A Visual Leap in CLIP Compositionality Reasoning through Generation of Counterfactual Sets
Appendix

A. Dataset Generation Details

In Section 3 of the main paper, we describe the process of obtaining counterfactual pairs using large language models and generative diffusion models. Here, we provide more detailed information on this process.

A.1. More generating examples.

Figure 1 presents additional examples of generated counterfactual sets.

A.2. More details of prompting.

Our use of GPT-4o with the following prompt template:

"You are a master of composition who excels at extracting key objects and their attributes from input text and supplementing the original text with more detailed imagination but no more than ten words, creating layouts that conform to human aesthetics. Your task is described as follows:

Extract the key entities and their corresponding attributes from the input text, and determine how many regions should be split.

For each key object identified in the previous step, perform one of the following random modifications:

1. Randomly delete one object.
2. Randomly add a new object by imitating other objects.
3. Randomly modify the attribute description of one object.

Use precise spatial imagination to assign each object to a specific area within the image and start numbering from 0. The area refers to dividing the entire image into different regions for a general layout. Each key entity is assigned to a region. And for each entity in the region, give it a more detailed description based on the original text. This layout should segment the image and strictly follow the method below:

1. a. Determine if the image needs to be divided into multiple rows (It should be noted that a single entity should not be split into different rows, except when describing different parts of a person like the head, clothes/body, and lower garment):
   (a) If so, segment the image into several rows and assign an identifier to each row from top to bottom (e.g., Row0, Row1, ...).
   (b) Specify the percentage of height each row occupies within the image (e.g., Row0 (height=0.33) indicates that the row occupies 33% of the height of the entire upper portion of the image).
2. Within each row, further assess the need for division into multiple regions (it should be noted that each region should contain only one entity):
   (a) If required, divide each row from left to right into several blocks and assign a number to each block (e.g., Region0, Region1, ...).
   (b) Specify the percentage of width each block occupies within its respective row (e.g., Region0 (Row0, width=0.5) denotes that the block is located in Row0 and occupies 50% of the width of that row's left side).
3. Output the overall ratio along with the regional prompts:
   (a) First, combine each row's height separated by semicolons.
   (b) Secondly, attach each row's regions' width after each row's height separated with comma.
   (c) If the row doesn't have more than one region, just continue to the next row.
   (d) It should be noted that we should use decimal representation in the overall ratio, and if there is only one row, just omit the row ratio.

The output should follow the format of the examples below:

Examples:

Caption: A young boy wearing a blue cap, a yellow t-shirt with a red star, and green shorts.

Key Entities Identification: We identify a boy with the following attributes:

Blue cap (head feature) Yellow t-shirt with a red star (clothes and body feature) Green shorts (lower garment) Plan the Structure Split for the Image: a. Rows Row0 (height=0.33): Top 33% of the image, which is the head of the boy with the blue cap. Row1 (height=0.33): Middle 33% of the image, the body of the boy with the yellow t-shirt. Row2 (height=0.33): Bottom 33% of the image, the lower body of the boy with the green shorts. Each row con-

tains only one region, so no further division within rows is necessary.

b. Regions with Rows: Region0 (Row0, width=1): A bright blue cap sits snugly on the boy's head, slightly tilted to the side, with a few strands of hair peeking out. Region1 (Row1, width=1): A cheerful yellow t-shirt adorned with a bold red star in the center, the fabric slightly wrinkled, fitting comfortably. Region2 (Row2, width=1): Loose-fitting green shorts, reaching just above the knees, with deep pockets and a drawstring waist. Final Split Ratio: 1;1;1

Regional Prompt: A bright blue cap sits snugly on the boy's head, slightly tilted to the side, with a few strands of hair peeking out BREAK A cheerful yellow t-shirt adorned with a bold red star in the center, the fabric slightly wrinkled, fitting comfortably BREAK Loose-fitting green shorts, reaching just above the knees, with deep pockets and a drawstring waist.

Compositionality Prompt: A bright blue cap sits snugly on the boy's head, slightly tilted to the side, with a few strands of hair peeking out is on the top of a cheerful yellow t-shirt adorned with a bold red star in the center, the fabric slightly wrinkled, fitting comfortably is near the Loose-fitting green shorts, reaching just above the knees, with deep pockets and a drawstring waist." .

