Use Case: Synthetize Data for Image-Text Datasets.

This notebook demonstrates how to use some Data-Juicer OPs to synthetize new dataset from a given seed dataset.

Synthetize new data is like replacing the old contents of samples with newly synthetic ones, so we use Mappers to acheive this goal.

Specifically, we will take `image_captioning_mapper` and `image_diffusion_mapper` as example OPs. The former one, `image_captioning_mapper`, generate new captions for images in each sample, and the latter one, `image_diffusion_mapper`, generate new images or edit contents of existing images according to their captions. These two OPs are empowered by some excellent models, such as BLIP-2 and Stable Diffusion. Users are allowed to replace with other models. More details about these two OPs can be found in the config_all.yaml file and their corresponding code implementations.

Now, let's begin the synthesis process to get a new dataset.

Dataset Preparation

Here we only consider a example dataset of two image-text pair samples. We write it to a `jsonl` file first.

The intermediate format of multimodal datasets in Data-Juicer is defined here.

In [1]:
```
import jsonlines as jl

ds = [
    {
        'text': '<__dj__image> a picture of prince and princess kate\'s m
        'images': ['imgs/img1.png'],
    },
    {
        'text': 'the setting sun in africa on a cloudy day stock photo ©
        'images': ['imgs/img2.png'],
    }
]

with jl.open('ds.jsonl', 'w') as writer:
    writer.write_all(ds)
```

And we download these two example images to `./imgs`.

In [2]:
```
!mkdir -p imgs && wget http://dail-wlcb.oss-cn-wulanchabu.aliyuncs.com/da
```

```
--2024-08-12 12:21:05--  http://dail-wlcb.oss-cn-wulanchabu.aliyuncs.com/data_juicer/tutorial_data/img1.png
Resolving dail-wlcb.oss-cn-wulanchabu.aliyuncs.com (dail-wlcb.oss-cn-wulanchabu.aliyuncs.com)... 39.101.35.6
Connecting to dail-wlcb.oss-cn-wulanchabu.aliyuncs.com (dail-wlcb.oss-cn-wulanchabu.aliyuncs.com)|39.101.35.6|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 232261 (227K) [image/png]
Saving to: ‘./imgs/img1.png’

./imgs/img1.png     100%[===================>] 226.82K  --.-KB/s    in 0.1s

2024-08-12 12:21:05 (2.10 MB/s) - ‘./imgs/img1.png’ saved [232261/232261]

--2024-08-12 12:21:05--  http://dail-wlcb.oss-cn-wulanchabu.aliyuncs.com/data_juicer/tutorial_data/img2.png
Resolving dail-wlcb.oss-cn-wulanchabu.aliyuncs.com (dail-wlcb.oss-cn-wulanchabu.aliyuncs.com)... 39.101.35.6
Connecting to dail-wlcb.oss-cn-wulanchabu.aliyuncs.com (dail-wlcb.oss-cn-wulanchabu.aliyuncs.com)|39.101.35.6|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 162649 (159K) [image/png]
Saving to: ‘./imgs/img2.png’

./imgs/img2.png     100%[===================>] 158.84K  --.-KB/s    in 0.1s

2024-08-12 12:21:05 (1.55 MB/s) - ‘./imgs/img2.png’ saved [162649/162649]
```

Visualization of Images

We can also prepare a function to visualize the sample in the dataset.

In [4]:

```python
from PIL import Image
import numpy as np
import matplotlib.pyplot as plt

%matplotlib inline
def vis(s):
    print(s['text'])
    img = Image.open(s['images'][0])
    plt.imshow(np.asarray(img))
    plt.show()

for s in ds:
    vis(s)
```

```
<__dj__image> a picture of prince and princess kate's mugs in a frame
```


```
the setting sun in africa on a cloudy day stock photo © monkeypox <__dj__i
mage>
```

Image Recaptioning

We can recaption an image with OP `image_captioning_mapper`. First we need to create a data recipe for this process.

In [3]:
```
recipe = '''
dataset_path: ds.jsonl
export_path: outputs/image_captioning_output/res.jsonl

process:
  - image_captioning_mapper:
      hf_img2seq: 'Salesforce/blip2-opt-2.7b'  # You can replace this pat
      keep_original_sample: false  # we only need the recaptioned caption
'''

with open('image_captioning.yaml', 'w') as fout:
    fout.write(recipe)
```

Then we can run the process program of Data-Juicer to process the dataset.

