CapsFusion: Rethinking Image-Text Data at Scale
Supplementary Material

6. More CapsFusion Examples

More examples of web-based raw captions, synthetic captions generated by BLIP, and their CapsFusion captions generated by CapsFus-LLaMA are provided in Fig. 8. CapsFusion captions can organically organize information from raw and synthetic captions.

7. Prompting Templates for Data Refining

The prompt for ChatGPT and CapsFus-LLaMA to integrate raw and synthetic captions is shown below.

```
Please merge and refine the information
    from the two given sentences.
Sentence 1 provides detailed real-world
knowledge, yet it suffers from flaws in
    sentence structure and grammar.
   Sentence 2 exhibits nice sentence
structure, but lacking in-depth real-world
details and may contain false information.
 Please combine them into a new sentence,
ensuring a well-structured sentence while
    retaining the detailed real-world
   information provided in Sentence 1.
Avoid simply concatenating the sentences.
       Sentence 1:  <raw caption>
     Sentence 2:  <synthetic caption>
```

8. Hyperparameters

Training hyperparameters of CapsFus-LLaMA and LMM are presented in Tabs. 6 and 7 respectively.

Configuration	CapsFusion Training
Model init	LLaMA-2-13B
Batch size	128
Data	1 million (raw, synthetic, fused) triplets from ChatGPT
Training Epoch	2
Peak Learning Rate	1e-5
End Learning Rate	0
Warmup Steps	500
LR Scheduler	cosine
Optimizer	AdamW [38]
Optimizer hyper-parameters	$\beta_1, \beta_2, \epsilon$ = 0.9, 0.95, 1e-8
Weight decay	0.0

Table 6. Summary of CapsFus-LLaMA training hyperparameters.

Configuration	Large Multimodal Model Training
Model init	LLaMA-2-13B
Batch size	8192
Data	10 / 50 / 100 million image-text pairs
Training Epoch	1
Peak Learning Rate	3e-4
End Learning Rate	3e-5
Warmup Steps	2000
LR Scheduler	cosine
Optimizer	AdamW [38]
Optimizer hyper-parameters	$\beta_1, \beta_2, \epsilon$ = 0.9, 0.98, 1e-6
Weight decay	0.0

Table 7. Summary of LMM training hyperparameters.

9. Details of SEED-Bench

SEED-Bench [32] incorporates 12 evaluation tasks including both the spatial and temporal comprehension to comprehensively assess the visual understanding capability of LMMs. We select 9 image-text tasks from them (the left 3 tasks are video-text tasks) for evaluation. The task details are introduced below.

- Scene Understanding. This task focuses on the global information in the image. Questions can be answered through a holistic understanding of the image.
- Instance Identity. This task involves the identification of a certain instance in the image, including the existence or category of a certain object in the image. It evaluates a model's object recognition capability.
- Instance Location. This task concerns the absolute position of one specified instance. It requires a model to correctly localize the object referred to in the question.
- Instance Attributes. This task is related to the attributes of an instance, such as color, shape or material. It assesses a model's understanding of an object's visual appearance.
- Instances Counting. This task requires the model to count the number of a specific object in the image. This requires the model to understand all objects, and successfully count the referred object's instances.
- Spatial Relation. This task asks an model to ground the two mentioned objects, and recognize their relative spa-

① <a gi-track='captionPersonalityLinkClicked' href=/galleries/search?phrase=Stefan+Raab&family=editorial&specificpeople=242927 ng-click='$event.stopPropagation()'>Stefan Raab and Sandra Riess hosts the 'Bundesvision Song Contest 2012' at the Max-Schmeling-Halle on September 28, 2012 in Berlin, Germany.

② The television studio with several monitors and people.

③ Stefan Raab and Sandra Riess hosted the 'Bundesvision Song Contest 2012' at the Max-Schmeling-Halle on September 28, 2012 in Berlin, Germany, in a television studio equipped with several monitors and filled with people.

① Actress <a gi-track='captionPersonalityLinkClicked' href=/galleries/search?phrase=Roxane+Mesquida&family=editorial&specificpeople=217749 ng-click='$event.stopPropagation()'>Roxane Mesquida, director Xan Cassavetes and actress Josephine de La Baume attend a screening of Magnolia Pictures' 'Kiss of the Damned' at ArcLight Cinemas on April 29, 2013 in Hollywood, California.

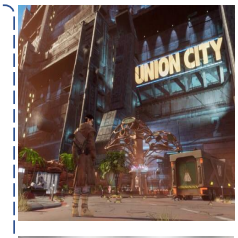
② Three women pose for the camera at an event.

③ Actress Roxane Mesquida, director Xan Cassavetes, and actress Josephine de La Baume were photographed posing for the camera at a screening of Magnolia Pictures' 'Kiss of the Damned' at ArcLight Cinemas on April 29, 2013.

① Minecraft Gets an 'Honest Game Trailer'

② an image of a man holding a creeper

③ An 'Honest Game Trailer' for Minecraft has been released, featuring an image of a man holding a creeper.



