

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 θ . 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 RETRIEVETOPK( $q_c, P, k$ )
4:    $E_q \leftarrow \theta.text\_to\_embedding(q_c)$ 
5:   for each  $I \in P$  do
6:      $E_i \leftarrow \theta.image\_to\_embedding(I)$ 
7:      $sim \leftarrow \cos(E_q, E_i)$  // Cosine similarity
8:     add ( $I, sim$ ) to  $U$ 
9:   end for
10:   $U \leftarrow$  sort  $U$  by  $sim$  in descending order
11:  return  $U[1 : k]$  // Top-k images
12: end function
// Step 2: Validate and Refine Textual Description
13:  $U \leftarrow$  RETRIEVETOPK( $q_c, P, k$ )
14: if  $U[1].image = I_{gt}$  then
15:    $q_f \leftarrow q_c$  // Top-1 image is ground truth
16: else
17:   if  $I_{gt} \in U.images$  then
18:     if HUMANVERIFICATION( $q_c, I_{gt}$ ) then
19:        $q_f \leftarrow q_c$  // Description is distinguishable
20:     else
21:        $q_f \leftarrow$  REFINEDESCRIPTION( $q_c, I_{gt}$ )
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  $q_f$ 

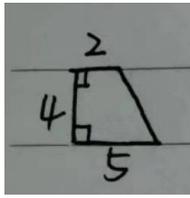
```

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 TARNs_VERIFY = (
```

1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334



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.

1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392

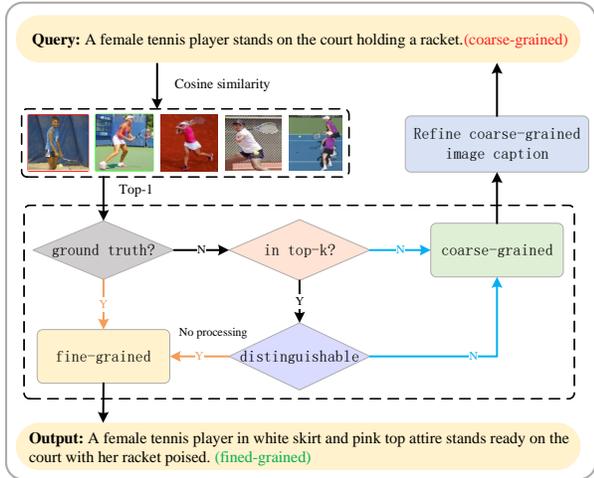


Figure 9: Building process of refining coarse-grained text descriptions

```
13 "The correct or error in the sentence represents the
14 original evaluation result of the sentence, "
15 "the following sentence is the original sentence, ---
16 the following sentence represents the modified
17 sentence "
18 "\n-----\n"
19 "The format for your review is as follows: \n"
20 "-----\n"
21 "<correct or error> <origin sentence> --- <revised
22 sentence>(Optional)\n"
23 "-----\n"
24 "For example:\n"
25 "-----\n"
26 "YES correct a man wearing a hat and a white shirt is
27 cleaning windows"
28 "YES correct a man in a white shirt stands high up on
29 scaffolding"
30 "NO error a man stands on boards on top of a huge
31 ladder --- a man is standing on boards on top of a
32 huge "
33 "ladder "
34 "YES correct a man in a white shirt and hat works on
35 top of scaffolding"
36 "YES correct a guy works on a building"
37 "-----\n"
38 "We have provided the English sentences that need to
39 be reviewed for English grammar as follows:\n"
40 "-----\n"
41 "{context_str}"
42 "\n-----\n"
43 "Based on the above information, Please complete the
44 English grammar verification and revised review.\n"
45 )
```

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
3 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"
1401 10 "The five descriptions of the image are:\n"
1402 11 "-----\n"
1403 12 "the man with pierced ears is wearing glasses and an
1404 orange hat on his head",
1405 "a man with glasses is wearing a beer can crocheted
1406 hat",
1407 "a man with pierced ears and glasses is wearing an
1408 orange hat on his head",
1409 "a man in an orange hat starring at something",
1410 "a man wears an orange hat and glasses"
1411 "-----\n"
1411 19 "Answer: yes\n"
1412 20
1413 21 "Question:Are the descriptions corresponding to the 5
1414 images related to the color?\n"
1415 22 "We provide five descriptions of the image are:\n"
1416 23 "-----\n"
1417 24 "{context_str}"
1418 25 "\n-----\n"
1419 26 "-----\n"
1419 27 "Answer: <yes or no>\n"
1420 28 "-----\n"
1421 29
1421 30 "Please complete the image caption classification
1422 based on the above information."
1423 31 )

```

D.2 Prompt for Image Position Classification

