# 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 guidede 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 (Radford 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

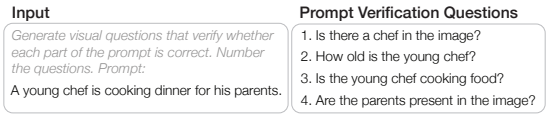
### 3.1 Prompt Verification

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. Dont 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."*

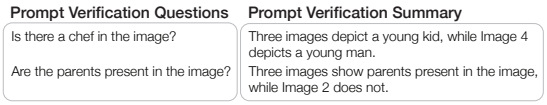
When using GPT-4's chat completion API, we set the role of the system as:

*"You are a helpful assistant that is describing images for people with visual impairment."*

| Category | Name | Question | Model |
|---|---|---|---|
| Content | Setting | What is the setting of the image? | BLIP-2 |
| | Subjects | What are the subjects of the image? | BLIP-2 |
| | Objects | What are the objects in this image? | Dectic |
| | Emotion | What is the emotion of the image? | BLIP-2 |
| | Usage | Where would this image likely be used? | BLIP-2 |
| Style & Errors | Medium | What is the medium of the image? | CLIP |
| | Lighting | What is the lighting in this image? | CLIP |
| | Perspective | What is the perspective of this image? | CLIP |
| | Colors | What are the main colors used in this image? | BLIP-2 |
| | Errors | What are the errors in this image? | CLIP |

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)).

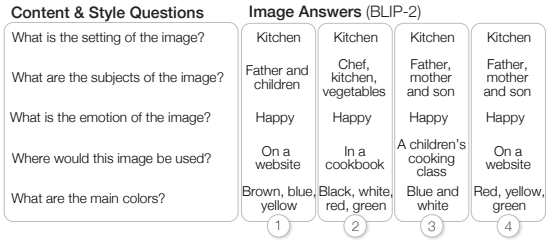
The temperature value was set to 0.8. The summaries either indicate that all images have the same answer (*e.g.*, *"All images have a chef in the image"*), or they alert users to differences:



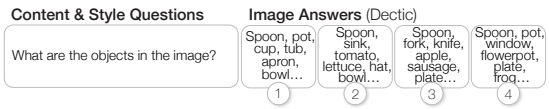

### 3.2 Visual Content & Style Extraction

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:

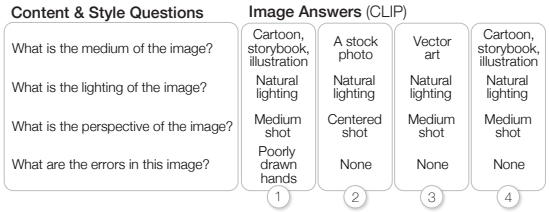



For our objects question, we used Dectic (Zhou et al., 2022), a state-of-the-art object detection model, with an open detection vocabulary and a confidence threshold of 0.3 to enable users to access all objects:




We used an *open-ended vocabulary* set to detect all objects, rather than only limiting the vocabulary set to objects mentioned in the prompt, to enable users to access additional objects that the text-to-image model added during generation.

For the remaining questions covering medium, lighting, perspective, and errors, we answer the question for each image candidate by using CLIP (Radford et al., 2021) to determine the similarity between the image and a limited set of answer choices (similar to CLIP interrogator (int, 2023)). To provide answers that could inform future prompts, we curated our answer choices for medium, lighting, and perspective from Midjourney's list of styles (mid, 2023c) and DALL-E's prompt book (dal, 2023). To address common image generation errors, we retrieved the answer choices for our errors question from prior work (Reddy et al., 2021; sta, 2023). For each question, our pipeline presents the top three answer choices with a similarity score between the answer choice embedding and the image embedding above a threshold of 0.18:




### 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).

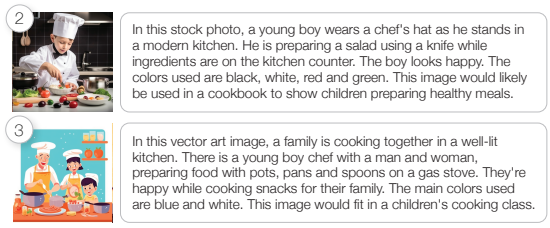



Figure 1: Per-image descriptions provided by our pipeline.

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**.


