FINECAPTION: Compositional Image Captioning Focusing on Wherever You Want at Any Granularity

Supplementary Material

8. More Quantitative Analysis of COMPOSITIONCAP

In this section, we present additional quantitative analyses of COMPOSITIONCAP to highlight the diversity and richness of compositional image captions.

Figure 6 shows a word cloud generated from the captions in COMPOSITIONCAP, illustrating the prominence of key descriptive terms across different regions in the dataset. Words like *front*, *white*, *right*, and *top* dominate the cloud, reflecting the dataset's emphasis on spatial positioning, attributes, and object relations. The diverse vocabulary extends to nuanced properties such as *shadow*, *material*, and *texture*, which underline the dataset's compositional expressiveness. This variety ensures that models trained on COMPOSITIONCAP are well-equipped to handle multi-faceted image regions with complex relationships.

Figure 7 provides a histogram of attribute counts in COMPOSITIONCAP, showcasing the distribution of entities across various descriptive aspects. The most frequent attributes include basic spatial and color properties, while less common features involve material and gesture-based descriptions. This balanced distribution supports comprehensive evaluation of model performance across common and rare descriptive cases.

Figure 8 depicts the distribution of mask size ratios in COMPOSITIONCAP, representing the ratio of the region mask size relative to the entire image. This distribution highlights a diverse range of region sizes, from small, focused objects to large, encompassing areas. The gradual decline in counts as mask size ratios increase demonstrates the dataset's emphasis on evaluating both fine-grained and holistic compositional capabilities.

The quantitative insights presented here reaffirm the utility of COMPOSITIONCAP as a robust benchmark for evaluating aspect-aware compositional reasoning in image captioning tasks.

9. Explanation of Attributes

Table 6 provides detailed explanations of the attributes used in the Attribute-Aware Regional Captioning task.

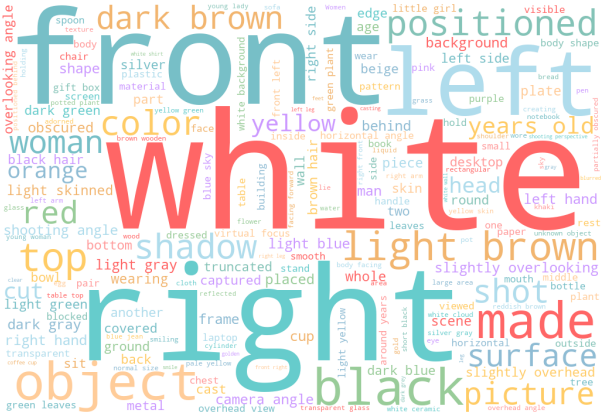


Figure 6. Word cloud of key terms from the captions in COMPOSITIONCAP, illustrating the diverse compositional region description of images.

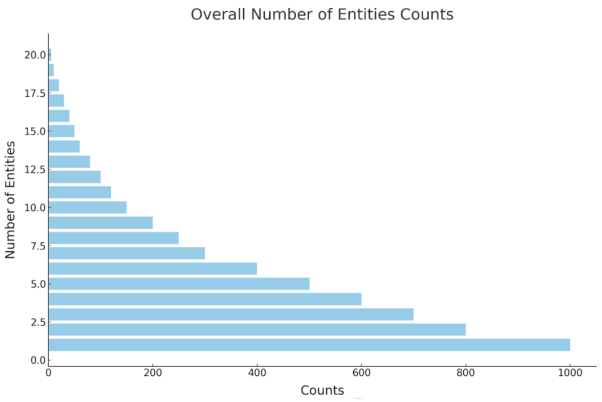


Figure 7. Distribution of entities in COMPOSITIONCAP.

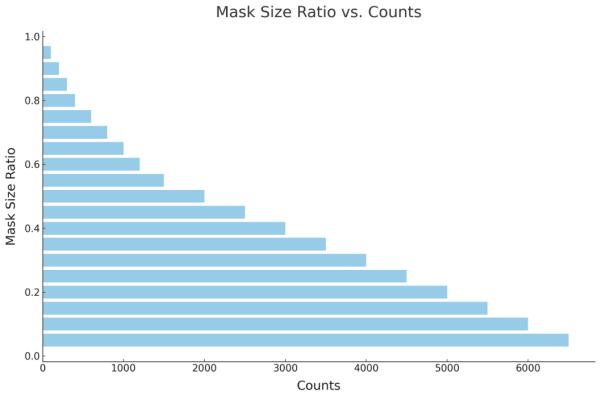


Figure 8. Distribution of mask ratio in COMPOSITIONCAP.