In [4]:
```
!dj-process --config image_captioning.yaml
```

2024-07-26 04:20:27.194 | INFO | data_juicer:setup_mp:67 - Setting multiprocess start method to 'forkserver'.
2024-07-26 04:20:34 | INFO | data_juicer.config.config:646 - Back up the input config file [/mnt/workspace/lielin.hyl/dj_synth_test/tutorials/image_captioning.yaml] into the work_dir [/mnt/workspace/lielin.hyl/dj_synth_test/tutorials/outputs/image_captioning_output]
2024-07-26 04:20:34 | INFO | data_juicer.config.config:668 - Configuration table:

key	values
config	[Path_fr(image_captioning.yaml, cwd=/mnt/workspace/lielin.hyl/dj_synth_test/tutorials)]
hpo_config	None
path_k_sigma_recipe	None
path_model_feedback_recipe	None
model_infer_config	None
model_train_config	None
data_eval_config	None
model_eval_config	None
data_probe_algo	'uniform'
data_probe_ratio	1.0
project_name	'hello_world'
executor_type	'default'

dataset_path	'/mnt/workspace/lielin.hyl/dj_synth_test/tutorials/ds.jsonl'
export_path	'/mnt/workspace/lielin.hyl/dj_synth_test/tutorials/outputs/image_captioning_output/res.jsonl'
export_shard_size	0
export_in_parallel	False
keep_stats_in_res_ds	False
keep_hashes_in_res_ds	False
np	4
text_keys	'text'
image_key	'images'
image_special_token	'<__dj__image>'
audio_key	'audios'
audio_special_token	'<__dj__audio>'
video_key	'videos'
video_special_token	'<__dj__video>'
eoc_special_token	'<|__dj__eoc|>'

```
| suffixes                  | []
|
| use_cache                 | True
|
| ds_cache_dir              | '/root/.cache/huggingface/datasets'
|
| cache_compress            | None
|
| use_checkpoint            | False
|
| temp_dir                  | None
|
| open_tracer               | False
|
| op_list_to_trace          | []
|
| trace_num                 | 10
|
| op_fusion                 | False
|
| process                   | [{'image_captioning_mapper': {'accelerato
r': 'cpu',                                                    |
|                           |                                 'audio_key':
'audios',                                                    |
|                           |                                 'batched_op':
False,                                                       |
|                           |                                 'caption_nu
m': 1,                                                       |
|                           |                                 'cpu_require
d': 1,                                                       |
|                           |                                 'hf_img2seq':
'/mnt/workspace/lielin.hyl/models/blip2-opt-2.7b',           |
|                           |                                 'image_key':
'images',                                                    |
|                           |                                 'keep_candida
te_mode': 'random_any',                                      |
|                           |                                 'keep_origina
l_sample': False,                                            |
```

```
│                              │                                        'mem_require
d': 0,                                                                        │
│                              │                                        'prompt': Non
e,                                                                            │
│                              │                                        'prompt_key':
None,                                                                         │
│                              │                                        'spec_numproc
s': 0,                                                                        │
│                              │                                        'text_key':
'text',                                                                       │
│                              │                                        'use_actor':
False,                                                                        │
│                              │                                        'video_key':
'videos'}}]                                                                   │
├──────────────────────────────┼──────────────────────────────────────────────────────
──────────────────────────────────────────────┤
│ percentiles                  │ []
│
├──────────────────────────────┼──────────────────────────────────────────────────────
──────────────────────────────────────────────┤
│ export_original_dataset      │ False
│
├──────────────────────────────┼──────────────────────────────────────────────────────
