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

Uğur Şahin*[1], Hang Li*[2,4], Qadeer Khan[1,3], Daniel Cremers[1,3], Volker Tresp[2,3]
[1]TU Munich, [2]LMU Munich [3]MCML, [4]Siemens AG

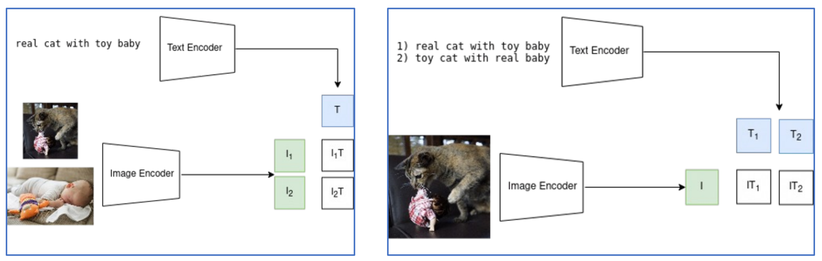
JAN 4-8 WACV 2024 WAIKOLOA HAWAII

1) ABSTRACT

Contemporary large-scale visual language models (VLMs) exhibit strong representation capacities, making them ubiquitous for enhancing image and text understanding tasks. They are often trained in a contrastive manner on a large and diverse corpus of images and corresponding text captions scraped from the internet. Despite this, VLMs often struggle with compositional reasoning tasks which require a fine-grained understanding of the complex interactions of objects and their attributes. This failure can be attributed to two main factors: 1) Contrastive approaches have traditionally focused on mining negative examples from existing datasets. However, the mined negative examples might not be difficult for the model to discriminate from the positive. An alternative to mining would be negative sample generation 2) But existing generative approaches primarily focus on generating hard negative texts associated with a given image. Mining in the other direction, i.e., generating negative image samples associated with a given text has been ignored. To overcome both these limitations, we propose a framework that not only mines in both directions but also generates challenging negative samples in both modalities, i.e., images and texts. Leveraging these generative hard negative samples, we significantly enhance VLMs' performance in tasks involving multimodal compositional reasoning.

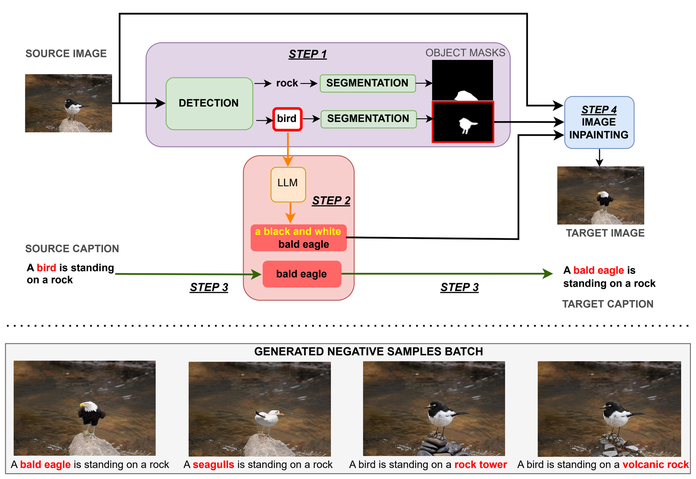
2) MOTIVATION

Visual language models often struggle with compositional reasoning tasks that require a fine-grained understanding of the complex interactions of objects and attributes. How to improve the compositional reasoning skills of visual-language networks?

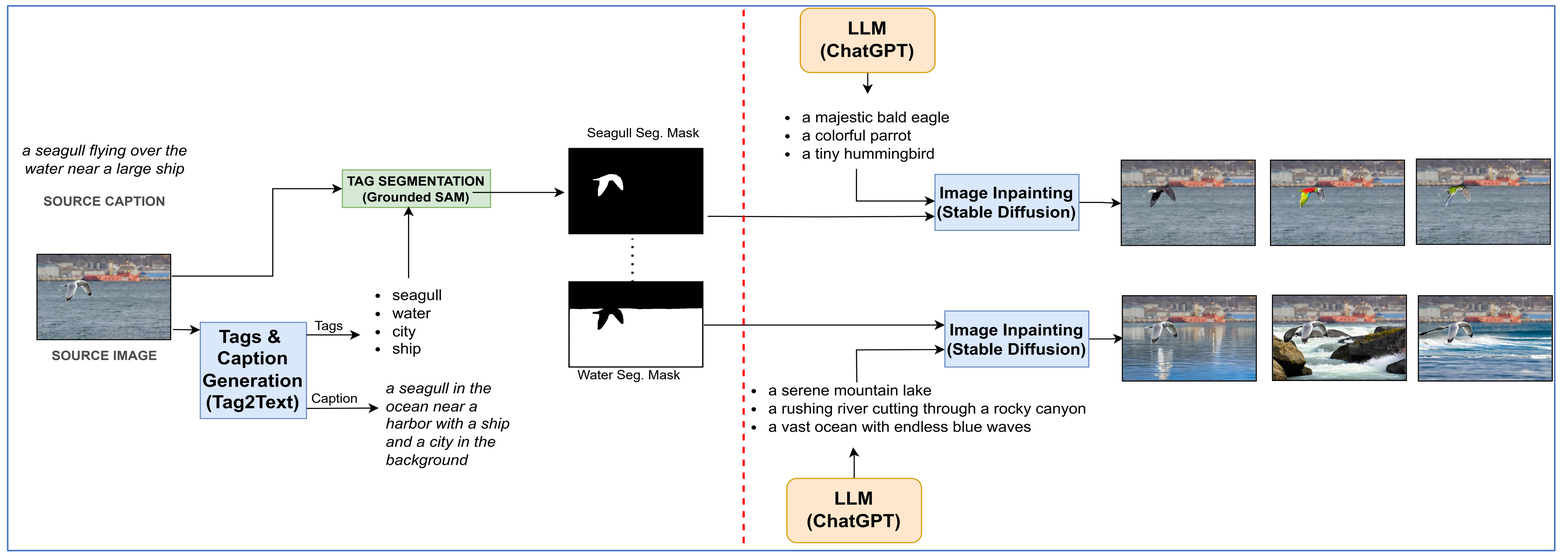


3) METHODOLOGY

Overview of our proposed generative approach for image-text synthesis from a given source image and a corresponding caption.



