A. More Implementation Details

The pre-trained OFA [58] is trained on about 20M image-text pairs. We further fine-tune OFA using examples synthesized from VQAv2 [14] on COCO 2014 training set, which contains 443757 examples. Notice that we haven't used COCO 2014 validation set to avoid data leakage.

For model training, the most important details are given in Section 3. We finetune PromptCap on the synthesized dataset for 2 epochs, which takes about 10 hours on 4 NVIDIA A40 GPUs. For all the experiments that use random in-context examples for GPT-3, we report the average result of 3 runs.

B. Common Q & A

Compared with OFA, does the performance gain of PromptCap come from more training data? No. As discussed in §5.3, OFA is a large-scale vision-language model pre-trained on 20M image-text pairs and 20 vision-language tasks, including many VQA tasks. PromptCap is fine-tuned on VQAv2 [14], which is included in OFA training data. The performance gain comes from the idea of **controlling image captions with natural language instructions.** Compared with OFA, we synthesize VQAv2 into prompt-guided captioning tasks via GPT-3 and use it to further fine-tune OFA. This way, OFA gains the ability of prompt-guided captioning with **the same amount of annotated data**, leading to significant performance gains.

Can PromptCap do generic captioning? Yes. We find that prompting PromptCap with the question “what does the image describe?" leads to high-quality generic captions. The CIDEr is 150.1 on the COCO validation set in the setting in §5.6.

C. Prompt Examples

In this section, we show examples of the prompts we used for prompting GPT-3.

C.1. Prompt for example synthesis

Below is the full version of the prompt for training example synthesis with GPT-3. We formulate the task into a summarization tasks, and include the human-written examples in the prompt. The question-answer pair synthesized is at the end of the prompt, and GPT-3 generates the prompt-guided caption example.

```
Summarize the context to help answer the
question

Original contexts: A very clean and well
decorated empty bathroom. A blue and white
bathroom with butterfly themed wall tiles.
A bathroom with a border of butterflies and
 blue paint on the walls above it.
Question: Is the sink full of water?
Answer: no
Summary: A bathroom with an empty sink.

Original contexts: Several metal balls sit
in the sand near a group of people.. People
 standing around many silver round balls on
 the ground.. Silver balls are lined up in
the sand as people mill about in the
background.. Silver balls on sand with
people walking around. . silver balls
laying on the ground around a smaller red
ball.
Question: What color are the round objects?
Answer: silver
Summary: People standing around many silver
 round balls on the ground.

Original contexts: A kitchen cart stands in
 the middle of a kitchen with wooden
cupboards.. A bright kitchen with hardwood
floors and wooden cupboards.. a kitchen
that is fully furnished with a cart in the
center of it.. The kitchen is brightly lit
from the window.. Brightly colored kitchen
with wooden floors, large images on wall,
and small island.
Question: What is the source of light in
this picture?
Answer: sun
Summary: A bright kitchen lit by sunlight.

Original contexts: An empty kitchen with
white and black appliances.. A refrigerator
 and stove are in a small kitchen area. .
Small kitchen in a personal home with dual
sinks.. A small kitchen with sink, stove
and refrigerator.. A small kitchen with
several appliances and cookware.
Question: How many cabinets in this room?
Answer: 4
Summary: A small kitchen with 4 cabinets.

Original contexts: Green tiled backsplash
highlighted by low overhead lighting.. A
kitchen counter is illuminated by a hood
light. A kitchen sink next to an empty
counter with a tiled wall.. A back splash
is added to the wall in the kitchen.. A
picture of a sink top with dim lighting.
Question: What material is the backsplash
made of?
Answer: tile
Summary: Green tiled backsplash highlighted
 by low overhead lighting.
```

Original contexts: A graffiti-ed stop sign across the street from a red car . A vandalized stop sign and a red beetle on the road. A red stop sign with a Bush bumper sticker under the word stop.. A stop sign that has been vandalized is pictured in front of a parked car.. A street sign modified to read stop bush.
Question: What season is it in this photo?
Answer: summer
Summary: A stop sign and a car on a street in summer.

