

# K-VQG: A Study on Knowledge-based Visual Question Generation

Ammar Hatiya

April 26, 2024

# Abstract

Extracting information from images is an increasingly prominent task due to its ever-expanding utility and applications. Generating knowledge-based questions (and answers) from images is a more novel approach of this that complements traditional image analysis techniques, offering a more interpretive method for understanding visual content. Knowledge-based questions are designed to prompt responses that extract relevant information from the subject image. In this study, we developed and examined various techniques for performing Knowledge-Based Visual Question Generation (K-VQG). These techniques included prompt engineering utilizing the latest large multimodal model (GPT-4 Vision) and employing sequence-to-sequence (Seq2Seq) models with semantic role labels (SRLs). Within the study, we employed various prompts aimed at providing context through both categorization and few-shot learning techniques. Alongside this, we implemented a sequence-to-sequence (Seq2Seq) model for textual question generation, anchored in semantic role labels, using LLM-generated captions of the images as its input. We observed that the proposed 2-shot learning method delivered the best results on the quantitative metrics (Semantic Similarity, METEOR, etc.).

**Keywords:** Visual Question Generation; Seq2seq; SRL; Knowledge-Based Question Generation; Visual Question Answering

# Acknowledgements

I would like to express my immense gratitude to my supervisor Dr. Kourosh Davoudi for his guidance, patience, and support in the accomplishment of my research objectives. I am deeply thankful to Ontario Tech University for providing the necessary resources and facilities to conduct this research. I would also like to extend special thanks to Dr. Aijun An, Seyed Nima Tayarani Bathaie, and Niloufar Beyranvand for their invaluable assistance and support throughout this endeavor. Finally, I express my profound gratitude to my family, whose unwavering support has been invaluable throughout the progression of my research. This endeavor would not have been achievable without their assistance.

# Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgment</b>	<b>ii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Application . . . . .	1
1.3 Research Context and Contributions . . . . .	3
1.4 Outline . . . . .	4
<b>2 Literature Review</b>	<b>5</b>
2.1 Natural Question Generation . . . . .	5
2.2 Class Acquisition . . . . .	6
<b>3 Methodology</b>	<b>7</b>
3.1 Seq2seq Model (empowered by semantic role labels) . . . . .	7
3.2 Vanilla Prompt Engineering Method . . . . .	9
3.3 1-Shot Prompt Engineering . . . . .	10
3.4 2-Shot Prompt Engineering . . . . .	12
3.5 1-Shot + Categorization Prompt Engineering . . . . .	15
3.5.1 Automating Categorization . . . . .	17
3.5.2 "Item Dictionary" Prompt . . . . .	18

3.5.3	"Procedural" Prompt . . . . .	20
3.5.4	"Other" Category Prompt . . . . .	22
<b>4</b>	<b>Evaluation</b>	<b>25</b>
4.1	Data Collection and Preprocessing . . . . .	25
4.1.1	Ground Truth . . . . .	26
4.2	Automatic Evaluation Metrics . . . . .	26
4.2.1	Scoring Functions . . . . .	27
4.3	Automatic Evaluation Results . . . . .	28
4.3.1	Automatic Categorization Results . . . . .	28
4.3.2	Question Generation Results . . . . .	28
4.4	Findings . . . . .	38
4.5	Summary . . . . .	39
<b>5</b>	<b>Conclusion</b>	<b>40</b>
5.1	Thesis Contribution Highlights . . . . .	41
5.2	Limitations . . . . .	41
5.3	Future Work . . . . .	42
	<b>Bibliography</b>	<b>42</b>

# List of Tables

1.1	Distinct Methods . . . . .	4
3.1	Categorization of Images . . . . .	16
4.1	Categorization Results . . . . .	28
4.2	Evaluation Metrics for Different Methods . . . . .	29
4.3	Questions . . . . .	29
4.4	Answers . . . . .	30
4.5	Q&A Pairs . . . . .	30
4.6	Question . . . . .	34
4.7	Answers . . . . .	34
4.8	Q&A Pairs . . . . .	34

# List of Figures

1.1	Multimodal application for K-VQG . . . . .	2
2.1	Examples of Natural Question Generation . . . . .	6
3.1	Sequence to Sequence Method Architecture . . . . .	8
3.2	Sequence to Sequence model Architecture [2] . . . . .	8
3.3	Vanilla Prompt Architecture . . . . .	9
3.4	1-Shot Prompt Architecture . . . . .	10
3.5	1-Shot + Categorization Prompt Architecture . . . . .	15
4.1	Automatic Evaluation Formulas . . . . .	27
4.2	Semantic Similarity Scores for Questions . . . . .	31
4.3	Semantic Similarity Scores for Answers . . . . .	32
4.4	Semantic Similarity Scores for Q&A Pairs . . . . .	33
4.5	METEOR Scores for Questions . . . . .	35
4.6	METEOR Scores for Answers . . . . .	36
4.7	METEOR Scores for Q&A Pairs . . . . .	37
4.8	Example of Generated Results (Image: WBA0027X) . . . . .	38

# Chapter 1

## Introduction

### 1.1 Motivation

The task of generating knowledge-based questions from images is extremely important, due to its potential applicability in numerous real-world systems. In its simplest form, it serves as a tool to extract data from images, which could otherwise be overlooked or not captured at all in different modes. Thus, knowledge-based visual question generation can further bridge between visual information and natural language. The old adage “a picture is worth a thousand words” serves as a testament to the fact that images can be extremely information-dense. Considering the remarkable abundance of images available on the internet, and the growing usage of these images in anything from billboards to social media, it can be said that there exists an untapped gold mine of data from within these images that we have yet to capture, or use.

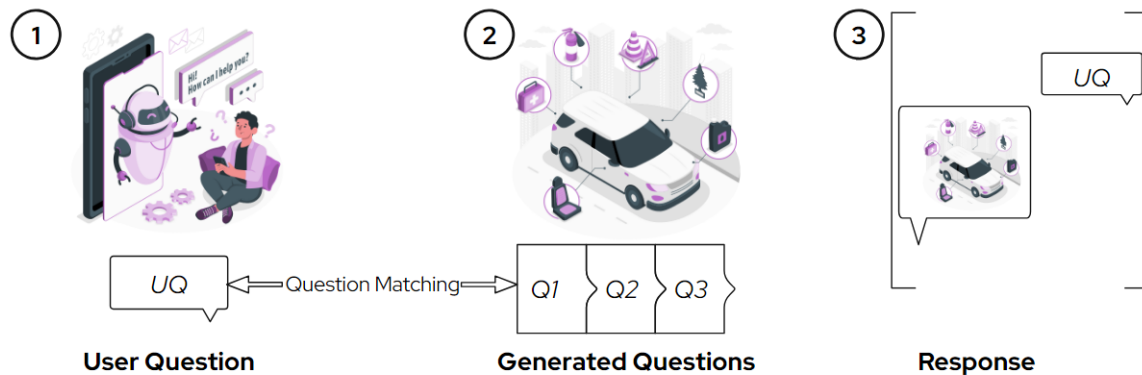
### 1.2 Application

Research conducted within this domain holds the potential to generate practical applications across various industries. For example, K-VQG could be used in educational content generation, where course content is replete with technical diagrams, illustrations



and graphics. In this case, a system equipped with K-VQG could automatically generate questions as well as answers to serve as learning assessments to gauge the learner's understanding of the subject matter. Another possible application of this could be to function as assistive technology for the visually impaired. AI systems can assist individuals with visual impairments by generating questions about images, providing a detailed understanding of visual scenes that may be challenging for them to perceive directly. Upon receiving a question from the individuals, it can be matched with a pre-existing generated question and answer set to help them understand the image better. With thorough development, another potential application for K-VQG could be in the form of a large multi-modal model that has the capability to aid an individual with respect to any questions they may have pertaining to a specific subject. Within the automotive industry for example, a user may have a query concerning the functionality of a component related to their car. The user would be able to pose the question to the model, which would then match the question to question(s) that were generated using visual content featured in an owner's manual or an online guide. Since the question is answerable via the image, giving the image to the user would suffice to answer their query. The flow of this process is outlined in Figure 1.1.

