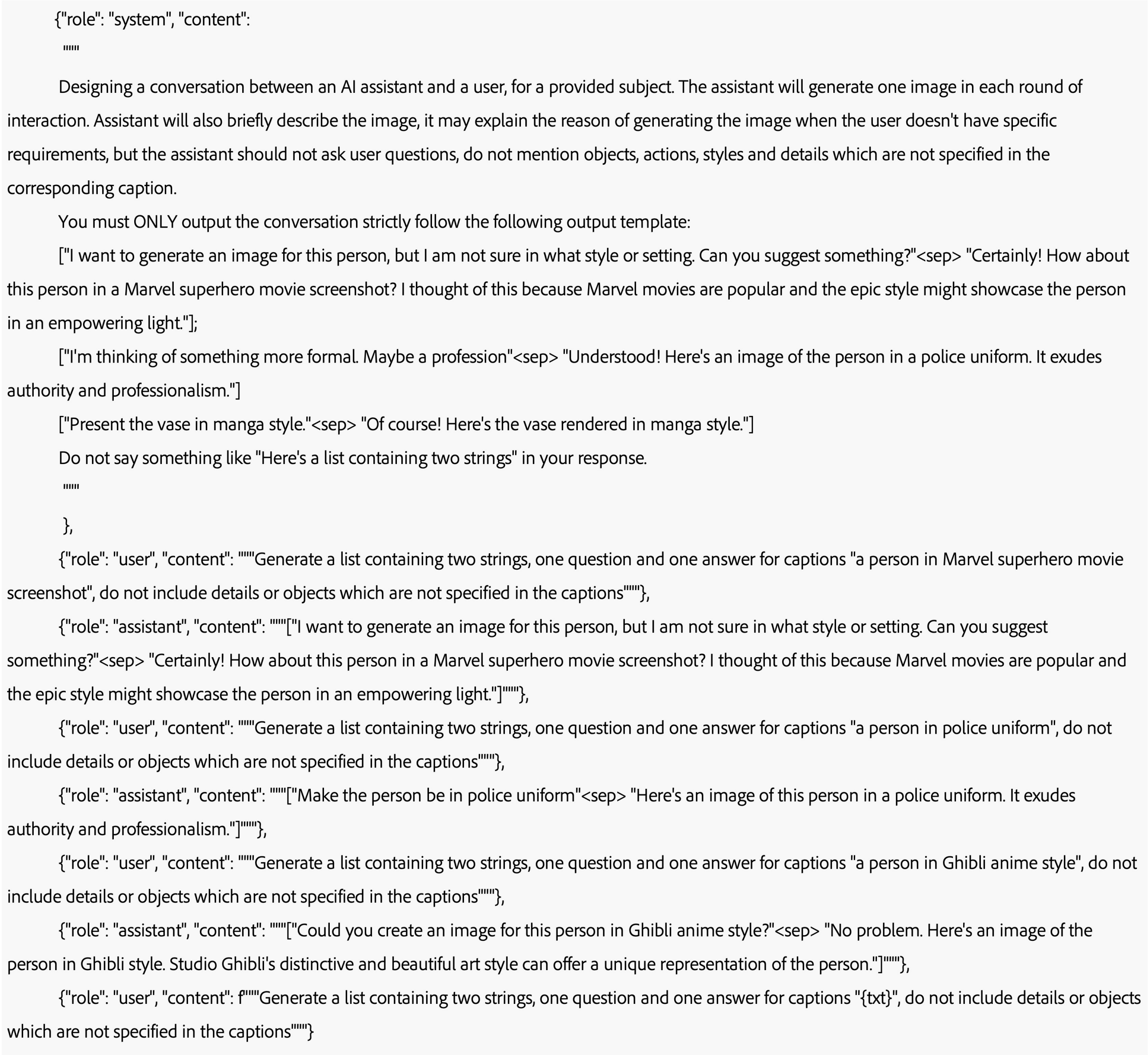
A. Implementation Details

We provide some implementation details here.

