

Supplementary Material: Enhancing Multimodal Compositional Reasoning of Visual Language Models with Generative Negative Mining

A. Filtering Strategies

A.1. Failure Modes

In our experiments, we conduct a manual evaluation of synthesized images and identify certain failure modes introduced below: 1. **object tag leads to wrong mask**: The prompted object is not correctly segmented from the context. This leads to an unusable generation or increased complexity as the original object remains unaffected. 2. **excessive segmentation**: In complex scenes, items are segmented towards part of the image that no longer contains the intended object. This degrades the image composition and may render the associated caption invalid. 3. **poor inpainting**: The performance of Stable Diffusion is affected by previous steps, including image complexity, portrayal quality. 4. **unusual state of the object**: In specific cases, objects appear in an unexpected positions and angles, making it challenging to inpaint the area. 5. **confusion due to multiple instances**: When multiple items are present in an image, the generation performance may be decreased. For example, in an image of multiple plates, painting one of them does not effectively change the meaning as expected. 6. **high complexity in the image**: Images with a variety of objects may cause small portions of the image left for each of them. 7. **small mask size**: In some cases, the identified object is so small that generation fails due to poor quality. 8. **lack of descriptiveness in portrayal**: ChatGPT may produce portrayals unsuitable for image generation. The lack of descriptiveness can lead to nearly identical images with minimal differences. 9. **animate objects**: Animals and humans are hard to portray because their posture dynamically changes with the action.

Figure 1 illustrates the failure modes in image generation, where original and generated versions of images are presented.

A.2. Implementation of Filters

To address these issues, we mainly use two filters. The first filter is the BLIP ITM head that returns a matching score between an image and text. The second filter calculates the variance within object-level generations.



Figure 1. Samples of detected failure modes. The left side shows original images whereas the middle and right sides show two generations.

ITM Filter BLIP’s ITM head outputs a confidence value if a given image and text pair match. As suggested in their original work, we use the decision threshold of 0 to pass a sample as valid. We utilize BLIP to compute two scores, the variation score that grades the inpainted image and patched caption, and the original score that grades the inpainted image and original caption. The variation score is used in our filtering pipeline. The original score is only used for statistical analysis, as it is not capable of determining fine-grained concept matching.

Variance Filter To filter out samples with small mask sizes and samples generated similarly, we calculate a variance score for images, as shown in Figure 2. First, we stack all the images and create a tensor of size (Batch Size, Channels, Height, Width). Then we calculate the standard deviation of the first dimension, in which we know the area

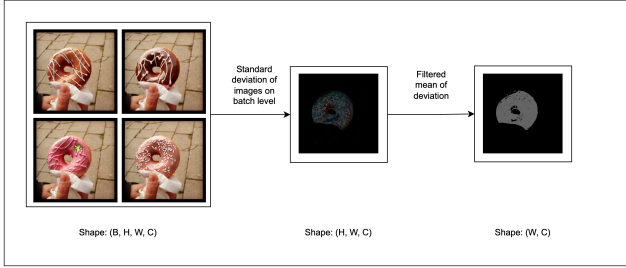


Figure 2. Calculation of Image Variation Area, visualized

outside of the mask should yield 0 since they have identical values. For the masked part, we further calculate the mean of the channels, reducing the tensor to shape (Height, Width). In this form, we apply a condition whether the average deviation is higher than a value ϵ . The final array is a boolean array, in which we average across two dimensions to calculate the latest value. The next section provides a deeper insight into the statistics of the filtering process.

A.3. Analysis of Generated Images

To gain deeper insights into generated variation images, we analyze the histogram of filter scores in Figure 3. As shown in the left histogram of the figure, the variation score ranges between -6 and 5. Our manual examination reveals that the samples with scores between -1 and 0 contain optimal images. Although this range may result in a loss of high-quality generations, we did not dive deeper into mining such samples. However, advanced filtering can be utilized to improve the generation. The middle histogram reveals that a significant amount of samples exhibit a small image variation area. Therefore, we use the median value and set the threshold for filtering to 14.

