Brush2Prompt: Contextual Prompt Generator for Object Inpainting

Supplementary Material

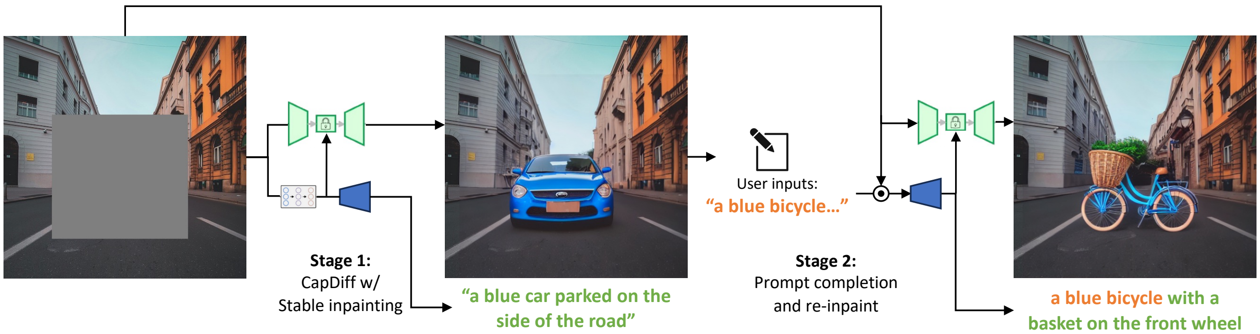


Figure 1. Pipeline visualization of our CapDiff and prompt completion pipeline.

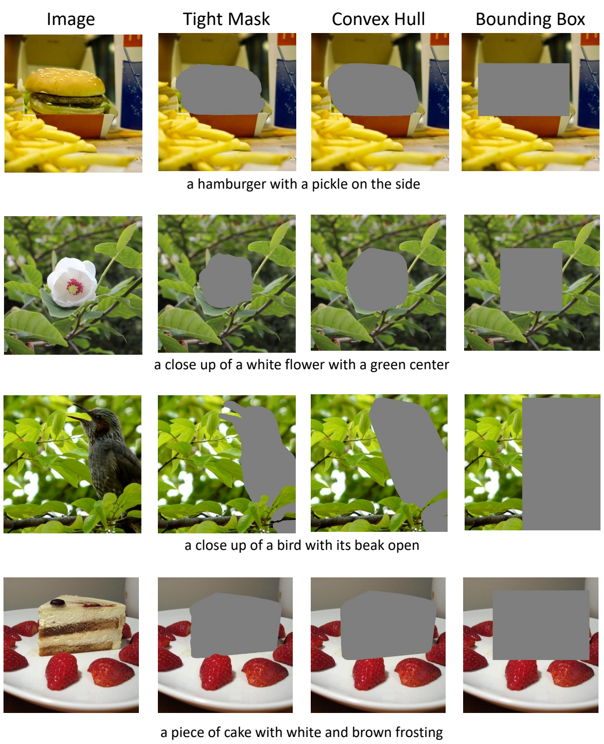


Figure 2. Visualization of the Brush2PromptBench dataset.

1. Additional benchmark dataset details

We provide additional details of our benchmark dataset Brush2PromptBench. As mentioned in the main paper, we select a subset of the OpenImages dataset where the images are larger than 512×512, we then apply a strict filter where we discard object masks where the locally generated BLIP [2] caption does not contain the category label. The resulting test set contains 4255 images. In our caption evaluations, we select the first instance by index, resulting in one mask-caption pair per image. We present visualizations of this benchmark dataset in Figure 2.

Mask shape	Method	BLEU ↑	BERT-Score ↑	Self-BLEU ↓	Div-4 ↑
Tight Mask	CapDiff	0.177	0.732	0.169	0.858
	CatDiff-Prompt	0.135	0.732	0.210	0.817
Convex Hull	CapDiff	0.162	0.724	0.152	0.876
	CatDiff-Prompt	0.117	0.719	0.184	0.844
Bounding box	CapDiff	0.149	0.715	0.141	0.885
	CatDiff-Prompt	0.097	0.709	0.150	0.877

Table 1. Comparison between CapDiff and CatDiff-Prompt.