──────────────────────────────────────────────┤
│ save_stats_in_one_file       │ False
│
├──────────────────────────────┼──────────────────────────────────────────────────────
──────────────────────────────────────────────┤
│ ray_address                  │ 'auto'
│
├──────────────────────────────┼──────────────────────────────────────────────────────
──────────────────────────────────────────────┤
│ debug                        │ False
│
├──────────────────────────────┼──────────────────────────────────────────────────────
──────────────────────────────────────────────┤
│ work_dir                     │ '/mnt/workspace/lielin.hyl/dj_synth_test/tu
torials/outputs/image_captioning_output'                                      │
├──────────────────────────────┼──────────────────────────────────────────────────────
──────────────────────────────────────────────┤
│ timestamp                    │ '20240726042033'
│
├──────────────────────────────┼──────────────────────────────────────────────────────
──────────────────────────────────────────────┤
│ dataset_dir                  │ '/mnt/workspace/lielin.hyl/dj_synth_test/tu
torials'                                                                      │
├──────────────────────────────┼──────────────────────────────────────────────────────
──────────────────────────────────────────────┤
│ add_suffix                   │ False
│
╘══════════════════════════════╧══════════════════════════════════════════════════════
══════════════════════════════════════════════╛
2024-07-26 04:20:34 | INFO     | data_juicer.core.executor:49 - Using cach
e compression method: [None]
2024-07-26 04:20:34 | INFO     | data_juicer.core.executor:54 - Setting up
data formatter...
2024-07-26 04:20:34 | INFO     | data_juicer.core.executor:76 - Preparing
exporter...
2024-07-26 04:20:34 | INFO     | data_juicer.core.executor:153 - Loading d
ataset from data formatter...
```

```
Setting num_proc from 4 back to 1 for the jsonl split to disable multiproc
essing as it only contains one shard.
Generating jsonl split: 2 examples [00:00, 358.87 examples/s]
num_proc must be <= 2. Reducing num_proc to 2 for dataset of size 2.
2024-07-26 04:20:35 | INFO     | data_juicer.format.formatter:185 - Unifyi
ng the input dataset formats...
2024-07-26 04:20:35 | INFO     | data_juicer.format.formatter:200 - There
are 2 sample(s) in the original dataset.
Filter (num_proc=2): 100%|##########| 2/2 [00:06<00:00,  3.04s/ examples]
num_proc must be <= 2. Reducing num_proc to 2 for dataset of size 2.
2024-07-26 04:20:41 | INFO     | data_juicer.format.formatter:214 - 2 samp
les left after filtering empty text.
2024-07-26 04:20:41 | INFO     | data_juicer.format.formatter:237 - Conver
ting relative paths in the dataset to their absolute version. (Based on th
e directory of input dataset file)
Map (num_proc=2): 100%|##########| 2/2 [00:06<00:00,  3.04s/ examples]
2024-07-26 04:20:47 | INFO     | data_juicer.format.mixture_formatter:137
- sampled 2 from 2
2024-07-26 04:20:47 | INFO     | data_juicer.format.mixture_formatter:143
- There are 2 in final dataset
2024-07-26 04:20:47 | INFO     | data_juicer.core.executor:159 - Preparing
process operators...
Loading checkpoint shards: 100%|##########| 2/2 [00:05<00:00,  2.73s/it]
2024-07-26 04:20:53 | INFO     | data_juicer.core.executor:166 - Processin
g data...
num_proc must be <= 2. Reducing num_proc to 2 for dataset of size 2.
2024-07-26 04:20:53 | WARNING  | data_juicer.utils.process_utils:36 - The
required cuda memory of Op[image_captioning_mapper] has not been specifie
d. Please specify the mem_required field in the config file, or you might
encounter CUDA out of memory error. You can reference the mem_required fie
ld in the config_all.yaml file.
Loading checkpoint shards: 100%|██████████| 2/2 [00:05<00:00,  2.6
5s/it]xamples/s]
Loading checkpoint shards: 100%|██████████| 2/2 [00:05<00:00,  2.7
2s/it]
image_captioning_mapper_process (num_proc=2): 100%|##########| 2/2 [00:15<
00:00,  7.57s/ examples]
2024-07-26 04:21:09 | INFO     | data_juicer.core.executor:255 - Op [image
_captioning_mapper] Done in 15.468(s). Left 2 samples.
2024-07-26 04:21:09 | INFO     | data_juicer.core.executor:259 - All Ops a
re done in 15.468(s).
2024-07-26 04:21:09 | INFO     | data_juicer.core.executor:262 - Exporting
dataset to disk...
2024-07-26 04:21:09 | INFO     | data_juicer.core.exporter:140 - Export da
taset into a single file...
Creating json from Arrow format: 100%|##########| 1/1 [00:00<00:00, 136.52
ba/s]
```

We can read the result dataset to check the differences before and after recaptioning.

