Description Enhancement of Generated Images via Automatic Visual Question Generation and Answering

Aneesh Shetty
University of Texas at Austin
aneeshks@cs.utexas.edu

Mina Huh
University of Texas at Austin
minahuh@cs.utexas.edu

Abstract
Advances in text-to-image generation models let creators generate multiple high-fidelity images based on a text description (i.e. prompt). Yet, for people with visual impairments, it is difficult to assess the content and quality of the generated images and compare them to choose one. We propose a pipeline to generate rich description of AI generated images to assist broader users to understand them. In our pipeline, we use a large language model (GPT-4) to generate visual questions, vision-language models (BLIP-2) to extract answers, and a large language model (GPT-4) to summarize the results into final description. We evaluate the efficacy of our pipeline in comparison with a baseline image-captioning model and human describers. To further improve the visual grounding and accuracy of the answering pipeline, we experiment using foundation image segmentation model as an oracle to aid in visual question Answering.

1 Introduction
Large-scale text-to-image generation models, such as DALL-E (Ramesh et al., 2021), Stable Diffusion (Rombach et al., 2021), and Midjourney (mid, 2023a), present an opportunity for creators with visual impairments to generate images directly from text descriptions (i.e., prompts). However, current text-to-image generation tools are inaccessible to creators with visual impairments, as creators must visually inspect the content and quality of the generated images to iteratively refine their prompt and select from multiple generated candidate images.

In this project, we look at the task of generating enriched descriptions for these images, and summarizing their similarities and differences using a pipeline of Large language model (LLM) and Visual Question Answering (VQA) models. Using this downstream task as a pivot, we explore the efficacy of GPT-4 (OpenAI, 2023) to generate visual questions based on the image caption, and BLIP-2 (Hu et al., 2023) as a VQA system to generate highly descriptive and informative image descriptions for AI-generated images.

To evaluate our pipeline, we created two sets of baseline descriptions for 80 images generated using text-to-image model: descriptions created by humans and descriptions created using BLIP-2 image captioning model. Our evaluation study revealed that our proposed pipeline generates descriptions that have comparable coverage of visual information to the human generated descriptions. We also measured the accuracy of the visual information in our descriptions, which we report in the later sections.

From the evaluation we observed that common reason for inaccurate information in our description is due to hallucinations in VQA, which happen when the visual question asks about objects not present in the image, or when the VQA model attends to other similar objects than the target objects when generating answers. To tackle this challenge, we propose an updated pipeline that use a promptable segmentation model (Liu et al., 2023; Kirillov et al., 2023) and generate a masked image that only shows the target object of the visual question to reduce hallucinations in VQA. We evaluate the updated pipeline with respect to BLIP-2 model, performing a small study on its biases and hallucinations.

2 Background
As a background, we reviewed relevant work in Image Captioning, VQG & VQA, and image segmentation models.

2.1 Image Captioning
Improving the accessibility of image generation systems involves not only ensuring access to all features but also ensuring that the produced content is accessible. A primary method for making images more accessible is representing them as text descriptions, such as image captions or alt texts (e.g., “A person walking on the street”). Early work achieved this using crowd workers (Von Ahn and Dabbish, 2004; Bigham et al., 2010), while recent research has developed machine-learning-based systems that automatically generate the descriptions (Xu et al., 2015; Vinyals et al., 2015; Li et al., 2023).

Yet, conventional image captioning models tend to generate generic and concise captions or even identical captions when input images are similar to each
other (Dai et al., 2017; Dai and Lin, 2017; Mao et al., 2022). Recent works have explored Distinctive Image Captioning using CLIP guided group optimization (Zhang et al., 2023), compare with reference images in attribute/object-level and scene level (Mao et al., 2022), and measured semantic distance between the captions of similar images (Wang et al., 2020).

In our work, we use automatically generated visual questions and answers to create rich visual information of individual images, then use a LLM (GPT-4) to create a summary description that highlights the similarities and differences of the images. Recently, using VQA to generate captions has also been proposed by Zhu et al. (Zhu et al., 2023). Yet, our pipeline is uniquely designed for AI-generated images and generate visual questions that are based on the original text-prompt, image captions, as well as image prompt guidelines.

