

## A SPECIFIC COMPLEX EXAMPLES OF RECOS

In Section 3.1 and Section 4.3, we have shown the twelve categories of data covered in ReCos and the accuracy of five different models in image-text retrieval in each category. Also, we make certain analysis with results. According to Figure 6 and Figure 7, it can be clearly concluded that existing large visual language models such as blip2 have excellent performance in both coarse-grained retrieval and fine-grained retrieval tasks. However, the results are unsatisfactory in areas such as code understanding, language translation, and numerical calculations. Therefore, we carefully selected examples of recognition errors on these three categories of BLIP2 and CLIP.

The example on the left is an image-text pair for numerical calculation, which requires the model to have certain abilities in mathematical graphics and master the calculation ability of polygon area. The difficulty is greatly improved compared with simple addition and subtraction calculations.

The middle example examines the translation ability of the model. Different from ordinary text translation, the model not only needs to be able to master Chinese and English translation at the same time, but also needs to have a certain logical understanding ability, because the text description itself is not a translation of the text in the picture.

The example on the right contains a simple code diagram and its text explanation. The existing model can generate corresponding code according to the user's needs, but it still lacks good interpretability of the specific code, and in the image and text retrieval task, the model It is necessary to accurately identify the code blocks in the picture first.

## B TEXT DESCRIPTION GENERATES PSEUDO CODE

In Section 3.2, we provided a detailed overview of the image annotation process. In this section, we present an alternative verification solution from an engineering perspective to ensure that the generated image descriptions are fine-grained. The primary goal is to ensure that the generated image descriptions match as closely as possible with only one image from the candidate pool, thereby ensuring the effectiveness of fine-grained retrieval.

In this Section, the pseudocode flow chart generated by five text descriptions corresponding to each image in ReCos will be provided. The flowchart corresponding to the pseudocode is shown in Figure 9. It will serve as a supplement to Section 3.2 and reflect the logic and rigor in the construction process of recos.

### B.1 Coarse-grained Text Refinement Algorithm

In the given pseudocode for the Coarse-grained text refinement (CTR), several parameters and functions play crucial roles. The input parameters include a coarse textual query  $q_c$ , an image candidate pool  $P$ , a specified number  $k$  of top images to retrieve, and the ground truth image  $I_{gt}$ . The output is a fine-grained textual description  $q_f$ . The core of the algorithm is encapsulated in three main steps:

- **Retrieve Top-k images.** This step involves the function `RetrieveTopK`, which computes text and image embeddings using the BLIP2 model denoted by  $\theta$ . It calculates cosine similarity between the query embedding and each image

embedding in the pool to identify the top  $k$  most relevant images.

- **Validate and Refine Textual Description.** Here, the algorithm checks if the top-1 image from the retrieved set matches the ground truth. If not, it proceeds to verify whether the true image is within the top  $k$  and whether the current description  $q_c$  can uniquely identify  $I_{gt}$ . If the description is not sufficiently detailed, it is refined through the `RefineDescription` function.
- **Output the refined description.** Finally, the algorithm outputs the refined description  $q_f$ , which should accurately describe the ground truth image in a detailed manner.

---

### Algorithm Coarse-grained Texts Refinement (CTR)

---

**Input:** Coarse textual query  $q_c$ , image candidate pool  $P$ , maximum number of top images  $k$ , ground truth image  $I_{gt}$ ;

**Output:** Fine-grained textual description  $q_f$

1:  $U \leftarrow$  empty set;  $U' \leftarrow P$ ;  $q_f \leftarrow q_c$  2:  $\theta \leftarrow$  initialize BLIP2 model with pretrained weights

// Step 1: Retrieve Top-k images using BLIP2

3: **function** `RETRIEVE_TOPK`( $q_c, P, k$ )

4:    $E_q \leftarrow \theta.\text{text\_to\_embedding}(q_c)$

5:   **for each**  $I \in P$  **do**

6:      $E_i \leftarrow \theta.\text{image\_to\_embedding}(I)$

7:      $\text{sim} \leftarrow \cos(E_q, E_i)$  // Cosine similarity

8:     add  $(I, \text{sim})$  to  $U$

9:   **end for**

10:    $U \leftarrow$  sort  $U$  by  $\text{sim}$  in descending order

11:   return  $U[1 : k]$  // Top-k images

12: **end function**

// Step 2: Validate and Refine Textual Description

13:  $U \leftarrow \text{RETRIEVE\_TOPK}(q_c, P, k)$

14: **if**  $U[1].\text{image} = I_{gt}$  **then**

15:    $q_f \leftarrow q_c$  // Top-1 image is ground truth

16: **else**

17:   **if**  $I_{gt} \in U.\text{images}$  **then**

18:     **if** `HUMAN_VERIFICATION`( $q_c, I_{gt}$ ) **then**

19:        $q_f \leftarrow q_c$  // Description is distinguishable

20:     **else**

21:        $q_f \leftarrow \text{REFINE\_DESCRIPTION}(q_c, I_{gt})$

22:        $q_c \leftarrow q_f$

23:       goto 13 // Repeat the refinement process

24:     **end if**

25:   **end if**

26: **end if**

// Step 3: Output the refined description

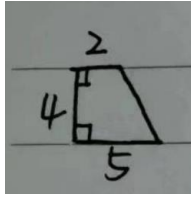
27: return  $q_f$

---

## C PROMPT FOR GRAMMAR CHECKING

The main purpose of this template is to perform a check for grammatical correctness on the generated text descriptions, ensuring their initial syntactic accuracy. Ultimately, human sampling checks will further ensure the correctness of grammar and the uniqueness of the text descriptions.