Figure 1.1: Multimodal application for K-VQG



## 1.3 Research Context and Contributions

Although the previously mentioned potential applications are promising, we encounter certain challenges during the research phase required for the development of K-VQG, and its subsequent applications. One such challenge is the availability of previous research regarding this subject matter. As we will discuss later, there exists work done in the field of question generation, question answering, knowledge-based class acquisition, and image analysis. However, there is by no means a handful of research papers regarding K-VQG, which can be likened to an amalgamation of the topics mentioned. As a result, obtaining a comprehensive overview of the required tasks and methodologies within the concept of K-VQG presents a challenge. Another challenge is that of obtaining datasets that specifically address the subject of knowledge-aware questions, being sourced from images. This presents a significant challenge because data collection is often quite expensive and time-consuming.

With our contributions, our study aims to serve as development and analysis for approaches in the field of knowledge-based visual question generation, and to provide some groundwork that can be used as a starting point for further research in this area. We also aim to provide a guideline to show how large multimodal datasets like GPT-4V can be used to accelerate this process via prompt engineering, and/or how one could create initial datasets. In our investigation, we use each of the question and answer prompts to explore the effects of distinct contexts, or parameters — herein referred to as distinct methods— on the resulting generated questions and answers. Alongside this, we’ve also employed a sequence-to-sequence (Seq2Seq) model for textual question generation, that makes use of semantic role labels. Due to this being a textual generation model, we used GPT-4V to generate a comprehensive description of the images examined, as input to this model. The resulting questions underwent analysis as outcomes of a distinct method. Table 1.1 outlines all of the distinct methods that were examined.

Table 1.1: Distinct Methods

Method	Description	Result Type(s)
Vanilla	Starting Point	Questions & Answers
1-Shot	1 Example	Questions & Answers
2-Shot	2 Examples	Questions & Answers
1-Shot + Categorization	Example + Label	Questions & Answers
Seq2Seq (w/ SRL)	Sequence-to-sequence model using semantic role labels	Questions

## 1.4 Outline

This report is organized in 4 Chapters as follows:

- In Chapter 2, we embark on an exploration of the existing literature in the realms of question generation and answering.
- Subsequently, in Chapter 3, a comprehensive overview of the methodologies utilized for conducting the experiments is provided, encompassing the development of distinct methods and an analysis of the Semantic Role Labeling (SRL-based) approach.
- Following this, in Chapter 4, a detailed exposition of the experimental procedures is presented, including the setup, data collection, testing and training splits, hardware specifications, and the acquired results.
- Finally, in Chapter 5, the conclusions are presented, offering objective reflections and a concise summary while addressing limitations and proposing avenues for future research.

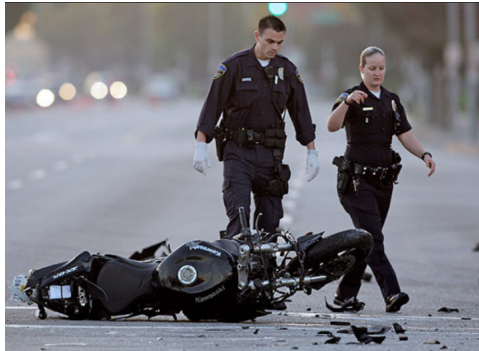
# Chapter 2

## Literature Review

### 2.1 Natural Question Generation

Within the space of question generation, many pieces of literature overlap in a way with the work that we have conducted. Oftentimes, the work is done to address the concept of creating natural and thought-provoking questions about an image [3]. Here, natural questions are defined as questions about what is inferred, rather than something literal. The emphasis is on moving beyond literal descriptions and exploring deeper connections between vision and language, as well as focusing on common sense inference and abstract events evoked by objects in the images. This cannot be classified as "knowledge-based" since the answers to these questions are not directly observable in the image itself. Figure 2.1 illustrates an example of the types of questions classified as natural questions. It should be noted that these questions cannot be answered solely by examining the image; rather, they require inference based on information derived from the image.

Figure 2.1: Examples of Natural Question Generation



Q1: Was anyone injured in the crash?

Q2: What caused this accident?

## 2.2 Class Acquisition

Some other projects with some overlap in our work, include those that deal with class acquisition of unknown objects within images [1]. This approach aims to gather information about object classes that have not been previously learned by the recognition model. Here, the main priority is that of finding unknown objects and being able to classify them. This could be viewed as a subset of our goal, since some of our generated questions would naturally deal with identifying some of the components inside an image. On the other hand, our work on knowledge-based question generation relates to generating questions that can be answered by analyzing the content of an image. While there may be some overlap in terms of generating questions, the emphasis of K-VQG is more on extracting meaningful information from images through questions that directly relate to the visual content.

# Chapter 3

## Methodology

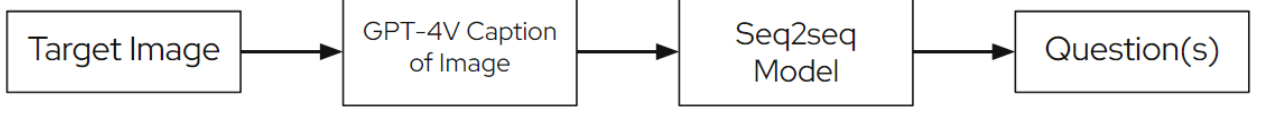
Given an image, we want to compare the results of generating questions through 5 methods. All GPT-based methods evoke a set of five questions and answers, while the SRL-based model generates as many as possible, given the GPT-generated caption. Each distinct method is outlined below.

Within this section, various prompts are presented to illustrate different methods. The placeholder image labeled '*Source Image for Prompt*' serves as a representation of the image that would accompany the prompt in actual implementation. This placeholder image is used to indicate the visual aspect of the prompt and is not a part of the finalized content.

### 3.1 Seq2seq Model (empowered by semantic role labels)

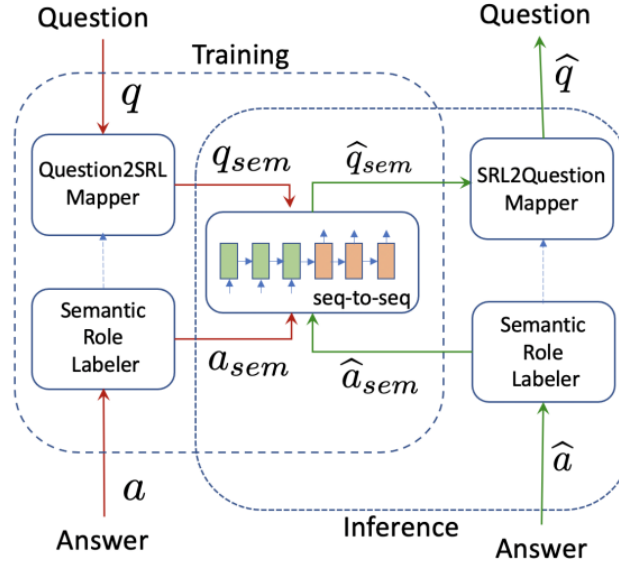
The following approach, which we outline in this section, draws upon a methodology proposed in the paper titled "Question Generation Using Sequence-to-Sequence Model with Semantic Role Labels" [2].