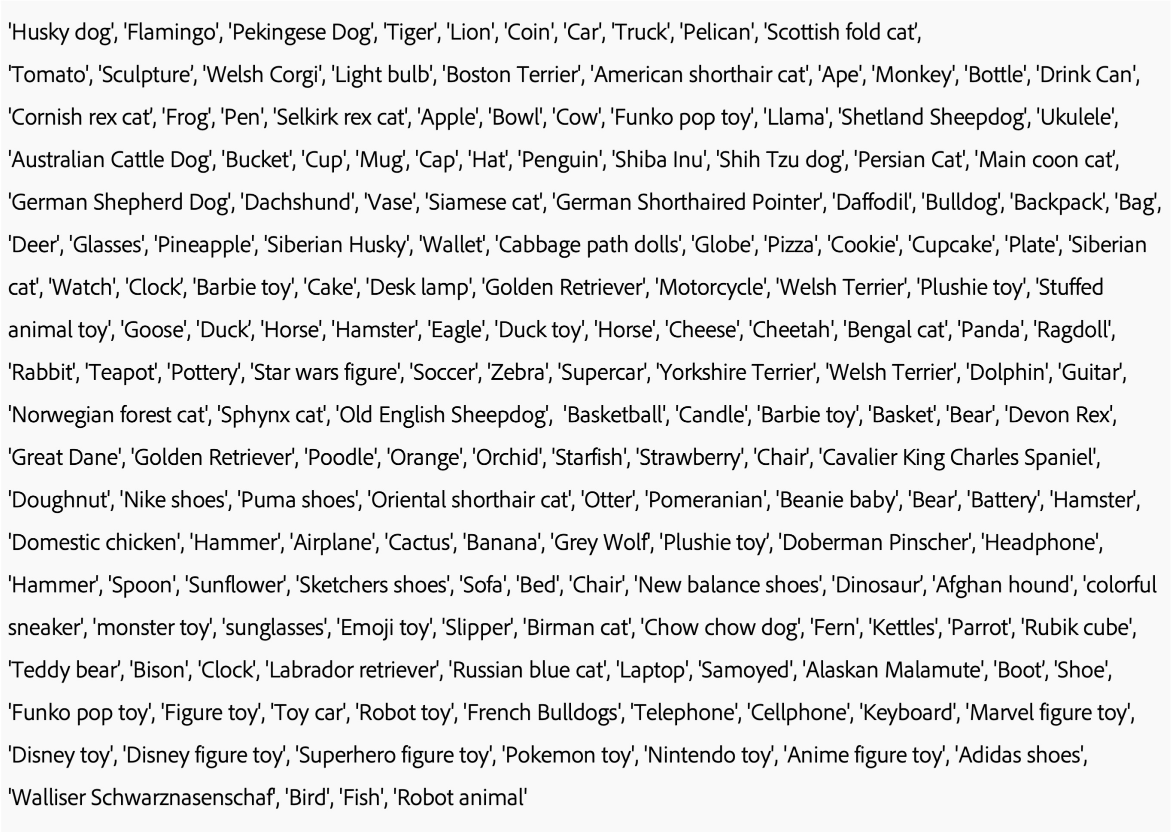
Text data generation We provide prompt example which we use to generate text data with the Llama-2-70B-chat model in Figure 8. We are able to obtain the expected output with this prompt in most cases, while it also leads to failure sometimes. Thus we apply an automatic filtering mechanism on the generated text data, where undesired outputs such as empty strings are filtered out.



```
{"role": "system", "content":
 """
 Designing a conversation between an AI assistant and a user, for a provided subject. The assistant will generate one image in each round of interaction. Assistant will also briefly describe the image, it may explain the reason of generating the image when the user doesn't have specific requirements, but the assistant should not ask user questions, do not mention objects, actions, styles and details which are not specified in the corresponding caption.
 You must ONLY output the conversation strictly follow the following output template:
 ["I want to generate an image for this person, but I am not sure in what style or setting. Can you suggest something?"<sep> "Certainly! How about this person in a Marvel superhero movie screenshot? I thought of this because Marvel movies are popular and the epic style might showcase the person in an empowering light."];
 ["I'm thinking of something more formal. Maybe a profession"<sep> "Understood! Here's an image of the person in a police uniform. It exudes authority and professionalism."]
 ["Present the vase in manga style."<sep> "Of course! Here's the vase rendered in manga style."]
 Do not say something like "Here's a list containing two strings" in your response.
 """
 },
 {"role": "user", "content": """Generate a list containing two strings, one question and one answer for captions "a person in Marvel superhero movie screenshot", do not include details or objects which are not specified in the captions"""},
 {"role": "assistant", "content": """["I want to generate an image for this person, but I am not sure in what style or setting. Can you suggest something?"<sep> "Certainly! How about this person in a Marvel superhero movie screenshot? I thought of this because Marvel movies are popular and the epic style might showcase the person in an empowering light."]"""},
 {"role": "user", "content": """Generate a list containing two strings, one question and one answer for captions "a person in police uniform", do not include details or objects which are not specified in the captions"""},
 {"role": "assistant", "content": """["Make the person be in police uniform"<sep> "Here's an image of this person in a police uniform. It exudes authority and professionalism."]"""},
 {"role": "user", "content": """Generate a list containing two strings, one question and one answer for captions "a person in Ghibli anime style", do not include details or objects which are not specified in the captions"""},
 {"role": "assistant", "content": """["Could you create an image for this person in Ghibli anime style?"<sep> "No problem. Here's an image of the person in Ghibli style. Studio Ghibli's distinctive and beautiful art style can offer a unique representation of the person."]"""},
 {"role": "user", "content": f"""Generate a list containing two strings, one question and one answer for captions "{txt}", do not include details or objects which are not specified in the captions"""}
```

