

Semantic Editing Increment Benefits Zero-Shot Composed Image Retrieval

Supplementary Material
Anonymous Authors

A PROMPT TEMPLATES

We harness the capabilities of LLM to seamlessly integrate the rich textual information extracted from reference images with the editing information or constraints provided by relative text. This integration enables the inference of natural language for the purpose of editing. By utilizing multiple prompts, the approach generates breadth-edited captions that capture diverse interpretations and perspectives, thereby facilitating breadth inference.

Formally, given a reference caption T_i^C of the reference image and relative text T , we design a simple prompt template $f(\cdot, \cdot)$, combining the reference caption and relative text to create a full prompt for LLM:

$$p_i = f(T_i^C, T), \quad (1)$$

where p_i represents the prompt that combines the reference image caption and relative text. The image caption serves as a content prompt prepended with "Image Content:", and the relative text serves as a modification instruction prepended with "Instruction:". We fill these two parts into the template to get the full prompt. Then, we input the prompt into the LLM for reasoning and obtain the generated edited caption T_i^E :

$$T_i^E = \text{LLM}(p_i). \quad (2)$$

We include only one example in each LLM query for compositional reasoning to generate a variety of breadth-edited captions efficiently and cost-effectively. The prompt templates $f(\cdot, \cdot)$ for CIRCO/CIRR and FashionIQ are shown as follows:

Prompt for CIRCO/CIRR

I have an image. Given an instruction to edit the image, carefully generate a description of the edited image. I will put my image content beginning with "Image Content:". The instruction I provide will begin with "Instruction:". The edited description you generate should begin with "Edited Description:". You just generate one edited description only beginning with "Edited Description:". The edited description needs to be as simple as possible and only reflect image content. Just one line.

An example:

Image Content: a man adjusting a woman's tie.
Instruction: has the woman and the man with the roles switched.
Edited Description: a woman adjusting a man's tie.
Image Content: {reference caption}
Instruction: {relative text}
Edited Description: {}

Prompt for FashionIQ

I have a {dress type} image. Given an instruction to edit the {dress type}, carefully generate a description of the edited {dress type}. I will put my image content beginning with "Image Content:". The instruction I provide will begin with "Instruction:". The edited description you generate should begin with "Edited Description:". The edited description needs to be as simple as possible and only reflects {dress type} content. The subject of Edited Description is the {dress type}. Just one line.

An example:

Image Content: woman in a blue floral dress.
Instruction: make it that is shiny and silver with shorter sleeves, and fit and flare.
Edited Description: a shiny, silver, fit and flare dress with short sleeves.
Image Content: {reference caption}
Instruction: {relative text}
Edited Description: {}

B QUALITATIVE RESULTS

In Figure 1, we display qualitative results for composed image retrieval, performed on the test set of CIRCO. Our results display a reference image, a relative text, breadth edited captions, and the top-1 retrieved image in each column. In the breadth edited captions, words lacking semantic relevance are highlighted in red, while semantically matching words are highlighted in green. The outcomes illustrate the efficacy of our SEIZE method in accurately retrieving the desired image. For example, in the first row, the model needs to preserve the semantic categories of the glasses and cellphone depicted in the reference image and position it near the water to successfully retrieve the target image. We have deliberately chosen a range of representative retrieval pairs to showcase the flexibility of our method across various editing scenarios. These alterations involve different semantic aspects like color, background, quantity, zoom, and content. This diversity enables us to demonstrate how SEIZE can adapt to alterations in the relative text from multiple viewpoints, emphasizing its versatility and resilience.

C RESULT REPLICATION

Our code and the associated datasets are openly available. Detailed instructions on how to download the datasets, set up the necessary environment, and replicate our experimental results can be found at the link provided below. Additionally, to facilitate the validation of our findings, we have made available pre-computed results that can be directly submitted to the test platforms.

	Query	Edited Breadth Captions			Target Image
Add, Content	 <i>has a body of water in the background</i>	<i>A man wearing glasses on a cellphone ...</i>	<i>A man talking on his cellphone ...</i>	<i>... A man standing on a rocky hill using a ...</i>	
Remove, Content	 <i>is only one and is seen from the side</i>	<i>One bottle seen from the side on a bench ...</i>	<i>A single beer bottle viewed from the side ...</i>	<i>... A brown bottle sitting on a bench, ...</i>	
Count, Background	 <i>is only one and is inside a display case</i>	<i>A single cell phone inside a display case</i>	<i>a single old style cell phone inside ...</i>	<i>... sits inside a display case on a speaker</i>	
Background	 <i>is shot on a beach and has one person next to it</i>	<i>... standing next to an elephant on a beach.</i>	<i>Two people sitting on the backs of the elephant ...</i>	<i>... One man and a young child sitting on the back ...</i>	
Color	 <i>is white and has a microwave oven next to it</i>	<i>A white refrigerator without a cover ...</i>	<i>... a microwave oven next to it on the counter</i>	<i>... with plastic wrapping over the handles</i>	
Zoom, Background	 <i>shows more trees, is zoomed in and the girl is not smiling</i>	<i>A serious young lady is sitting on ...</i>	<i>... sits on a wooden bench in a zoomed-in view</i>	<i>... A non-smiling woman in a white dress sits ...</i>	

Figure 1: Qualitative results on the test set of CIRCO. Within the breadth edited captions, red words are semantically irrelevant to the retrieved image and green words semantically match the retrieved image.

The code and accompanying resources are publicly accessible at the following URL: <https://anonymous.4open.science/r/SEIZE-11BC>.