Figure 2: Image comparison descriptions.

| Category | Name | Question | Model |
|---|---|---|---|
| Content | Setting | What is the setting of the image? | BLIP-2 |
| | Subjects | What are the subjects of the image? | BLIP-2 |
| | Objects | What are the objects in this image? | Dectic |
| | Emotion | What is the emotion of the image? | BLIP-2 |
| | Usage | Where would this image likely be used? | BLIP-2 |
| Style & Errors | Medium | What is the medium of the image? | CLIP |
| | Lighting | What is the lighting in this image? | CLIP |
| | Perspective | What is the perspective of this image? | CLIP |
| | Colors | What are the main colors used in this image? | BLIP-2 |
| | Errors | What are the errors in this image? | CLIP |

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-

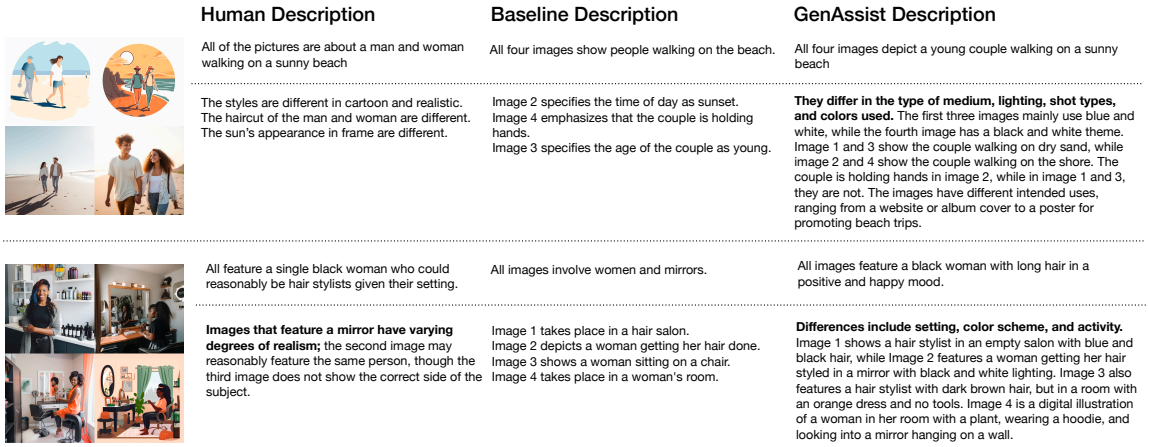



Figure 3: Two image sets and the descriptions of the similarities and differences used in the pipeline coverage evaluation (each image set described by a different human describer).

| Category | Sub-category | Accuracy (%) | Total (#) |
|---|---|---|---|
| Prompt-verification | | 92.82 | 418 |
| Content | Setting | 97.53 | 81 |
| | Subjects | 98.60 | 143 |
| | Objects | 82.86 | 1243 |
| | Emotion | 87.5 | 80 |
| | Usage | 97.50 | 80 |
| Style | Medium | 82.76 | 174 |
| | Lighting | 94.33 | 141 |
| | Perspective | 71.83 | 142 |
| | Colors | 99.1 | 221 |
| Errors | Errors | 60.00 | 5 |

Table 3: Results of the pipeline on 20 sets of images.

sual information categories (Table 2). We counted only the correct information in the descriptions. One of the researchers annotated the descriptions and the other researcher reviewed the annotations. To compute the accuracy of the detailed visual information in pipeline, one of the researchers examined the 20 sets of images with the three tables generated by the pipeline (prompt-verification table, visual content table, and visual style table) and counted the number of correct and incorrect answers in each table.

***Results.***

#### 4.0.1 Coverage

We summarize our coverage evaluation results in Table 4. Overall, pipeline's summary descriptions covered more types of information (both in the similarities and differences) than the human describers. The coverage of pipeline's individual image descriptions were comparable to human describers. In the coverage of differences, we can see that pipeline spotted more than twice the number of total differences than the human describes (4.55 vs. 2.25). pipeline captured more information about the content and styles than the human-generated descriptions. Yet, we can see that human describers outperformed pipeline in providing the error information in the description. For instance, one human describer specified in the similarities description *"...All of the images have some AI generation error with fingers or clothing. "*. While pipeline and the baseline used the same GPT-4 prompt to extract the similarities and differences, we can see that the baseline's descriptions of the differences did not summarize the information while those generated by human and pipeline did (Figure 3). Because the baseline did not capture a lot of visual information with the BLIP-2 caption, the descriptions of the differences often repeated the original caption without describing the trend.