Figure 3.1: Sequence to Sequence Method Architecture



As shown in figure 3.2, the approach incorporates several components, including a Semantic Role Labeler (SRLer), a Seq2Seq model, and two semantic mappers (Question2SRL and SRL2Question). Initially, the SRLer extracts semantic representations (SRL labels) from answers in the training dataset, while Question2SRL maps questions to their corresponding semantic representations. Following this, a Seq2Seq model is trained using these semantic representations to transform an SRL representation of an answer into that of a question. During the inference stage, the Semantic Role Labeler extracts semantic representations of an answer, which are then converted to an SRL representation of a question by the newly trained Seq2Seq model. Finally, SRL2Question converts the SRL representation of the question into a natural language question.

Figure 3.2: Sequence to Sequence model Architecture [2]



To implement this approach into our study, we first needed comprehensive descriptions of the images involved, which would serve as input to the Seq2seq [2]. This was

done by prompting GPT-4V to generate descriptions using the following prompt format:

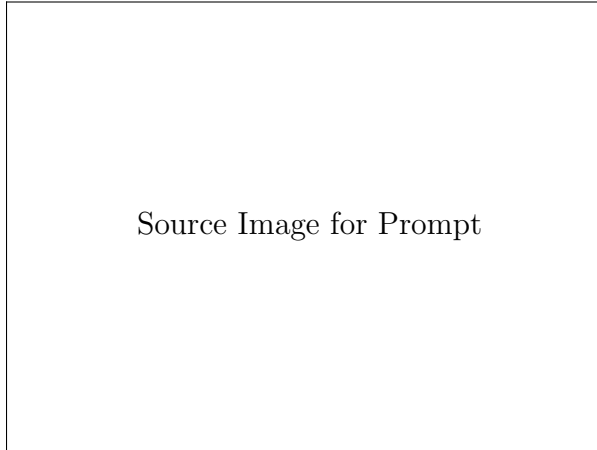
---

Provide a comprehensive description of the attached image.

Nothing else should be in your response.

Start the response with "Response:" if successful, and with "E:" otherwise.

For example: Response: ....



---

## 3.2 Vanilla Prompt Engineering Method

This prompt functions as a foundational reference for comparison. It will serve as a foundation for examining the impacts of various methods that were solely based on GPT-4V.

Figure 3.3: Vanilla Prompt Architecture

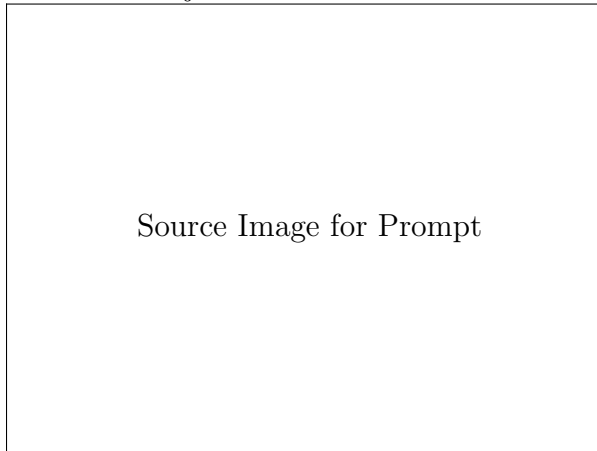


The prompt is formatted as follows:

---



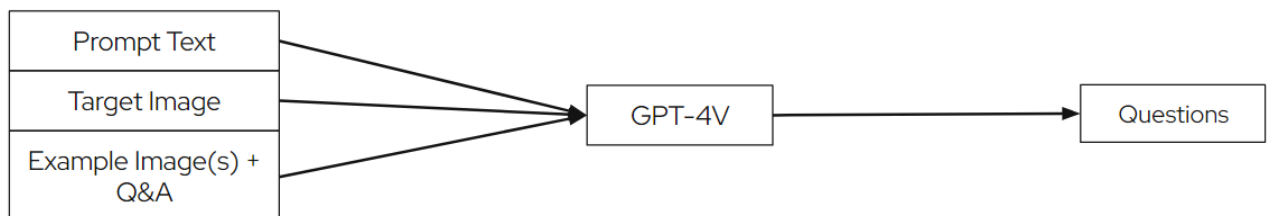
Generate 5 questions and answers about the attached image, which can be answered by the image. Format the questions and answers as a dictionary where keys are questions, and answers are values. For example: {"Q1": "A1", "Q2": "A2", "Q3": "A3", "Q4": "A4", "Q5": "A5"} No other text/values should be present in the response and the response shouldn't be json.



### 3.3 1-Shot Prompt Engineering

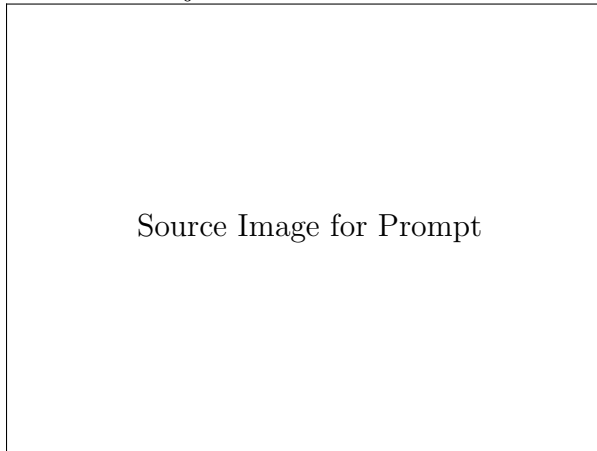
This prompt extends from our initial foundational prompt (3.1) by incorporating a single example, intended to serve as a reference for GPT.

Figure 3.4: 1-Shot Prompt Architecture



The format of the prompt is as follows:

Generate 5 questions and answers about the attached image, which can be answered by the image. Format the questions and answers as a dictionary where keys are questions, and answers are values. For example: {"Q1": "A1", "Q2": "A2", "Q3": "A3", "Q4": "A4", "Q5": "A5"} No other text/values should be present in the response and the response shouldn't be json.

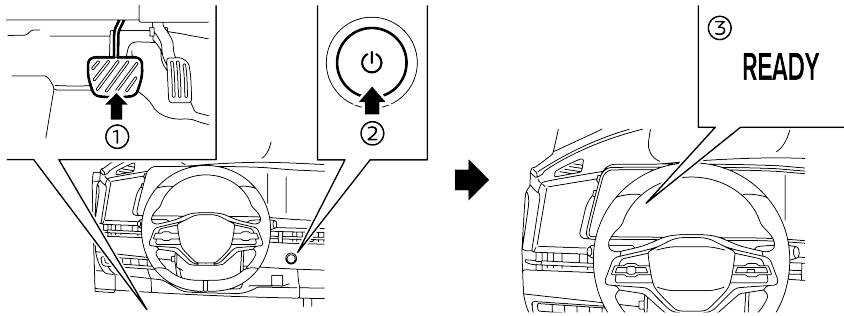


Here is an example for you to follow (Image with questions and answers):

- **How do I start my car?:** To start your car, first make sure your foot is firmly pressing the brake pedal. Then, press the start/stop engine button typically located on the dashboard or near the steering wheel.
- **What does the 'READY' indicator mean on my dashboard?:** The 'READY' indicator on your dashboard signifies that your car's engine is turned on and ready to drive. For electric or hybrid vehicles, this may not always be accompanied by the sound of an engine starting as these cars can operate silently.
- **Do I need to keep pressing the brake pedal while starting the car?:** Yes, you should keep pressing the brake pedal while starting the car. This is a safety feature to ensure that the vehicle doesn't move unexpectedly when starting.
- **Why won't my car start even when I press the brake and the start button?:** If your car doesn't start, check to ensure that the key fob is inside the vehicle and the battery is not depleted. Also, make sure that the gear is in the

'Park' or 'Neutral' position. If these conditions are met and the car still won't start, consult the vehicle's manual or a professional, as there may be an issue with the vehicle's electrical system or the start button itself.