No.	Aspect Explanation
1	**Category Name**: The general label or classification identifying the main subject in the image region, such as ”dog,” ”tree,” or ”car.”
2	**Body Shape**: The form or outline of a living being’s physique, including size, proportions, and overall build.
3	**Skin Texture and Color**: The appearance of the skin’s surface, detailing aspects like smoothness, roughness, and pigmentation.
4	**Clothing, Shoes, Accessories**: The garments, footwear, and additional items worn or carried by a person, reflecting style or function.
5	**Interaction with Other Objects**: How the subject is engaging with surrounding items, such as holding, sitting on, or leaning against something.
6	**Body Pose/Gesture**: The positioning and movement of the subject’s body parts, indicating action or posture.
7	**Other Attributes**: Additional characteristics not covered by other aspects, like patterns, markings, or unique features.
8	**Relative Location with Other Objects**: The spatial relationship between the subject and other elements in the scene, indicating proximity or arrangement.
9	**Color**: The hues and shades present in the subject, contributing to its visual appearance.
10	**Materials/Texture**: The substance an object is made of and the feel of its surface, such as metal, wood, smooth, or rough.
11	**Camera Viewpoint**: The angle and perspective from which the image is captured, like frontal, side, aerial, or close-up views.
12	**Associative Visual Effect**: Visual elements that create specific impressions or moods, such as shadows, reflections, or blurs.
13	**Shape**: The external form or outline of an object, defining its geometry and structure.
14	**Facial Expression**: The look on a person’s face conveying emotion, like smiling, frowning, or surprised.
15	**Hair**: The style, color, length, and texture of hair on a person or animal.
16	**Age Ranging**: An estimation of the subject’s age group, such as infant, child, teenager, adult, or elderly.
17	**Object Pose for Deformable Object**: The positioning and form of objects that can change shape, like a twisted rope or crumpled paper.
18	**Style**: The distinctive appearance or design of the subject, reflecting artistic trends, fashion, or aesthetic elements.

Table 6. Explanation of 18 Aspects in Attribute-Aware Regional Captioning task of FINECAPTION.

10. More Details and Cases

In this section, we provide the prompt used for the GPT-4-as-a-Judge evaluation method. Additionally, we present more examples of our collected data in COMPOSITIONCAP. Figures 9 to 12 illustrate the diversity and richness of our dataset. Figure 13 to Figure 16 showcase the predictions of FINECAPTION for the Regional Dense Captioning task.

Prompt for GPT4-as-a-Judge

Evaluator Instructions:
You are an evaluator tasked with assessing the reasonableness of a model-generated caption for a specific attribute in a masked region of an image.
You will be provided with:
An image with a masked region (region of interest).
A model-predicted caption.
A reference description.

Important Notes:
The model's prediction does not need to exactly match the reference; it is acceptable as long as it reasonably describes the region and the attribute.
The reference description serves as a suggestion or one possible answer, not an exact target.
This is an open-ended generation task.
Example: If the attribute relates to a person's age, and the prediction is "40-50 years old" while the reference is "45-50 years old," the prediction is considered reasonable.

Your Task:
Determine if the caption accurately and reasonably describes the expected attribute of the region of interest.
Provide a binary answer ("Yes" or "No") based solely on whether the attribute description is reasonable.
Please return "Yes" or "No" only, without any additional information. Please carefully examine all compositional details within the mask region!!

Category Name:
This is a young lady.
Body Pose/Gesture:
She is sitting with her right leg on her sitting leg, with a person on her shoulder and her back to the camera.
Body Shape:
She has a normal body shape.
Skin texture and color:
She has light skin.
Clothing shoes accessories:
She was wearing a light brown sleeveless top on her upper body, blue denim shorts on her lower body, and a light brown wide brimmed hat on her head.
Age ranging:
She is around 30-40 years old.
Relative Location with Other Objects:
She is in front of the man to the right, above the wooden plane, in front of the sky and trees.
Hair:
She has golden hair.
Interaction with Other Objects:
She sat on a wooden surface, with a man resting on her right shoulder.
Associative visual effect:
There are some shadows behind her that reflect on the wooden surface.

Figure 9. Example for Attribute-Aware Regional Captioning task in COMPOSITIONCAP.

Category Name:
This is a rabbit.
Color:
Its hair is gray in color and its eyes are black.
Materials/Texture:
The hair material of the rabbit is soft and shiny, and the texture is delicate and layered.
Camera Viewpoint:
It was shot from a horizontal perspective.
Interaction with Other Objects:
It lies on the grass, with its chin resting on the blue egg.
Relative Location with Other Objects:
It is to the right of the pink egg, to the left of the red egg, to the rear of the right of the egg, above the blue egg, and above the grass.
Object pose for deformable object:
It is lying down, ears erect, and looks very vigilant.

Figure 10. Example for Attribute-Aware Regional Captioning task in COMPOSITIONCAP.