B. Dataset Stitching Details

B.1. More Stitching examples.

Figure 2 presents additional examples of merged counterfactual sets.

B.2. More details of prompting.

Our use of GPT-4o with the following prompt template:

"Using the reference collage, describe the relative positions of objects and their attributes in the four images that compose the larger picture. Do not describe the image itself, but only the objects and their relationships. For example, 'a young person left to a white dog, and on the right of the dog is a small cat' or '2 stopped cars above an old man, and a bus parked nearby the old man.' Your description should be between 60 to 75 tokens and output in a single, long sentence without breaking it into segments. Use the list of reference captions for each object, with the order in the list being top-left, top-right, bottom-left, and bottom-right."

C. Innovation of the Proposed Loss Function

Traditional vision-language models (VLMs) often use contrastive loss functions, to align visual and textual representations. These loss functions typically consider one positive example per anchor and rely on a large number of negative examples drawn from the batch or a memory bank. This dependence on numerous negative samples can lead to inefficiencies in training, including increased computational overhead and sensitivity to batch composition.

To address these inefficiencies, we introduce a novel loss function that incorporates multiple positive examples and emphasizes pairwise relationships within structured data groups, which we refer to as *counterfactual sets*. By transforming the learning paradigm from a one-to-many framework to a many-to-many framework, our approach leverages the richer information inherent in these counterfactual sets.

C.1. Leveraging Multiple Positive Examples from Counterfactual Sets

In our method, each *counterfactual set* comprises several images and texts that represent variations of a common theme or concept, differing in specific attributes or relationships. Rather than treating only one image-text pair as positive and the rest as negatives, we consider all correct image-text pairings within the set as positive examples. For any given image or text, there may be multiple corresponding positives within the same set.

By incorporating multiple positive examples, the model learns more generalized and robust representations that capture the underlying semantics shared among the positive examples. This approach contrasts with traditional methods that often overlook these nuances by focusing solely on individual positive-negative pairs.

C.2. Reformulating Contrastive Learning with Pairwise Loss

To effectively utilize multiple positive examples, we reformulate the loss function to focus on pairwise comparisons within each counterfactual set. Our loss function operates over all possible image-text pairs within a set, assigning labels based on whether or not the pairs match.

Given a counterfactual set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{m}$, where each x_i is an image and y_i is the corresponding text, we define the pairwise loss $\mathcal{L}_{\text{pair}}$ as:

$$\mathcal{L}_{\text{pair}} = \sum_{i=1}^{m} \sum_{j=1}^{m} \ell\left(\mathcal{I}(x_i, y_j), l_{ij}\right), \tag{1}$$

Here,

- $\mathcal{I}(x_i, y_j)$ denotes the similarity score between image x_i and text y_j,
- $l_{ij} = 1$ if x_i and y_j form a matching pair (positive), and $l_{ij} = -1$ if they do not (negative),
- ℓ is a loss function that penalizes incorrect matches and rewards correct ones.

A practical choice for ℓ is the binary cross-entropy loss with a sigmoid activation function:

$$\ell\left(\mathcal{I}(x_i, y_j), l_{ij}\right) = -\log\left(\sigma\left(l_{ij} \cdot \left(\mathcal{I}(x_i, y_j) - b\right)\right)\right), \tag{2}$$

where σ is the sigmoid function, and b is a bias term.

This formulation naturally incorporates multiple positive pairs and captures the relational structure within the set. By considering all pairwise relationships, the model learns to distinguish fine-grained differences and similarities among the images and texts.

C.3. Achieving Enhanced Performance with Reduced Computational Cost

Traditional contrastive learning methods often require large batch sizes to obtain a sufficient number of negative samples, resulting in high GPU memory usage and increased computational costs. These requirements can pose practical limitations, especially when computational resources are constrained.

By leveraging the inherent structure of counterfactual sets, we reduce the need for extensive negative sampling. Negative examples are effectively provided by the mismatched pairs within each set, which are semantically relevant yet distinct. This approach allows us to use smaller batch sizes without compromising the diversity and effectiveness of negative samples.