1 TARNNS\_VERIFY = (



奶茶的香浓味道

```
my_list = ['Hello', 'FGCE']
my_str = ' '.join(my_list)
print(my_str)
```

The right-angled trapezoid has an area calculated by averaging its bases of 2 and 5 units, and multiplying by the 4-unit height, totaling 14 square units.

The Chinese content in the plain text image is '奶茶的香浓味道'

The code demonstrates the use of the join() method to merge list elements into a single string with a space.

Figure 8: Samples from cognition-based retrieval tasks where both BLIP2 and CLIP struggle.

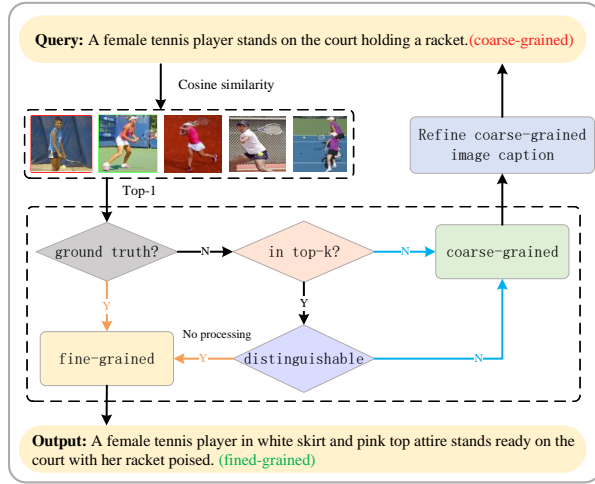


Figure 9: Building process of refining coarse-grained text descriptions

```
"I hope you will act as an English grammar reviewer.
Your task is to check and review whether the
grammatical "
"modifications of the following independent English
sentences are correct.\n"
"If the sentence has no grammatical errors, don't
make any changes!\n"
"If there are no grammatical errors in the sentence's
grammar, please output YES at the beginning of the
"
"sentence. otherwise output NO.\n "
"The sentence format to be reviewed is as follows:\n
"
"-----\n"
"<correct or error> <origin sentence> --- <revised
sentence>(Optional)\n"
"\n-----\n"
"The specific meaning and explanation are as follows
:\n"
"-----\n"
```

```
"The correct or error in the sentence represents the
original evaluation result of the sentence, "
"the following sentence is the original sentence, ---
the following sentence represents the modified
sentence "
"\n-----\n"
"The format for your review is as follows: \n"
"-----\n"
"<correct or error> <origin sentence> --- <revised
sentence>(Optional)\n"
"-----\n"
"For example:\n"
"-----\n"
"YES correct a man wearing a hat and a white shirt is
cleaning windows"
"YES correct a man in a white shirt stands high up on
scaffolding"
"NO error a man stands on boards on top of a huge
ladder --- a man is standing on boards on top of a
huge "
"ladder "
"YES correct a man in a white shirt and hat works on
top of scaffolding"
"YES correct a guy works on a building"
"-----\n"
"We have provided the English sentences that need to
be reviewed for English grammar as follows:\n"
"-----\n"
"{context_str}"
"\n-----\n"
"Based on the above information, Please complete the
English grammar verification and revised review.\n"
)
```

## D PROMPT FOR IMAGE CLASSIFICATION

The primary purpose of this template is to assist humans in roughly categorizing images into 12 subtasks, but the final determination of the images and their categories is made by humans. In this section, we provide several representative templates.

### D.1 Prompt for Image Color Classification

```
1 IMAGE_COLOR_CLASSIFY = (
2     "I hope you will act as an image description color
    recognition expert, that is, "
```

```
1393 3 "you can accurately determine whether 5 image
1394 descriptions are related to color, "
1395 4 "such as: white, black, green, etc.\n"
1396 5 'As long as one of the 5 descriptions involves color,
then you answer yes. '
1397 6 'If none of the five descriptions involve color,
1398 please answer no.\n'
1399 7
1400 8 "For example:\n"
1401 9 "-----\n"
1402 10 "The five descriptions of the image are:\n"
1403 11 "-----\n"
1404 12 "the man with pierced ears is wearing glasses and an
1405 orange hat on his head",
1406 "a man with glasses is wearing a beer can crocheted
1407 hat",
1408 "a man with pierced ears and glasses is wearing an
1409 orange hat on his head",
1410 "a man in an orange hat starring at something",
1411 "a man wears an orange hat and glasses"
1412 16 "-----\n"
1413 17 "Answer: yes\n"
1414 18
1415 19
1416 20
1417 21 "Question:Are the descriptions corresponding to the 5
1418 images related to the color?\n"
1419 22 "We provide five descriptions of the image are:\n"
1420 23 "-----\n"
1421 24 "{context_str}"
1422 25 "\n-----\n"
1423 26 "-----\n"
1424 27 "Answer: <yes or no>\n"
1425 28 "-----\n"
1426 29
1427 30 "Please complete the image caption classification
1428 based on the above information."
1429 31 )
1430 32
```

