:

Absolute Size reflects how large an object is in comparison to the entire image. We categorize this into three level: large, medium, or small, and define them following:

$$\text{Size}_{abs.}(x) = \begin{cases} \texttt{small}, & P \leq 0.2 \\ \texttt{medium}, & 0.4 \leq P \leq 0.6 \\ \texttt{large}, & P \geq 0.8, \end{cases} \tag{1}$$

where P denotes the proportion of the space occupied by the object relative to the area of the entire image. A safety

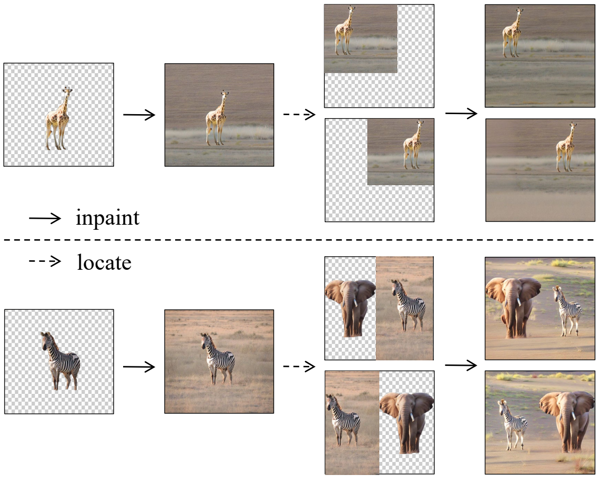


Figure 4. **Consistent background inpainting strategy.** We first generate an initial background shared by all candidate images. Then, we expand around this region, ensuring consistency in the backgrounds of different images.

threshold is deliberately introduced between these three levels to prevent ambiguity.

Relative Size focuses on the size relationship between two objects. It is categorized as follows: object A is [smaller than, equal to, larger than] object B, and we measure this following:

$$\text{Size}_{rea.}(A, B) = \begin{cases} A\,\texttt{smaller than}\,B, & R \leq 0.5 \\ A\,\texttt{equal to}\,B, & 0.9 \leq R \leq 1.1 \\ A\,\texttt{larger than}\,B, & R \geq 2, \end{cases} \tag{2}$$

where $R = \frac{S_A}{S_B}$ is the ratio of the areas of object A and object B.

Absolute Position signifies the location of an object relative to the image. We partition the image into a 3×3 grid, defining nine possible positions: top-left, top, top-right, left, center, right, bottom-left, bottom, and bottom-right. The absolute position of an object is determined based on the grid in which its center point resides.

Relative Position describes the spatial relationship between two objects. We consider four common spatial relationships: A is [to the left of, to the right of, above, below] B. The position relationship between objects is defined based on the relative positions of their center points.

Existence indicates whether an object appears in a given image, expressed using existential quantifiers: There is [no, at least one] object in the image.

Count represents the number of occurrences of an object, providing a metric for the model's quantitative understanding. Due to potential occlusion issues with a large number of objects, we restrict our consideration to the range of 1 to 9: “there are [one, two, $\cdots$, nine] object(s) in the image”.

Benchmarks	Visual HN	Scalability	Text-realistic	Photo-realistic	#Candidates
VALSE [21]	✗	✓	✓	✓	2
ARO [40]	✗	✓	✗	✓	2
Winoground [31]	✓	✗	✓	✓	2
Eqben [33]	✓	✓	✓	✓✗	2
SPEC(Ours)	✓	✓	✓	✓	2-9

Table 1. Comparison of SPEC with other fine-grained benchmarks.

Data Format. The basic unit of SPEC is an individual test case, wherein each test case comprises two components: an image candidate set, which differs only in certain visual aspects, and a text candidate set, which differs only in the corresponding language descriptions. We formally represent a test case as:

$$\mathbb{T} : (\mathcal{I} = \{I_1, \cdots, I_K\}, \mathcal{T} = \{T_1, \cdots, T_K\}), \tag{3}$$

where $\mathcal{I}$ and $\mathcal{T}$ represent the image and text candidate sets, respectively. The i-th image I_i is paired with the i-th text T_i, *i.e.*, they mutually describe each other. K is the semantic cardinality of the test case, determined by the definition of each subset. For instance, if a test case belongs to the absolute size subset, then $K = 3$ (representing the three semantics: large, medium, small). In Fig. 5, we present examplar test cases for each subset, and more examples can be found in the supplementary.