#### 4.0.2 Accuracy

Table 3 summarizes the results of the accuracy evaluation. Prompt verification, content, and style categories all achieved over 90% accuracy except for medium, perspective and emotion. The pipeline's prompt verification and content information extracted using BLIP-2 (setting, subject, emotion, usage) all achieved high accuracy. (from 87.5%-98.6%). (92.82%) and content information (setting: 97.53%, subjects: 98.60%) extracted using BLIP-2, showed high accuracy. In the purpose category, we show that emotion accuracy is 87.5% and usage is high at 97.5%. In the style categories, the accuracy of the color and lighting information was high, but medium and perspective accuracy was lower. In the 80 images in the dataset, pipeline only detected five images as having errors, and detected the correct error types in three of them. The most common errors made in our pipeline were from perspective, medium, and error categories which are all extracted using the CLIP score. For perspective and medium, the majority of the errors were due to CLIP matching images to common style expressions (*e.g.,* natural lighting, centered-shot) which likely reflects prevalence of these expressions in the training data. In the incorrect output of errors, pipeline detected cartoon or sketch images as *'poorly drawn faces'* errors. One reason for the relatively low accuracy of object detection results is that we empirically set the output threshold of pipeline's object detec-

| (Correct Only) | | Total Content (#) | | | Total Style (#) | | | Total Error (#) | | | Total (#) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Human | Baseline | pipeline | Human | Baseline | pipeline | Human | Baseline | pipeline | Human | Baseline | pipeline |
| Similarities | $\mu$ | 1.5 | 1.65 | **2.45** | 0.70 | 0.00 | **0.80** | **0.10** | 0.00 | 0.00 | 2.35 | 1.65 | **3.25** |
| | $\sigma$ | 0.61 | 0.59 | 1.10 | 0.80 | 0.00 | 0.83 | 0.31 | 0.00 | 0.00 | 0.83 | 0.85 | 1.29 |
| Differences | $\mu$ | 1.50 | 1.95 | **2.35** | 0.65 | 0.35 | **2.20** | **0.05** | 0.00 | 0.00 | 2.25 | 2.30 | **4.55** |
| | $\sigma$ | 0.69 | 0.39 | 0.49 | 0.75 | 0.49 | 1.01 | 0.22 | 0.00 | 0.00 | 0.84 | 0.93 | 1.26 |
| Per Image Descriptions | $\mu$ | **1.71** | 0.69 | **1.71** | **0.71** | 0.04 | 0.68 | **0.05** | 0.00 | 0.01 | **2.47** | 0.73 | 2.41 |
| | $\sigma$ | 0.39 | 0.10 | 0.26 | 0.22 | 0.07 | 0.30 | 0.05 | 0.00 | 0.03 | 0.74 | 0.33 | 0.75 |

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

tion (Detic) as 0.3 to present diverse objects to users in addition to information about the main subject extracted by BLIP-2 in our pipeline.

## 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 image and/or the question. Due to the project timeline and the small number of incorrect/hallucinating examples we observed in our collected dataset, the improvements we propose are tested over a small set.