2. Automatic prompt completion formulation

In Section 5 of the main paper, we discussed an extension of our pipeline for prompt completion. We define our complete pipeline in two stages. In the first stage, given only a masked image, our CapDiff model automatically suggests several prompts that describe appropriate object descriptions for concept insertion. In the second stage, the user enters a partial prompt based on the suggestions from CapDiff, which is then input into the text decoder together with the masked image embedding to generate the updated recommendation. We illustrate the complete prompt generation and completion pipeline in Figure 1.

Based on the above pipeline, we can also generate full captions by using CatDiff to suggest object categories, then feeding the category as prefix for prompt completion (e.g. "A horse..."). We call this alternative **CatDiff-Prompt** and quantitatively compare it with CapDiff in Table 1. The results suggest that CapDiff achieves both higher caption quality and better sentence diversity, we believe this is due to the fact that enforcing a fixed prefix ("A <category>...") leads to lower diversity and worse alignment from the ground truth captions compared to directly predicting the sentences.

Mask shape	Method	BLEU ↑	ROUGE ↑	BERTScore ↑	Dist-1 ↑	Self-BLEU ↓	Div-4 ↑
Tight Mask	BLIP-VQA [2]	0.005	0.071	0.537	0.998	0.600	0.038
	InstructBLIP [1]	0.071	0.308	0.704	0.882	0.466	0.570
	LLaVA-Resample [3]	0.031	0.275	0.684	0.955	0.286	0.715
	LLaVA-5-Prompt [3]	0.023	0.269	0.656	0.998	0.163	0.785
	CapDiff (Ours)	0.177	0.427	0.732	0.845	0.169	0.858
Convex Hull	BLIP-VQA [2]	0.004	0.059	0.537	0.999	0.607	0.028
	InstructBLIP [1]	0.061	0.283	0.692	0.892	0.441	0.591
	LLaVA-Resample [3]	0.025	0.246	0.667	0.954	0.271	0.730
	LLaVA-5-Prompt [3]	0.017	0.244	0.641	0.998	0.134	0.812
	CapDiff (Ours)	0.162	0.405	0.724	0.841	0.152	0.876
Bounding Box	BLIP-VQA [2]	0.004	0.053	0.526	0.999	0.627	0.025
	InstructBLIP [1]	0.060	0.280	0.691	0.881	0.438	0.597
	LLaVA-Resample [3]	0.026	0.254	0.678	0.970	0.312	0.694
	LLaVA-5-Prompt [3]	0.020	0.245	0.648	0.998	0.128	0.881
	CapDiff (Ours)	0.149	0.383	0.715	0.838	0.141	0.885

Table 2. Caption generation results on the OpenImages dataset including Convex Hull results.

3. Additional results on CapDiff evaluation

We present additional CapDiff evaluations including Convex Hull numbers in Table 2. We observe a similar trend with CatDiff, where caption accuracies decrease while sentence diversities increase as we loosen the shape constraint. In all three mask shape settings, our method achieves the best accuracy-diversity trade-off across generic visual-language models.

4. Additional visualizations

We provide additional visualizations of our CapDiff pipeline in Figure 3, Figure 4 and Figure 5. We provide visualizations for some failure cases in Figure 6. We also provide more visualizations regarding the significance of our shape awareness design in Figure 7, and a pair-wise comparison with generic visual-language models in Figure 8.

5. User Study

We conducted a mini user study with 50 images using Appen, and each image is assigned to 5 users. We recruited 98 users in total. We assessed our CapDiff model against the stronger baseline, LLaVA. Participants were shown masked images and the predictions from both methods. They were asked to make binary choices to indicate their preference for fidelity and diversity. Preferences were counted when one method received over 3 votes (out of 5); otherwise, the result was considered a tie. The results demonstrated the user preference for CapDiff over LLaVA in terms of fidelity and diversity. The results are presented in Table 3.

	Prefer LLaVA	Tie	Prefer CapDiff
Fidelity	12%	58%	**30%**
Diversity	6%	68%	**26%**

Table 3. User study. CapDiff generally achieves better fidelity and diversity.