Comparing SPEC with other benchmarks. In Tab. 1, we conduct a comprehensive comparison of SPEC with four similar benchmarks. **Visual HN** indicates the presence of image hard negatives, which are essential for the text-to-image matching task. Notably, VALSE [21] and ARO [40] concentrate solely on text hard negatives, neglecting their visual counterparts. **Scalability** assesses whether the data construction method can be scaled up. For instance, Winoground [31] is limited to 400 examples due to high costs for manual collection. Additionally, **Text-realistic** evaluates the grammatical correctness of the texts, where ARO involves directly swapping the positions of two words without considering grammar. Similarly, **Image-realistic** indicates the realism of the images. Eqben [33] incorporates images rendered using a virtual engine, compromising their quality. In contrast, images from SPEC, while synthesized, leverage advanced image generation models and our effective data construction pipeline, resulting in photo-realistic images. Finally, we compare the number of **Candidates** in each test case. In all datasets except SPEC, each example features only two candidates, *i.e.*, identifying the correct item from two candidates, which is relatively straightforward. In contrast, tasks in SPEC are more challenging,

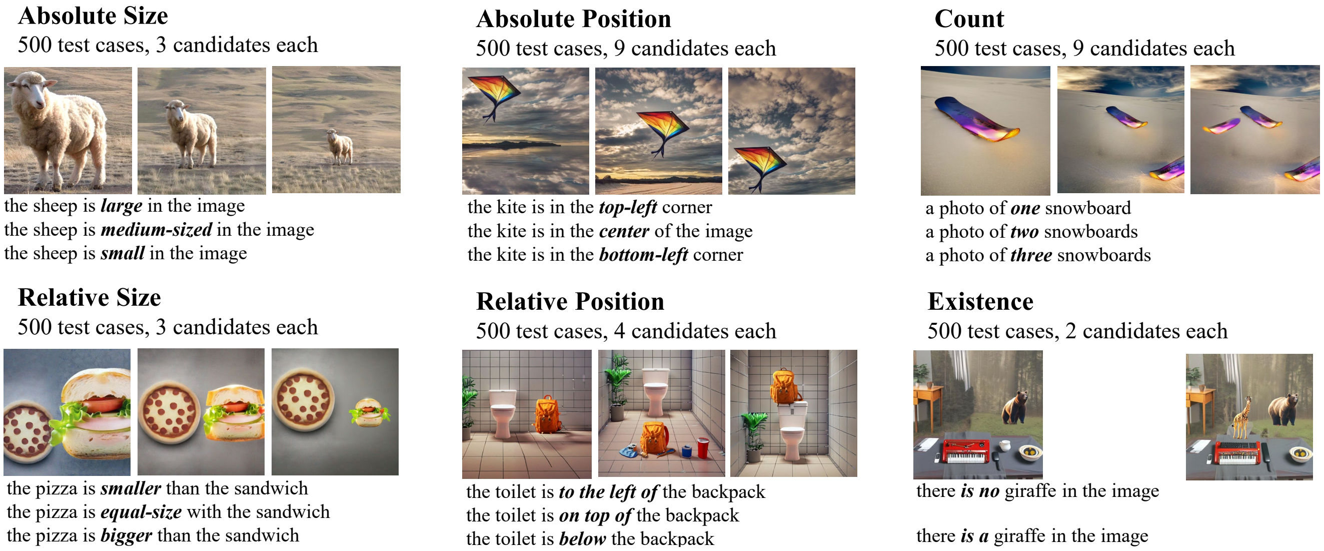

Absolute Size
500 test cases, 3 candidates each
the sheep is ***large*** in the image
the sheep is ***medium-sized*** in the image
the sheep is ***small*** in the image
Absolute Position
500 test cases, 9 candidates each
the kite is in the ***top-left*** corner
the kite is in the ***center*** of the image
the kite is in the ***bottom-left*** corner
Count
500 test cases, 9 candidates each
a photo of ***one*** snowboard
a photo of ***two*** snowboards
a photo of ***three*** snowboards
Relative Size
500 test cases, 3 candidates each
the pizza is ***smaller*** than the sandwich
the pizza is ***equal-size*** with the sandwich
the pizza is ***bigger*** than the sandwich
Relative Position
500 test cases, 4 candidates each
the toilet is ***to the left of*** the backpack
the toilet is ***on top of*** the backpack
the toilet is ***below*** the backpack
Existence
500 test cases, 2 candidates each
there ***is no*** giraffe in the image
there ***is a*** giraffe in the image