```

1427 1 IMAGE_POSITION_CLASSIFY = (
1428 2 "I hope you will act as an image description '
1429 orientation description' recognition expert, that is
, "
1430 3 "you can accurately determine whether 5 image
1431 descriptions are related to orientation or position,
"
1432 "such as: behind , In front of..., to the
1433 right of... "
1434 5 "below..., above..., east of..., north of..., on top
1435 of..., etc.\n"
1436 6 'As long as one of the 5 descriptions involves
1437 position or orientation, then you answer yes. '
1438 7 'If none of the five descriptions involve position,
1439 please answer no.\n'
1440 8
1440 9 "For example:\n"
1441 10 "-----\n"
1442 11 "The five descriptions of the image are:\n"
1443 12 "-----\n"
1444 13 "a woman with a black shirt and tan apron is standing
1445 behind a counter in a restaurant",
1446 "a female barista dressed in black is making coffee
1447 behind a pink counter",
1448 "a girl in a black shirt is smiling as she works
1449 behind a bar",
1450 "a happy woman in a black shirt and standing in front
of a counter",

```

```

17 "a young woman in a black shirt and stands in front
1451 of a counter full of coffee cups"
1452 "-----\n"
1453 "Answer: yes\n"
1454
1455
1456 "Question:Are the descriptions corresponding to the 5
1457 images related to the position or orientation?\n"
1458 "We provide five descriptions of the image are:\n"
1459 "-----\n"
1460 "{context_str}"
1461 "\n-----\n"
1462 "-----\n"
1463 "Answer: <yes or no>\n"
1464 "-----\n"
1465
1466 "Please complete the image caption classification
1467 based on the above information."
1468
1469 )

```

D.3 Prompt for Image Count Classification

```

1469 1 IMAGE_COUNT_CLASSIFY = (
1470 2 "I hope you will act as an image description 'count'
1471 recognition expert, that is, "
1472 3 "you can accurately determine whether 5 image
1473 descriptions are related to count, "
1474 4 "such as: one, two, three, many, few, several, some,
1475 a couple of, a few, a dozen, "
1476 5 "a bunch, a lot of, a bunch of, a couple, etc.\n"
1477 6 "As long as one of the 5 descriptions involves 'count'
1478 ', then you answer yes. "
1479 7 "If none of the five descriptions involve 'count',
1480 please answer no.\n"
1481 8
1482 9 "For example:\n"
1483 10 "-----\n"
1484 11 "The five descriptions of the image are:\n"
1485 12 "-----\n"
1486 13 "six people ride mountain bikes through a jungle
1487 environment",
1488 "men surrounded by nature are riding mountain bikes
1489 on a trail",
1490 "there are six men mountain biking in a forest
1491 terrain",
1492 "six people riding bikes on a trail in the forest",
1493 "a group of people is bike riding in the woods"
1494 "-----\n"
1495 "Answer: yes\n"
1496
1497
1498 "Question:Are the descriptions corresponding to the 5
1499 images related to the 'count'?\n"
1500 "We provide five descriptions of the image are:\n"
1501 "-----\n"
1502 "{context_str}"
1503 "\n-----\n"
1504 "-----\n"
1505 "Answer: <yes or no>\n"
1506 "-----\n"
1507
1508 "Please complete the image caption classification
1509 based on the above information."
1510
1511 )

```

D.4 Prompt for Image Action Classification

```

1512 1 IMAGE_ACTION_CLASSIFY = (
1513 2 "I hope you will act as an image description 'action'
1514 recognition expert, that is, "
1515
1516 )

```