- Step 1: The source image is first passed through a detection and segmentation algorithm to identify all the relevant objects in the scene (bird and rock) and also create independent masks of these objects. The remaining steps in this figure focus on the bird object.
- Step 2: A large language model (LLM) then takes the detected objects to create 1) an alternate representation of that object (bald eagle) 2) A more fine-grained and descriptive representation of the same object (a black and white bald eagle)
- Step 3:The source caption is replaced with an alternate representation to produce the target caption.
- Step 4: The original mask of the object and the descriptive alternate caption are fed to an inpainting algorithm to replace the bird with a black and white bald eagle in the source image to produce the target image.



➢ Image Analysis and Tag Extraction

To accurately identify the regions of an image that need modification, we utilize a comprehensive annotation approach to decompose the scene into its constituent parts. Firstly, we utilize off-the-shelf image-to-text models, specifically Tag2Text, for object detection and caption generation. As shown on the left side of above figure, the Tag2Text model outputs a list of detected object labels in the image, along with a descriptive caption summarizing the entire scene. The descriptive caption is needed to ensure that all the identified objects have been covered in the caption.

Modify each of the given words in three different ways.
You can add colors, shapes and material information whenever necessary.
You can change the word with semantically similar words.
Add a short description of the portrayal up to three words in parenthesis.

Input:
Objects: bread, bag
Context: a person holding a piece of bread with bananas and a bag

Output:
Bread:
1) A warm, freshly baked loaf (freshly baked loaf)
2) A crumbly gluten-free muffin (gluten-free muffin)
Bag:
1) A leather satchel with intricate patterns (leather satchel)
2) A sturdy canvas backpack with patches (canvas backpack)

Input:
Objects: seagull, water, city, ship
Context: a seagull flying over the water near a large ship

Output:

ChatGPT

Seagull:
1) a majestic bald eagle (bald eagle)
2) a colorful parrot (parrot)
Water:
1) a serene mountain lake (mountain lake)
2) a rushing river cutting through a rocky canyon (rocky canyon)
...

➢ Concept Augmentation Using LLM

A detected object is transformed into a similar concept. Our aim focuses on modifying the object's appearance, attributes, and categories while keeping other things and the overall context the same. This includes transforming an object into a more fine-grained instance with richer attributes (e.g., transforming a house to a Victorian one with a wooden entrance), or modifying the background into different environments (transforming the sky into rocky mountains).

➢ Caption Editing

We replace the object in the original source caption with the newly generated phrase. We label the generated image with our edited caption as a ground-truth image-text pair.

➢ Image Editing

To enable fine-grained modification of an image region, we adopt the concept of image inpainting for transforming the original object in the image into the target object. In this scenario, image inpainting involves removing a specific region and filling it with content that seamlessly integrates with the image's context while considering the input information.

4) EXPERIMENTS

Setting: Our training dataset is generated based on the COCO dataset, which has a training split with 110k image-text pairs. In our experiments, we created variations for 12.656 unique images, where for each image, we selected approximately three objects on average and generated four text variations for each object. After filtering out low-quality generations, we ended up with 82.010 image and text pairs. We generate our test set from the COCO Karpathy test split.

We evaluate our model on composition-oriented benchmarks of different scales and different compositional aspects.

1) Winoground is a hand-crafted dataset of 800 image-text pairs, for each set of two texts, the texts have exactly the same words but with different word orders, the texts are mapped to two visually distinct images.
2) ARO has more than 50,000 test images paired with automatically built text examples with changed attributes, relationships, and word order, leveraging VG, COCO, and Flickr.
3) CREPE introduces new negative texts for existing images in CC-12M, YFCC-5M, LAION-400M, where the number of changed words in the text is gradually increased, treated as different levels of complexity.

	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%

Model	Text Score	Image Score	Group Score
CLIP	30.75	11.0	8.75
Ours-CLIP	**34.25**	**12.5**	**10.0**
Relative Gains	+11.1%	+13.6%	+14.2%

	A-Color	A-Material	A-Size	A-State	A-Action	Avg All
CLIP	71	**73.3**	**68**	53.3	62.7	65.66
Ours	**76.6**	68.7	56.23	**57.99**	**72.62**	**66.4**

5) CONCLUSION

Our work tackles the limitations of existing visual language models in terms of compositional reasoning between text and images. We proposed a data generation pipeline that leveraged generative models to introduce challenging negative examples required for contrastive learning. Our proposed method effectively improves the compositionality and discriminative capabilities of VLMs. Experimental results demonstrate that training with our method consistently outperforms existing VLMs on various compositional reasoning benchmark datasets. This was done by addressing the scarcity of hard negative examples for both the image and text modalities. Our work highlights the importance of generative approaches in advancing the field of visual language understanding and bridging the gap between humans and VLMs on compositional reasoning tasks.