With smaller batch sizes, the memory footprint per training iteration decreases, directly reducing GPU memory consumption. This reduction enables training on hardware with limited memory capacity and allows for more models or experiments to be conducted simultaneously. Additionally, smaller batches lead to faster data loading times and lower latency in computation, contributing to a more efficient training process.

D. Impact of Image Background in Counterfactual data

We investigate how incorporating diverse backgrounds in our synthetic images affects compositional reasoning performance and model robustness. Rather than generating isolated objects against plain backgrounds, we systematically introduce varied contextual backgrounds (e.g., changing "a red cube beside a blue sphere" to "a red cube beside a blue sphere on a wooden table/in a grassy field/against a brick wall").

Models are fine-tuned on 300K images with and without our background augmentation technique. The results demonstrate consistent improvements across multiple benchmarks when background augmentation is applied:

On the ARO-A benchmark, models trained with background-augmented images achieve 73.4% accuracy, representing a solid improvement of 1.8 percentage points over the 71.6% baseline without background augmentation. For ARO-R, adding diverse backgrounds yields a more substantial 2.1 percentage point improvement (85.7% vs. 83.6%).

On the VL-Checklist benchmark, models trained with background diversity show a 1.5 percentage point improvement (85.3% vs. 83.8%). Most impressively, on the challenging Winoground dataset, background-augmented training leads to a 2.3 percentage point improvement (20.8% vs. 18.5%).

These consistent improvements across all benchmarks (ranging from +1.5% to +2.3%) demonstrate that incorporating diverse backgrounds serves as an effective data augmentation strategy. The approach introduces beneficial visual variations that help models better distinguish between compositional relationships regardless of context.

Our analysis suggests background augmentation enhances model robustness by forcing the vision encoder to focus on essential object relationships across varying contexts. This makes models less sensitive to background artifacts and more attentive to the spatial relationships that define compositional understanding. Additionally, the variety of backgrounds prevents models from forming spurious correlations between specific backgrounds and object configurations.

These findings extend our understanding of effective data augmentation for VLMs, showing that while object relationships remain central to compositional reasoning, contextual diversity through background variation contributes meaningfully to model robustness. This insight is incorporated into our final GCS model training pipeline, contributing to the state-of-the-art results.

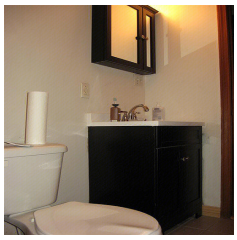
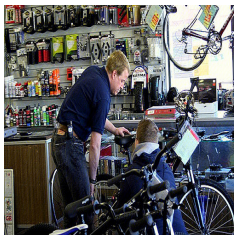
COCO	Counterfactual pair1	Counterfactual pair2	Counterfactual pair3
			
This is a simple styled bathroom. A white toilet is positioned on the left side. On top of its tank, there's a roll of toilet paper. Adjacent to the toilet, on the right, is a black bathroom vanity. Above the vanity, there's a mirrored medicine cabinet. Below the cabinet is the sink, with a faucet and some liquid soap placed beside it. The light colored walls and the dark tiled floor complete the setup, creating a neat and clean space.	A modern bathroom with a minimalist design. The focal point is a sleek, white pedestal sink with clean lines. Above the sink is a rectangular mirror with a backlit frame, providing a soft, ambient light. On the sink, there are a few neatly arranged items, including a soap dispenser, a toothbrush holder, and a small tray with additional toiletries. To the right of the sink, a towel ring holds a neatly folded white towel. The walls and floor are covered in light-colored tiles, giving the space a clean and cohesive look.	A modern, minimalist bathroom. The bathroom features a white toilet and a sleek white vanity with a vessel sink and a contemporary faucet. Above the sink is a large mirror with a wooden frame. To the left of the toilet, there is a potted plant adding a touch of greenery to the space. The walls and floor are light-colored, contributing to the clean and airy feel of the room. There is a window on the right side, allowing natural light to enter the space. The overall design is simple and elegant.	A modern, minimalist bathroom. The bathroom features a white toilet and a sleek white vanity with a vessel sink and a contemporary faucet. Above the sink is a large mirror with a wooden frame. To the left of the toilet, there is a potted plant adding a touch of greenery to the space. The walls and floor are light-colored, contributing to the clean and airy feel of the room. There is a window on the right side, allowing natural light to enter the space. The overall design is simple and elegant.
			