In [5]:
```
with jl.open('outputs/image_captioning_output/res.jsonl') as reader:
    for s in reader:
        vis(s)
```

```
<__dj__image> the royal wedding mug is shown on red background
 <|__dj__eoc|>
```

<__dj__image> lone lone acacia against a dramatic sunrise in africa
<|__dj__eoc|>

As we can see, this OP recaption these two images, remove some noisy information from the texts and add some details to tem to improve their qulity, such as adding "red background" and removing "© monkeypox". It suggests that recaption the images with more excellent models could synthesize better image-text samples with better cross-modality alignment.

Image Synthesis

We can generate a new image or edit the content of the original image with OP `image_diffusion_mapper`. Similarly, we need to create a data recipe for this process.

In [6]:
```
recipe = '''
dataset_path: ds.jsonl
export_path: outputs/image_diffusion_output/res.jsonl

process:
  - image_diffusion_mapper:
      hf_diffusion: 'CompVis/stable-diffusion-v1-4'  # You can replace th
      keep_original_sample: false  # we only need the recaptioned caption
      caption_key: 'text'
'''

with open('image_diffusion.yaml', 'w') as fout:
    fout.write(recipe)
```

Then we can run the process program of Data-Juicer to process the dataset.

In [7]:
```
!dj-process --config image_diffusion.yaml
```

2024-07-26 04:22:31.685 | INFO | data_juicer:setup_mp:67 - Setting multiprocess start method to 'forkserver'.
2024-07-26 04:22:39 | INFO | data_juicer.config.config:646 - Back up the input config file [/mnt/workspace/lielin.hyl/dj_synth_test/tutorials/image_diffusion.yaml] into the work_dir [/mnt/workspace/lielin.hyl/dj_synth_test/tutorials/outputs/image_diffusion_output]
2024-07-26 04:22:39 | INFO | data_juicer.config.config:668 - Configuration table:

key	values
config	[Path_fr(image_diffusion.yaml, cwd=/mnt/workspace/lielin.hyl/dj_synth_test/tutorials)]
hpo_config	None
path_k_sigma_recipe	None
path_model_feedback_recipe	None
model_infer_config	None
model_train_config	None
data_eval_config	None
model_eval_config	None
data_probe_algo	'uniform'
data_probe_ratio	1.0
project_name	'hello_world'
executor_type	'default'

dataset_path	'/mnt/workspace/lielin.hyl/dj_synth_test/tutorials/ds.jsonl'
export_path	'/mnt/workspace/lielin.hyl/dj_synth_test/tutorials/outputs/image_diffusion_output/res.jsonl'
export_shard_size	0
export_in_parallel	False
keep_stats_in_res_ds	False
keep_hashes_in_res_ds	False
np	4
text_keys	'text'
image_key	'images'
image_special_token	'<__dj__image>'
audio_key	'audios'
audio_special_token	'<__dj__audio>'
video_key	'videos'
video_special_token	'<__dj__video>'
eoc_special_token	'<|__dj__eoc|>'