Figure 8. Prompt example we used to generate data samples with Llama-2-70B-chat model.

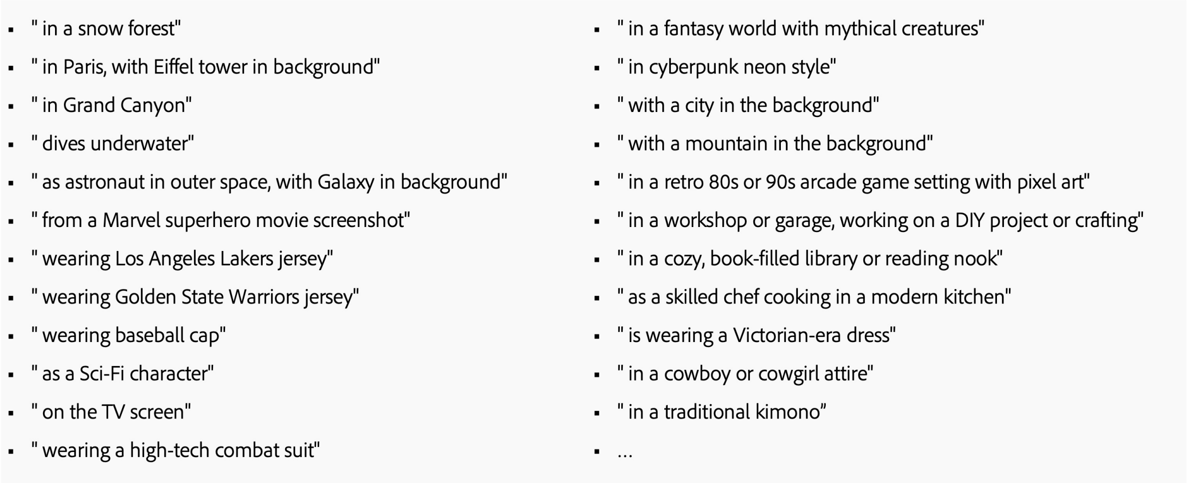
Image data generation As mentioned in Section 4, original images $\{\mathbf{x}_i\}$ in object domain are generated first generated with pre-trained text-to-image generation model (Stable Diffusion). Let c_i be a class name, we use prompt "A realistic photo picture of c_i, ultra quality, sharp focus, tack sharp, dof, film grain, Fujifilm XT3, crystal clear, 8K UHD, highly detailed" to generate the image. c_i is randomly selected from the classes shown in Figure 9.