Figure 5. **An overview of the SPEC benchmark.** SPEC consists of six distinct subsets, distributed across the dimensions of Size, Position, Existence and Count. Each test case consists of an image candidate set, which differs only in certain visual concept, and a text candidate set, which differs only in corresponding language concept. Due to space constraints, we present a maximum of three images and texts here, however, more comprehensive test cases are available in the supplementary material.

and the semantic space covered by the candidate set is more comprehensive. For example, in the case of relative position, the candidates include all four semantic directions (up, down, left, right), and in absolute position and count, it extends to nine candidates. As will be discussed below, more confusing candidates can be readily used as hard negatives to improve current VLMs.

4. Diagnose: Probing VLMs on SPEC

We conduct a systematical evaluation of four state-of-the-art VLMs: CLIP [23], BLIP [17], FLAVA [29] and CoCa [39] using our newly proposed SPEC benchmark, aiming to uncover that to what extent and in what aspect are they excelling or suffering.

4.1. Evaluation Task and Protocol

Symmetric image-text matching task. We conduct evaluations using the image-text matching task, where each test case $\mathbb{T}$=$(\mathcal{I},\mathcal{T})$ comprises an image set $\mathcal{I}$ and a corresponding text set $\mathcal{T}$, as outlined in Eq. (3). The evaluation process is symmetric, considering both visual and textual perspectives. In the image-to-text matching task, when querying with an image I_i, the model is required to accurately identify T_i from the candidate set $\mathcal{T}$. Similarly, in the text-to-image matching task, the goal is to find I_i for a given query text T_i. In practice, we accomplish the matching process using the similarity $s(I,T)$ between a given image I and text T. A candidate will be selected if its similarity with the query ranks first among the entire candidate set.

Evaluation protocols. We measure the performance on SPEC using two metrics: I2TACC and T2IACC, representing the accuracy of image-to-text and text-to-image matching task, respectively:

$$\text{I2T}\textsc{acc} = \frac{1}{|D|}\sum_{(\mathcal{I}_i,\mathcal{T}_i)\in D}\frac{1}{|\mathcal{I}_i|}\sum_{I_j\in\mathcal{I}_i} h(I_j,\mathcal{T}_i) \tag{4}$$

$$\text{T2I}\textsc{acc} = \frac{1}{|D|}\sum_{(\mathcal{I}_i,\mathcal{T}_i)\in D}\frac{1}{|\mathcal{T}_i|}\sum_{T_j\in\mathcal{T}_i} g(T_j,\mathcal{I}_i), \tag{5}$$

where D contains $|D|$ test cases, $h(I_j,\mathcal{T}_i)$ equals to 1 if and only if I_j correctly find its matched text T_j from the candidate set $\mathcal{T}_i$, otherwise it is set to 0. Similarly, $g(T_j,\mathcal{I}_i)$ equals 1 if and only if T_j correctly finds its matched image I_j from the candidate set $\mathcal{I}_i$, otherwise, it is set to 0.

4.2. Key Insights from SPEC Results

We evaluate four VLMs using the SPEC benchmark, and their results are summarized in Tab. 2. We find that all the models exhibit a limited accuracy close to random chance, from which we gain the following insights:

Even state-of-the-art VLMs perform at chance level. From the results, we surprisingly find that even the most advanced VLMs achieve only a marginal advantage compared to random chance, which sharply contrasts with their impressive performance on common tasks. For instance, CLIP [23] demonstrates a mere 33.4% T2IACC for relative