Original contexts: Lady carrying a purse walking along side a man.. A city sidewalk is lined with lamp posts. A man and a woman stand on the sidewalk lined with street lights.. A city sidewalk with storefronts on the right.. Two people leaving a building to walk down the street.
Question: Which item in this picture helps people see after dark?
Answer: streetlight
Summary: A city sidewalk lit by streetlight.

Original contexts: A sink and a toilet inside a small bathroom.. White pedestal sink and toilet located in a poorly lit bathroom.. Clean indoor bathroom with tiled floor and good lighting.. a bathroom with toilet and sink and blue wall. a blue bathroom with a sink and toilet.
Question: How many rolls of toilet paper are on the shelves above the toilet?
Answer: 0
Summary: A bathroom with a toilet and a sink. There is no toile paper on the shelves.

Original contexts: A couple enjoying beverages and a snack on a sunny day. Showing a doughnut while holding drinks near a car.. A man and woman sharing apple cider and a doughnut. Two people are standing in front of an open car trunk holding drinks and a doughnut. . A man and a woman eating donuts and having drinks.. A man holding beer and a woman holding a pastry and beer.
Question: How do we know this guy is not likely to have packed a razor?
Answer: has beard
Summary: A man with beard and a woman are eating donuts and having drinks.

Original contexts: Woman riding a bicycle down an empty street.. A woman in green is riding a bike.. a woman wearing a bright green sweater riding a bicycle. A woman on a bicycle is going down the small town street.. A woman bikes down a one way street.
Question: What kind of fruit is the helmet supposed to be?
Answer: watermelon
Summary: A woman with a watermelon style helmet riding a bicycle.

Original contexts: A panoramic view of a kitchen and all of its appliances. A panoramic photo of a kitchen and dining room A wide angle view of the kitchen work area multiple photos of a brown and white kitchen. A kitchen that has a checkered patterned floor and white cabinets.
Question: Is the counter curved?
Answer: no
Summary: A photo of a kitchen with a counter that is not curved.

Original contexts: A woman is walking a dog in the city.. A woman and her dog walking down a sidewalk next to a fence with some flowers. . A woman walking her dog on the sidewalk.. A woman walks her dog along a city street.. A woman walks her dog on a city sidewalk.
Question: What color vehicle is closest to the mailbox?
Answer: silver
Summary: A silver vehicle next to a mailbox on the sidewalk.

Original contexts: some pancakes cover with bananas, nuts, and some whipped cream . Two pancakes on top of a white plate covered in whipped cream, nuts and a banana .. Pancakes with bananas, nuts and cream, covered in syrup. . Pancakes topped with bananas, whipped cream and walnuts.. Pancakes topped with bananas, nuts, and ice cream.
Question: What restaurant was this dish cooked at?
Answer: ihop
Summary: Pancakes with banans, nuts, and cream, cooked at ihop.

Original contexts: The two people are walking down the beach.. Two people carrying surf boards on a beach.. Two teenagers at a white sanded beach with surfboards.. A couple at the beach walking with their surf boards.. A guy and a girl are walking on the beach holding surfboards.

Question: What is on the man’s head?
Answer: hat
Summary: A man and a woman walking on the beach with surfboards. The man is wearing a hat.

Original contexts: A sink and a toilet inside a small bathroom.. White pedestal sink and toilet located in a poorly lit bathroom.. Clean indoor bathroom with tiled floor and good lighting.. a bathroom with toilet and sink and blue wall. a blue bathroom with a sink and toilet.
Question: Is there natural light in this photo?
Answer: no
Summary: A photo of a small bathroom in artificial light.