6. Details for captions generation with LLMs

As mentioned in 4.1, we compare with baseline models by prompting them to generate caption suggestions. Since various input prompts can yield different results in these models, here we provide some comparisons of various prompts we tried to justify our current setting. We present the prompting results of these models in Figure 9. As can be seen, when prompted inappropriately, these visual-language models can generate invalid responses (BLIP-VQA, InstructBLIP) or over-complicated answers (LLAVA). While improvements in model responses can potentially be achieved by significant efforts in prompt engineering, it nevertheless causes inconveniences for the user compared to our proposed automatic prompt recommendation pipeline.

References

[1] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1, 2

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

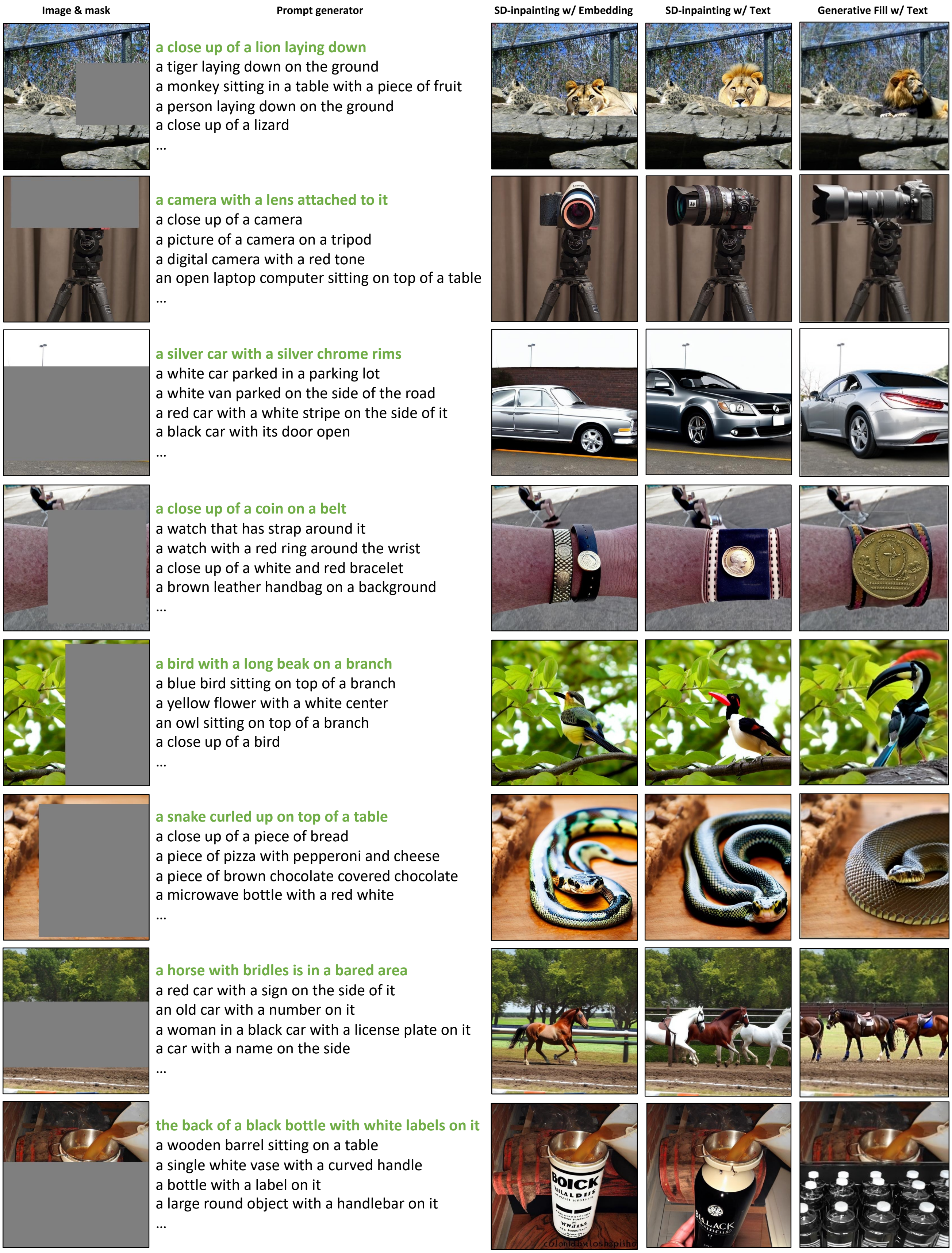


Figure 3. More visualizations of our CapDiff pipeline.