```

1509 3      "you can accurately determine whether 5 image
1510 4      descriptions are related to action, "
1511 5      "such as: cycling, skate, play basketball, play
1512 6      table tennis, play billiards, play tennis, "
1513 7      "climb mountains, surf, taekwondo, wrestling, etc.\n"
1514 8      "As long as one of the 5 descriptions involves '
1515 9      action', then you answer yes. "
1516 10     "If none of the five descriptions involve 'action',
1517 11     please answer no.\n"
1518 12     "For example:\n"
1519 13     "-----\n"
1520 14     "The five descriptions of the image are:\n"
1521 15     "-----\n"
1522 16     "six people ride mountain bikes through a jungle
1523 17     environment",
1524 18     "men surrounded by nature are riding mountain bikes
1525 19     on a trail",
1526 20     "there are six men mountain biking in a forest
1527 21     terrain",
1528 22     "six people riding bikes on a trail in the forest",
1529 23     "a group of people is bike riding in the woods"
1530 24     "-----\n"
1531 25     "Answer: yes\n"
1532 26     "-----\n"
1533 27     "Question:Are the descriptions corresponding to the 5
1534 28     images related to the 'action'? \n"
1535 29     "We provide five descriptions of the image are:\n"
1536 30     "-----\n"
1537 31     "{context_str}"
1538 32     "\n-----\n"
1539 33     "-----\n"
1540 34     "Answer: <yes or no>\n"
1541 35     "-----\n"
1542 36     "Please complete the image caption classification
1543 37     based on the above information."
1544 38 )

```

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 "For example:\n"
1560 17 "-----\n"
1561 18 "The five descriptions of the image are:\n"
1562 19 "-----\n"
1563 20 "woman in a white dress standing with a tennis racket
1564 21 and two people in green behind her",
1565 22 "young pretty blond women holding a tennis racket
1566 23 dressed as if to begin a tennis match",
1567 24 "a young woman in a fashionable tennis racket",
1568 25 "slim blond woman in white dress holds tennis racket"
1569 26 ,
1570 27 "a young lady in white holding a tennis racket behind
1571 28 her"
1572 29 "-----\n"
1573 30 "Answer: yes\n"
1574 31 "-----\n"
1575 32 "Question:Are the descriptions corresponding to the 5
1576 33 images related to the 'object'? \n"
1577 34 "We provide five descriptions of the image are:\n"
1578 35 "-----\n"
1579 36 "{context_str}"
1580 37 "\n-----\n"
1581 38 "-----\n"
1582 39 "Answer: <yes or no>\n"
1583 40 "-----\n"
1584 41 "Please complete the image caption classification
1585 42 based on the above information."
1586 43 )

```

```

1567 18     "a young lady in white holding a tennis racket behind
1568 19     her"
1569 20     "-----\n"
1570 21     "Answer: yes\n"
1571 22     "-----\n"
1572 23     "Question:Are the descriptions corresponding to the 5
1573 24     images related to the 'figure'? \n"
1574 25     "We provide five descriptions of the image are:\n"
1575 26     "-----\n"
1576 27     "{context_str}"
1577 28     "\n-----\n"
1578 29     "-----\n"
1579 30     "Answer: <yes or no>\n"
1580 31     "-----\n"
1581 32     "Please complete the image caption classification
1582 33     based on the above information."
1583 34 )

```

D.6 Prompt for Image Object Classification

```

1584 1  IMAGE_OBJECT_CLASSIFY = (
1585 2  "I hope you will act as an image description 'object'
1586 3  recognition expert, that is, "
1587 4  "you can accurately determine whether 5 image
1588 5  descriptions are related to object,"
1589 6  "such as: teacher grading papers, tennis player
1590 7  playing tennis, artist painting a masterpiece,"
1591 8  "a man holding a book, a woman in front of the pizza,
1592 9  the motorcycle that the man is riding, etc.\n"
1593 10 "As long as one of the 5 descriptions involves '
1594 11 object', then you answer yes. "
1595 12 "If none of the five descriptions involve 'object',
1596 13 please answer no.\n"
1597 14 "For example:\n"
1598 15 "-----\n"
1599 16 "The five descriptions of the image are:\n"
1600 17 "-----\n"
1601 18 "woman in a white dress standing with a tennis racket
1602 19 and two people in green behind her",
1603 20 "young pretty blond women holding a tennis racket
1604 21 dressed as if to begin a tennis match",
1605 22 "a young woman in a fashionable tennis racket",
1606 23 "slim blond woman in white dress holds tennis racket"
1607 24 ,
1608 25 "a young lady in white holding a tennis racket behind
1609 26 her"
1610 27 "-----\n"
1611 28 "Answer: yes\n"
1612 29 "-----\n"
1613 30 "Question:Are the descriptions corresponding to the 5
1614 31 images related to the 'object'? \n"
1615 32 "We provide five descriptions of the image are:\n"
1616 33 "-----\n"
1617 34 "{context_str}"
1618 35 "\n-----\n"
1619 36 "-----\n"
1620 37 "Answer: <yes or no>\n"
1621 38 "-----\n"
1622 39 "Please complete the image caption classification
1623 40 based on the above information."
1624 41 )