2.2 Visual Question Generation & Answering

Recent work in Visual Question Generation (VQG) propose models and metrics for good questions, like mutual information between image, generated question and answer category (Krishna et al., 2019), knowledge-aware question generation (Uehara and Harada, 2023), and generating single sub-question to answer a main question based on information gain (Uehara et al., 2022). For generating contextually relevant visual questions, recent works have leveraged image captions to generate visual questions (Changpinyo et al., 2022) and used multiple conversational interactions between the ChatGPT to generate visual questions (Zhu et al., 2023).

Visual Question Answering (VQA) has been a central research topic in vision-language tasks. Recent work utilizes Vision-language pre-trained models (Rafford et al., 2021; Li et al., 2021) and utilize them on various downstream tasks including VQA. For end-to-end Vision Language-Pretraining Different architectures have been proposed like encoder-decoder (Chen et al., 2022b), and unified transformer architectures (Li et al., 2022). End-to-end pre-training using large-scale image-text pairs can be computationally expensive. The other method, Modular vision-language pre-training methods leverage off-the-shelf pre-trained models and keep them frozen during VLP, such as freezing the image encoder (Zhai et al., 2022) or language model (Chen et al., 2022a). These methods present challenges in aligning visual features to the text space. We intend to utilize BLIP-2 (Li et al., 2023) within our pipeline as an answer generator, to improve information on the image.

2.3 Image Segmentation

Foundational segmentation model (e.g., SAM) (Kirillov et al., 2023) has opened research on many image grounding tasks and automatic dataset labelling. Particularly, many annotation and image grounding tasks can be tackled using Segmentation Models in conjunction with large open-set object detectors like GroundingDINO (Liu et al., 2023) as oracles. In this work, we aim to use this setup to segment the image using text prompting (Gro, 2023) and use it as an explicit signal to guide visual question answering, while also describing situations where it is applicable and where it can be detrimental.

3 Pipeline - Generating Descriptions

3.1 Prompt Verification

While the text-to-image model generates output images based on the prompt, the generated image often does not reflect the specifications in the prompt, especially if the prompt is long, complicated or ambiguous (Hu et al., 2023). To help users assess how well their generated images adhered to their prompt, our pipeline provides prompt verification.

To perform prompt verification, we first use GPT-4 (OpenAI, 2023) to generate visual questions that verify each part of the prompt. We input the prompt verification text instruction:

“Generate visual questions that verify whether each part of the prompt is correct. Number the questions.”

followed by the user’s prompt. GPT-4 outputs a series of questions as shown.

We use BLIP-2 model with ViT-G Flan-T5-XXL setup (Li et al., 2023) to generate answers to the visual prompt verification questions for each of the four generated candidate images.

For each generated image and prompt verification question, we instruct the BLIP-2 model with the starting sequence:

“Answer the given question. Don’t imagine any contents that are not in the image.”

to reduce hallucinations with non-existent information:

To help users quickly find which images do or do not adhere to the prompt, we use GPT-4 to summarize the responses to each question using the following prompt:

“Below are the answers of four similar images to one visual question. Write one sentence summary that captures the similarities and differences of these results. The summary should fit within 250 character limit.”

When using GPT-4’s chat completion API, we set the role of the system as:

“You are a helpful assistant that is describing images for people with visual impairment.”
3.2 Visual Content & Style Extraction

Generated image candidates often feature similarities or differences that are not present in the original prompt. For example, the prompt “A young chef is cooking dinner for his parents” does not specify the style such that the resulting images include three illustrations and one photo. To enable access to image content and style details that were not specified in the prompt, we extract the visual content and visual style of the generated image candidates. To surface content and style similarities and differences that are important for improving image generation prompts, we used text-to-image prompt guidelines (mid, 2023b,c; dal, 2023) to inform our approach.