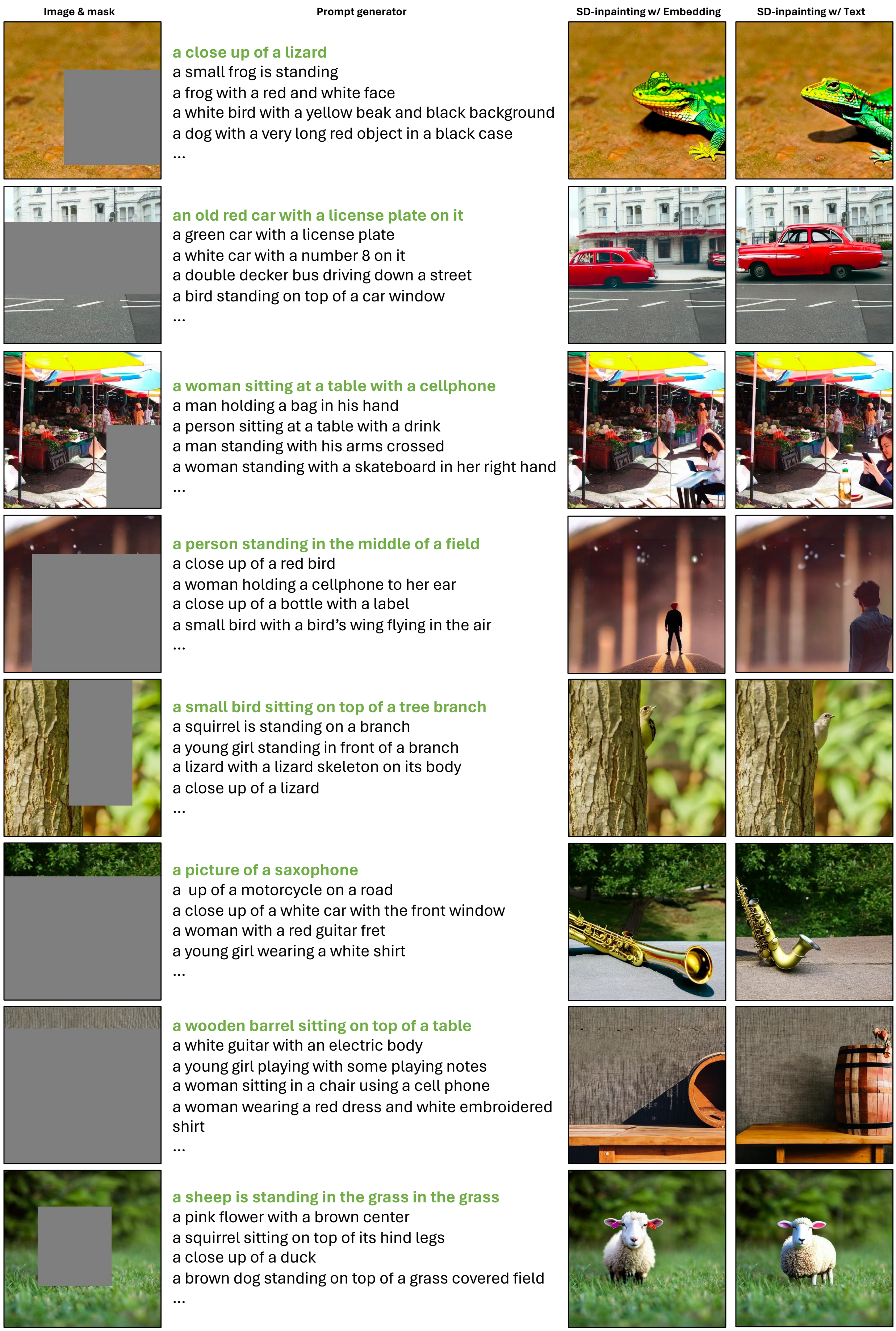


Figure 4. More visualizations of our CapDiff pipeline. (Cont.)

