Di3PO - Diptych Diffusion DPO for Targeted Improvements in Image Generation

Supplementary Material

A. Prompt Templates for Data Generation and Filtering

A.1. Diptych Generation Prompt

To generate the diptych images, we first use Gemini with a metaprompt instructing it to act as a graphic designer and generate a detailed background.

```
"""You are a graphic designer responsible
for high-quality text based graphics. You need to write a prompt for
a text to image model to generate backgrounds for text to be generated on.
The prompt needs to be as detailed as possible. Describe the background
on which you expect the text to appear. Include specific details of the
background, such as its color, shape, material, texture, or style.
Make sure your description is very clear and creative. Do not include
{right_input} or {misspelling} in the background description.

Then output an answer in the following format:

generated_background: the background for the image to render, make sure
not to include the word {right_input} or the word {misspelling} in the
background description.

Here is an example of a good answer:

generated_background: The font of the text is an elegant font, such as a
delicate script or a bold serif. The text should be the focal point of the
image, with a size and placement that commands attention. For the background,
create a scene of a serene, misty forest at dawn. The color palette
should be soft and ethereal, with pale blues, greens, and pinks blending
seamlessly. Sunlight filters through the mist, creating beams of light that
illuminate the scene. Dewdrops cling to leaves and spiderwebs, catching the
light and adding a touch of sparkle. The overall atmosphere should evoke a
sense of tranquility and wonder. The word should be rendered in a color that
complements the background, perhaps a warm gold or a shimmering silver.
It should appear as if it's part of the scene, perhaps nestled among the
leaves or rising with the mist. The text should have a subtle texture,
perhaps resembling dewdrops or etched metal, to further integrate it with the
background. This image should capture the essence of beauty in both its
visual and textual form, creating a captivating and memorable graphic."""
```

The output from Gemini (`generated_background`) is then inserted into the following T2I prompt template to create the final prompt sent to the image generation model.

```
"""Two images with a left and right panel, placed
side by side. Both images are fundamentally identical in terms of their
backdrop, lighting, and color palettes.
```

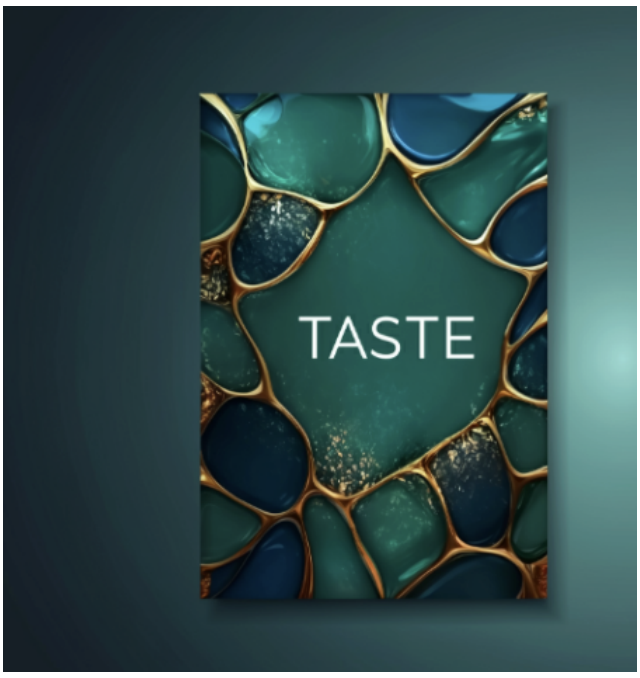
```
The left and right panel have this background.

{first_orientation} Image: Create an image with this background below.
On this image render the word {right_input}. The word {right_input} is
placed on the same background as the second image.

{second_orientation} Image: Create an identical image to the first image
with the exact same background. The word {misspelling} is placed on the
same background as the first image. It is extremely important to spell
the word as **{misspelling}**.

Background: {generated_background}
"""
```