| suffixes | []
| use_cache | True
| ds_cache_dir | '/root/.cache/huggingface/datasets'
| cache_compress | None
| use_checkpoint | False
| temp_dir | None
| open_tracer | False
| op_list_to_trace | []
| trace_num | 10
| op_fusion | False
| process | [{'image_diffusion_mapper': {'accelerator': 'cpu', 'audio_key': 'audios', 'aug_num': 1, 'batched_op': False, 'caption_key': 'text', 'cpu_required': 1, 'guidance_scale': 7.5, 'hf_diffusion': '/mnt/workspace/lielin.hyl/models/stable-diffusion-v1-4', 'hf_img2seq': 'Salesforce/blip2-opt-2.7b',

```
│                                 │                                 'image_key':
'images',                                                                       │
│                                 │                                 'keep_original
_sample': False,                                                                │
│                                 │                                 'mem_require
d': 0,                                                                           │
│                                 │                                 'revision': 'm
ain',                                                                           │
│                                 │                                 'spec_numproc
s': 0,                                                                           │
│                                 │                                 'strength': 0.
8,                                                                              │
│                                 │                                 'text_key': 't
ext',                                                                           │
│                                 │                                 'torch_dtype':
'fp32',                                                                         │
│                                 │                                 'use_actor': F
alse,                                                                           │
│                                 │                                 'video_key':
'videos'}}]                                                                      │
├─────────────────────────────────┼─────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────────────────┤
│ percentiles                     │ []
│
├─────────────────────────────────┼─────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────────────────┤
│ export_original_dataset         │ False
│
├─────────────────────────────────┼─────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────────────────┤
│ save_stats_in_one_file          │ False
│
├─────────────────────────────────┼─────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────────────────┤
│ ray_address                     │ 'auto'
│
├─────────────────────────────────┼─────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────────────────┤
│ debug                           │ False
│
├─────────────────────────────────┼─────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────────────────┤
│ work_dir                        │ '/mnt/workspace/lielin.hyl/dj_synth_test/tu
torials/outputs/image_diffusion_output'                                         │
├─────────────────────────────────┼─────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────────────────┤
│ timestamp                       │ '20240726042237'
│
├─────────────────────────────────┼─────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────────────────┤
│ dataset_dir                     │ '/mnt/workspace/lielin.hyl/dj_synth_test/tu
torials'                                                                        │
├─────────────────────────────────┼─────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────────────────┤
│ add_suffix                      │ False
│
╘═════════════════════════════════╧═════════════════════════════════════════════
═════════════════════════════════════════════════════════════════════════════════╛
2024-07-26 04:22:39 | INFO     | data_juicer.core.executor:49 - Using cach
e compression method: [None]
```