- **Question\_5:** Answer\_5



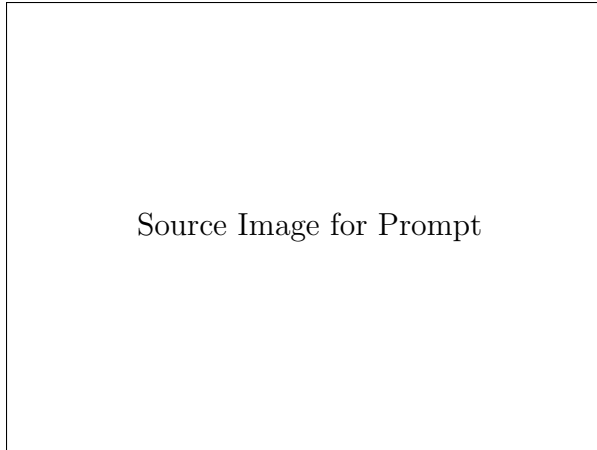
### 3.4 2-Shot Prompt Engineering

This prompt extends from our initial foundational prompt (3.1) by incorporating 2 examples, intended to serve as a reference for GPT. The prompt architecture mirrors that of the 1-shot prompt configuration (Figure 3.4), with the distinction of incorporating two examples instead of one.

The format of the prompt is as follows:

---

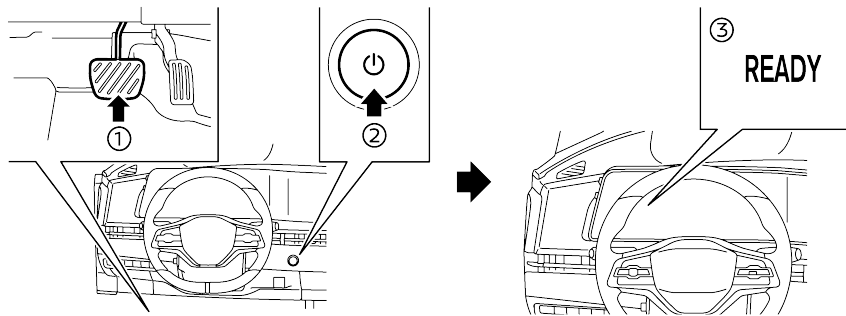
Generate 5 questions and answers about the attached image, which can be answered by the image. Format the questions and answers as a dictionary where keys are questions, and answers are values. For example: {"Q1": "A1", "Q2": "A2", "Q3": "A3", "Q4": "A4", "Q5": "A5"} No other text/values should be present in the response and the response shouldn't be json.



Here are 2 examples for you to follow (Images with questions and answers): **Example Image 1:**

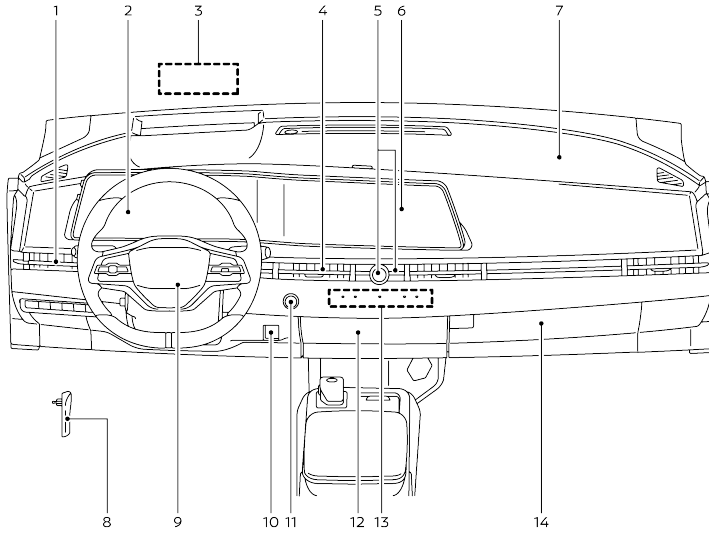
- **How do I start my car?:** To start your car, first make sure your foot is firmly pressing the brake pedal. Then, press the start/stop engine button typically located on the dashboard or near the steering wheel.
- **What does the 'READY' indicator mean on my dashboard?:** The 'READY' indicator on your dashboard signifies that your car's engine is turned on and ready to drive. For electric or hybrid vehicles, this may not always be accompanied by the sound of an engine starting as these cars can operate silently.
- **Do I need to keep pressing the brake pedal while starting the car?:** Yes, you should keep pressing the brake pedal while starting the car. This is a safety feature to ensure that the vehicle doesn't move unexpectedly when starting.
- **Why won't my car start even when I press the brake and the start button?:** If your car doesn't start, check to ensure that the key fob is inside the vehicle and the battery is not depleted. Also, make sure that the gear is in the 'Park' or 'Neutral' position. If these conditions are met and the car still won't start, consult the vehicle's manual or a professional, as there may be an issue with the vehicle's electrical system or the start button itself.

- **Question\_5: Answer\_5**



**Example Image 2:**

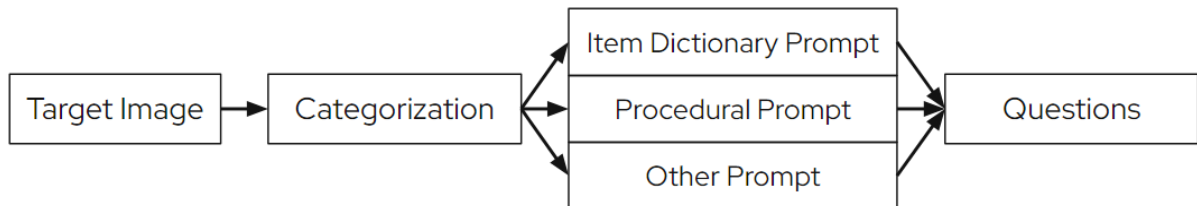
- **Where is the side ventilator located in my vehicle?:** The side ventilator is located at the end of the dashboard on the driver and passenger sides, near the doors.
- **Where can I find the meters and gauges in my vehicle?:** The meters and gauges are positioned directly in front of the steering wheel in the driver's line of sight for easy monitoring.
- **Where is the Head-Up Display (HUD) located?:** The Head-Up Display (HUD), if your vehicle is equipped with it, projects important information onto the windshield directly in the driver's view.
- **Where is the push-button power switch located?:** The push-button power switch, used to start and stop the engine, is located on the dashboard or the center console, depending on the vehicle model.
- **Where is the hood release handle found in my vehicle?:** The hood release handle is typically positioned inside the vehicle, on the lower left side of the dashboard or near the kick panel.