***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.

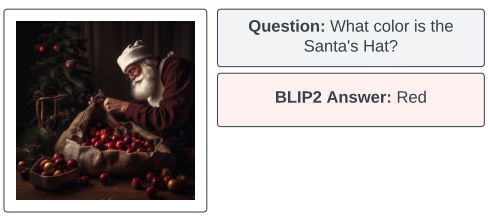



Figure 4: Hallucination from question text

***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.

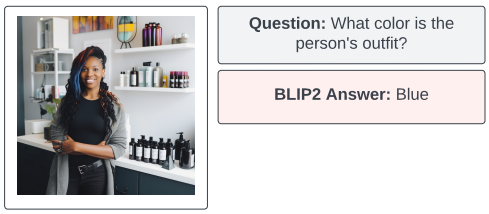



Figure 5: Hallucination from dominant aspects and incorrect attention

### 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:

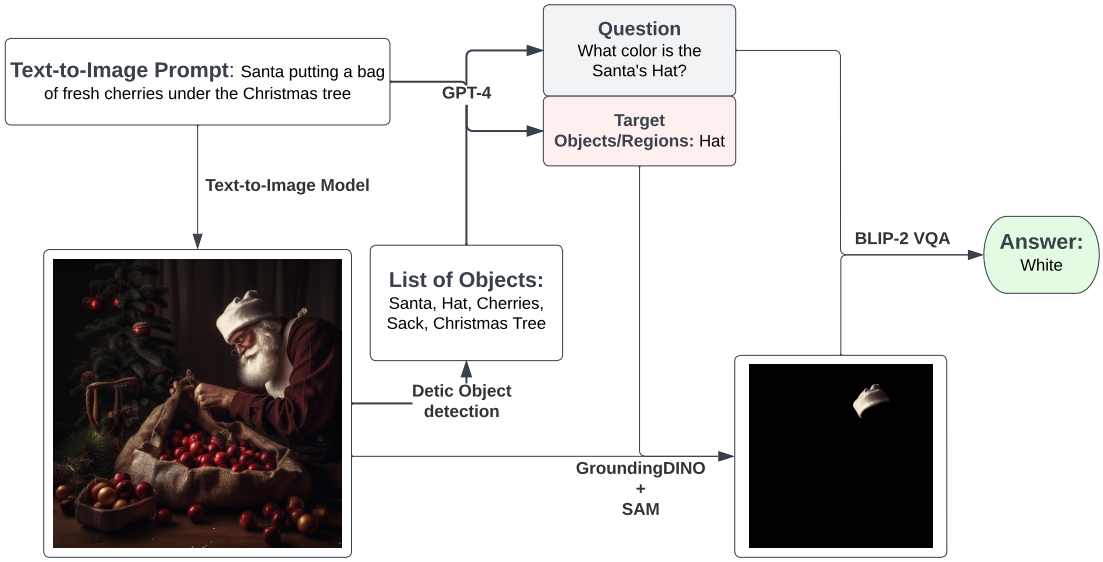



Figure 6: Natural language segmentation augmented question answering

*" You are a helpful assistant who will be describing images for people with visual impairment. You will be given a caption and a list of objects that are possibly present in an image. Generate a set of 15 visual questions about the image with short answers. For each visual question, if possible give a list of target objects or regions in the image that should be focused on while answering that question. "*

An initial idea to get the regions of interest was to syntactically parse the question and the prompt to get nouns, objects and other features, and output the relevant portions. But we quickly figured out that GPT-4 was capable of producing these regions of interest in a zero-shot fashion just based on the question it generates.

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 correct after segmentation out of total questions correctly answered by original image VQA).

| Error Type | Improvement (#/#) | Agreement (#/#) |
|---|---|---|
| Incorrect Attention | 10 / 17 | 13 / 15 |
| Bias Towards Question | 7 / 15 | 11 / 12 |

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.

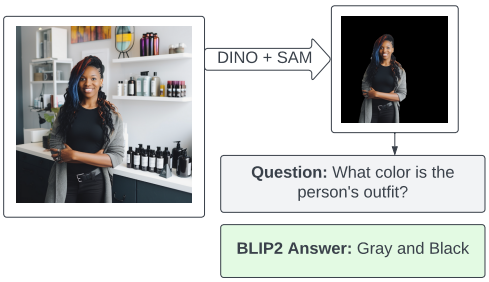



Figure 7: Correcting hallucinations from incorrect attention


Figure 8: Correcting hallucinations from question bias

## 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

2023. Clip interrogator.

2023. Dall-e2 propmt book.

2023. Grounded segment anything.

2023. Guide to image descriptions.

2023a. Midjourney.

2023b. Midjourney propmt guidelines.

2023c. Midjourney styles and keywords.

2023. Stable diffusion negative prompts.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Rémi Cadène, Corentin Dancette, Hédi Ben-Younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases in visual question answering. *CoRR*, abs/1906.10169.

Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. All you may need for vqa are image captions. *arXiv preprint arXiv:2205.01883*.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022a. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022b. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE international conference on computer vision*, pages 2970–2979.

Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. *Advances in Neural Information Processing Systems*, 30.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2008–2018.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection.

Yangjun Mao, Long Chen, Zhihong Jiang, Dong Zhang, Zhimeng Zhang, Jian Shao, and Jun Xiao. 2022. Rethinking the reference-based distinctive image captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4374–4384.

OpenAI. 2023. Gpt-4 technical report.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Mr D Murahari Reddy, Mr Sk Masthan Basha, Mr M Chinnaiahgari Hari, and Mr N Penchalaiah. 2021. Dall-e: Creating images from text. *UGC Care Group I Journal*, 8(14):71–75.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.

Kohei Uehara, Nan Duan, and Tatsuya Harada. 2022. Learning to ask informative sub-questions for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4681–4690.

Kohei Uehara and Tatsuya Harada. 2023. K-vqg: Knowledge-aware visual question generation for commonsense acquisition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4401–4409.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. 2020. Compare and reweight: Distinctive image captioning using similar images sets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 370–386. Springer.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133.

Youyuan Zhang, Jiuniu Wang, Hao Wu, and Wenjia Xu. 2023. Distinctive image captioning via clip guided group optimization. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 223–238. Springer.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. In *ECCV*.

Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. 2023. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*.