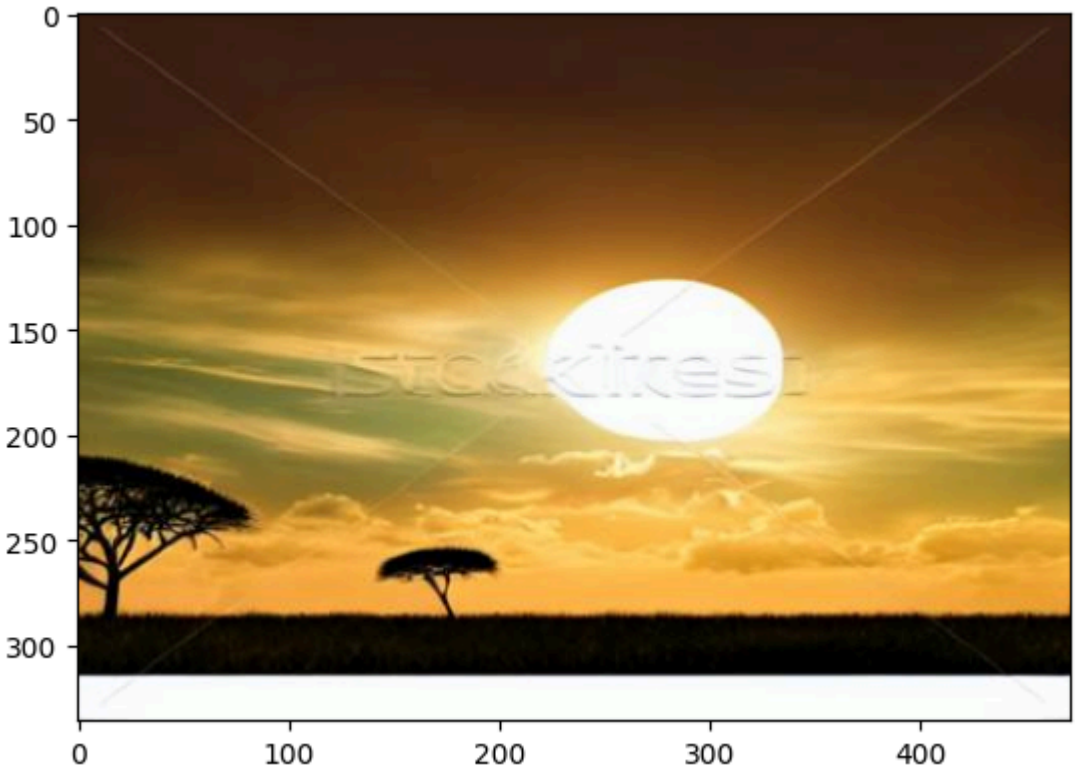
```
2024-07-26 04:22:39 | INFO     | data_juicer.core.executor:54 - Setting up data formatter...
2024-07-26 04:22:39 | INFO     | data_juicer.core.executor:76 - Preparing exporter...
2024-07-26 04:22:39 | INFO     | data_juicer.core.executor:153 - Loading dataset from data formatter...
num_proc must be <= 2. Reducing num_proc to 2 for dataset of size 2.
2024-07-26 04:22:40 | INFO     | data_juicer.format.formatter:185 - Unifying the input dataset formats...
2024-07-26 04:22:40 | INFO     | data_juicer.format.formatter:200 - There are 2 sample(s) in the original dataset.
num_proc must be <= 2. Reducing num_proc to 2 for dataset of size 2.
2024-07-26 04:22:40 | INFO     | data_juicer.format.formatter:214 - 2 samples left after filtering empty text.
2024-07-26 04:22:40 | INFO     | data_juicer.format.formatter:237 - Converting relative paths in the dataset to their absolute version. (Based on the directory of input dataset file)
2024-07-26 04:22:40 | INFO     | data_juicer.format.mixture_formatter:137 - sampled 2 from 2
2024-07-26 04:22:40 | INFO     | data_juicer.format.mixture_formatter:143 - There are 2 in final dataset
2024-07-26 04:22:40 | INFO     | data_juicer.core.executor:159 - Preparing process operators...
Loading pipeline components...: 100%|##########| 7/7 [00:24<00:00,  3.51s/it]
2024-07-26 04:23:05 | INFO     | data_juicer.core.executor:166 - Processing data...
num_proc must be <= 2. Reducing num_proc to 2 for dataset of size 2.
2024-07-26 04:23:05 | WARNING  | data_juicer.utils.process_utils:36 - The required cuda memory of Op[image_diffusion_mapper] has not been specified. Please specify the mem_required field in the config file, or you might encounter CUDA out of memory error. You can reference the mem_required field in the config_all.yaml file.
Loading pipeline components...: 100%|██████████| 7/7 [00:00<00:00,  7.26it/s]amples/s]
Loading pipeline components...: 100%|██████████| 7/7 [00:00<00:00,  7.21it/s]
100%|██████████████████████████████████████████| 40/40 [00:04<00:00,  8.14it/s]
100%|██████████████████████████████████████████| 40/40 [00:04<00:00,  8.01it/s]
image_diffusion_mapper_process (num_proc=2): 100%|##########| 2/2 [00:28<00:00, 14.37s/ examples]
2024-07-26 04:23:34 | INFO     | data_juicer.core.executor:255 - Op [image_diffusion_mapper] Done in 29.125(s). Left 2 samples.
2024-07-26 04:23:34 | INFO     | data_juicer.core.executor:259 - All Ops are done in 29.125(s).
2024-07-26 04:23:34 | INFO     | data_juicer.core.executor:262 - Exporting dataset to disk...
2024-07-26 04:23:34 | INFO     | data_juicer.core.exporter:140 - Export dataset into a single file...
Creating json from Arrow format: 100%|##########| 1/1 [00:00<00:00, 164.41ba/s]
```

We can read the result dataset to check the differences before and after synthesis.

In [9]:
```
with jl.open('outputs/image_diffusion_output/res.jsonl') as reader:
    for s in reader:
        vis(s)
```

<__dj__image> a picture of prince and princess kate's mugs in a frame

the setting sun in africa on a cloudy day stock photo © monkeypox <__dj__image>



As we can see, this OP synthesize new images for this two captions, also try to remove extra and noisy vision information from the images. For example, most texts and watermarks from both synthsized images are removed, and the conceptions of "mugs" and "sun" in these two images are enhanced. Therefore, these two synthesized images can be considered useful for training of modality alignment.

Conclusion

In this notebook, we learn how to synthesize new samples with Data-Juicer OPs through a use case on image-text datasets and check their quality changes in the synthesized results.