### 3.5 1-Shot + Categorization Prompt Engineering

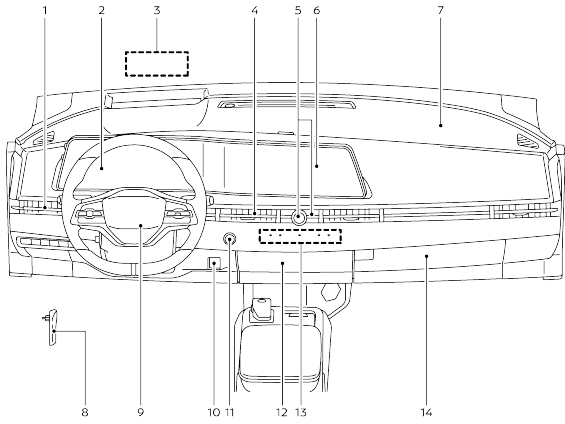
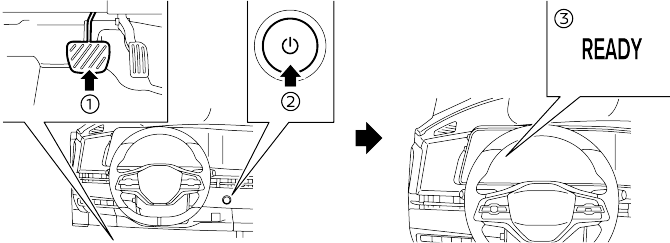
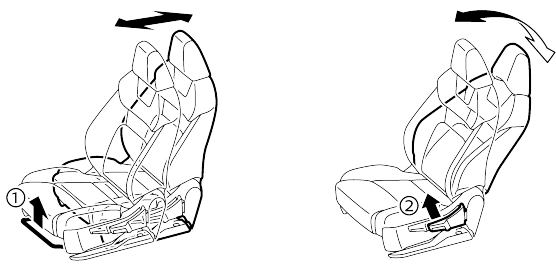
This prompt introduces the concept of categorization, or labeling, to the dataset.

Figure 3.5: 1-Shot + Categorization Prompt Architecture



The images in the dataset are categorized as 1 of 3 categories, shown in table 3.1. The categorization prompt is further enhanced by including an example image corresponding to that category, along with five questions and answers. This is done to tailor the resulting questions and answers to a specific category’s ideal inquiries. For instance, in an image labeled "Procedural," a question might pertain to the necessary steps for executing the depicted task. The subsection 3.5.1 outlines how the images were categorized, while the rest of Section 3.5 discusses how we used the categorization labels to perform K-VQG.

Table 3.1: Categorization of Images

Category	Description
Item Dictionary	<p>Visually annotated representation illustrating the components or features of a specific item, aiding in comprehension and reference. For example:</p> 
Procedural	<p>Presents a step-by-step sequence of actions or processes, offering guidance on performing a specific task or procedure. For example:</p> 
Other	<p>Diagram of the structure or workings of an object. Does not clearly belong to Item Dictionary or Procedural Categories. For example:</p> 

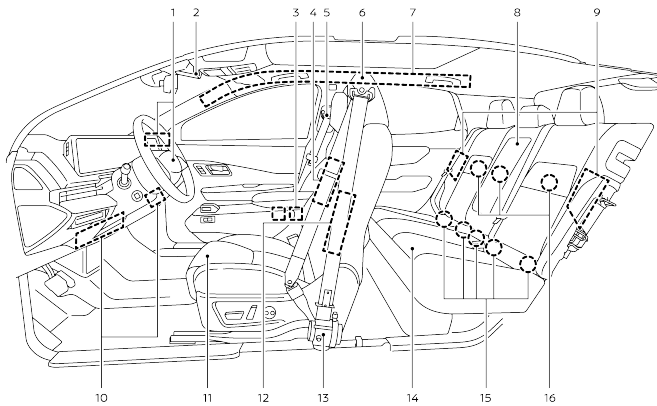
### 3.5.1 Automating Categorization

The categorization of images can be automated using GPT-4V. We did this using the following prompt:

---

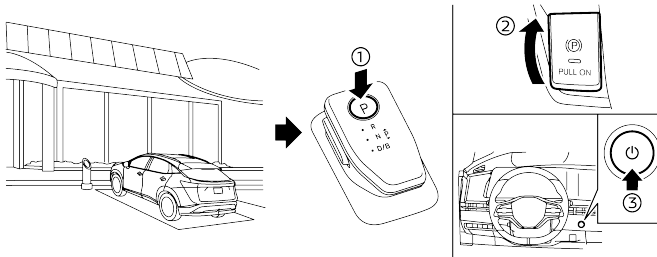
Label the following image as 1 of the following 3 categories:

1. "Diagram-ItemDictionary": If the image outlines parts of an item. These images are always numbered to differentiate the different parts. Here is an example of a "Diagram-ItemDictionary" image:



2. "Diagram-Procedural": If the image is of a clearly labelled step-by-step, numbered, procedure. The Procedure outlined should always be more than 1 step.

Here is an example of a "Diagram-Procedural" image:

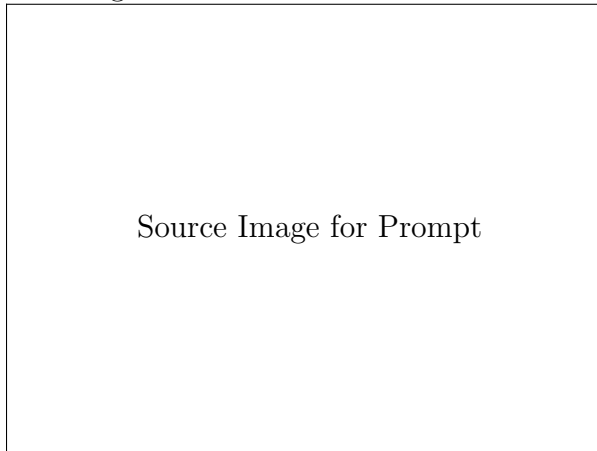


3. "Diagram-Other": If the image doesn't fall into the any of the other categories (i.e. everything else).

Do not compel the categorization of the image into the first and second categories. Your response should only contain the category, nothing else and no Punctuation. Below is



the image that needs to be labelled:



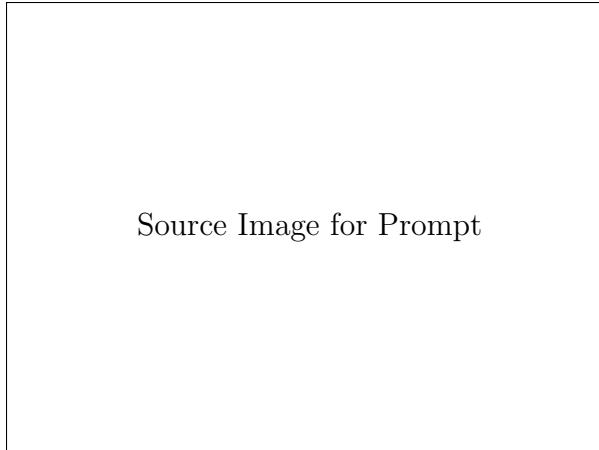
---

### 3.5.2 "Item Dictionary" Prompt

The following is an outline of the prompt given to images categorized as "Item Dictionary" images:

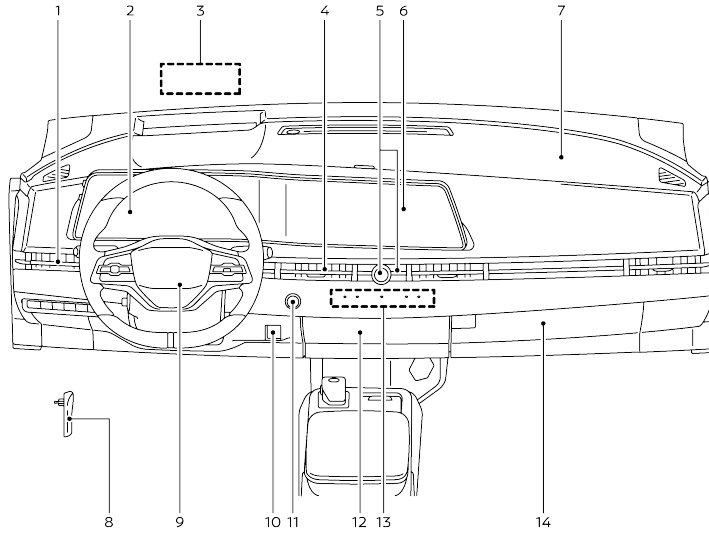
---

Generate 5 questions and answers about the attached image, which can be answered by the image. *The given image is an Item dictionary image, meaning that it numerically labels different parts present in the image.* Format the questions and answers as a dictionary where keys are questions, and answers are values. For example: {"Q1": "A1", "Q2": "A2", "Q3": "A3", "Q4": "A4", "Q5": "A5"} No other text/values should be present in the response and the response shouldn't be json.



