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

Supplementary Material

## 1. More Details About SPEC Benchmark

We provide additional details regarding the construction of the SPEC benchmark and present additional examples for visualization.

### 1.1. Consistent Background Filling Strategy

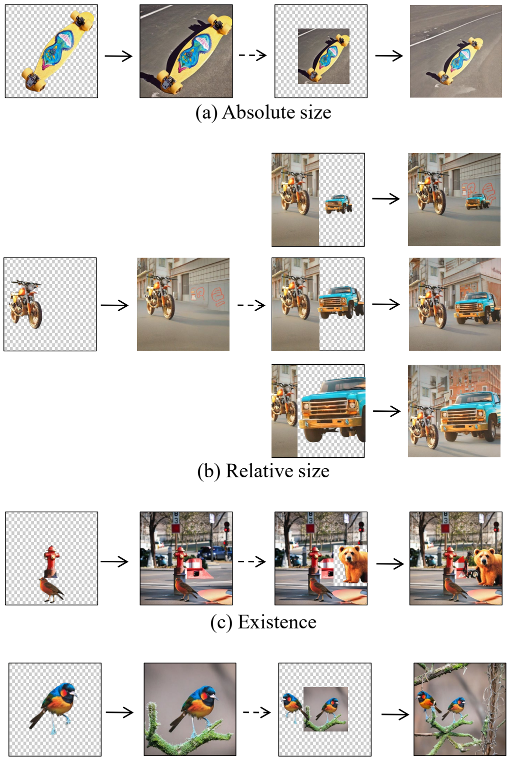
In Sec. 3.2.4, we provide examples illustrating the background filling strategy, specifically focusing on absolute position and relative position. Here, we will further elaborate on the remaining aspects, namely absolute size, relative size, existence and count.

**Absolute size.** As shown in Fig. 6 (a), we produce images of skateboards with varying absolute sizes through the following steps: First, a skateboard is positioned on an empty canvas, and a corresponding background is generated, resulting in an image featuring a "large" skateboard. Subsequently, the image is resized and placed on a new canvas. The remaining background is then generated, yielding an image of a "medium-sized" skateboard. Then we repeat this process to attain an image of a "small" skateboard.

**Relative size.** As illustrated in Fig. 6 (b), the process of generating images featuring motorcycles and trucks with different relative size relationships involves the following steps: First, we first generate a background for the motorcycle, then the truck is introduced onto the canvas, with adjustment to the size to achieve the desired relative size relationship with the motorcycle. Finally, the remaining empty background is filled in to complete the composition.

**Existence.** As depicted in Fig. 6 (c), we manipulate the existence of an object within the image through the following steps: First, we place the foreground elements (*i.e.,* the bird and fire hydrant in the illustration) that are common to both images on the canvas, and generate a suitable background, resulting in an image without a bear. Subsequently, a region is cropped from the images, the foreground of the bear is pasted to this space, and the background is seamlessly filled in. This process yields an image with a bear.

**Count.** As illustrated in Fig. 6 (d), we generate images with different count of objects by following these steps: First, a single object is positioned on a blank canvas, and an initial background is generated around it, producing an image containing one object. Subsequently, the image is resized and centered on a new canvas. The foreground of the object is then duplicated, pasted into the vacant space, and the remaining background is filled in, thereby creating an image with two objects. This process is iteratively applied, enabling the progressive generation of images with an increasing number of objects.


(a) Absolute size

(b) Relative size

(c) Existence

(d) Count

Figure 6. More examples of consistent background filling strategies, including absolute size, relative size, existence, and count. These are supplementary to Fig. 4.