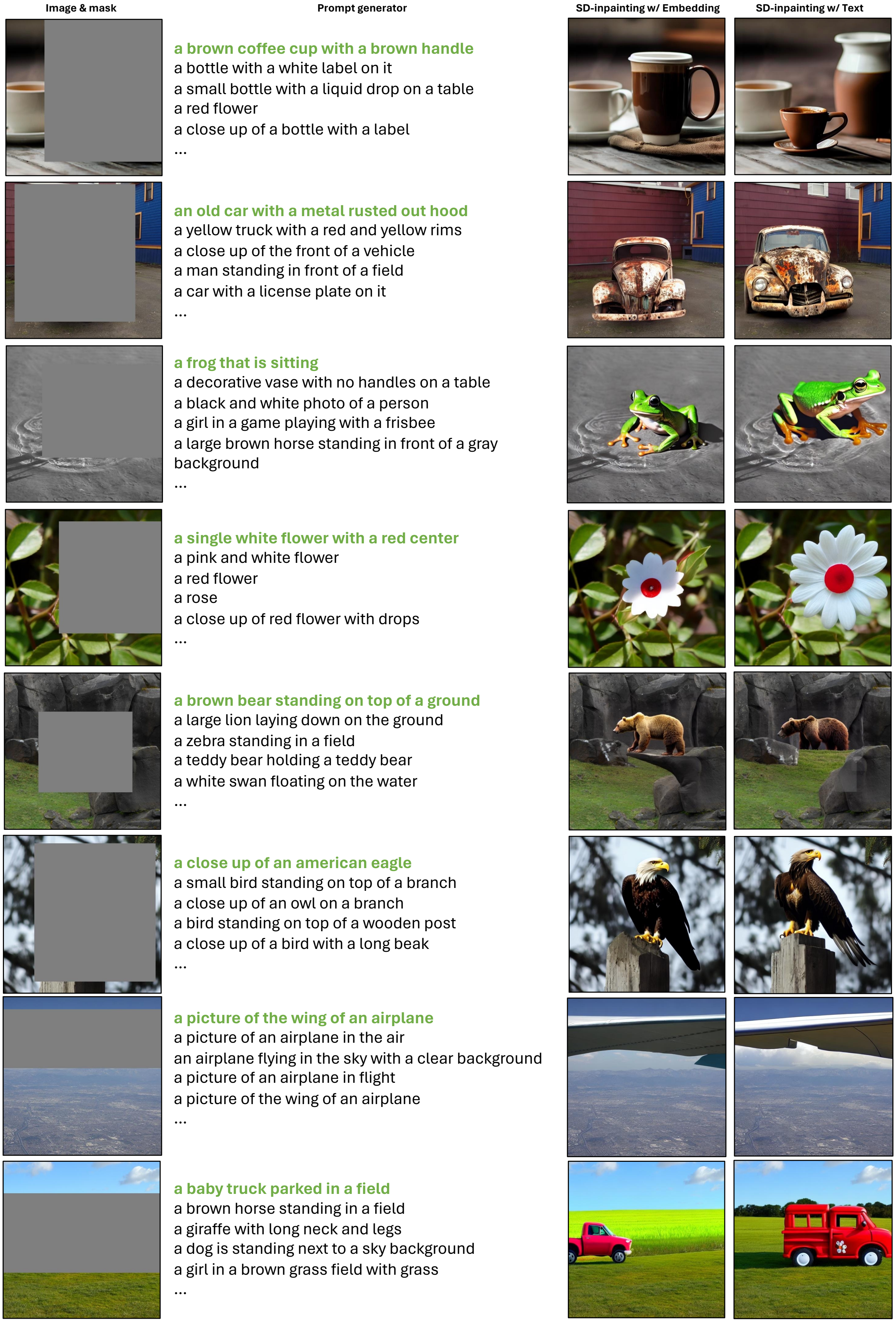


Figure 5. More visualizations of our CapDiff pipeline. (Cont.)