Here is an example for you to follow (Image with questions and answers):

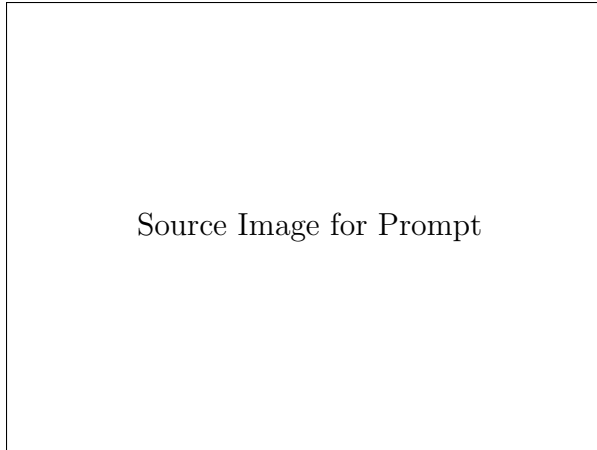
- **Where is the side ventilator located in my vehicle?:** The side ventilator is located at the end of the dashboard on the driver and passenger sides, near the doors.
- **Where can I find the meters and gauges in my vehicle?:** The meters and gauges are positioned directly in front of the steering wheel in the driver's line of sight for easy monitoring.
- **Where is the Head-Up Display (HUD) located?:** The Head-Up Display (HUD), if your vehicle is equipped with it, projects important information onto the windshield directly in the driver's view.
- **Where is the push-button power switch located?:** The push-button power switch, used to start and stop the engine, is located on the dashboard or the center console, depending on the vehicle model.
- **Where is the hood release handle found in my vehicle?:** The hood release handle is typically positioned inside the vehicle, on the lower left side of the dashboard or near the kick panel.



### 3.5.3 "Procedural" Prompt

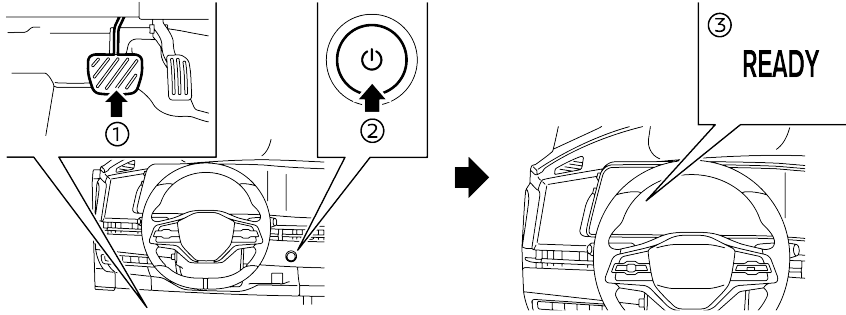
The following is an outline of the prompt given to images categorized as "Procedural" images:

Generate 5 questions and answers about the attached image, which can be answered by the image. *The given image is an Item dictionary image, meaning that it numerically labels different parts present in the image.* Format the questions and answers as a dictionary where keys are questions, and answers are values. For example: {"Q1": "A1", "Q2": "A2", "Q3": "A3", "Q4": "A4", "Q5": "A5"} No other text/values should be present in the response and the response shouldn't be json.



Here is an example for you to follow (Image with questions and answers):

- **How do I start my car?:** To start your car, first make sure your foot is firmly pressing the brake pedal. Then, press the start/stop engine button typically located on the dashboard or near the steering wheel.
- **What does the 'READY' indicator mean on my dashboard?:** The 'READY' indicator on your dashboard signifies that your car's engine is turned on and ready to drive. For electric or hybrid vehicles, this may not always be accompanied by the sound of an engine starting as these cars can operate silently.
- **Do I need to keep pressing the brake pedal while starting the car?:** Yes, you should keep pressing the brake pedal while starting the car. This is a safety feature to ensure that the vehicle doesn't move unexpectedly when starting.
- **Why won't my car start even when I press the brake and the start button?:** If your car doesn't start, check to ensure that the key fob is inside the vehicle and the battery is not depleted. Also, make sure that the gear is in the 'Park' or 'Neutral' position. If these conditions are met and the car still won't start, consult the vehicle's manual or a professional, as there may be an issue with the vehicle's electrical system or the start button itself.
- **Question\_5:** Answer\_5



### 3.5.4 "Other" Category Prompt

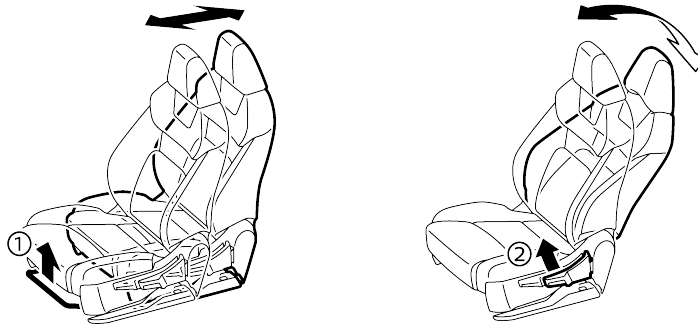
The following is an outline of the prompt given to images that belonged to neither the "Item Dictionary", nor the "Procedural" categories:

Generate 5 questions and answers about the attached image, which can be answered by the image. *The given image is a diagram, meaning it is a schematic representation of the structure or workings of an object.* Format the questions and answers as a dictionary where keys are questions, and answers are values. For example: {"Q1": "A1", "Q2": "A2", "Q3": "A3", "Q4": "A4", "Q5": "A5"} No other text/values should be present in the response and the response shouldn't be json.

Source Image for Prompt

Here is an example for you to follow (Image with questions and answers):

- **How do I adjust the position of my car seat forward and backward?:** To adjust your seat position, pull up the adjusting lever marked with a "1" in the diagram referenced as WBB0014X. Then, slide the seat to your desired position and release the lever to lock the seat in place.
- **Can I recline the front seat of my car for more comfort?:** Yes, you can recline your front seat for added comfort. Lift the adjusting lever indicated by a "2" in image WBB0014X, tilt the seatback to your desired angle, and then release the lever to secure it in the new position.
- **Is it possible to adjust the seat while the vehicle is in motion?:** For safety reasons, it is recommended to adjust your seat only when the vehicle is parked. If you must adjust it while the vehicle is in motion, ensure it is done quickly and carefully, referring to the adjustment levers shown in WBB0014X.
- **How do I know if my car seat is adjusted correctly for driving?:** A properly adjusted car seat will allow you to reach all controls comfortably and see the road clearly. Your knees should have ample room, and you should be able to press the pedals without stretching. Refer to WBB0014X for the controls that adjust the seat.
- **Why can't I adjust my seat forward or backward?:** If your seat won't adjust, check that nothing is obstructing the seat tracks. Refer to image WBB0014X and ensure you are using the correct lever, labeled "1" for this adjustment. If the lever is raised and there are no obstructions, but the seat still won't move, there may be an issue with the seat mechanism that requires professional attention.



