Brush2Prompt: Contextual Prompt Generator for Object Inpainting

Supplementary Material

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

<table>
<thead>
<tr>
<th>Mask shape</th>
<th>Method</th>
<th>BLEU ↑</th>
<th>BERT-Score ↑</th>
<th>Self-BLEU ↓</th>
<th>Div-4 ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tight Mask</td>
<td>CapDiff</td>
<td>0.177</td>
<td>0.732</td>
<td>0.169</td>
<td>0.858</td>
</tr>
<tr>
<td></td>
<td>CatDiff-Prompt</td>
<td>0.155</td>
<td>0.732</td>
<td>0.210</td>
<td>0.817</td>
</tr>
<tr>
<td>Convex Hull</td>
<td>CapDiff</td>
<td>0.162</td>
<td>0.724</td>
<td>0.152</td>
<td>0.876</td>
</tr>
<tr>
<td></td>
<td>CatDiff-Prompt</td>
<td>0.117</td>
<td>0.719</td>
<td>0.184</td>
<td>0.844</td>
</tr>
<tr>
<td>Bounding box</td>
<td>CapDiff</td>
<td>0.149</td>
<td>0.715</td>
<td>0.141</td>
<td>0.885</td>
</tr>
<tr>
<td></td>
<td>CatDiff-Prompt</td>
<td>0.097</td>
<td>0.709</td>
<td>0.150</td>
<td>0.877</td>
</tr>
</tbody>
</table>

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.

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 $512 \times 512$, 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.
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.

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.

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

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

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

<table>
<thead>
<tr>
<th>Preference</th>
<th>Prefer LLaVA</th>
<th>Tie</th>
<th>Prefer CapDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fidelity</td>
<td>12%</td>
<td>58%</td>
<td>30%</td>
</tr>
<tr>
<td>Diversity</td>
<td>6%</td>
<td>68%</td>
<td>26%</td>
</tr>
</tbody>
</table>

References


Figure 3. More visualizations of our CapDiff pipeline.
a close up of a lizard
a small frog is standing
a frog with a red and white face
a white bird with a yellow beak and black background
a dog with a very long red object in a black case

an old red car with a license plate on it
a green car with a license plate
a white car with a number 8 on it
a double decker bus driving down a street
a bird standing on top of a car window

a woman sitting at a table with a cellphone
a man holding a bag in his hand
a person sitting at a table with a drink
a man standing with his arms crossed
a woman standing with a skateboard in her right hand

a person standing in the middle of a field
a close up of a red bird
a woman holding a cellphone to her ear
a close up of a bottle with a label
a small bird with a bird’s wing flying in the air

a small bird sitting on top of a tree branch
a squirrel is standing on a branch
a young girl standing in front of a branch
a lizard with a lizard skeleton on its body
a close up of a lizard

a picture of a saxophone
a up of a motorcycle on a road
a close up of a white car with the front window
a woman with a red guitar fret
a young girl wearing a white shirt

a wooden barrel sitting on top of a table
a white guitar with an electric body
a young girl playing with some playing notes
a woman sitting in a chair using a cell phone
a woman wearing a red dress and white embroidered shirt

a sheep is standing in the grass in the grass
a pink flower with a brown center
a squirrel sitting on top of its hind legs
a close up of a duck
a brown dog standing on top of a grass covered field

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 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.
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.
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.