We first created a list of visual questions about the image based on existing prompt guidelines, i.e. prompt guideline questions. The prompt guideline questions consist of questions about the content of the image (subjects, setting, objects), the purpose of the image (emotion, likely use), the style of the image (medium, lighting, perspective, color), and an additional question about errors in the image to surface distortions in the generated images such as blurring or unnatural human body features (Table 2). To answer our prompt guideline questions for each image, we answered 5 questions (setting, subjects, emotion, likely use, colors) using Visual Question Answering with BLIP-2, similar to our prompt verification approach:

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Question</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>Setting</td>
<td>What is the setting of the image?</td>
<td>BLIP-2</td>
</tr>
<tr>
<td>Subjects</td>
<td>What are the subjects of the image?</td>
<td>BLIP-2</td>
<td></td>
</tr>
<tr>
<td>Objects</td>
<td>What are the objects in this image?</td>
<td>Dectic</td>
<td></td>
</tr>
<tr>
<td>Emotion</td>
<td>What is the emotion of the image?</td>
<td>BLIP-2</td>
<td></td>
</tr>
<tr>
<td>Usage</td>
<td>Where would this image likely be used?</td>
<td>BLIP-2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Style</th>
<th>Medium &amp; Errors</th>
<th>What is the medium of the image?</th>
<th>CLIP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>What is the lighting in this image?</td>
<td>CLIP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>What is the perspective of this image?</td>
<td>CLIP</td>
<td></td>
</tr>
<tr>
<td>Colors</td>
<td>What are the main colors used in this image?</td>
<td>BLIP-2</td>
<td></td>
</tr>
<tr>
<td>Errors</td>
<td>What are the errors in this image?</td>
<td>CLIP</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Our prompt guideline questions including the question category, question name, and question, along with the model we used to answer the question (BLIP-2 (Li et al., 2023), CLIP (Radford et al., 2021), or Dectic (Zhou et al., 2022)).

The temperature value was set to 0.8. The summaries either indicate that all images have the same answer (e.g., “All images have a chef in the image”), or they alert users to differences:

- Prompt Verification Questions: Is there a chef in the image? Are the parents present in the image?
- Prompt Verification Summary: Three images depict a young kid, while Image 4 depicts a young man. Three images show parents present in the image, while Image 2 does not.

For our objects question, we used Dectic (Zhou et al., 2022), a state-of-the-art object detection model, with an open detection vocabulary and a confidence threshold of 0.3 to enable users to access all objects:

3.3 Description Summarization

To enable users to quickly assess their image results, we summarize the results from our pipeline to create a per image description for each image and a summary of image similarities and differences.

To generate per image descriptions, we first obtain the BLIP-2 caption for each image that provides a concise overview of the image content (e.g., “A family preparing food in the kitchen with a window.”). Then, we obtain additional detail about the image by generating questions about the BLIP-2 caption with GPT-4 with the prompt: “Given the caption, generate 10 visual questions that are likely to be asked by an audience..."
with visual impairments". Unlike the other questions in our pipeline that are common across all images, this step enables the pipeline a chance to ask image-specific questions to add detail (e.g., “What is the view outside the window?” is only asked for Image 4). We generate the answers to these questions using BLIP-2.

We create individual image descriptions by first aggregating all information acquired in our pipeline for each image including the prompt verification, prompt guideline, and caption-detail question-answer pairs for each image. Then, we guide GPT-4 with the aggregated visual information and the prompt:

“Below is the information of an image. Write a description of this image for the audience with visual impairment. Describe the medium first. Your response should fit within 250 character limit. Do not add additional information that was not provided. Do not describe parts that are not clear or cannot be determined from the given information.”

GPT-4 generates rich descriptions for each image (Figure 1).

![Image 1](https://via.placeholder.com/150)

In this stock photo, a young boy wears a chef’s hat as he stands in a modern kitchen. He is preparing a salad using a knife while ingredients are on the kitchen counter. The boy looks happy. The colors used are black, white, red and green. This image would likely be used in a cookbook to show children preparing healthy meals.

![Image 2](https://via.placeholder.com/150)

In this vector art image, a family is cooking together in a well-lit kitchen. There is a young boy chef with his mother, preparing food with pots, pans and a spoon on a gas stove. They are happy while cooking snacks for their family. The main colors used are blue and white. This image would fit in a children’s cooking class.

Figure 1: Per-image descriptions provided by our pipeline.

We generate the comparison description by similarly providing all of the information extracted from our pipeline to GPT-4 with the prompt:

“Below is the information for four images. Write one paragraph about the similarities between the four images and one paragraph about the differences between the four images. The summary should be concise.”.