## D.2 Prompt for Image Position Classification

```
1431 1 IMAGE_POSITION_CLASSIFY = (
1432 2 "I hope you will act as an image description '
1433 orientation description' recognition expert, that is
1434 , "
1435 3 "you can accurately determine whether 5 image
1436 descriptions are related to orientation or position,
1437 "
1438 4 "such as: behind , In front of..., to the
1439 right of... "
1440 5 "below..., above..., east of..., north of..., on top
1441 of..., etc.\n"
1442 6 'As long as one of the 5 descriptions involves
1443 position or orientation, then you answer yes. '
1444 7 'If none of the five descriptions involve position,
1445 please answer no.\n'
1446 8
1447 9 "For example:\n"
1448 10 "-----\n"
1449 11 "The five descriptions of the image are:\n"
1450 12 "-----\n"
1451 13 "a woman with a black shirt and tan apron is standing
1452 behind a counter in a restaurant",
1453 "a female barista dressed in black is making coffee
1454 behind a pink counter",
1455 "a girl in a black shirt is smiling as she works
1456 behind a bar",
1457 "a happy woman in a black shirt and standing in front
1458 of a counter",
1459 16
```

```
1460 17 "a young woman in a black shirt and stands in front
1461 of a counter full of coffee cups"
1462 18 "-----\n"
1463 19 "Answer: yes\n"
1464 20
1465 21
1466 22 "Question:Are the descriptions corresponding to the 5
1467 images related to the position or orientation?\n"
1468 23 "We provide five descriptions of the image are:\n"
1469 24 "-----\n"
1470 25 "{context_str}"
1471 26 "\n-----\n"
1472 27 "-----\n"
1473 28 "Answer: <yes or no>\n"
1474 29 "-----\n"
1475 30
1476 31 "Please complete the image caption classification
1477 based on the above information."
1478 32 )
1479 33
```

## D.3 Prompt for Image Count Classification

```
1480 1 IMAGE_COUNT_CLASSIFY = (
1481 2 "I hope you will act as an image description 'count'
1482 recognition expert, that is, "
1483 3 "you can accurately determine whether 5 image
1484 descriptions are related to count, "
1485 4 "such as: one, two, three, many, few, several, some,
1486 a couple of, a few, a dozen, "
1487 5 "a bunch, a lot of, a bunch of, a couple, etc.\n"
1488 6 "As long as one of the 5 descriptions involves 'count
1489 ', then you answer yes. "
1490 7 "If none of the five descriptions involve 'count',
1491 please answer no.\n"
1492 8
1493 9 "For example:\n"
1494 10 "-----\n"
1495 11 "The five descriptions of the image are:\n"
1496 12 "-----\n"
1497 13 "six people ride mountain bikes through a jungle
1498 environment",
1499 "men surrounded by nature are riding mountain bikes
1500 on a trail",
1501 "there are six men mountain biking in a forest
1502 terrain",
1503 "six people riding bikes on a trail in the forest",
1504 "a group of people is bike riding in the woods"
1505 16 "-----\n"
1506 17 "Answer: yes\n"
1507 18
1508 19
1509 20
1510 21 "Question:Are the descriptions corresponding to the 5
1511 images related to the 'count'?\n"
1512 22 "We provide five descriptions of the image are:\n"
1513 23 "-----\n"
1514 24 "{context_str}"
1515 25 "\n-----\n"
1516 26 "-----\n"
1517 27 "Answer: <yes or no>\n"
1518 28 "-----\n"
1519 29
1520 30 "Please complete the image caption classification
1521 based on the above information."
1522 31 )
1523 32
```

## D.4 Prompt for Image Action Classification

```
1524 1 IMAGE_ACTION_CLASSIFY = (
1525 2 "I hope you will act as an image description 'action'
1526 recognition expert, that is, "
1527 3
```

```

1509 3      "you can accurately determine whether 5 image
1510      descriptions are related to action, "
1511 4      "such as: cycling, skate, play basketball, play
1512      table tennis, play billiards, play tennis, "
1513 5      "climb mountains, surf, taekwondo, wrestling, etc.\n"
1514 6      "As long as one of the 5 descriptions involves '
1515      action', then you answer yes. "
1516 7      "If none of the five descriptions involve 'action',
1517      please answer no.\n"
1518 8
1519 9      "For example:\n"
1520 10     "-----\n"
1521 11     "The five descriptions of the image are:\n"
1522 12     "-----\n"
1523 13     "six people ride mountain bikes through a jungle
1524 14     environment",
1525 15     "men surrounded by nature are riding mountain bikes
1526 16     on a trail",
1527 17     "there are six men mountain biking in a forest
1528 18     terrain",
1529 19     "six people riding bikes on a trail in the forest",
1530 20     "a group of people is bike riding in the woods"
1531 21     "-----\n"
1532 22     "{context_str}"
1533 23     "\n-----\n"
1534 24     "-----\n"
1535 25     "Answer: <yes or no>\n"
1536 26     "-----\n"
1537 27
1538 28     "Please complete the image caption classification
1539 29     based on the above information."
1540 30
1541 31 )