```

D.7 Prompt for Image Scene Classification

```

1624 1  IMAGE_SCENE_CLASSIFY = (

```

```

1625 2 "I hope you will act as an image description 'scene'
1626 recognition expert, that is, "
1627 3 "you can accurately determine whether 5 image
1628 descriptions are related to scene,"
1629 4 "such as: hospital, swimming pool, construction site,
1630 roadside, seaside, school, supermarket, "
1631 5 "restaurant, theater, library, park, office, cafe,
1632 gym, train station, airport, shop/store, "
1633 6 "swimming pool, museum, residential area, cinema,
1634 post office, pet store, meadow, amusement park, etc
1635 .\n"
1636 7 "As long as one of the 5 descriptions involves 'scene'
1637 ', then you answer yes. "
1638 8 "If none of the five descriptions involve 'scene',
1639 please answer no.\n"
1640 9
1641 10 "For example:\n"
1642 11 "-----\n"
1643 12 "The five descriptions of the image are:\n"
1644 13 "-----\n"
1645 14 "three girls wearing goggles are jumping into a
1646 swimming pool together",
1647 15 "children jumping in a blue pool surrounded by blue
1648 pool chairs and huts",
1649 16 "a woman and three children are jumping into a
1650 swimming pool",
1651 17 "children jump off the edge into a pool",
1652 18 "three girls jump into a pool"
1653 19 "-----\n"
1654 20 "Answer: yes\n"
1655 21
1656 22
1657 23 "Question:Are the descriptions corresponding to the 5
1658 images related to the 'scene'?\n"
1659 24 "We provide five descriptions of the image are:\n"
1660 25 "-----\n"
1661 26 "{context_str}"
1662 27 "\n-----\n"
1663 28 "-----\n"
1664 29 "Answer: <yes or no>\n"
1665 30 "-----\n"
1666 31
1667 32 "Please complete the image caption classification
1668 based on the above information."
1669 33 )

```

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.

```

1683 1 DEFAULT_IMAGE_CAPTION_GENERATION = (
1684 2 "I want you to play an image description expert.Your
1685 task is to generate 5 descriptions as detailed as "
1686 3 "possible based on the image. "
1687 4 "The requirements are as follows:\n"
1688 5 "1. Each description is one sentence.\n"
1689 6 "2. The focus of each description is color, position
1690 and action.\n"
1691 7 "3. The grammar is required to be correct and logical
1692 .\n"
1693 8 "4. Each description generated should be as detailed
1694 and matter-of-fact as possible.\n"
1695 9 "5. Dont generate meaningless, empty descriptions
1696 .\n"
1697 10
1698 11 "Additional descriptive information about the
1699 uploaded image is:\n",
1700 12 "======"
1701 13 "{note_text}"
1702 14 "======"
1703 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.

```

1705 1 IMAGE_CAPTION_GENERATION_WITH_EXTRA = (
1706 2 "I want you to act as an image description expert.
1707 Your task is to generate 5 descriptions
1708 corresponding"
1709 3 " to the given picture based on the picture and 5
1710 descriptions of similar pictures. The description
1711 should be"
1712 4 " as close as possible to the 5 descriptions I gave
1713 you. It should be as similar as possible and the
1714 number of"
1715 5 " modified words should be as few as possible, but
1716 the generated description should be different from
1717 the 5"
1718 6 " descriptions provided previously.\n"
1719 7
1720 8 "You can refer to the following angles: color,
1721 position, existence, action, count, figure, object,
1722 scene, etc."

```