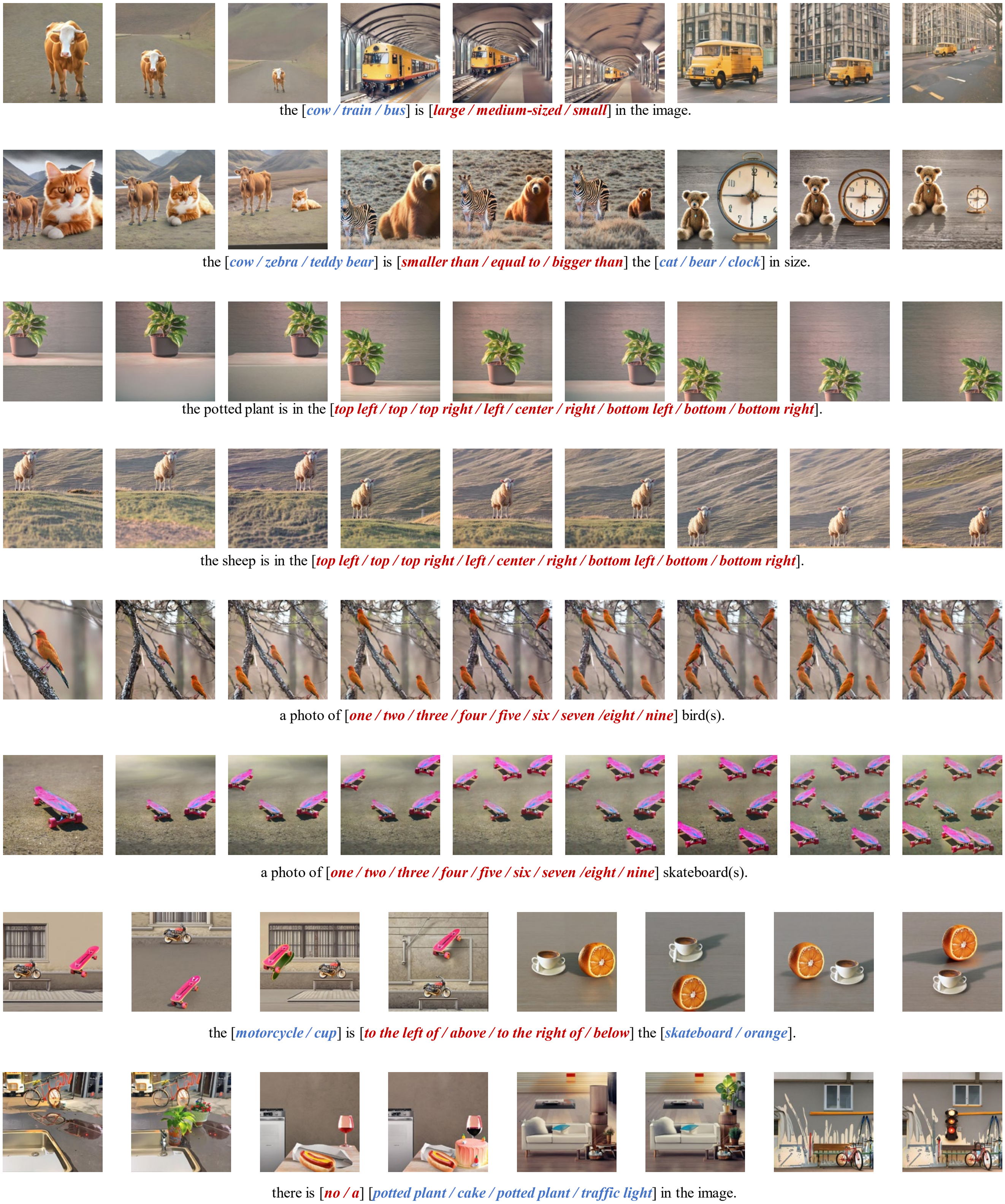
### 1.2. More Examples of SPEC

As a complement to Fig. 5, we show more test cases in Fig. 7. We can observe that each test case in SPEC has 2 to 9 image and text candidates. In contrast to benchmarks with only two image-text pairs, having more candidates within a test case makes the matching task more challenging.

## 2. More Details About Implementation

### 2.1. Baseline Models

In the main paper, we evaluate four state-of-the-art VLMs using the SPEC benchmark. This section delves into a more



the [***cow / train / bus***] is [***large / medium-sized / small***] in the image.

the [***cow / zebra / teddy bear***] is [***smaller than / equal to / bigger than***] the [***cat / bear / clock***] in size.

the potted plant is in the [***top left / top / top right / left / center / right / bottom left / bottom / bottom right***].

the sheep is in the [***top left / top / top right / left / center / right / bottom left / bottom / bottom right***].

a photo of [***one / two / three / four / five / six / seven /eight / nine***] bird(s).

a photo of [***one / two / three / four / five / six / seven /eight / nine***] skateboard(s).

the [***motorcycle / cup***] is [***to the left of / above / to the right of / below***] the [***skateboard / orange***].

there is [***no / a***] [***potted plant / cake / potted plant / traffic light***] in the image.

Figure 7. Visualization of more test cases from SPEC benchmark.

|  | 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 |
| negCLIP [39] | 44.0 | 43.1 | 91.3 | 31.6 | 32.4 | 95.1 | 11.4 | 12.6 | 92.3 | 28.5 | 28.1 | 88.8 | 64.6 | 50.0 | 86.7 | 33.7 | 34.3 | 81.0 |
| TSVLC [7] | 40.3 | 34.9 | 92.5 | 33.8 | 34.0 | 96.5 | 12.6 | 12.4 | 91.6 | 26.2 | 26.4 | 90.3 | 62.2 | 50.7 | 90.4 | 21.8 | 21.1 | 85.9 |
| 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 5. **More evaluation results on SPEC.** We evaluated the performance of two additional VLMs, negCLIP [39] and TSVLC [7], on SPEC. Both of these models are finetuned for fine-grained understanding.

detailed description of these four pretrained models, including their network architectures and pretrained checkpoints.