```

## D.5 Prompt for Image Figure Classification

```

1544 1 IMAGE_FIGURE_CLASSIFY = (
1545 2     "I hope you will act as an image description 'figure'
1546 3     recognition expert, that is, "
1547 4     "you can accurately determine whether 5 image
1548 5     descriptions are related to figure, "
1549 6     "such as: boys, man, woman, women, children, students
1550 7     , parents, teachers, teacher, friends, "
1551 8     "strangers, villains, artist, athletes, scholars,
1552 9     captains, employees, bosses, singer,"
1553 10    "singers, actors, authors, detectives, chefs,
1554 11    cleaners, children, child, boy, etc.\n"
1555 12    "As long as one of the 5 descriptions involves '
1556 13    figure', then you answer yes. "
1557 14    "If none of the five descriptions involve 'figure',
1558 15    please answer no.\n"
1559 16
1560 17    "For example:\n"
1561 18    "-----\n"
1562 19    "The five descriptions of the image are:\n"
1563 20    "-----\n"
1564 21    "woman in a white dress standing with a tennis racket
1565 22    and two people in green behind her",
1566 23    "young pretty blond women holding a tennis racket

```

```

18      "a young lady in white holding a tennis racket behind
19      her"
20      "-----\n"
21      "Answer: yes\n"
22
23      "Question:Are the descriptions corresponding to the 5
24      images related to the 'figure'? \n"
25      "We provide five descriptions of the image are:\n"
26      "-----\n"
27      "{context_str}"
28      "\n-----\n"
29      "-----\n"
30      "Answer: <yes or no>\n"
31      "-----\n"
32
33      "Please complete the image caption classification
34      based on the above information."
35
36 )

```

## D.6 Prompt for Image Object Classification

```

1 IMAGE_OBJECT_CLASSIFY = (
2     "I hope you will act as an image description 'object'
3     recognition expert, that is, "
4     "you can accurately determine whether 5 image
5     descriptions are related to object,"
6     "such as: teacher grading papers, tennis player
7     playing tennis, artist painting a masterpiece,"
8     "a man holding a book, a woman in front of the pizza,
9     the motorcycle that the man is riding, etc.\n"
10    "As long as one of the 5 descriptions involves '
11    object', then you answer yes. "
12    "If none of the five descriptions involve 'object',
13    please answer no.\n"
14
15    "For example:\n"
16    "-----\n"
17    "The five descriptions of the image are:\n"
18    "-----\n"
19    "woman in a white dress standing with a tennis racket
20    and two people in green behind her",
21    "young pretty blond women holding a tennis racket
22    dressed as if to begin a tennis match",
23    "a young woman in a fashionable tennis racket",
24    "slim blond woman in white dress holds tennis racket"
25    ,
26    "a young lady in white holding a tennis racket behind
27    her"
28    "-----\n"
29    "Answer: yes\n"
30
31
32    "Question:Are the descriptions corresponding to the 5
33    images related to the 'object'? \n"
34    "We provide five descriptions of the image are:\n"
35    "-----\n"
36    "{context_str}"
37    "\n-----\n"
38    "-----\n"
39    "Answer: <yes or no>\n"
40    "-----\n"
41
42    "Please complete the image caption classification
43    based on the above information."
44
45 )

```

## D.7 Prompt for Image Scene Classification

```

1 IMAGE_SCENE_CLASSIFY = (

```

```

2      "I hope you will act as an image description 'scene'
3      recognition expert, that is, "
4      "you can accurately determine whether 5 image
5      descriptions are related to scene,"
6      "such as: hospital, swimming pool, construction site,
7      roadside, seaside, school, supermarket, "
8      "restaurant, theater, library, park, office, cafe,
9      gym, train station, airport, shop/store, "
10     "swimming pool, museum, residential area, cinema,
11     post office, pet store, meadow, amusement park, etc
12     .\n"
13     "As long as one of the 5 descriptions involves 'scene'
14     ', then you answer yes. "
15     "If none of the five descriptions involve 'scene',
16     please answer no.\n"
17
18     "For example:\n"
19     "-----\n"
20     "The five descriptions of the image are:\n"
21     "-----\n"
22     "three girls wearing goggles are jumping into a
23     swimming pool together",
24     "children jumping in a blue pool surrounded by blue
25     pool chairs and huts",
26     "a woman and three children are jumping into a
27     swimming pool",
28     "children jump off the edge into a pool",
29     "three girls jump into a pool"
30     "-----\n"
31     "Answer: yes\n"
32
33
34     "Question:Are the descriptions corresponding to the 5
35     images related to the 'scene'? \n"
36     "We provide five descriptions of the image are:\n"
37     "-----\n"
38     "{context_str}"
39     "\n-----\n"
40     "-----\n"
41     "Answer: <yes or no>\n"
42     "-----\n"
43
44     "Please complete the image caption classification
45     based on the above information."
46 )