GPT-4 briefly summarizes the image similarities and differences (Figure 2). To help users quickly assess whether to revise their prompt or continue exploring, we keep both the comparison description and per-image description.

![Comparison Images](https://via.placeholder.com/150)

Figure 2: Image comparison descriptions.

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Question</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>Setting</td>
<td>What is the setting of the image?</td>
<td>BLIP-2</td>
</tr>
<tr>
<td>Subjects</td>
<td>What are the subjects of the image?</td>
<td>BLIP-2</td>
<td></td>
</tr>
<tr>
<td>Objects</td>
<td>What are the objects in this image?</td>
<td>DecTic</td>
<td></td>
</tr>
<tr>
<td>Emotion</td>
<td>What is the emotion of the image?</td>
<td>BLIP-2</td>
<td></td>
</tr>
<tr>
<td>Usage</td>
<td>Where would this image likely be used?</td>
<td>BLIP-2</td>
<td></td>
</tr>
<tr>
<td>Style &amp; Errors</td>
<td>Medium</td>
<td>What is the medium of the image?</td>
<td>CLIP</td>
</tr>
<tr>
<td>Dies</td>
<td>Lighting</td>
<td>What is the lighting in this image?</td>
<td>CLIP</td>
</tr>
<tr>
<td>Perspective</td>
<td>What is the perspective of this image?</td>
<td>CLIP</td>
<td></td>
</tr>
<tr>
<td>Colors</td>
<td>What are the main colors used in this image?</td>
<td>BLIP-2</td>
<td></td>
</tr>
<tr>
<td>Errors</td>
<td>What are the errors in this image?</td>
<td>CLIP</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Prompt guideline questions including the question category, question name, and question, along with the model used to answer the question (BLIP-2 (Li et al., 2023), CLIP (Radford et al., 2021), or DecTic (Zhou et al., 2022)).

4 Evaluation

We measured the coverage of the descriptions generated by the pipeline and the accuracy of the information presented in pipeline’s tables. We compare the coverage of pipeline-generated caption with the human-generated caption and the caption generated by a state-of-the-art image captioning model BLIP-2 (Li et al., 2023).

**Method.** We selected 20 image sets (20 prompts x 4 generated images for each prompt = 80 total images) from Midjourney’s community feed spanning different prompt lengths, content types, and styles. We recruited two people with experience describing images to provide descriptions for 10 randomly selected image sets each. For each image set, the describers provided descriptions of each individual image, and the similarities and differences between the images. We provided describers with prompt guidelines (mid, 2023b), image description guidelines (ima, 2023), an example set of descriptions created by pipeline, and the prompt for each image set to inform their descriptions. Both describers spent 3.5 hours to create descriptions for the 10 sets of images — or around 21 minutes per image set.

We compared the coverage of pipeline-generated descriptions to those generated by a baseline captioning tool (BLIP-2) and human describers. For comparison, we annotated the similarities and differences descriptions for all 20 sets of images and annotated the individual descriptions for 10 sets of images. We chose the 10 sets with the longest human descriptions to compare pipeline with the highest quality descriptions. Because BLIP-2 cannot take multiple images as input to extract similarities and differences, we generated captions of the 4 images using BLIP-2, then prompted GPT-4 with the same prompt we used in our system to generate summary descriptions:

“Below is the information of four images. Write one paragraph about the similarities of the images and one paragraph about the differences between the four images.”

We tallied whether the descriptions contained details about the image in each of our set of pre-defined vi-
Figure 3: Two image sets and the descriptions of the similarities and differences used in the pipeline coverage evaluation (each image set described by a different human describer).

<table>
<thead>
<tr>
<th>Category</th>
<th>Sub-category</th>
<th>Accuracy (%)</th>
<th>Total (#)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prompt-verification</td>
<td>Setting</td>
<td>97.53</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Subjects</td>
<td>98.60</td>
<td>143</td>
</tr>
<tr>
<td></td>
<td>Objects</td>
<td>82.86</td>
<td>1243</td>
</tr>
<tr>
<td></td>
<td>Emotion</td>
<td>87.5</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Usage</td>
<td>97.50</td>
<td>80</td>
</tr>
<tr>
<td>Style</td>
<td>Medium</td>
<td>82.76</td>
<td>174</td>
</tr>
<tr>
<td></td>
<td>Lighting</td>
<td>94.33</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td>Perspective</td>
<td>71.83</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>Colors</td>
<td>99.1</td>
<td>221</td>
</tr>
<tr>
<td>Errors</td>
<td>Errors</td>
<td>60.00</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3: Results of the pipeline on 20 sets of images.
We see some evidence of these errors in using BLIP-VQA models, constructing a few examples for each bias towards Language over Visual Information.