① Beyond a Steel Sky on consoles on November 30 .

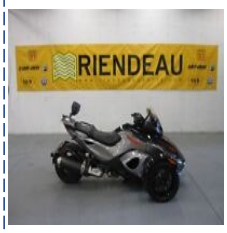
② The video game,'Union City, is shown.

③ On November 30, the video game "Beyond a Steel Sky" will be released on consoles, showcasing the immersive world of "Union City.

① Bota Caterpillar Second Shift Boot + Cinto Couro - Pull up

② The caterpy boots are brown and have yellow laces.

③ The Bota Caterpillar Second Shift Boot + Cinto Couro is a brown boot with yellow laces.



① 2012 Can-Am Spyder RS-S SE5

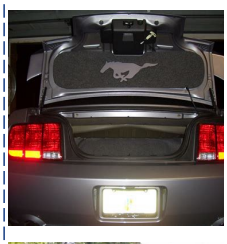
② A motorcycle is parked in front of a banner

③ In front of a banner, there is a 2012 Can-Am Spyder RS-S SE5 motorcycle parked.

① le restaurant : Cristal Room

② The dining room is set for dinner with chandeliers.

③ The Cristal Room is a restaurant that is elegantly decorated with chandeliers, creating a sophisticated ambiance for dinner.



① trunk lid cosmetic cover saleen

② The back end of a car with its trunk open.

③ The Saleen trunk lid cosmetic cover enhances the appearance of the back end of a car when the trunk is open.

① Plus Size Black Color Butterfly Long Sleeve Shirt Tops

② A woman in black shirt and jeans with colorful butterfly print.

③ A woman is wearing a plus-size black shirt with long sleeves, featuring a colorful butterfly print, paired with jeans.

① 1951 Ford Other Pickups

② An old red truck parked in the parking lot.

③ In the parking lot, there is a 1951 Ford Other Pickups, an old red truck.'

Figure 8. Examples of ① raw captions (from LAION-2B), ② synthetic captions (from LAION-COCO, generated by BLIP), and their corresponding ③ CAPSFUSION captions. Knowledge from raw captions (in blue) and information from synthetic captions (in yellow) are organically fused into integral CAPSFUSION captions. The dataset will be released and more examples can be found there.

tial relation within the image.

- Instance Interaction. This task requires the model to recognize the state relation or interaction relations between two humans or objects.
- Visual Reasoning. This task evaluates if a model is able to reason based on the visual information. This requires the

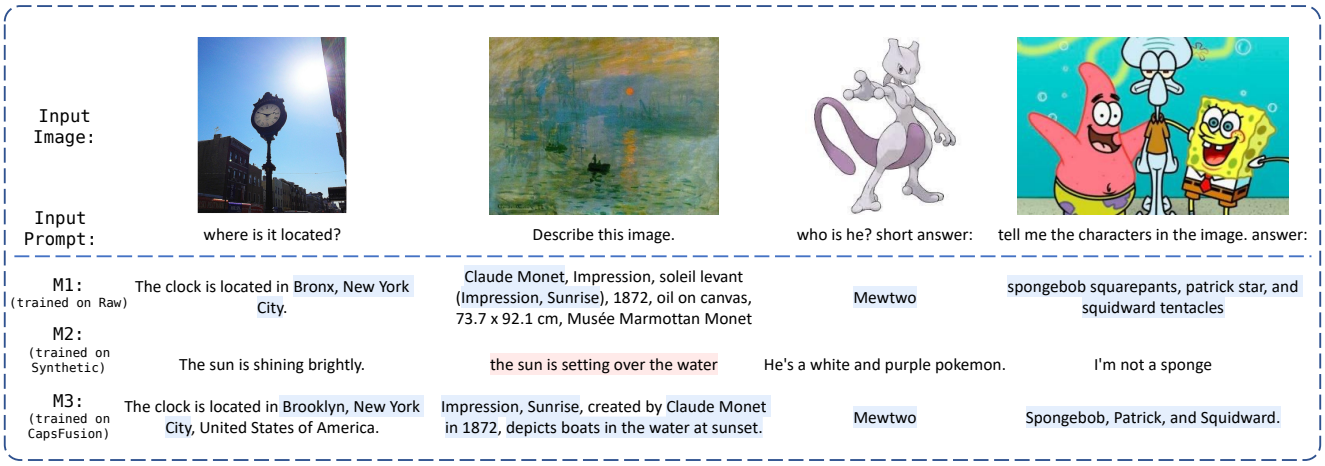


Figure 9. Outputs of models trained with different caption datasets. Models trained on raw and CapsFusion captions (M1 and 3) possess strong world knowledge (in blue), while the model trained on synthetic captions (M2) can only generate generic concepts (in red).

model to fully understand the image and utilize its commonsense knowledge to correctly answer the questions.

- Text Understanding. For this task, the model should answer question about the textual elements in the image.