```

## E PROMPT FOR IMAGE CAPTION GENERATION

Regarding the templates for generating image descriptions, we can roughly divide them into the following three categories:

- (1) **Default Image Description Generation Templates.** These templates are suitable for generating descriptions that do not require additional prompts, such as simple descriptions of image color categories.
- (2) **Description Generation Templates with Additional Information.** These templates are used to provide additional information manually when generating description. For example, descriptions regarding quantities. It should be noted that GPT-4 exhibits weaker capability in generating descriptions about quantities, especially concerning the quantities between closely positioned objects. Therefore, it is necessary for us to manually provide additional quantity information.
- (3) **Similar Image Generation Templates with Additional Information.** These templates generate descriptions similar

to the original image description but tailored to the features of the current image based on both the description of the original image and the characteristics of the current image. It can be understood as textual confusion.

### E.1 Prompt for Default Image Caption Generation

For text generation templates, we lists several requirements for text generation to reduce the probability of GPT-4 generating irrelevant text. We also hand-write the corresponding examples, requiring the model to refer to the examples for generation, and make corresponding adjustments based on specific categories of data. Finally, a manual verification step is included to ensure that the generated description can uniquely identify the corresponding image and maintain consistency with the category.

```

1  DEFAULT_IMAGE_CAPTION_GENERATION = (
2      "I want you to play an image description expert.Your
3      task is to generate 5 descriptions as detailed as "
4      "possible based on the image. "
5      "The requirements are as follows:\n"
6      "1. Each description is one sentence.\n"
7      "2. The focus of each description is color, position
8      and action.\n"
9      "3. The grammar is required to be correct and logical
10     .\n"
11     "4. Each description generated should be as detailed
12     and matter-of-fact as possible.\n"
13     "5. Dont generate meaningless, empty descriptions
14     .\n"
15 )

```

### E.2 Prompt for Image Caption Generation with Extra Information.

For complex subcategories or subtasks, such as quantity and OCR (Optical Character Recognition), this template can assist the model in generating more accurate image descriptions by manually providing prompts related to the image description. This approach helps to improve the quality and accuracy of the descriptions.

```

1  IMAGE_CAPTION_GENERATION_WITH_EXTRA = (
2      "I want you to act as an image description expert.
3      Your task is to generate 5 descriptions
4      corresponding"
5      " to the given picture based on the picture and 5
6      descriptions of similar pictures. The description
7      should be"
8      " as close as possible to the 5 descriptions I gave
9      you. It should be as similar as possible and the
10     number of"
11     " modified words should be as few as possible, but
12     the generated description should be different from
13     the 5"
14     " descriptions provided previously.\n"
15
16     "You can refer to the following angles: color,
17     position, existence, action, count, figure, object,
18     scene, etc."
19 )

```



1741 9 "Each description generated can be modified from  
1742 these perspectives according to the corresponding "  
1743 10 "description, requiring the number of words to be as  
1744 small as possible and consistent with the real  
1745 situation "  
1746 11 "of the given picture.\n"  
1747 12  
1748 13 "Please provide 5 detailed descriptions from the  
1749 perspective of the category above.Each description  
1750 should be "  
1751 14 "as short as possible but as detailed as possible.  
1752 Each description should be a grammatically correct  
1753 and "  
1754 15 "complete sentence. Do not give any irrelevant  
1755 explanation other than description.\n"  
1756 16  
1757 17 "To provide you with an idea for generating  
1758 descriptions:\n"  
1759 18 "You should check the descriptions I provide you one  
1760 by one, and then replace the attributes or keywords  
1761 in "  
1762 19 "the description one by one based on the image  
1763 information and the reference categories provided to  
1764 you. "  
1765 20 "For example: first, replace the color of the  
1766 description according to the picture content; if the  
1767 color "  
1768 21 "is not suitable , you can try to replace the  
1769 quantity; if the quantity is also inappropriate, you  
1770 can try "  
1771 22 "to replace the position, and so on. When replacing,  
1772 try to replace only keywords as much as possible. "  
1773 23 "If other non-core words can remain unchanged, they  
1774 should be left unchanged as much as possible. "  
1775 24 "In other words, the replaced description should be  
1776 very similar to the provided description, "  
1777 25 "preferably a description that replaces certain  
1778 keywords and conforms to the image characteristics.\n  
1779 n"  
1780 26  
1781 27 "=====\n"  
1782 28 "For example1: 5 descriptions of similar pictures are  
1783 :\n"  
1784 29 "the man with pierced ears is wearing glasses and an  
1785 orange hat on his head"  
1786 30 "a man with glasses is wearing a beer can crocheted  
1787 hat"  
1788 31 "a man with pierced ears and glasses is wearing an  
1789 orange hat on his head"  
1790 32 "a man in an orange hat starring at something"  
1791 33 "a man wears an orange hat and glasses"  
1792 34  
1793 35 "Based on the given image content, 5 similar image  
1794 descriptions and the reference category, "  
1795 36 "the new description generated is:\n"  
1796 37 "'the man is wearing glasses and a green-striped hat  
1797 on his head'"  
1798 38 "'a man with glasses is wearing a green-striped hat'"  
1799 39 "'there's a man with a hat striped in shades of green  
1800 and eye wear perched on his nose'"  
1801 40 "'a man in an green-striped hat starring at  
1802 something'"  
1803 41 "'a man wears an green-striped hat and glasses'"  
1804 42 "=====\n"  
1805 43  
1806 44 "=====\n"  
1807 45 "For example2: 5 descriptions of similar pictures are  
1808 :\n"