	Absolute Size			Relative Size			Absolute Position			Relative Position			Existence			Count		
	I2T	T2I	CLS	I2T	T2I	CLS	I2T	T2I	CLS	I2T	T2I	CLS	I2T	T2I	CLS	I2T	T2I	CLS
Random	33.3	33.3	1.3	33.3	33.3	2.5	11.1	11.1	1.3	25.0	25.0	2.5	50.0	50.0	5.0	11.1	11.1	1.3
FlAVA [29]	37.3	37.2	89.3	32.9	32.3	84.1	13.1	15.7	88.8	25.5	26.7	84.0	57.9	51.9	74.4	14.4	21.2	75.8
BLIP [17]	43.3	42.7	97.7	33.2	32.5	98.2	12.2	11.0	97.9	30.5	29.7	97.6	55.4	50.1	96.0	37.4	37.1	93.8
CoCa [39]	39.1	36.1	95.9	33.6	33.3	95.4	11.7	11.8	96.1	30.3	28.8	92.2	50.9	50.0	83.7	36.5	35.6	89.6
CLIP [23]	42.5	36.1	92.1	34.4	33.4	94.9	12.6	12.3	92.8	28.0	26.6	90.1	58.3	51.2	84.0	25.1	23.2	83.9
Ours	68.9	60.7	96.3	40.3	44.1	97.3	30.6	34.2	96.9	46.6	46.9	96.2	83.4	53.1	92.5	55.6	57.8	92.5

Table 2. **Evaluation results on SPEC.** We extensively benchmark four state-of-the-art VLMs on SPEC to investigate their comprehension of fine-grained visual-linguistic concepts. Our evaluation employ two metrics, I2TAcc and T2IAcc, and we report the detailed performance on the each subset. We also report the classification accuracy, CLS, to highlight the capability in recognizing object categories. The first row indicates accuracy at chance level, serving as a baseline for comparison.

size recognition, while the chance-level accuracy is 33.3%. While BLIP performs reasonably well in absolute size, surpassing random level by around 9.7%, the I2TAcc on relative size is 0.7% lagged behind. CoCa [39] and FLAVA [29] also exhibit significant weaknesses in performance. The last row presents the performance of our improved model, demonstrating significant advancements across all metrics (as will be introduced in Sec. 5).

The challenge arises from the task itself, not the data. One might attribute the poor performance of VLMs to the data quality or distribution. To address this concern, we conduct an additional experiment. Specifically, we perform classification experiments using the SPEC dataset to assess the model's understanding of nouns or object categories. In this context, the models exhibit impressive performance, achieving approximately 90% Top-1 accuracy in the 80-class classification task (denoted as CLS in Tab. 2). This aligns well with earlier findings [40] that VLMs struggle in compositional reasoning while excelling in object category recognition. The remarkable accuracy of the models in the object classification task confirms the high quality of SPEC data. This also validates that the challenges faced by VLMs stem from the tasks which require fine-grained recognition rather than issues with the data itself.

4.3. Discussion on Model Limitations

We attribute the poor performance of vision and language models on SPEC to to their pretraining methods, specifically, the inherent limitation in standard contrastive loss. The conventional contrastive learning involves randomly sampling batches of images and texts, requiring the model to identify matching pairs within the batch. This task is intended to facilitate alignment between text and image spaces. However, as highlighted in prior studies [2, 40], the substantial differences between items in a randomly sampled batch allows the model to effortlessly complete this task. It can easily achieves this by focusing solely on nouns in the text and object categories in the images through a shortcut [8]. This leads the model biased towards noun concepts, neglecting other finer-grained concepts. This is the reason why these models demonstrate poor performance on fine-grained tasks that demanding understanding concepts beyond nouns.

5. Optimize: A Simple but Effective Remedy

We experiment with CLIP [23] and propose a remedy to enhance its performance in fine-grained understanding.

5.1. Method

CLIP [23] consists of an visual encoder to extract image embedding: $e_I = \mathcal{E}_I(I)$ and a textual encoder to extract text embedding: $e_T = \mathcal{E}_T(T)$. The similarity score between an image I and a text T is computed following:

$$s(I,T) = \exp\left(\frac{\tau e_I^T e_T}{\|e_I\|^2 \|e_T\|^2}\right), \tag{6}$$

where τ is a learnable temperature.

In order to guide CLIP to focus on fine-grained visual-linguistic concepts, we incorporate confusing images and text as hard negatives within the same batch. This requires CLIP to pull positives closer and push hard negatives away, thereby enhancing its ability to discern nuanced visual and textual differences. Specifically, we introduce an hard negative aware contrastive loss $\mathcal{L}_{hn} = \mathcal{L}_{hn}^{\text{I2T}} + \mathcal{L}_{hn}^{\text{T2I}}$, which comprises an image-to-text and a text-to-image term:

$$\mathcal{L}_{hn}^{\text{I2T}} = -\sum_i \log \frac{s(I_i, T_i)}{\sum\limits_{T_j \in \mathcal{T}} s(I_i, T_j) + \sum\limits_{T_k^{hn} \in \mathcal{T}^{hn}} s(I_i, T_k^{hn})} \tag{7}$$