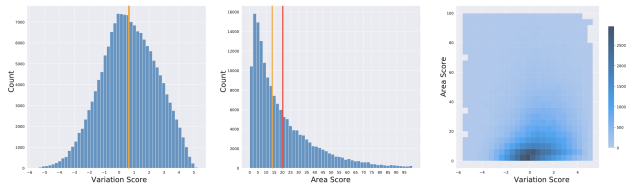


Figure 3. 2D histogram of Variation Score and Area Score. The red line indicates the mean and the orange line indicates the median. We observe normal distribution on ITM Score and log-normal distribution on the area score. We prefer ITM Score > 0 and Area Score > 14

Another metric related to the variation area is the delta in the mask. Within each variation group that replaces the same object, we calculate a delta score inside the boundaries of the mask. The delta in mask calculation is similar to image variation calculation. However, we skip the step to check if the value is larger than the epsilon value to mea-

sure the amount of difference. We employ this value for our statistical analysis. Items with high delta and high masking percentages tend to have generations aligned with their portrayals. Objects with greater size in physical life are likely to have higher masking percentages. However, there are also items like pizza which is relatively small in physical life compared to other objects but still covers a higher area in images.

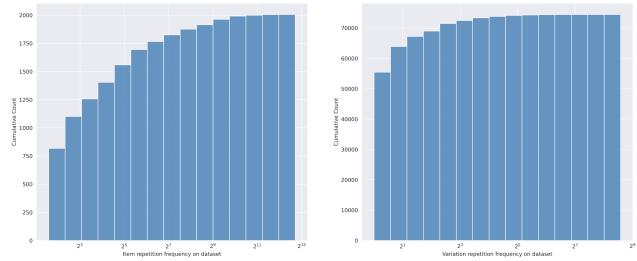


Figure 4. Repetition frequency of items and variations

An analysis of filters on the object level provides further insight into the statistics of the objects. We visualized the most commonly portrayed items after filtering in Figure 5.

The training dataset includes 164,021 variations generated for 41,003 objects identified in 12,656 images. There are 2006 different items and 74,460 different phrases. ChatGPT utilizes a total number of 9,464 unique tokens (identified by the CLIP tokenizer) and 12,413 unique words. Objects like heads and pictures have the most diverse set of portrayals. Common objects categories such as keyboards, desks, trains, and kitchens have the most repetitive representations. We observe a uniqueness drop with respect to the total count. The examination of highly repetitive objects that the words used in portrayals are not descriptive enough. For example, "minimalist kitchen" and "modern kitchen" are two frequent responses ChatGPT has produced for the kitchen. Nevertheless, minimalist or modern does not express a niche vision but a general concept.

Last, we check item frequency and uniqueness of short descriptions. As presented in the left histogram of Figure 4, the count of items identified only once is 818. A glance over singular items reveals their infrequent existence. The majority of singular objects are uncommon in everyday environments: asparagus, octopus, orchid, etc. A fractional subset of the singular objects are phrases: a herd of giraffes, mountain goat, car mirror. This indicates that the COCO dataset doesn't have a diverse set of objects present in the pictures. The right side of Figure 4 presents the frequency histogram of short descriptions. The count of variations identified only once is 55,458. However, there are more than 200 occurrences of short descriptions like a cobblestone street, a sleek glass table, and a majestic oak tree. Such high repetitions indicate ChatGPT is repetitive in terms of variation genera-

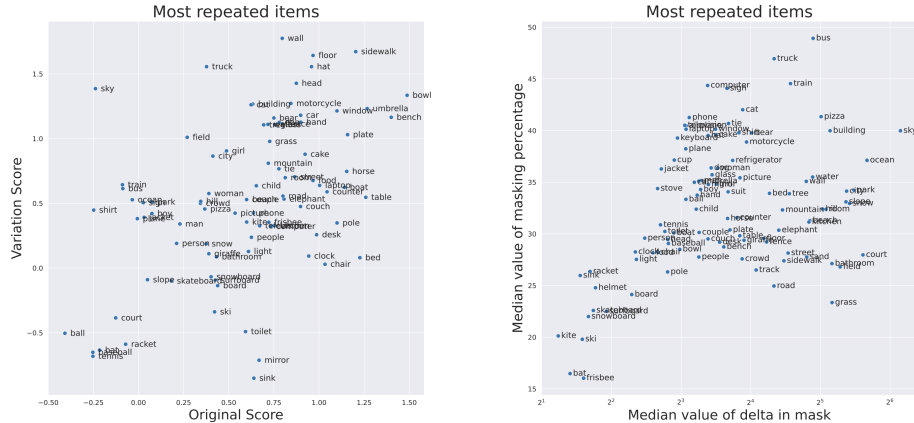


Figure 5. Most common items by their average scores of filters. In the left plot, the x-axis represents the median value of the original score. The y axis represents the median value of variation scores. In the right plot, the x-axis represents the median value of the delta in the mask while y-axis represents the median value of the masking percentage.

tions.