46 "a black and white dog is running in a grassy garden  
1799 surrounded by a green fence",  
1800 47 "a boston terrier is running on lush green grass in  
1801 front of a brown fence",  
1802 48 "a black and white dog is running through the grass  
1803 in the background",  
1804 49 "a dog runs on the green grass near a white wooden  
1805 fence",  
1806 50 "a brown terrier is running in the grass"  
1807 51  
1808 52 "Based on the given image content, 5 similar image  
1809 descriptions and the reference category, "  
1810 53 "'a black and white dog is standing in a grassy  
1811 garden surrounded by a white fence'",  
1812 54 "'a bicolor canine is standing on lush green grass in  
1813 front of a white fence'",  
1814 55 "'a black and white dog is standing through the grass  
1815 in the background'",  
1816 56 "'a dog stands on the green grass near a white fence'  
1817 ",  
1818 57 "'a brown terrier is standing in the grass'"  
1819 "=====\n"  
1820 58  
1821 59 "The requirements for the generated image description  
1822 are as follows:\n"  
1823 60 "=====\n"  
1824 61 "{context\_str}"  
1825 62 "=====\n"  
1826 63  
1827 64 "The supplementary information for the uploaded image  
1828 is: \n"  
1829 65 "=====\n"  
1830 66 "{extra\_information}"  
1831 67 "=====\n"  
1832 68  
1833 69 "Based on the given image content, 5 similar image  
1834 descriptions and the reference category, "  
1835 70 "the new description generated is:\n"  
1836 71 )  
1837 72 )

### E.3 Prompt for Similar Image Caption Generation With Extra Information.

The main purpose of this template is to generate perturbed text for the original image, thereby further increasing the difficulty of image retrieval text tasks. Specifically, the template is designed to produce text descriptions for images similar to the original image, which resemble the text of the original image but also align with the features of the similar images. Several representative examples will be provided.

#### 1. Prompt for similar image caption generation about color with extra information.

```
1 SIM_COLOR_CAPTION_WITH_EXTRA = (  
2     "I want you to act as an image description generation  
3     expert, and your task is to generate corresponding"  
4     " image descriptions based on images. Specific  
5     requirements are as follows:\n"  
6     "1. I will provide you with 5 descriptions of  
7     pictures similar to this image, and you will use the  
8     "  
9     "descriptions to generate a description corresponding  
10    "to the image I provide you sentence by sentence.\n"  
11    "2. Please generate a description of the image from a  
12    "color perspective.\n"  
13    "3. Each description should be very similar to the  
14    "referenced description and should have "
```

1857 8 "as few modified words as possible.\n"  
1858 9 "4. Each description must be a sentence that is  
1859 10 grammatically correct and logically strict, and is "  
1860 11 "required to be as detailed as possible and cannot be  
1861 12 the same as the description provided.\n"  
1862 13  
1863 14 "Don't give any extraneous information other than a  
1864 15 description.\n"  
1865 16  
1866 17 "=====\n"  
1867 18 "For example1: 5 descriptions of similar pictures are  
1868 19 :\n"  
1869 20 "the man with pierced ears is wearing glasses and an  
1870 21 orange hat on his head"  
1871 22 "a man with glasses is wearing a beer can crocheted  
1872 23 hat"  
1873 24 "a man with pierced ears and glasses is wearing an  
1874 25 orange hat on his head"  
1875 26 "a man in an orange hat starring at something"  
1876 27 "a man wears an orange hat and glasses"  
1877 28  
1878 29 "Based on the given image content, 5 similar image  
1879 30 descriptions and the reference category, "  
1880 31 "the new description generated is:\n"  
1881 32 "'the man is wearing glasses and a green-striped hat  
1882 33 on his head'"  
1883 34 "'a man with glasses is wearing a green-striped hat'"  
1884 35 "'there's a man with a hat striped in shades of green  
1885 36 and eye wear perched on his nose'"  
1886 37 "'a man in an green-striped hat starring at  
1887 38 something'"  
1888 39 "'a man wears an green-striped hat and glasses'"  
1889 40 "=====\n"  
1890 41  
1891 42 "=====\n"  
1892 43 "For example2: 5 descriptions of similar pictures are  
1893 44 :\n"  
1894 45 "a black and white dog is running in a grassy garden  
1895 46 surrounded by a green fence",  
1896 47 "a boston terrier is running on lush green grass in  
1897 48 front of a brown fence",  
1898 49 "a black and white dog is running through the grass  
1899 50 in the background",  
1900 51 "a dog runs on the green grass near a white wooden  
1901 52 fence",  
1902 53 "a brown terrier is running in the grass"  
1903 54  
1904 55 "Based on the given image content, 5 similar image  
1905 56 descriptions and the reference category, "  
1906 57 "'a black and white dog is standing in a grassy  
1907 58 garden surrounded by a white fence'",  
1908 59 "'a bicolor canine is standing on lush green grass in  
1909 60 front of a white fence'",  
1910 61 "'a black and white dog is standing through the grass  
1911 62 in the background'",  
1912 63 "'a dog stands on the green grass near a white fence'  
1913 64 ",  
1914 65 "'a brown terrier is standing in the grass'"  
1915 66 "=====\n"  
1916 67  
1917 68 "5 descriptions of similar pictures are:\n"  
1918 69 "=====\n"  
1919 70 "{context\_str}"  
1920 71 "=====\n"  
1921 72  
1922 73 "The supplementary information for the uploaded image  
1923 74 is: \n"  
1924 75 "=====\n"  
1925 76 "{extra\_information}"  
1926 77