**CLIP:** We use the 'ViT/B-32' variant of CLIP [23] with weights resumed from the checkpoint released in [23].

**BLIP:** We use the 'ViT-B' variant of BLIP [17] with weights resumed from the checkpoint released in [17], which is finetuned on COCO [18] for image-text retrieval.

**FLAVA:** We use the 'full' version of FLAVA [29] with weights resumed from the checkpoint released in [29].

**CoCa:** We used the 'ViT/B-32' variant of CoCa [38] with weights resumed from the checkpoint pretrained on LAION-2B dataset [28].

### 2.2. Benchmarks

**Eqben [33]:** Eqben comprises a total of 250k image-text pairs. For ease of testing, the maintainer also provide a subset containing 25k randomly sampled data points. We evaluate our model on this subset and report three metrics: image score, text score and group score.

**ARO [39]:** ARO comprises four subsets: Visual Genome Attribution, which emphasizes object attributes, Visual Genome Relation, which centers on inter-object relationships, as well as COCO Order and Flickr Order, which focus on word ordering. In Tab. 4, we present the average accuracy of COCO Order and Flickr Order in the last column named 'Order'.

**Zero-shot Benchmark:** To test the zero-shot capability of our model, we conduct experiments on 9 common classification and retrieval datasets and report their average accuracy in Tab. 3. These datasets include CIFAR10, CIFAR100, ImageNet1K, STL10, Flowers102, OxfordPets, Caltech101, Flickr30k and COCO.

### 2.3. Training Data

To improve the model in fine-grained visual-linguistic understanding, we introduce a loss term that is sensitive to hard negatives. We utilize the pipeline outlined in Sec. 3.2 to generate training data. Specifically, for each of the visual-linguistic concepts, namely absolute size, relative size, absolute position, relative position, existence and count, we generate nearly 20k training samples for the training process. Each training sample consists of two image-text pairs, denoted as $(I_0, T_0)$ and $(I_1, T_1)$, where $I_0$ and $I_1$, as well as $T_0$ and $T_1$, serve as hard negative for each other.

| Config |  |  | SPEC |  | Zero-shot |
|---|---|---|---|---|---|
| $\mathcal{L}_{clip}$ | $\mathcal{L}_{hn}^{\text{I2T}}$ | $\mathcal{L}_{hn}^{\text{T2I}}$ | I2T | T2I | Accuracy |
| ✓ |  |  | 32.2 | 31.5 | 69.4 |
| ✓ | ✓ |  | 50.5 | 45.8 | 69.1 |
| ✓ |  | ✓ | 50.1 | 46.8 | 68.9 |
| ✓ | ✓ | ✓ | 53.3 | 49.4 | 68.7 |

Table 6. To investigate the importance of training data from both modalities, we conducted experiments separately using only text hard negatives and only image hard negatives.

## 3. More Experimental Results

We present more experimental results in this section.

**Evaluating more models on SPEC benchmarks.** We evaluate two additional VLMs, negCLIP [39] and TSVLC [7], on SPEC. These models are finetuned to enhance their capability for comprehending fine-grained visual-linguistic concepts, contrasting with the models outlined in Tab. 2, which are pretrained for general purpose. The experimental results in Tab. 5 indicate an enhancement in the performance of these models on SPEC. However, the improvement remains limited, possibly because hard negatives were introduced only to the language part during their finetuning process. Benchmarks such as SPEC, which evaluate models from two modalities, also necessitate visual enhancements.

**The significance of hard negatives from two modalities.** Constrained by the challenges in manipulating visual data, previous efforts [7, 39] primarily construct textual hard negatives. In response, we tackle this challenge by employing a progressive approach to generate visual training data. To assess the importance of hard negatives for both modalities, we conduct separate experiments to evaluate the model's performance on SPEC with only text hard negatives (utilizing $\mathcal{L}_{hn}^{\mathrm{I2T}}$ from Eq. (7)) and with only image hard negatives (utilizing $\mathcal{L}_{hn}^{\mathrm{T2I}}$ from Eq. (8)). From Tab. 6, we observe that the improvement is limited when using hard negatives from a single modality alone. Simultaneously incorporating data from both visual and linguistic modalities yields the best results, demonstrating the effectiveness of utilizing hard negatives from both modalities.

**Limitations** We note that when constructing images involving multiple objects, such as relative size, relative position, and existence, it is necessary to sample multiple object instances from the library and arrange them within the same scene. In our implementation, we employ random sampling without considering the co-occurrence relationships between objects. This may result in some less plausible object combinations, such as a zebra and a toilet appearing in the same scene. However, we maintain effective control over attributes of interest, such as the size and position of objects, allows us to use this data to assess the models' understanding of specific visual-linguistic concepts.