$$\mathcal{L}_{hn}^{\text{T2I}} = -\sum_i \log \frac{s(I_i, T_i)}{\sum\limits_{I_j \in \mathcal{I}} s(I_j, T_i) + \sum\limits_{I_k^{hn} \in \mathcal{I}^{hn}} s(I_k^{hn}, T_i)}, \tag{8}$$

where $\mathcal{I}$ and $\mathcal{T}$ represent the trivial images and texts, respectively, while $\mathcal{I}^{hn}$ and $\mathcal{T}^{hn}$ denote non-trivial hard negatives. In our implementation, these hard negative examples are constructed using the data pipeline described in Sec. 3.2, and more details are in the supplementary.

To preserve the inherent zero-shot capability of CLIP, we also leverage the conventional image-text pairs from LAION-400M [27]. We apply the standard contrastive loss of CLIP [23] to these data, introducing an additional loss term $\mathcal{L}_{clip}$. The overall loss consists of two terms:

$$\mathcal{L} = \mathcal{L}_{clip} + \lambda \mathcal{L}_{hn}, \tag{9}$$

where λ is a hyperparameter that balancing these two terms. Training on this multi-task loss enables improving the performance of CLIP in fine-grained understanding while maintaining its zero-shot capability.

5.2. Experiments

Training details. We experiment with the ViT-B/32 variant of CLIP, and resume from the OpenAI pretrained checkpoint [23]. We finetune for 1,000 steps using a cosine schedule with an initial learning rate of $1e$-6 and use 800 steps for warm up. The batch size of LAION data is set to 2048, and the batch size of hard negative data is set to 768. The weight λ of the hard negative aware loss is set to 0.2.

Main results. We utilize the SPEC benchmark to assess the understanding of model in fine-grained concepts. In Tab. 3, we present the average I2TACC and T2IACC on all subsets of SPEC. To assess the general performance of the model, we also utilize the toolkit from ELEVATER [16] to evaluate the zero-shot performance on 9 classification and retrieval datasets, and report the average accuracy. Compared to CLIP [23] , our model demonstrates remarkable advancements with a 19.8% boost in I2TACC, an 18.9% improvement in T2IACC on SPEC, and a noteworthy 1.2% enhancement in zero-shot accuracy. We also conduct ablation on different training configurations. From the results in Tab. 3, it can be observed that the introducing of $\mathcal{L}_{hn}$ significantly improves the performance on SPEC. Moreover, the $\mathcal{L}_{clip}$ plays a crucial role in preserving zero-shot performance. Without it, we observe a decline in accuracy by 5.1%. With the combination of these two losses, we achieve substantial improvement in SPEC while maintaining the original zero-shot capability.

Validation on other fine-grained benchmarks. To assess whether our approach aids the model in acquiring fundamental visual-linguistic understanding or merely leads to overfitting on SPEC, we conduct evaluations on two additional benchmarks which also focus on the assessment of fine-grained concepts. ARO [40] explores three aspects of vision-language understanding: object attributes, inter-object relations, and word ordering. Eqben [33] focuses on minimal visual semantic changes, aiming to diagnose VLMs in understanding fine-grained concepts such as counting and location. In Tab. 4, we present the experimental results, demonstrating a clear improvement compared to CLIP [23] on both datasets, For example, compared to CLIP, our method shows an average improvement of 2% on Eqben and respective enhancements of 3.2%, 9.8%, and 7.4% on the three subsets of ARO. The consistent improvement on these datasets demonstrates that our approach has facilitated the model in acquiring transferable fine-grained understanding capabilities.

	Config		SPEC		Zero-shot
	$\mathcal{L}_{clip}$	$\mathcal{L}_{hn}$	I2T	T2I	Accuracy
CLIP			33.5	30.5	67.5
+ $\mathcal{L}_{hn}$		✓	64.3	60.8	62.4
+ $\mathcal{L}_{clip}$	✓		32.2	31.5	69.4
+ $\mathcal{L}_{hn}$+$\mathcal{L}_{clip}$ (ours)	✓	✓	53.3	49.4	68.7

Table 3. **Main Results:** The first row represents the pretrained checkpoint without fine-tuning. We sequentially introduce two additional loss terms to investigate their impact for the performance.