55 "=====\n"  
56 "Please add additional information to each  
57 description generated.\n"  
58  
59 "Based on the given image content and its  
60 supplementary information, as well as the  
description of 5 "  
"similar images,the new description generated is:\n"  
)  
**2. Prompt for similar image caption generation about count with extra information.**  
1 SIM\_COUNT\_CAPTION\_WITH\_EXTRA = (  
2 "I want you to act as an image description generation  
3 expert, and your task is to generate corresponding"  
4 " image descriptions based on images.Specific  
5 requirements are as follows:\n"  
6 "1. I will provide you with 5 descriptions of  
7 pictures similar to this image, and you will use the  
8 "  
9 "descriptions to generate a description corresponding  
10 to the image I provide you sentence by sentence.\n"  
11 "2. Please generate a description of the image from a  
12 count perspective.\n"  
13 "3. Each description should be very similar to the  
14 referenced description and should have "  
15 "as few modified words as possible.\n"  
16 "4. Each description must be a sentence that is  
17 grammatically correct and logically strict, and is "  
18 "required to be as detailed as possible and cannot be  
19 the same as the description provided.\n"  
20  
21 "Don't give any extraneous information other than a  
22 description.\n"  
23  
24 "=====\n"  
25 "For example1: 5 descriptions of similar pictures are  
26 :\n"  
27 "Three giraffes are standing in a field with patches  
28 of trees in the distance."  
29 "A trio of giraffes is spotted across a savanna  
30 landscape."  
31 "There are three giraffes positioned amongst tall  
32 grass in a natural setting."  
33 "A group of three giraffes maintains a distance from  
one another in their habitat."  
"The savanna hosts a count of three giraffes under a hazy sky."  
"Based on the given image content, 5 similar image descriptions and the reference category, "  
"the new description generated is:\n"  
"Four giraffes are standing in a field with patches of trees in the distance."  
"A quartet of giraffes is spotted across a savanna landscape."  
"There are four giraffes positioned amongst tall grass in a natural setting."  
"A group of four giraffes maintains a distance from one another in their habitat."  
"The savanna hosts a count of four giraffes under a hazy sky."  
"=====\n"  
"=====\n"  
"For example2: 5 descriptions of similar pictures are :\n"  
"Four elephants are gathered together in a natural setting.",  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972

1973 34 "quartet of elephants stands side by side on the  
1974 grassy land.",  
1975 35 "There are four elephants visible, with one appearing  
1976 smaller than the others.",  
1977 36 "A group of four elephants is present, with some  
1978 standing and one seemingly resting.",  
1979 37 "Four elephants, potentially a family unit, are  
1980 spotted in a grassy area."  
1981 38  
1982 39 "Based on the given image content, 5 similar image  
1983 descriptions and the reference category, "  
1984 40 "Five elephants are gathered together in a natural  
1985 savanna setting."  
1986 41 "A quintet of elephants stands in a line on the arid  
1987 grassland."  
1988 42 "There are five elephants in view, with the smallest  
1989 amongst them at the front."  
1990 43 "A group of five elephants is present, with four  
1991 standing and the smallest one leading the way."  
1992 44 "Five elephants, potentially a family group, are  
1993 spotted traversing a grassy expanse."  
1994 45 "=====\n"  
1995 46  
1996 47 "5 descriptions of similar pictures are:\n"  
1997 48 "=====\n"  
1998 49 "{context\_str}"  
1999 50 "=====\n"  
2000 51  
2001 52 "The supplementary information for the uploaded image  
2002 is: \n"  
2003 53 "=====\n"  
2004 54 "{extra\_information}"  
2005 55 "=====\n"  
2006 56 "Please add additional information to each  
2007 description generated.\n"  
2008 57  
2009 58 "Based on the given image content, 5 similar image  
2010 descriptions and the reference category, "  
2011 59 "the new description generated is:\n"  
2012 60 )  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030