B. Evaluation on Winoground Splits

As mentioned in Section 5.1, we evaluate our model’s performance on the splits of Winoground that test different aspects of reasoning capabilities. From Tab. 1, we see that our model outperforms the baseline in the compositional reasoning task which requires detailed descriptions of visual scenes. However, our model fails when comprehending certain difficult texts, especially when the meaning of the text becomes challenging. For example, in the case of the phrase *the brave in the face of fear*, the image depicts a small cub confronting a fierce lion, while the model needs an in-depth understanding of the word *brave* to associate it to the cub. During the finetuning process, our model may demonstrate the phenomenon of “catastrophic forgetting”, if the quality, diversity, and scale of our dataset do not match with the original pretraining dataset. In particular, the presence of repetitive text samples in our augmented dataset may impede the performance of the text encoder.

C. Examples from Our Dataset

In Fig. 6, we showcase a few examples from our generated dataset. Our approach is advantageous in that we can generate a diverse dataset with challenging negative examples. For instance, the images in the first row depict scenarios that are highly unlikely in the real world, since an ice cream cart will never appear at an airport for aircraft maintenance. These examples serve as a true test of the model’s understanding of the cart concept.

Furthermore, some examples in our dataset differ in fine-grained details that can be challenging even for humans. An example of this can be observed in the last row. The model

	Compositional (171)			Complex (78)		
CLIP	31.58	11.70	9.36	23.08	6.41	3.85
Ours	38.01	14.62	10.53	29.49	8.97	6.41
Gains	+22.5%	+27.2%	+12.5%	+23.9%	+39.9%	+66.5%
	Unusual Image (56)			Unusual Text (50)		
CLIP	26.79	8.93	5.36	34.0	14.0	10.0
Ours	28.57	8.93	8.93	30.0	10.0	10.0
Gains	+6.7%	0.0%	+66.3%	-11.8%	-28.5%	0.0%
	Ambiguous(46)			Visually Difficult(38)		
CLIP	30.43	15.22	15.22	15.79	0.00	0.00
Ours	26.09	8.70	8.70	18.42	2.63	2.63
Gains	-14.2%	-43.8%	-43.8%	+16.6%	+2.63%	+2.63%
	Non compositional(30)					
CLIP	76.67	36.67	33.33			
Ours	70.00	40.00	36.67			
Gains	-8.7%	+9.0%	+10.0%			

Table 1. Comparison of models on Winoground subsets that evaluate distinct reasoning abilities. The numbers in parentheses represent the sample count for each split. Our model excels in compositional reasoning tasks that demand a detailed description of the scene. However, it struggles when it comes to understanding subtle differences in the text that may require background knowledge, e.g., unusual text.

needs to analyze the specific type of grass in order to make an accurate prediction.

D. Comparison with SOTA Method

To obtain a comprehensive understanding of the compositional reasoning ability of our approach, we conduct a comparative analysis with a state-of-the-art method, TSVLC [1], on two established benchmarks, Winoground and VL-Checklist. Table 2 presents the evaluation results on Winoground, where our model significantly outperforms TSVLC in terms of text score, despite lower image score and group score. For VL-Checklist, we present the evalu-






















	ice cream cart		golf course cart		grocery cart		farm cart
	modern metal frame		vintage wooden frame		rustic barnwood frame		ornate golden frame
	serene willow-lined canal		narrow winding canal		tranquil countryside canal		bustling urban canal
	mossy brick wall		graffiti brick wall		crumbling brick wall		modern brick wall
	muddy puddle		city street puddle		rippling puddle		reflecting puddle
	leather-bound book		stack of children's books		paperback novel		old tattered book
	starry night sky		stormy sky		dramatic sunset sky		clear blue sky
	cozy knitted blanket		fluffy faux fur blanket		colorful patchwork quilt		crisp white bedsheet
	soft dinner roll		crispy baguette		rustic sourdough loaf		whole-grain seed bread
	thatched cottage		modern house		log cabin		Victorian mansion
	vast grassy plains		floral meadow plains		grazing bison plains		golden wheat plains