Original contexts: Fog is in the air at an intersection with several traffic lights.. An intersection during a cold and foggy night.. Empty fog covered streets in the night amongst traffic lights.. City street at night with several stop lights.. It is a foggy night by a traffic light.
Question: Which direction is okay to go?
Answer: straight
Summary: A traffic light in a foggy night, showing it is okay to go straight.

Original contexts: A graffiti-ed stop sign across the street from a red car . A vandalized stop sign and a red beetle on the road. A red stop sign with a Bush bumper sticker under the word stop.. A stop sign that has been vandalized is pictured in front of a parked car.. A street sign modified to read stop bush.
Question: What color is the car driving north?
Answer: red
Summary: A stop sign and a red car driving north.

Original contexts: A man in a wheelchair and another sitting on a bench that is overlooking the water.. Two people sitting on dock looking at the ocean.. Two older people sitting down in front of a beach.. An old couple at the beach during the day.. A person on a bench, and one on a wheelchair sitting by a seawall looking out toward the ocean.
Question: What is the person on the left sitting on?
Answer: bench
Summary: A person sit on a bench on the left, and another sitting in a wheelchair on the right, all looking at the ocean.

Original contexts: A parked motor scooter sitting next to a bicycle.. A picture of a motorbike and two pedal bicycles.. A motor scooter that has an advertisment on the back next to a bicycle.. A grey moped parked by building next to a bicycle.. a motor bike parked next to a bike by a building.
Question: Which model of bike is shown in this picture?
Answer: vespa
Summary: A vespa bike parking next to a bicycle.

Original contexts: People standing around a park bench next to a bicycle.. A group of women are describing a new setup for a building plan. a group of people in a field of grass near a building. Several people standing in an area with picnic tables looking at a board.. A woman giving a presentation in a park.
Question: What is the woman in the blue jacket standing on?
Answer: bench
Summary: A woman in blue jacket standing on a bench, with a group of people around her.

Original contexts: A orange tabby cat laying down on a black car. An orange cat laying on the hood on a car.. A cat sits on top of a black car.. A cat that is sitting on top of a black car.. A yellow cat sleeping on the hood of a black car parked in the garage.
Question: What brand of car is this?
Answer: subaru
Summary: An orange cat laying on top of a black subaru.

Original contexts: A bicycle parked in front of a building next to a pile of garbage.. Black and white photograph of a homeless person under their many belongings. Two people huddle on a bench under their belongings.. A homeless person is bundled within a pile of belongings.. an image of two homeless people laying under debris on a bench
Question: How is the bike affixed to the pole?
Answer: chain
Summary:

```
-----Prompt Ends Here-----
LM completion: A bicycle chained to a pole,
 with a pile of garbage next to it.
```

C.2. GPT-3 In-Context Learning for VQA

Here we show an example of solving VQA with GPT-3 in-context learning on OK-VQA [38]. We use the same prompt template as PICa [67]. This example contains 8 closest examples retrieved from the training set.

```
Please answer the question according to the
 above context.

===
Context: a bowl of broccoli and lemon
slices on a table
===
Q: How do you make that?
A: steam

===
Context: an orange tree with oranges behind
 a fence
===
Q: What fruit is that?
A: orange

===
Context: a bowl of oranges and limes on a
table
===
Q: What types of fruit are these?
A: orange and lime

===
Context: two oranges and a green leaf on a
white table
===
Q: What fruits are those?
A: orange

===
Context: a basket of oranges sitting on a
wooden table
===
Q: What family of fruits is shown?
A: citrus

===
Context: a green apple sitting on top of a
bunch of bananas
===
Q: What type of fruit is this?
A: apple

===
Context: a bowl filled with oranges sitting
 on top of a table
===
Q: Where can this fruit be found?
A: tree

===
Context: an orange cut in half on a white
plate
===
Q: What fruit is this?
A: orange

===
Context: a glass bowl filled with fruit on
top of a table
===
Q: What is the fruit bowl made of?
A:
-----Prompt Ends Here-----
LM completion: glass
```