**3. Prompt for similar image caption generation about object with extra information.**

```
SIM_OBJECT_CAPTION_WITH_EXTRA = (  
    "I want you to act as an image description generation  
    expert, and your task is to generate corresponding"  
    " image descriptions based on images. Specific  
    requirements are as follows:\n"  
    "1. I will provide you with 5 descriptions of  
    pictures similar to this image, and you will use the  
    "  
    "descriptions to generate a description corresponding  
    to the image I provide you sentence by sentence.\n"  
    "2. Please generate a description of the image from a  
    object perspective.\n"  
    "3. Each description should be very similar to the  
    referenced description and should have "  
    "as few modified words as possible.\n"  
    "4. Each description must be a sentence that is  
    grammatically correct and logically strict, and is "  
    "required to be as detailed as possible and cannot be  
    the same as the description provided.\n"  
    "  
    "Don't give any extraneous information other than a  
    description.\n"  
    "  
    "=====\n"  
    "For example1: 5 descriptions of similar pictures are  
    :\n"
```

16 "Four tall glasses stand in a row, with the first two  
showcasing amber and yellow hues.\n",  
17 "A quartet of vertically oriented glasses is arrayed  
against a window, beginning with a brown and a green  
"  
18 "glass.\n",  
19 "There are four decorative glasses on display, each  
featuring a unique color or transparency.\n",  
20 "On the sill, a collection of four glasses captures  
the light, with the initial pair tinted and the rest  
"  
21 "clear.\n",  
22 "A sequence of four glasses with embossed patterns is  
presented, the first colored in bronze, the second  
in "  
23 "lime, and the last two transparent.\n "  
24  
25 "Based on the given image content, 5 similar image  
descriptions and the reference category, "  
26 "the new description generated is:\n"  
27 "Four ceramic vessels stand in a row, with the first  
showcasing a speckled grey and the second featuring  
"  
28 "vibrant red and blue hues.\n",  
29 "A quartet of ceramic pots is arrayed against a  
window, beginning with a textured grey and a  
multicolored "  
30 "striped one.\n",  
31 "There are four decorative pots on display, each  
featuring a unique color scheme and texture.\n",  
32 "On the sill, a collection of four handcrafted  
pottery pieces captures the light, with the initial  
ones "  
33 "boasting a mix of red, blue, and white glazes.\n",  
34 "A sequence of four artisanal pottery items with  
glazed patterns is presented, the first colored in  
mottled "  
35 "grey, the second in bold stripes, and the subsequent  
ones in blue and orange tones.\n "  
36 "=====\n"  
37  
38 "=====\n"  
39 "For example2: 5 descriptions of similar pictures are  
:\n"  
40 "A vibrant assortment of fresh produce, including  
three apples, is artfully arranged on a kitchen  
counter.\n",  
41 "Among the colorful vegetables, three apples add a  
touch of sweetness to the selection of healthy "  
42 "ingredients.\n",  
43 "Three red apples nestle amongst a variety of  
vegetables, including leafy greens and carrots, in a  
well-lit "  
44 "kitchen scene.\n",  
45 "A fresh food display features three apples alongside  
an array of vegetables, ready for a nutritious meal  
"  
46 "preparation.\n",  
47 "The picture displays a delightful assortment of  
produce, including three red apples and two oranges  
that "  
48 "offer a sweet contrast to the various vegetables  
present.\n "  
49  
50 "Based on the given image content, 5 similar image  
descriptions and the reference category, "  
51 "A diverse assortment of fresh produce, including  
three oranges, is neatly arranged on a wooden  
surface.\n",

2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088



2089	52	"Among the array of green vegetables, three oranges	63	"=====\\n"	2147
2090		introduce a vibrant citrus element to the mix of "	64	"{context_str}"	2148
2091	53	"nutritious ingredients.\\n",	65	"=====\\n"	2149
2092	54	"Three oranges are positioned near an assortment of	66		2150
2093		vegetables, including leafy greens and a leek, "	67	"The supplementary information for the uploaded image	2151
2094	55	"on a well-lit wooden counter.\\n",		is: \\n"	2152
2095	56	"A healthful array of produce is displayed, featuring	68	"=====\\n"	2153
2096		three oranges in line with a variety of vegetables,	69	"{extra_information}"	2154
2097		"	70	"=====\\n"	2155
2098	57	"signaling readiness for meal creation.\\n",	71	"Please add additional information to each	2156
2099	58	"The image showcases a pleasing selection of		description generated.\\n"	2157
2100		vegetables, with three oranges and two grapefruits	72		2158
2101	59	adding a "	73	"Based on the given image content, 5 similar image	2159
2102	60	"juicy contrast to the greens and earthy roots.\\n "		descriptions and the reference category, "	2160
2103	61	"=====\\n"	74	"the new description generated is:\\n"	2161
2104	62		75	)	2162
2105		"5 descriptions of similar pictures are:\\n"			2163
2106					2164
2107					2165
2108					2166
2109					2167
2110					2168
2111					2169
2112					2170
2113					2171
2114					2172
2115					2173
2116					2174
2117					2175
2118					2176
2119					2177
2120					2178
2121					2179
2122					2180
2123					2181
2124					2182
2125					2183
2126					2184
2127					2185
2128					2186
2129					2187
2130					2188
2131					2189
2132					2190
2133					2191
2134					2192
2135					2193
2136					2194
2137					2195
2138					2196
2139					2197
2140					2198
2141					2199
2142					2200
2143					2201
2144					2202
2145					2203
2146					2204