'Husky dog', 'Flamingo', 'Pekingese Dog', 'Tiger', 'Lion', 'Coin', 'Car', 'Truck', 'Pelican', 'Scottish fold cat', 'Tomato', 'Sculpture', 'Welsh Corgi', 'Light bulb', 'Boston Terrier', 'American shorthair cat', 'Ape', 'Monkey', 'Bottle', 'Drink Can', 'Cornish rex cat', 'Frog', 'Pen', 'Selkirk rex cat', 'Apple', 'Bowl', 'Cow', 'Funko pop toy', 'Llama', 'Shetland Sheepdog', 'Ukulele', 'Australian Cattle Dog', 'Bucket', 'Cup', 'Mug', 'Cap', 'Hat', 'Penguin', 'Shiba Inu', 'Shih Tzu dog', 'Persian Cat', 'Main coon cat', 'German Shepherd Dog', 'Dachshund', 'Vase', 'Siamese cat', 'German Shorthaired Pointer', 'Daffodil', 'Bulldog', 'Backpack', 'Bag', 'Deer', 'Glasses', 'Pineapple', 'Siberian Husky', 'Wallet', 'Cabbage path dolls', 'Globe', 'Pizza', 'Cookie', 'Cupcake', 'Plate', 'Siberian cat', 'Watch', 'Clock', 'Barbie toy', 'Cake', 'Desk lamp', 'Golden Retriever', 'Motorcycle', 'Welsh Terrier', 'Plushie toy', 'Stuffed animal toy', 'Goose', 'Duck', 'Horse', 'Hamster', 'Eagle', 'Duck toy', 'Horse', 'Cheese', 'Cheetah', 'Bengal cat', 'Panda', 'Ragdoll', 'Rabbit', 'Teapot', 'Pottery', 'Star wars figure', 'Soccer', 'Zebra', 'Supercar', 'Yorkshire Terrier', 'Welsh Terrier', 'Dolphin', 'Guitar', 'Norwegian forest cat', 'Sphynx cat', 'Old English Sheepdog', 'Basketball', 'Candle', 'Barbie toy', 'Basket', 'Bear', 'Devon Rex', 'Great Dane', 'Golden Retriever', 'Poodle', 'Orange', 'Orchid', 'Starfish', 'Strawberry', 'Chair', 'Cavalier King Charles Spaniel', 'Doughnut', 'Nike shoes', 'Puma shoes', 'Oriental shorthair cat', 'Otter', 'Pomeranian', 'Beanie baby', 'Bear', 'Battery', 'Hamster', 'Domestic chicken', 'Hammer', 'Airplane', 'Cactus', 'Banana', 'Grey Wolf', 'Plushie toy', 'Doberman Pinscher', 'Headphone', 'Hammer', 'Spoon', 'Sunflower', 'Sketchers shoes', 'Sofa', 'Bed', 'Chair', 'New balance shoes', 'Dinosaur', 'Afghan hound', 'colorful sneaker', 'monster toy', 'sunglasses', 'Emoji toy', 'Slipper', 'Birman cat', 'Chow chow dog', 'Fern', 'Kettles', 'Parrot', 'Rubik cube', 'Teddy bear', 'Bison', 'Clock', 'Labrador retriever', 'Russian blue cat', 'Laptop', 'Samoyed', 'Alaskan Malamute', 'Boot', 'Shoe', 'Funko pop toy', 'Figure toy', 'Toy car', 'Robot toy', 'French Bulldogs', 'Telephone', 'Cellphone', 'Keyboard', 'Marvel figure toy', 'Disney toy', 'Disney figure toy', 'Superhero figure toy', 'Pokemon toy', 'Nintendo toy', 'Anime figure toy', 'Adidas shoes', 'Walliser Schwarznasenschaf', 'Bird', 'Fish', 'Robot animal'

Figure 9. Classes we used to generate original images for object domain.

With the original images, manually designed instructions are used to generate the target images $\{\tilde{\mathbf{x}}_i\}$ with ProFusion. Examples of our designed instructions are shown in Figure 10



- " in a snow forest"
- " in Paris, with Eiffel tower in background"
- " in Grand Canyon"
- " dives underwater"
- " as astronaut in outer space, with Galaxy in background"
- " from a Marvel superhero movie screenshot"
- " wearing Los Angeles Lakers jersey"
- " wearing Golden State Warriors jersey"
- " wearing baseball cap"
- " as a Sci-Fi character"
- " on the TV screen"
- " wearing a high-tech combat suit"
- " in a fantasy world with mythical creatures"
- " in cyberpunk neon style"
- " with a city in the background"
- " with a mountain in the background"
- " in a retro 80s or 90s arcade game setting with pixel art"
- " in a workshop or garage, working on a DIY project or crafting"
- " in a cozy, book-filled library or reading nook"
- " as a skilled chef cooking in a modern kitchen"
- " is wearing a Victorian-era dress"
- " in a cowboy or cowgirl attire"
- " in a traditional kimono"
- …

Figure 10. Examples of instructions we used in generating target images.