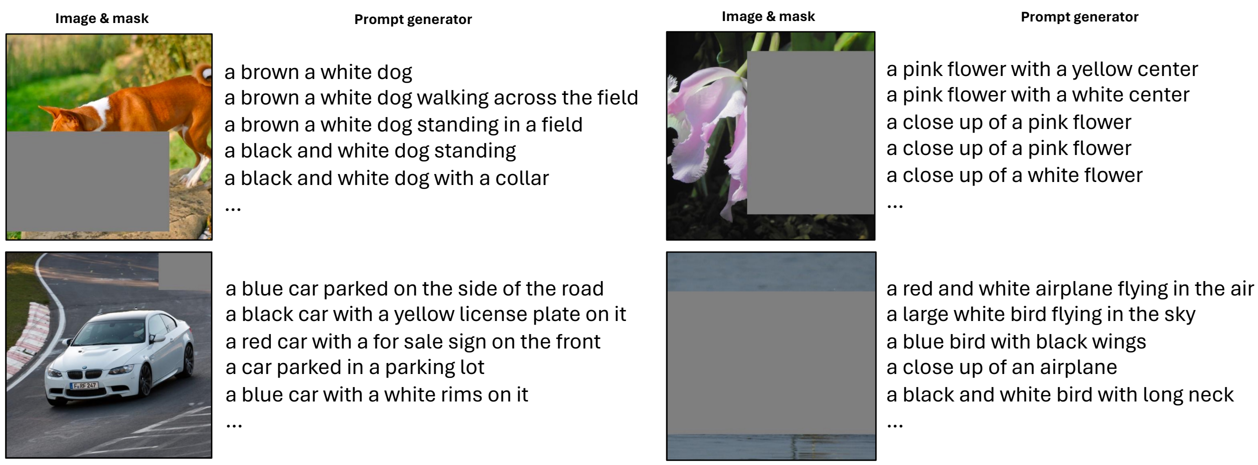


Figure 6. Failure cases of our model. When image context is too strong or the background is ambiguous, our model sometimes fails to to recommend sensible objects while maintaining diversity.