A man in a dark blue shirt and jeans is bending over, seemingly engaged in showing or discussing something related to a bicycle with another person who is sitting or crouching. They are in the middle - foreground of the image. The shop is filled with various cycling - related items. On the wall behind them, there are shelves stocked with bottles and other small products. Above the shelves, a bicycle is hanging from the ceiling.	A man standing in a bicycle shop. He is positioned on the right side of the image, looking at the bicycles. The shop has two rows of bicycles: one row is mounted on the wall at eye level, and the other row is on the floor below. The wall-mounted row has bicycles hanging by their wheels, while the floor row has bicycles standing upright. The shop is well-lit with overhead lighting, and the background includes more bicycles and cycling gear.	A male customer intently inspecting the bicycle's handlebars, his curiosity evident in his focused gaze is on the left of another male, possibly a store employee, gesturing towards the bike's gears, engaging in technical explanation is above the lower section showcasing the sturdy build of the bike and the attentive stance of the two males, emphasizing their interest in the bike's design.	Two people standing with a bicycle. The person on the left is wearing a blue jacket and is holding the bicycle by the handlebars. The person on the right is wearing a green hoodie and is leaning on the bicycle, with one foot on the ground and the other on a pedal. The background is a plain, textured wall, and the ground appears to be a paved surface. The bicycle is silver with black wheels.
			
A person wearing a red shirt and a hairnet is in the foreground on the right side, using a knife to cut a large, light yellow cake that is placed on a surface in the lower left of the image. In the background, it appears to be a food processing or kitchen area. There are stainless steel equipment on the left side, a trash bin near the middle, and a door further back	A woman with a delicate net covering her hair, her eyes concentrated on the precise action of slicing through the cake, the net adding a touch of elegance is on the right of the cake, adorned with frosting and decorations, as the knife smoothly cuts a piece, capturing the anticipation of the celebration.	A cake on a white cake stand in the center. A hand is holding the stand from the right side. On top of the cake is a strawberry with white drizzle cascading down. To the left of the cake, there is a slice of cake on a small white plate, garnished with strawberries. In the background, there is a wooden shelf with jars and containers. Below the shelf, two pans are hanging on the wall.	A person holding a cake on a metal cake stand behind the cake has a smooth, creamy frosting with a caramel drizzle on top. One hand holding the stand and the other holding a fork. The background is a neutral, light-colored setting, keeping the focus on the cake.

Figure 1. The collection of counterfactuals generated by diffusion models. First, a large language model (LLM) is used to identify and diversify core entities from complex text descriptions, embedding them into detailed scenes and generating corresponding individual images. Next, these images are positioned and spliced based on calculated coordinates. The spliced image serves as a reference input for the generation model, while the text description acts as a prompt to help generate an image that closely matches the description.

In the top left of the collage, a white dog on a green grassy field runs with a yellow frisbee in its mouth. Directly to its right, in the top right, there's a wooden cart loaded with suitcases beside a body of water. Below the dog, in the bottom left, an old fashioned clock tower with a balcony stands against the sky. Adjacent to it on the right, in the bottom right, is the long neck and head of a giraffe with a colorful building in the background.

In the top left, an orange wheeled cart laden with boxes is positioned beside water. To its immediate right, in the top right corner, there's a clock tower featuring a clock and a balcony. Directly beneath the cart, in the bottom left, a giraffe's head is in close-up view with a colorful building backdrop. Adjacent to the giraffe, towards the bottom right, a white dog is seen running on a grassy field, gripping a yellow frisbee in its mouth.

Figure 2. The counterfactual collection, created by merging real images, is showcased. A large language model (LLM) is utilized to synthesize descriptions corresponding to each image. The positions of various entities within the composite image are determined through the specified coordinates. These positional relationships are then integrated into more detailed and precise descriptions, enhancing the coherence between the visual content and its textual representation.