When we generate data samples with the assistant trained on automatically filtered data. We first generate images $\mathbf{w}$ with pre-trained Stable Diffusion, then apply Equation (4) with $\alpha = 0.7$.

B. Generated Examples

Some generated examples are provided below. We also provide examples where the proposed method is applied in multi-round generation and image editing task.

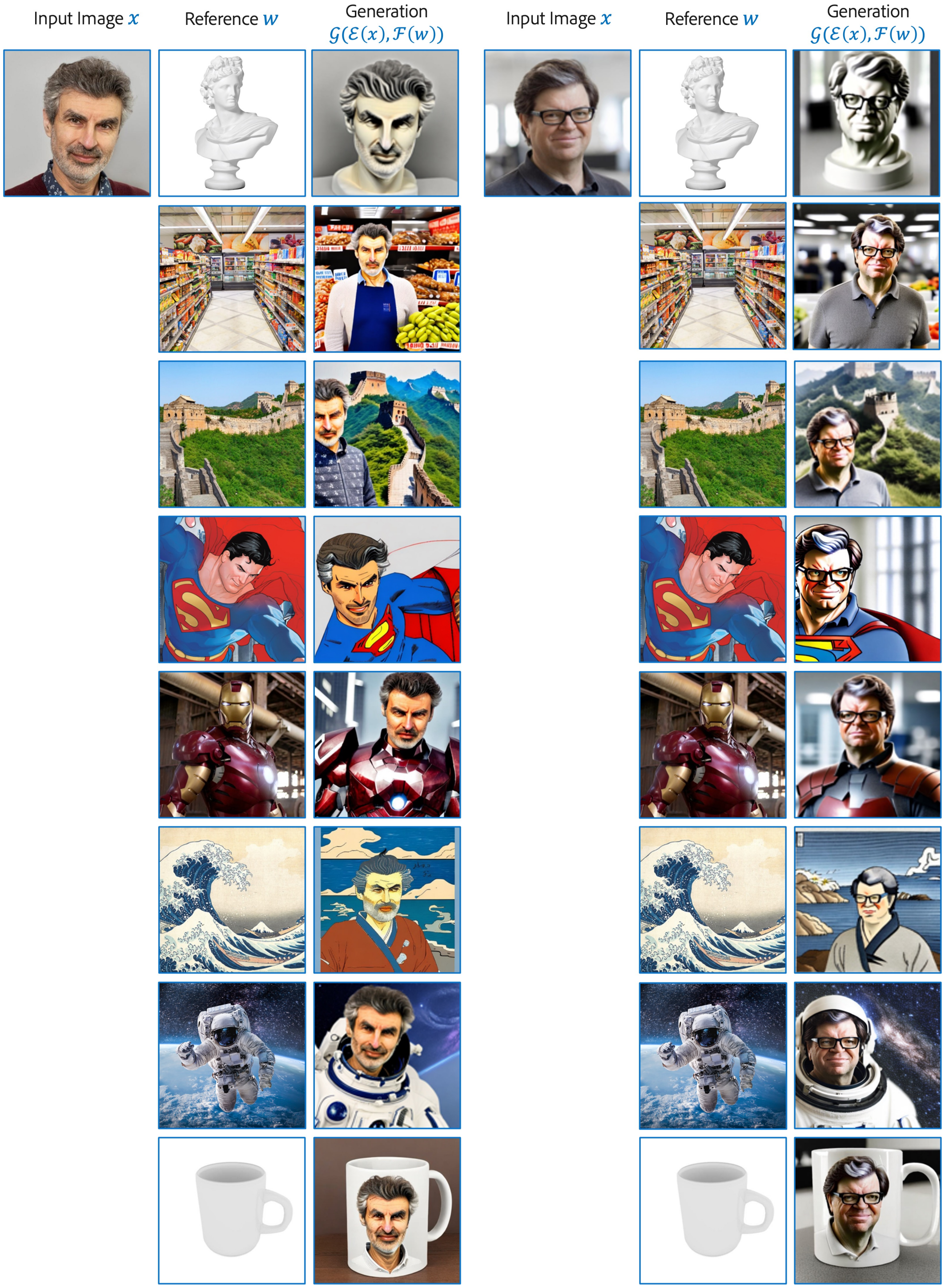
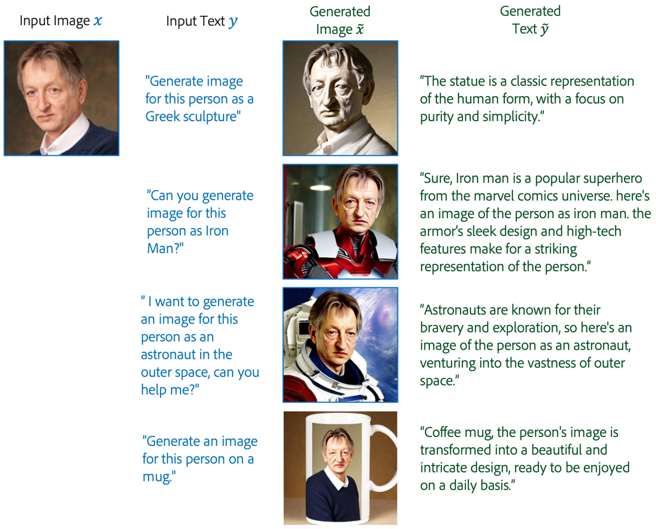