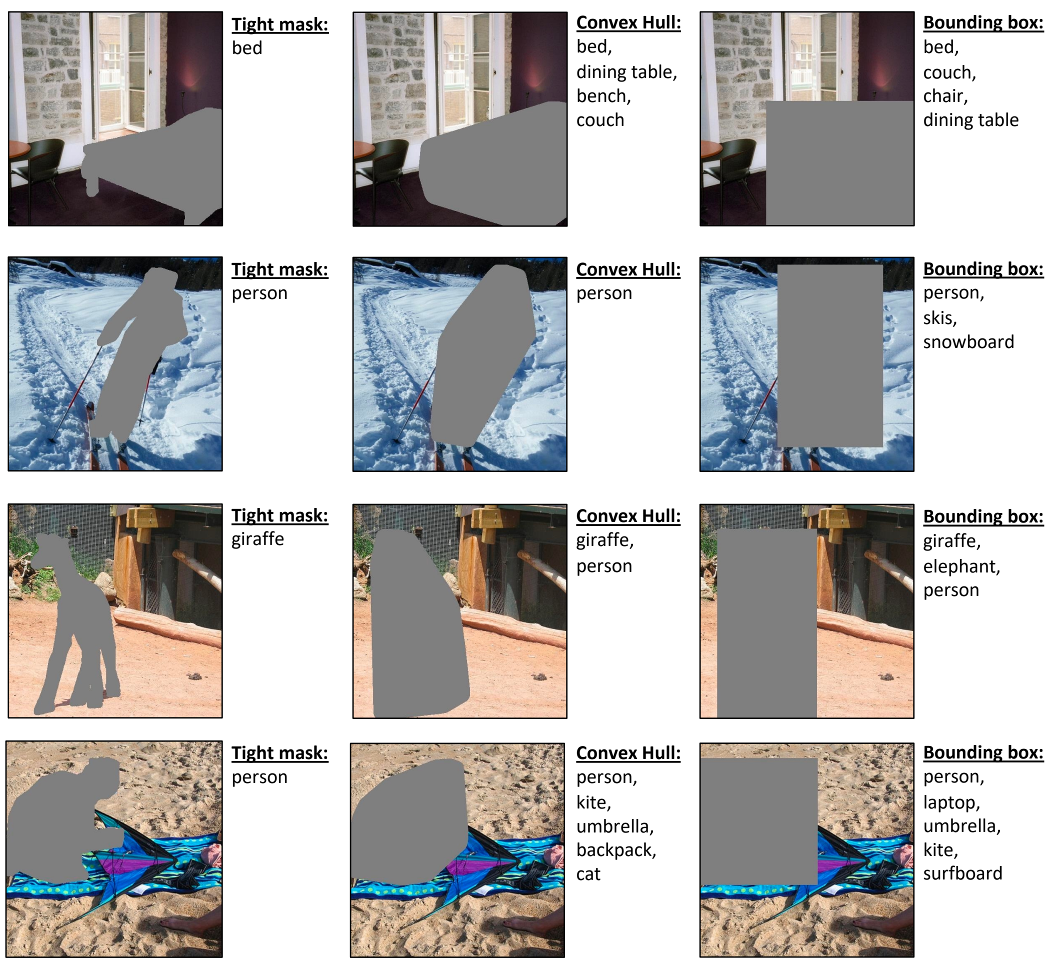
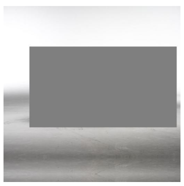


Figure 7. We demonstrate the significance of our shape awareness design with CatDiff on the COCO dataset. Augmenting mask shapes during training allows for precise control of diversity.

InstructBLIP
1. a man is smiling while wearing a gray hat
2. the boy is wearing a gray hat with his eyes closed
3. a man is smiling with his eyes closed while wearing a hat
4. the man is wearing a gray hat and has his eyes closed
5. the boy is smiling with his eyes closed

Ours
1. a cat is standing on its hind legs
2. a gray and white cat laying on the ground
3. a squirrel eating a piece of food in it's mouth
4. a close up of a fox
5. a cat that is laying down on the ground

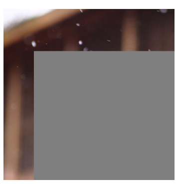


InstructBLIP
1. a gray square on a concrete floor with a white wall in the background
2. a gray square in the middle of a white room
3. an empty room with a large gray square on the floor
4. a gray background with an empty space in the middle
5. a large white room with a gray wall in the foreground

Ours
1. a white motorcycle with a red body
2. a blue car with a license on it
3. a pink car with a name tag on it
4. a silver car with silver rims
5. a black and brown car parked in a parking lot

LLAVA-5-Prompt
1. A cat sitting on a fence
2. A cat sitting on a tree bench
3. A cat sitting on a wooden log
4. A cat sitting on a rock
5. A cat sitting on a grassy area

Ours
1. a cat is standing on its hind legs
2. a gray and white cat laying on the ground
3. a squirrel eating a piece of food in it's mouth
4. a close up of a fox
5. a cat that is laying down on the ground

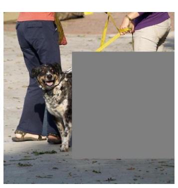


LLAVA-5-Prompt
1. A large wooden door
2. A window with a view of the rain
3. A rustic wooden fence
4. A barn with a slanted roof
5. A brick chimney on the roof

Ours
1. a close up of a brown dog with a brown background
2. a large brown horse with a blue saddle
3. a close up of a flower
4. a white dog with a collar around
5. a young girl holding a toothbrush in her right hand

Ours (w/o CapDiff)
1. a red and black train car is on a track
2. a red and black train traveling down train tracks
3. a red and blue train car sitting on top of a train track
4. a blue and white train traveling down train tracks
5. a red train car sitting next to a red fence