# Chapter 4

## Evaluation

We conducted an investigation on exploring and evaluating methods for extracting information from images through the generation of knowledge-based questions and answers. We examined the effect of different context modes on our prompt engineering methods, as well as how prompt engineering methods compare to the latest Seq2Seq question generation models (which employ SRLs).

### 4.1 Data Collection and Preprocessing

The dataset utilized in our study was meticulously compiled from images sourced directly from the owner’s Manual of a 2023 Nissan Ariya, encompassing a total of 451 images. Initially available in Scalable Vector Graphics (SVG) format, a vector image format derived from XML, the images underwent a transformation process to conform to the PNG format, aligning with the image specifications compatible with GPT-4 Vision models. This conversion facilitated integration and analysis within our research framework, ensuring functionality with the developed methods.



### 4.1.1 Ground Truth

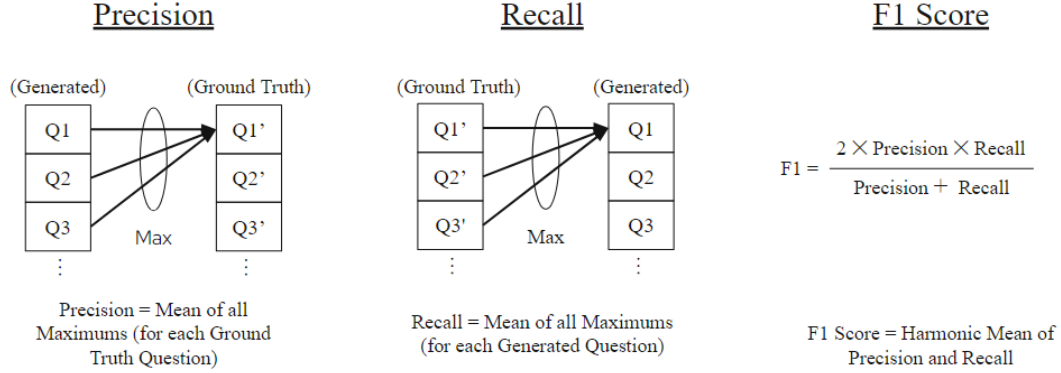
In our investigation, we diligently compiled a comprehensive set of 132 knowledge-based questions and corresponding answers, meticulously tailored to serve as a benchmark for our study. These questions and answers underwent meticulous curation from a subset of 17 images, thoughtfully selected from the original dataset of 451 images. Additionally, this dataset underwent a rigorous validation process, including expert review, to ensure the coherence of the questions and the accuracy of the answers. The questions were formulated to be answerable solely through visual inspection of the image content, without reliance on external references/knowledge. Moreover, careful consideration was given to formulating questions that align with the natural inquiries a car owner might pose, independent of any instructional manual guidance. Alongside these questions and answers, we’ve also categorized each image according to its function, based on the labels mentioned in section 3.5.

## 4.2 Automatic Evaluation Metrics

When conducting automatic evaluation, we calculated precision, recall and F-measure to quantify the quality of generated questions, answers, and the average of the pair. Part of the method was proposed in (Naeiji et al., 2023) [2]. To calculate the precision of a method using a given scoring function ( $S$ ) for an arbitrary image ( $I$ ), we compare the scores received from  $S$  as shown in Figure 4.1. For each ground truth sentence we have for image  $I$ , we compute the score  $s$ , for every generated sentence in a method. We then compute these scores for every ground truth sentence, and collect the maximum scores. The mean value of these maximum scores gives us the precision for that method, on image  $X$ . To determine recall, the procedure mirrors that of precision, with the exception of swapping the roles of ground truth and generated sentences. To compute the F1 score for a method on a given image  $X$ , we compute the harmonic mean (Figure 4.1) between

the precision and recall scores.

Figure 4.1: Automatic Evaluation Formulas



### 4.2.1 Scoring Functions

To quantitatively evaluate our generated results, we opted for two scoring functions that allowed us to offer a detailed representation of the results, facilitating a nuanced understanding of the outcomes.

The first, was spaCy's semantic similarity score. This metric quantifies the degree of semantic equivalence between two text sentences, providing insight into their contextual similarity. Specifically, the library uses pre-trained word vectors to compute the similarity scores, and captures relationships in meaning within the text. As well, the function employs vector averaging to compute similarity, which could yield averaged representations insensitive to word order and advanced phrase meanings. So while semantic similarity scores provide a numerical comparison between two phrases, their interpretation should be viewed judiciously.

The second score we considered, was the METEOR (Metric for Evaluation of Translation with Explicit Ordering) score. It considers both exact word matches and matches with synonyms and paraphrases, and also uses alignment information to handle word order variations in a sentence.

## 4.3 Automatic Evaluation Results

### 4.3.1 Automatic Categorization Results

The results shown in Table 4.1 display the precision, recall, F1-Score, and the accuracy of the automatic categorization which we used for the 1-Shot + Categorization method. Upon supplying the GPT prompt with two examples, one from the "Procedural" category and another from the "Item Dictionary" category, we observed a 9.9 percent increase in accuracy.

Table 4.1: Categorization Results

Model	Precision	Recall	F1 Score	Accuracy
Zero-Shot (No example)	0.631	0.686	0.521	0.565
2-Shot	0.652	0.744	0.586	0.663

This implies that GPT-4 Vision works better when we use a few-shot learning approach to guide it to better results (increased precision, recall, and accuracy).

### 4.3.2 Question Generation Results

This section contains results for the K-VQG approaches that we developed, followed by a discussion of findings, and an evaluation of methodology in the subsequent sections. Table 4.2 contains the results of performing evaluation using the BLEU and ROUGE-L metrics, and computing the mean across all ground truth questions. Tables 4.3, 4.4, 4.5 and figures 4.2, 4.3, 4.4 consist of semantic similarity (spaCy) score results for questions, answers, and Q&A Pairs. Tables 4.6, 4.7, 4.8 and figures 4.5, 4.6, 4.7 contains METEOR score results for questions, answers, and Q&A Pairs.

Table 4.2: Evaluation Metrics for Different Methods

Method	Bleu-4	Rouge-L
Vanilla	Precision: 0.0 Recall: 0.0 F1: 0.0	Precision: 23.61 Recall: 21.19 F1: 22.12
1-Shot + Categorization	Precision: 0.0 Recall: 0.0 F1: 0.0	Precision: 27.63 Recall: 27.5 F1: 27.16
1-Shot and 2-Shot (Mean)	Precision: 0.85 Recall: 1.21 F1: 0.99	Precision: 27.31 Recall: 26.79 F1: 26.69
Seq2Seq (w/ SRL)	Precision: 1.03 Recall: 0.91 F1: 0.92	Precision: 25.57 Recall: 26.9 F1: 25.76

### Semantic Similarity Scores

Table 4.3: Questions

Method	Precision	Recall	F1-score
1-Shot	0.9049	0.90469	0.9047
2-Shot	0.9048	0.9092	0.9070
Vanilla	0.8940	0.8942	0.8940
1-Shot + Categorization	0.8991	0.9018	0.9004
Seq2seq w/SRL	0.87445	0.88734	0.8808

Table 4.4: Answers

Method	Precision	Recall	F1-score
1-Shot	0.9230	0.9250	0.9239
2-Shot	0.9121	0.9208	0.9163
Vanilla	0.8438	0.8785	0.8601
1-Shot + Categorization	0.9195	0.9218	0.9206