Input Image x
Reference w
Generation $\mathcal{G}(\mathcal{E}(x), \mathcal{F}(w))$
Input Image x
Reference w
Generation $\mathcal{G}(\mathcal{E}(x), \mathcal{F}(w))$

Figure 11. Generated examples from the proposed CAFE.



"Generate image for this person as a Greek sculpture"

"The statue is a classic representation of the human form, with a focus on purity and simplicity."

"Can you generate image for this person as Iron Man?"

"Sure, Iron man is a popular superhero from the marvel comics universe. here's an image of the person as iron man. the armor's sleek design and high-tech features make for a striking representation of the person."

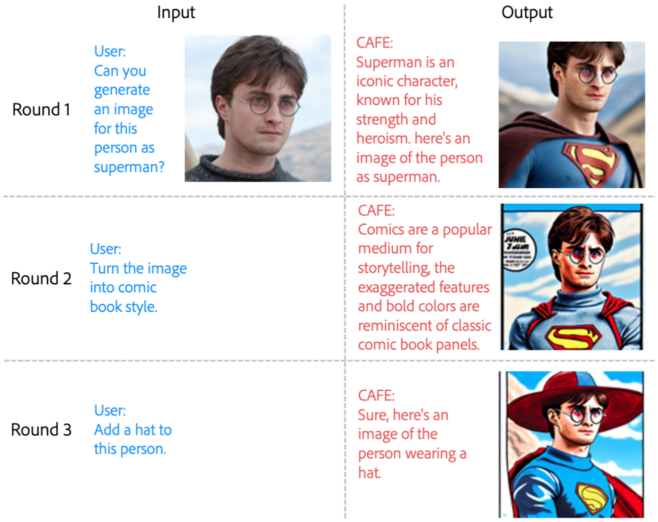
" I want to generate an image for this person as an astronaut in the outer space, can you help me?"

"Astronauts are known for their bravery and exploration, so here's an image of the person as an astronaut, venturing into the vastness of outer space."

"Generate an image for this person on a mug."

"Coffee mug, the person's image is transformed into a beautiful and intricate design, ready to be enjoyed on a daily basis."

Figure 12. Generated examples from the proposed CAFE.



Round 1

User: Can you generate an image for this person as superman?

CAFE: Superman is an iconic character, known for his strength and heroism. here's an image of the person as superman.

Round 2

User: Turn the image into comic book style.

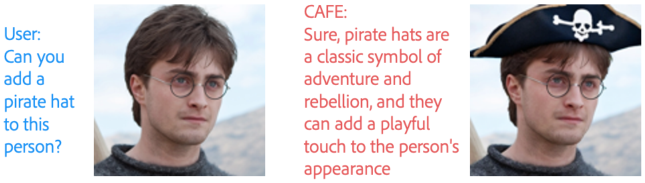
CAFE: Comics are a popular medium for storytelling, the exaggerated features and bold colors are reminiscent of classic comic book panels.

Round 3

User: Add a hat to this person.

CAFE: Sure, here's an image of the person wearing a hat.

Figure 13. Multi-round generation examples from the proposed CAFE.



User: Can you add a pirate hat to this person?

CAFE: Sure, pirate hats are a classic symbol of adventure and rebellion, and they can add a playful touch to the person's appearance

Figure 14. Image editing examples from the proposed CAFE.