Ours
1. an old fashioned white train is on the tracks
2. a train car is sitting on the tracks
3. a train car compartment in the floor
4. a red and white train traveling down train tracks
5. a blue train car sitting on the track



Ours (w/o CapDiff)
1. a white dog standing on top of a white floor
2. a brown and white dog wearing a green shirt
3. a black dog wearing a necklace and a purple scarf
4. a brown dog standing on top of a wooden floor
5. a white dog with a brown collar looking up

Ours
1. a brown and white dog sitting on a white floor
2. a white and brown dog standing in a field
3. a black and white dog with a collar on
4. a black and white dog standing in a gray background
5. a close up of a dog with a red collar

Figure 8. Pairwise comparison between recent visual-language models and our method with and without CapDiff. Our proposed CapDiff yields much more desirable results while maintaining high diversity.

BLIP-VQA: What is a description of a reasonable object to be inserted in the gray area? (repeated 5 times)

1. elephant's trunk
2. elephant's trunk
3. elephant's trunk
4. elephant's trunk
5. elephant's trunk

BLIP-VQA: Write a text prompt that describes a reasonable object to be inserted in the gray area (repeated 5 times)

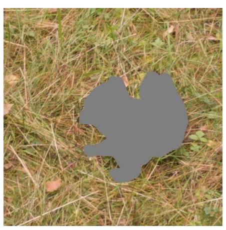
1. no idea
2. no idea
3. no idea
4. no idea
5. no idea

InstructBLIP: Write a text prompt that describes a reasonable object to be inserted in the gray area (repeated 5 times)

1. a silhouette of a dog in the grass
2. a dog is walking through a grassy field with a frisbee in its mouth
3. a dog is standing in the middle of a grassy field
4. a dog is standing in the grass with a frisbee in its mouth
5. a silhouette of a dog in the grass

InstructBLIP: Write five text prompts that describe reasonable objects to be inserted in the gray area

1. a dog's silhouette in the grass



LLAVA-Resample: Write a text prompt that describes a reasonable object to be inserted in the gray area (repeated 5 times)

1. A leaf is lying on the ground in the grass
2. A small, round, and black object, such as a leaf or a small stone, can be inserted into the gray area in the grassy field
3. A small, black, and white cat is sitting on the grass
4. A small, black, and white cat figurine
5. A leaf is lying on the ground in the grassy area

LLAVA: Write five detailed text prompts that describe reasonable objects to be inserted in the gray area

1. A small, shiny rock: The gray area in the image could be a rock that has been partially covered by the grass
2. It may be a leaf from a tree in the background or a leaf from a plant in the foreground
3. A small, black insect: The gray area might be a small insect, such as a grasshopper or a beetle, that is partially hidden by the grass
4. It may be a wildflower or a garden flower, depending on the surrounding environment
5. A small, green plant: The gray area might be a small, green plant that is partially hidden by the grass

LLAVA-5-Prompt: Write five text prompts that describe reasonable objects to be inserted in the gray area

1. A small bird perched on a leaf
2. A green leaf with a brown spot
3. A small insect crawling on a leaf
4. A small spider on a leaf
5. A leaf with a small hole in it

LLAVA: What is an object that can be inserted in the gray area? (repeated 5 times)

1. A small toy lizard, such as a stuffed animal or a plastic replica, can be inserted into the gray area where the lizard is depicted in the …
2. A small toy lizard or a plastic lizard figurine can be inserted in the gray area where the lizard is missing. This would create a more …
3. A small toy lizard or a plastic lizard figurine can be inserted in the gray area where the lizard is depicted in the image. This would …
4. A small toy lizard or a plastic lizard figurine can be inserted in the gray area where the lizard's body is missing. This would create …
5. A small toy lizard or a lizard figurine can be inserted into the gray area where the lizard's body is missing in the image. This would …

Figure 9. We show various responses from visual-language models we compared with different prompts. We find that inappropriate prompting leads to over-complicated or undesirable responses. Green indicates the model/setting is used for our main paper comparison, red indicates unused prompts.