	Eqben			ARO		
	Image	Text	Group	Attribute	Relation	Order
CLIP	17.6	21.4	10.1	63.2	63.9	53.3
CLIP$_{FT}$	18.1	23.7	11.1	65.1	68.0	54.1
Ours	19.5	24.0	11.7	66.4	73.7	60.7

Table 4. **Cross-dataset evaluation results on Eqben [33] and ARO [40].** To demonstrate that the improvement comes from the negative loss, rather than training on more data, we also report CLIP$_{FT}$, which is also finetuned but without negative samples.

6. Conclusion

In this study, we explored the comprehension abilities of Visual Language Models (VLMs) with respect to fine-grained visual-linguistic concepts. We first established an efficient pipeline to synthesize candidate images that exclusively differ in a particular visual attribute. Leveraging this pipeline, we created the SPEC benchmark to diagnose the comprehension proficiency of VLMs in terms of object size, position, existence, and count. Upon evaluating four leading VLMs using SPEC, we uncovered substantial performance limitations. To address this, we introduced an enhancement strategy that effectively optimizes the model for fine-grained understanding, while maintaining its original zero-shot capability.

Acknowledgement This project was supported by NSFC under Grant No. 62102092.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *ICCV*, 2019. 1

[2] Paola Cascante-Bonilla, Khaled Shehada, James Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gül Varol, Aude Oliva, Vicente Ordonez, Rogério Schmidt Feris, and Leonid Karlinsky. Going beyond nouns with vision & language models using synthetic data. In *ICCV*, 2023. 3, 7

[3] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel M. Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2022. 1

[4] Yuxin Chen, Zongyang Ma, Ziqi Zhang, Zhongang Qi, Chunfen Yuan, Ying Shan, Bing Li, Weiming Hu, Xiaohu Qie, and Jianping Wu. Vilem: Visual-language error modeling for image-text retrieval. In *CVPR*, 2023. 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 3

[7] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogério Schmidt Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *CVPR*, 2023. 3

[8] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2020. 7

[9] Golnaz Ghiasi, Yin Cui, A. Srinivas, Rui Qian, Tsung-Yi Lin, Ekin Dogus Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *CVPR*, 2020. 4

[10] Shashank Goel, Hritik Bansal, Sumit Kaur Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. In *NeurIPS*, 2022. 2

[11] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2015. 2

[12] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings*, 2021. 3

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 2

[14] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *ArXiv*, 2023. 2

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *ICCV*, 2023. 2, 4

[16] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS*, 2022. 8

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1, 2, 6, 7, 3

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 4, 3

[19] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *TACL*, 2023. 2, 3

[20] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, 2023. 4

[21] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *ACL*, 2021. 1, 2, 3, 5

[22] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, 2023. 2, 4

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 6, 7, 8, 3

[24] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2021*. 2, 4

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 4

[26] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, 2022. 3

[27] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *ArXiv*, 2021. 8

[28] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 3

[29] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022. 1, 2, 6, 7, 3

[30] James Smith, Paola Cascante-Bonilla, Assaf Arbelle, Donghyun Kim, Rameswar Panda, David Cox, Diyi Yang, Zsolt Kira, Rogério Schmidt Feris, and Leonid Karlinsky. Construct-vl: Data-free continual structured vl concepts learning*. In *CVPR*, 2023. 3

[31] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022. 1, 2, 5

[32] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. In *NeurIPS*, 2022. 2

[33] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *ICCV*, 2023. 1, 2, 3, 5, 8

[34] Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, and Yu-Gang Jiang. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *ICML*, 2023. 2

[35] Zuxuan Wu, Zejia Weng, Wujian Peng, Xitong Yang, Ang Li, Larry S Davis, and Yu-Gang Jiang. Building an open-vocabulary video clip model with better architectures, optimization and data. *TPAMI*, 2024. 2

[36] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*, 2023. 4

[37] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. In *CVPR*, 2024. 4

[38] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 1

[39] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022. 2, 6, 7, 3

[40] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bag-of-words models, and what to do about it? *ICLR*, 2023. 1, 2, 3, 5, 7, 8

[41] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *ICML*, 2021. 1, 2

[42] Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, Weiming Zhang, and Neng H. Yu. X-paste: Revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. In *ICML*, 2022. 4

[43] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *ArXiv*, 2022. 1, 2, 3