```

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 "as short as possible but as detailed as possible.
1752 Each description should be a grammatically correct
1753 and "
1754 "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 "the description one by one based on the image
1763 information and the reference categories provided to
1764 you. "
1765 "For example: first, replace the color of the
1766 description according to the picture content; if the
1767 color "
1768 "is not suitable , you can try to replace the
1769 quantity; if the quantity is also inappropriate, you
1770 can try "
1771 "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 "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
47 surrounded by a green fence",
48 "a boston terrier is running on lush green grass in
49 front of a brown fence",
50 "a black and white dog is running through the grass
51 in the background",
52 "a dog runs on the green grass near a white wooden
53 fence",
54 "a brown terrier is running in the grass"
55
56 "Based on the given image content, 5 similar image
57 descriptions and the reference category, "
58 "'a black and white dog is standing in a grassy
59 garden surrounded by a white fence'",
60 "'a bicolor canine is standing on lush green grass in
61 front of a white fence'",
62 "'a black and white dog is standing through the grass
63 in the background'",
64 "'a dog stands on the green grass near a white fence'
65 ",
66 "'a brown terrier is standing in the grass'"
67 "=====\n"
68
69 "The requirements for the generated image description
70 are as follows:\n"
71 "=====\n"
72 "{context_str}"
73 "=====\n"
74
75 "The supplementary information for the uploaded image
76 is: \n"
77 "=====\n"
78 "{extra_information}"
79 "=====\n"
80
81 "Based on the given image content, 5 similar image
82 descriptions and the reference category, "
83 "the new description generated is:\n"
84 )

```

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 "

```

1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856

```

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}"

```

```

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
34 one another in their habitat."
35 "The savanna hosts a count of three giraffes under a
36 hazy sky."
37
38 "Based on the given image content, 5 similar image
39 descriptions and the reference category, "
40 "the new description generated is:\n"
41 "Four giraffes are standing in a field with patches
42 of trees in the distance."
43 "A quartet of giraffes is spotted across a savanna
44 landscape."
45 "There are four giraffes positioned amongst tall
46 grass in a natural setting."
47 "A group of four giraffes maintains a distance from
48 one another in their habitat."
49 "The savanna hosts a count of four giraffes under a
50 hazy sky."
51 "=====\n"
52
53 "=====\n"
54 "For example2: 5 descriptions of similar pictures are
55 :\n"
56 "Four elephants are gathered together in a natural
57 setting.",