Table 4: Comparison of the coverage of pipeline-generated descriptions to those generated by a baseline captioning e.g. human baseline pipeline. The pipeline consistently captured more similarities and differences than the human describers.

<table>
<thead>
<tr>
<th>Description</th>
<th>Total Content (#)</th>
<th>Total Style (#)</th>
<th>Total Error (#)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Similarities</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>μ</strong></td>
<td>1.25</td>
<td>0.90</td>
<td>0.35</td>
</tr>
<tr>
<td><strong>σ</strong></td>
<td>0.90</td>
<td>0.70</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>Differences</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>μ</strong></td>
<td>1.35</td>
<td>0.85</td>
<td>0.50</td>
</tr>
<tr>
<td><strong>σ</strong></td>
<td>0.80</td>
<td>0.50</td>
<td>0.30</td>
</tr>
<tr>
<td><strong>Per Image</strong></td>
<td><strong>μ</strong></td>
<td>1.71</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>σ</strong></td>
<td>0.39</td>
<td>0.10</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Attending to incorrect information in the Image.

Another set of errors encountered from the VQA model are when it attends to the wrong object while answering questions. This could occur due to various reasons, some of which we noted as follows:

- Since text-to-image generated images are stochastic and noisy by nature, objects similar to one another in vision (like man/woman, cats/dogs) are mis-attended while answering questions.
  
  This is more of an issue with images generated to look realistic compared to stylized or cartoon images, since those are the ones which have more distortions that could make objects look visually similar.

- Questions about visual properties are often answered using the dominant object/feature of the image, instead of focusing on the relevant objects. Dominant can be in various aspects, size of the image, common or brighter colors in the image (e.g., see Figure 5). This can hinder the performance on questions with more targeted requirements.

5 Reducing Hallucinations from VQA Model

Even with the starting sequence prompt to BLIP-2 to not hallucinate information which is not-existent in the image, we observed that the VQA model still suffers hallucinations. As an extension, we investigate the different types of hallucinations that we observe in the VQA models, constructing a few examples for each from the collected dataset, and conduct a small ablation study on VQA models and evaluations on methods to improve inference on them by modification of the pipeline. The complete pipeline in shown in Figure 5.

Bias towards Language over Visual Information.

There has been past research in reducing unimodal bias in VQA models (Cadène et al., 2019). Particularly, models utilize shortcuts and answer the question based on textual hints without being grounded in the image. We see some evidence of these errors in using BLIP-2, particularly when the object or context in question likely has a large correlation in answers in standard text (e.g., see Figure 4, where Santa Hats are usually red in color). This can mainly be an artifact of BLIP-2 and other large VQA models being trained on general knowledge data acquired from the internet, and thus, lacks in visual reasoning in counter-intuitive or surprising setting like when images are generated from text2image models, whose outputs may not conform to standard norms.

Attending to incorrect information in the Image.

Another set of errors encountered from the VQA model are when it attends to the wrong object while answering questions. This could occur due to various reasons, some of which we noted as follows:

- Since text-to-image generated images are stochastic and noisy by nature, objects similar to one another in vision (like man/woman, cats/dogs) are mis-attended while answering questions.

  This is more of an issue with images generated to look realistic compared to stylized or cartoon images, since those are the ones which have more distortions that could make objects look visually similar.

- Questions about visual properties are often answered using the dominant object/feature of the image, instead of focusing on the relevant objects. Dominant can be in various aspects, size of the image, common or brighter colors in the image (e.g., see Figure 5). This can hinder the performance on questions with more targeted requirements.