Category Name:
This is a young woman.
Body Shape:
Her physique is normal.
Facial expression:
She is smiling.
Skin texture and color:
Her skin is yellow.
Body Pose/Gesture:
She sat with her left arm on her right thigh, her right leg on her left leg, her left hand naturally hanging down, her right hand leaning on her chin, her head slightly leaning to the right, and her eyes looking forward.
Clothing shoes accessories:
She wore a light grey short-sleeved shirt, dark blue jeans, a silver necklace around her neck, and a silver earring on her right ear.
Age ranging:
She is around 30-40 years old.
Relative Location with Other Objects:
Below her is the sofa, behind it is the wall and sofa backrest, behind it is the sofa backrest, behind it is the sofa backrest which is cut off from the picture in a large area.
More Description:
The woman's knee below is cut off outside the frame.
Hair:
She has long straight hair that gradually changes from linen brown to linen yellow, reaching the length of her chest

Figure 11. Distribution of attributes in COMPOSITIONCAP.


Category Name:

This is a bunch of grass with colorful flowers.

Color:

The color of the bouquet is purplish red, white and light purple, and the color of the grass is green.

Camera Viewpoint:

She is smiling.

Interaction with Other Objects:

Her skin is yellow.

Relative Location with Other Objects:

Behind it is a person, and below it is sand and cloth.

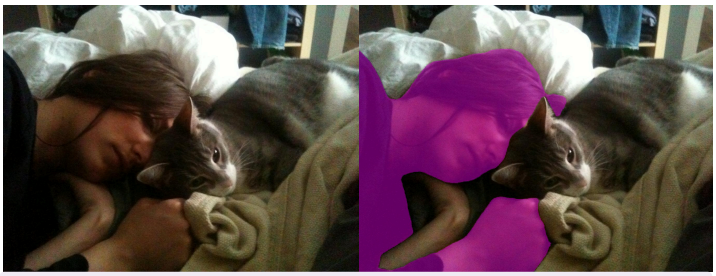
Figure 12. Example for Attribute-Aware Regional Captioning task in COMPOSITIONCAP.

Instruction:
Please generate a dense caption describing the masked contents in the image.

A white book with black text and a light gray pattern, made of paper, is placed on a shelf. The book is shot from a slightly overlooking angle, positioned above the shelf, to the right of another book.

Ground Truth:
A grey book with a white label and black text, encased in clear plastic, is photographed from the front, revealing its rectangular shape. It is placed inside a wooden cabinet, positioned to the right of another book and below a third book.

Figure 13. Case study for Regional Dense Captioning task for FINECAPTION.



Instruction:
Please generate a dense caption describing the masked contents in the image.

A young lady, aged between 25 and 35, with long brown hair and light skin, lies on her side with her head facing forward and her right hand resting on an object. She wears a dark gray long-sleeved shirt, and her normal body shape is partially covered by a blanket and a cat. Her shadow is cast on the bed beneath her, and she is positioned in front of the cat and the wall.

Ground Truth:
A woman around 30 years old with long dark brown hair lay on her side, her left hand curled up against her chin and her right hand clenched in front of it. She wore a black long-sleeved sweater, and her expression was in a daze. Her yellowish complexion was rosy, and her body shape appeared normal despite her amputated shoulder. Positioned to the left of a cat, she lay at the bottom left of a white quilt and the top left of a beige blanket.

Figure 14. Case study for Regional Dense Captioning task for FINECAPTION.

Instruction:
Draft a dense description for the parts of the image identified by the mask.

A young, light-skinned woman with long brown hair, dressed in a white long-sleeved top, grey trousers, and yellow-brown boots, sits on a bench with her feet on the ground. She faces the right rear, her right arm resting on her right leg, while her head turns to the right front. Her shadow is cast on the ground below her, and she appears to be between the ages of 25 and 35.

Ground Truth:
A young woman, aged between 20-30 years, with yellowish-brown, breast-length hair and white skin, sat cross-legged on a platform, her hands resting on her left thigh. She wore a grey and off-white striped top, blue jeans, and yellow-brown boots, and a smile graced her face. Her shadow was cast on the platform below, where, to her front right, stood a lady with black hair, and to her rear right, a sculpture.

Figure 15. Case study for Regional Dense Captioning task for FINECAPTION.

Instruction:
Generate a comprehensive caption for the masked portions of the image.

A young, light-skinned man, aged between 25 and 35, stands with his body facing the right front, his head twisted to the left rear, and his right arm naturally drooping. He has short black hair and is dressed in a white T-shirt and black trousers. Positioned behind another man to the right and in front of a background, he presents a composed and attentive demeanor.

Ground Truth:
A light-skinned man with short black hair, around 25-35 years old, stands facing forward with a smile, his right and left hands naturally drooping. He wears white clothes adorned with black and gray patterns on his chest. The man is positioned against a background, slightly to the right and in front of another background element, and to the left of an unknown object.

Figure 16. Case study for Regional Dense Captioning task for FINECAPTION.