```

1973	34	"quartet of elephants stands side by side on the grassy land.",	16	"Four tall glasses stand in a row, with the first two showcasing amber and yellow hues.\n",	2031
1974					2032
1975	35	"There are four elephants visible, with one appearing smaller than the others.",	17	"A quartet of vertically oriented glasses is arrayed against a window, beginning with a brown and a green "	2033
1976	36	"A group of four elephants is present, with some standing and one seemingly resting.",			2034
1977			18	"glass.\n",	2035
1978	37	"Four elephants, potentially a family unit, are spotted in a grassy area."	19	"There are four decorative glasses on display, each featuring a unique color or transparency.\n",	2036
1979					2037
1980	38		20	"On the sill, a collection of four glasses captures the light, with the initial pair tinted and the rest "	2038
1981	39	"Based on the given image content, 5 similar image descriptions and the reference category, "			2039
1982	40	"Five elephants are gathered together in a natural savanna setting."	21	"clear.\n",	2040
1983			22	"A sequence of four glasses with embossed patterns is presented, the first colored in bronze, the second in "	2041
1984	41	"A quintet of elephants stands in a line on the arid grassland."			2042
1985	42	"There are five elephants in view, with the smallest amongst them at the front."	23	"lime, and the last two transparent.\n "	2043
1986			24		2044
1987	43	"A group of five elephants is present, with four standing and the smallest one leading the way."	25	"Based on the given image content, 5 similar image descriptions and the reference category, "	2045
1988					2046
1989	44	"Five elephants, potentially a family group, are spotted traversing a grassy expanse."	26	"the new description generated is:\n"	2047
1990	45	"=====\n"	27	"Four ceramic vessels stand in a row, with the first showcasing a speckled grey and the second featuring "	2048
1991	46				2049
1992	47	"5 descriptions of similar pictures are:\n"	28	"vibrant red and blue hues.\n",	2050
1993	48	"=====\n"	29	"A quartet of ceramic pots is arrayed against a window, beginning with a textured grey and a multicolored "	2051
1994	49	"{context_str}"			2052
1995	50	"=====\n"	30	"striped one.\n",	2053
1996	51		31	"There are four decorative pots on display, each featuring a unique color scheme and texture.\n",	2054
1997	52	"The supplementary information for the uploaded image is: \n"			2055
1998	53	"=====\n"	32	"On the sill, a collection of four handcrafted pottery pieces captures the light, with the initial ones "	2056
1999	54	"{extra_information}"			2057
2000	55	"=====\n"	33	"boasting a mix of red, blue, and white glazes.\n",	2058
2001	56	"Please add additional information to each description generated.\n"	34	"A sequence of four artisanal pottery items with glazed patterns is presented, the first colored in mottled "	2059
2002	57				2060
2003	58	"Based on the given image content, 5 similar image descriptions and the reference category, "	35	"grey, the second in bold stripes, and the subsequent ones in blue and orange tones.\n "	2061
2004	59	"the new description generated is:\n"			2062
2005	60)	36	"=====\n"	2063
2006			37		2064
2007		3. Prompt for similar image caption generation about object with extra information.	38	"=====\n"	2065
2008			39	"For example2: 5 descriptions of similar pictures are :\n"	2066
2009	1	SIM_OBJECT_CAPTION_WITH_EXTRA = (40	"A vibrant assortment of fresh produce, including three apples, is artfully arranged on a kitchen counter.\n",	2067
2010	2	"I want you to act as an image description generation expert, and your task is to generate corresponding "			2068
2011	3	" image descriptions based on images. Specific requirements are as follows:\n"	41	"Among the colorful vegetables, three apples add a touch of sweetness to the selection of healthy "	2069
2012					2070
2013	4	"1. I will provide you with 5 descriptions of pictures similar to this image, and you will use the "	42	"ingredients.\n",	2071
2014			43	"Three red apples nestle amongst a variety of vegetables, including leafy greens and carrots, in a well-lit "	2072
2015	5	"descriptions to generate a description corresponding to the image I provide you sentence by sentence.\n"			2073
2016			44	"kitchen scene.\n",	2074
2017	6	"2. Please generate a description of the image from a object perspective.\n"	45	"A fresh food display features three apples alongside an array of vegetables, ready for a nutritious meal "	2075
2018					2076
2019	7	"3. Each description should be very similar to the referenced description and should have "	46	"preparation.\n",	2077
2020	8	"as few modified words as possible.\n"	47	"The picture displays a delightful assortment of produce, including three red apples and two oranges that "	2078
2021	9	"4. Each description must be a sentence that is grammatically correct and logically strict, and is "			2079
2022	10	"required to be as detailed as possible and cannot be the same as the description provided.\n"	48	"offer a sweet contrast to the various vegetables present.\n "	2080
2023					2081
2024	11		49		2082
2025	12	"Don't give any extraneous information other than a description.\n"	50	"Based on the given image content, 5 similar image descriptions and the reference category, "	2083
2026					2084
2027	13	"=====\n"	51	"A diverse assortment of fresh produce, including three oranges, is neatly arranged on a wooden surface.\n",	2085
2028	14	"For example1: 5 descriptions of similar pictures are :\n"			2086
2029	15				2087
2030					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",	68	is: \\n"	2152
2095	56	"A healthful array of produce is displayed, featuring	69	"=====\\n"	2153
2096		three oranges in line with a variety of vegetables,	70	"{extra_information}"	2154
2097	57	"	71	"=====\\n"	2155
2098		"signaling readiness for meal creation.\\n",	72	"Please add additional information to each	2156
2099	58	"The image showcases a pleasing selection of	73	description generated.\\n"	2157
2100		vegetables, with three oranges and two grapefruits	74	"Based on the given image content, 5 similar image	2158
2101	59	adding a "	75	descriptions and the reference category, "	2159
2102	60	"juicy contrast to the greens and earthy roots.\\n "		"the new description generated is:\\n"	2160
2103	61	"=====\\n")	2161
2104	62	"5 descriptions of similar pictures are:\\n"			2162
2105					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