Figure 6. Examples from our generated dataset. Each row demonstrates four variations generated using a COCO image-text pair. To highlight the differences among the images, we only provide descriptions for the modified part, instead of the caption for the entire image.

ation results on detailed data subsets¹ in Table 3, 4, 5, and 6. We obtained the evaluation scores of TSVLC from their published paper. From the results presented in the table, we find that our model performs comparably to TSCVL (The average of all individual metrics in Table 3-6 yields the following results: CLIP: 70.57, Ours: 72.37, TSVLC: 75.71). Note that TSCVL is trained on 3 million image-text pairs, while our approach is finetuned on a much smaller scale of approximately 100k images. In addition, TSCVL incorporates more sophisticated negative sampling strategies and curated loss functions. In contrast, we utilize the naive CLIP architecture and the simple contrastive loss function. Furthermore, it is worth noting that our method still outperforms CLIP on average, consistent with the observation on other datasets presented in the main text.

Model	Text Score	Image Score	Group Score
CLIP	30.75	11.0	8.75
TSVLC	26.0	15.75	11.0
Ours	34.25	12.5	10.0

Table 2. Comparison of our method with CLIP and TSVLC on Winoground benchmarks. We report the text score, image score, and group score which measure if the model can correctly match a text for an input image, or vice versa.

References

- [1] Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023. 3

¹We were unable to download the full HAKE dataset due to a server failure. Nevertheless, the evaluation of existing datasets provides representative insights into the compositional reasoning capabilities of our model.

	O-Large	O-Medium	O-Small	O-Center	O-Mid	O-Margin	Avg O
CLIP	86.95	77.75	72.75	85.5	80.5	70.6	79.00
TSVLC	90.5	81.95	77.6	89.75	83.8	73.35	82.82
Ours	92.04	84.97	77.52	92.01	85.95	77.99	85.08

Table 3. Evaluation on VG Object subset of VL-Checklist. TSVLC refers to the finetuned model on CC3M. Our approach utilizes CLIP and is further finetuned on our augmented dataset. Our model outperforms both the CLIP and TSVLC approaches.

	A-Color	A-Material	A-Size	A-State	A-Action	R-action	R-spatial	Avg A+R
CLIP	68.9	65.4	72.1	69.3	72.37	62.4	54.0	66.35
TSVLC	79.9	78	76.8	68.7	74.18	61.9	63.2	71.81
Ours	73.07	72.51	64.38	67.91	75.53	57.86	49.69	65.85

Table 4. Evaluation on VG Attribute and Relationship subset of VL-Checklist. TSVLC refers to the finetuned model on CC3M. Our approach utilizes CLIP and is further finetuned on our augmented dataset.

	O-Large	O-Medium	O-Small	O-Center	O-Mid	O-Margin	Avg All
CLIP	76.98	73.28	59.41	78.075	74.63	64.49	71.76
TSVLC	83.5	80.05	71.70	84.02	81.17	75.01	78.24
Ours	81.11	75.04	68.82	81.45	78.00	70.53	75.82

Table 5. Evaluation on SWIG subset of VL-Checklist. TSVLC refers to the finetuned model on CC3M. Our approach utilizes CLIP and is further finetuned on our augmented dataset. Though our model underperforms TSVLC, it still exhibits significant improvement over CLIP.

	A-Color	A-Material	A-Size	A-State	A-Action	Avg All
CLIP	71	73.3	68	53.3	62.7	65.66
TSVLC	75	76.7	69.9	55.9	64.6	68.42
Ours	76.6	68.7	56.23	57.99	72.62	66.4

Table 6. Evaluation on VAW subset of VL-Checklist. TSVLC is the final model which is finetuned on CC3M. Our approach utilizes CLIP and is further finetuned on our augmented dataset.