A.2. Verification Prompt

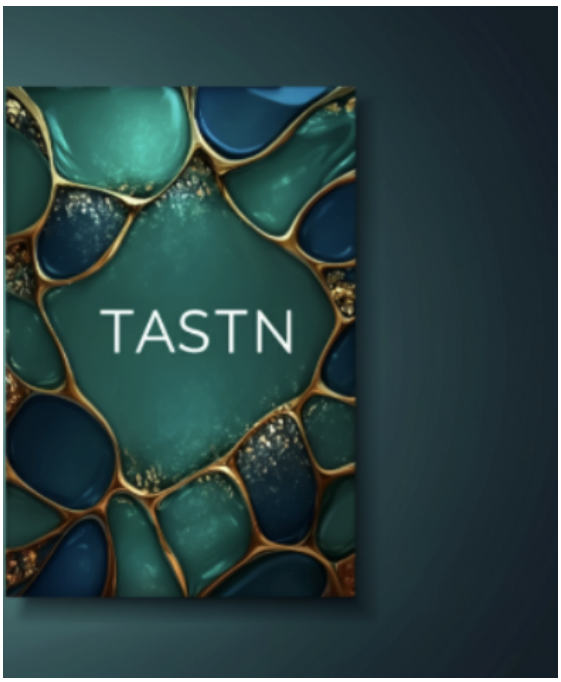
For the filtering stage, we use Gemini with the following prompt to verify the quality and correctness of the generated pairs.

```
"""You are a rigorous human rater for text on
image rendering graphics company. You are given two images, and you need
to verify that they are identical. The first image should be showcasing
text in the same background as the second image. You must check to make
sure that the background is identical, and that the text is rendered in
the same background as the second image. You must carefully attend to
even tiny details to make sure every single detail of the background,
such as color, shape, design, and style, is the same. You must also
check to make sure that the text in the first image is only slightly
different from the text in the second image. Both images should have
text in them. But, the text should not be the same in both images.

Then output an answer in the following format:

explanation: thought process and statement to justify your decision
passing: true or false indicating whether both checks are passed or not
confidence: a confidence score in your above decision of passing, out of 100

Some examples of the output given two images are as follows:
explanation: "The text on both the images is the different but there are
minor differences in the background. The background has slightly
different color."
passing: true
confidence: 80

explanation: "The text in the images are the same but the background is
different."
passing: false
confidence: 10

explanation: "The text in the second image is completely missing."
passing: false
confidence: 0

explanation: "The text in both the images is different. The backgrounds
are the same and the text in both images has been rendered clearly."
passing: true
```

Figure 7. The original Diptych generated by Imagen.



(a) Winning Image (x_w)



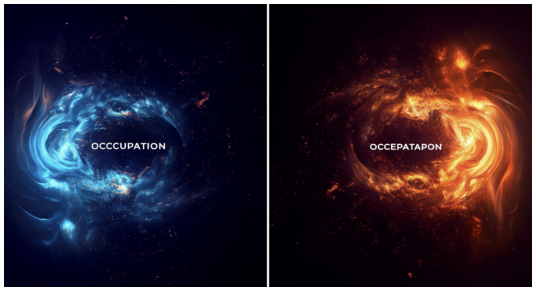
(b) Losing Image (x_l)

Figure 8. The split winning and losing images. The left image contains the correctly spelled word "TASTE", while the right contains the misspelling "TASTN".

```
explanation: "Both images have the same background. The text is slightly
different; the first image says \"TASTE\" while the second image says
\"TASTN\". The difference in text is minor, and the rendering of the
text on the same background is consistent."
passing: true
confidence: 100
```

Figure 9. Verifier response from a pair of passing images.

(a) Failure: No Text Rendered. The model failed to render any text in either panel. *Explanation: "Both images contain a picture of a cracked clay pot... There is no text present in either image."*



(b) Failure: Inconsistent Backgrounds. Although the style is similar, the colors and details of the backgrounds differ significantly. *Explanation: "The background, while sharing a similar style... has significantly different colors."*

(c) Failure: Identical Text. The model rendered the exact same text in both panels, failing to create a preference pair. *Explanation: "The text 'REIGN' is present in both... The text itself is identical in both images..."*

Figure 10. Examples of generated pairs that were correctly identified as failures by our automated verification pipeline.