![Figure 4: Hallucination from question text](image)

![Figure 5: Hallucination from dominant aspects and incorrect attention](image)

5.1 Augmenting a Segmentation Model

To incorporate additional visual supervision from the given question, we experiment with augmenting recent Foundational Models for Object Segmentation (Kirillov et al., 2023) to the original structure of our pipeline. The complete pipeline in shown in Figure 6. Similar to before, we use Detic to detect a list of objects in the image, and pass the image caption and the list of objects to GPT-4 to generate questions using the following prompt:
We evaluate the pipeline on Question Answering with Visual Question Answering Task by extracting relevant was capable of producing these regions of interest in previously used. This provides explicit supervision for the relevant portions. But we quickly figured out that GPT-4 and the masked images. We checked the improvement answers to these questions for both the original image ages using GPT-4. Then, using BLIP-2 we generated Section 20 images (one image randomly selected per set of four regions as a preprocessing step. For each image-question pair, we first use GroundingDINO (Liu et al., 2023), an Open Set Object Detector, to detect bounding boxes over the regions of interest. Next we use Segment Anything over these bounding boxes and mask out all the other information to get a masked image. We use this masked image and question pair and query BLIP-2 in the same setting as previously used. This provides explicit supervision for the Visual Question Answering Task by extracting relevant regions as a preprocessing step.

5.2 Evaluation of Segmentation

We evaluate the pipeline on Question Answering with 20 images (one image randomly selected per set of four images) from our previous dataset. Using the prompt in Section 5.1, we generated 15 questions each for 20 images using GPT-4. Then, using BLIP-2 we generated answers to these questions for both the original image and the masked images. We checked the improvement on the VQA of the visual questions that were first incorrectly answered with the original image. Table 5 summarizes the result. We selected the questions according to their categories based on the types of hallucinations discussed. Here, we report how much the masking approach improved VQA by having correct answers. We also check the agreement between the answers before and after masking the image in these categories for the questions answered correctly (number of questions correctly answered by original image VQA).

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Improvement (#/#)</th>
<th>Agreement (#/#)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect Attention</td>
<td>10 / 17</td>
<td>13 / 15</td>
</tr>
<tr>
<td>Bias Towards Question</td>
<td>7 / 15</td>
<td>11 / 12</td>
</tr>
</tbody>
</table>

Table 5: Shows Improvement and Agreement for Baseline vs Segmentation Augmented QA

Using the segmentation model as an oracle helped our pipeline to have more correct answers to different types of visual questions. We believe this is due to the fact that most questions asked were single-object questions, and thus a naive localization was an effective strategy (See Figure 7).

The pipeline helps in mitigating the bias towards questions in some cases (Figure 8) but not extremely efficient. Particularly, those questions are now correctly answered where the bias can be detected using just local visual information (color, pattern, etc) of one object. More nuanced biases cannot be handled just by denoising the image, and we believe that can only be tackled at a pre-training or fine-tuning stage.
was evaluated with respect to BLIP-2 model, which provided insights into biases and hallucinations in VQA models to generate enriched image descriptions. To address this challenge, we proposed an updated pipeline that uses a promptable segmentation model to reduce hallucinations in VQA. The updated pipeline was evaluated with respect to BLIP-2 model, which provided insights into biases and hallucinations in VQA models.

Overall, our study highlights the potential of using large language models and VQA models in generating enriched image descriptions and the importance of addressing hallucinations in VQA to improve the accuracy of the generated descriptions.

6 Conclusions

In conclusion, we proposed a pipeline of LLM and VQA models to generate enriched image descriptions and summarize similarities and differences between images. Our study evaluated the efficacy of GPT-4 and BLIP-2 models in generating visual questions and informative image descriptions, respectively. The evaluation demonstrated that our proposed pipeline generates descriptions that have a much better coverage of visual information to human-generated descriptions.

We also identify inaccuracies in information in our descriptions arising due to hallucinations in VQA. To address this challenge, we proposed an updated pipeline that uses a promptable segmentation model to reduce hallucinations in VQA. The updated pipeline was evaluated with respect to BLIP-2 model, which provided insights into biases and hallucinations in VQA models.

Overall, our study highlights the potential of using large language models and VQA models in generating enriched image descriptions and the importance of addressing hallucinations in VQA to improve the accuracy of the generated descriptions.

References

2023a. Midjourney.
2023b. Midjourney propmt guidelines.
2023c. Midjourney styles and keywords.
2023. Stable diffusion negative prompts.