Table 4.5: Q&amp;A Pairs

Method	Precision	Recall	F1-score
1-Shot	0.9103	0.9111	0.9107
2-Shot	0.9039	0.9096	0.9067
Vanilla	0.8642	0.8787	0.8712
1-Shot + Categorization	0.9047	0.9069	0.9058

Figure 4.2: Semantic Similarity Scores for Questions

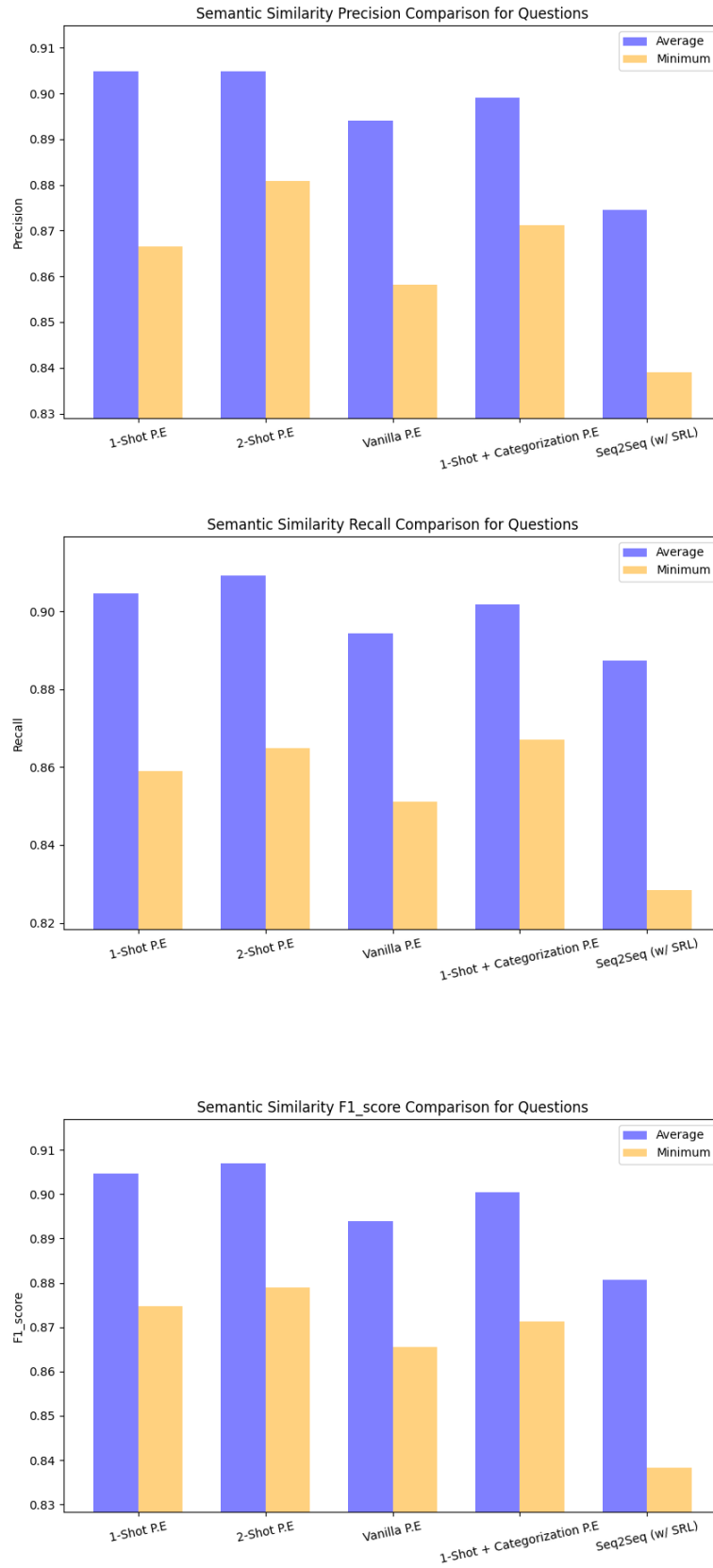


Figure 4.3: Semantic Similarity Scores for Answers

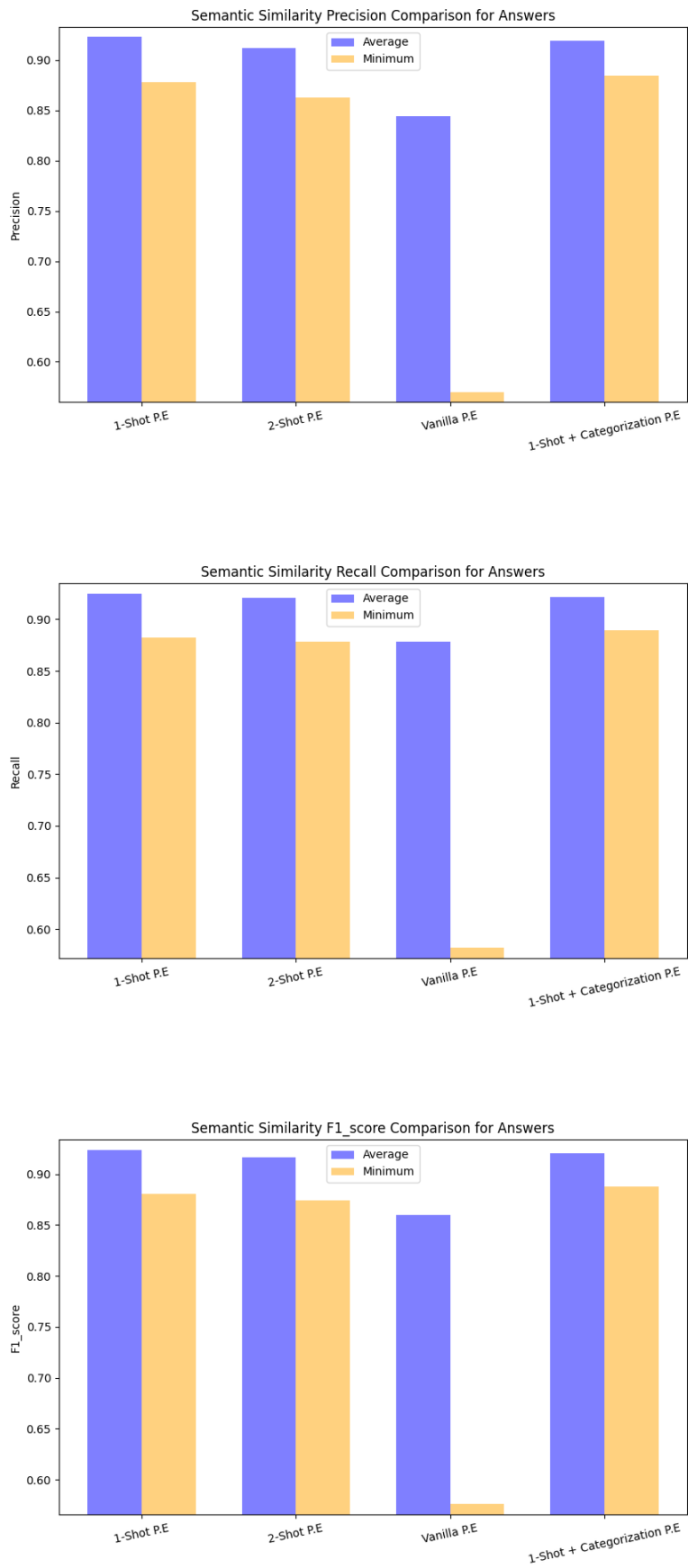
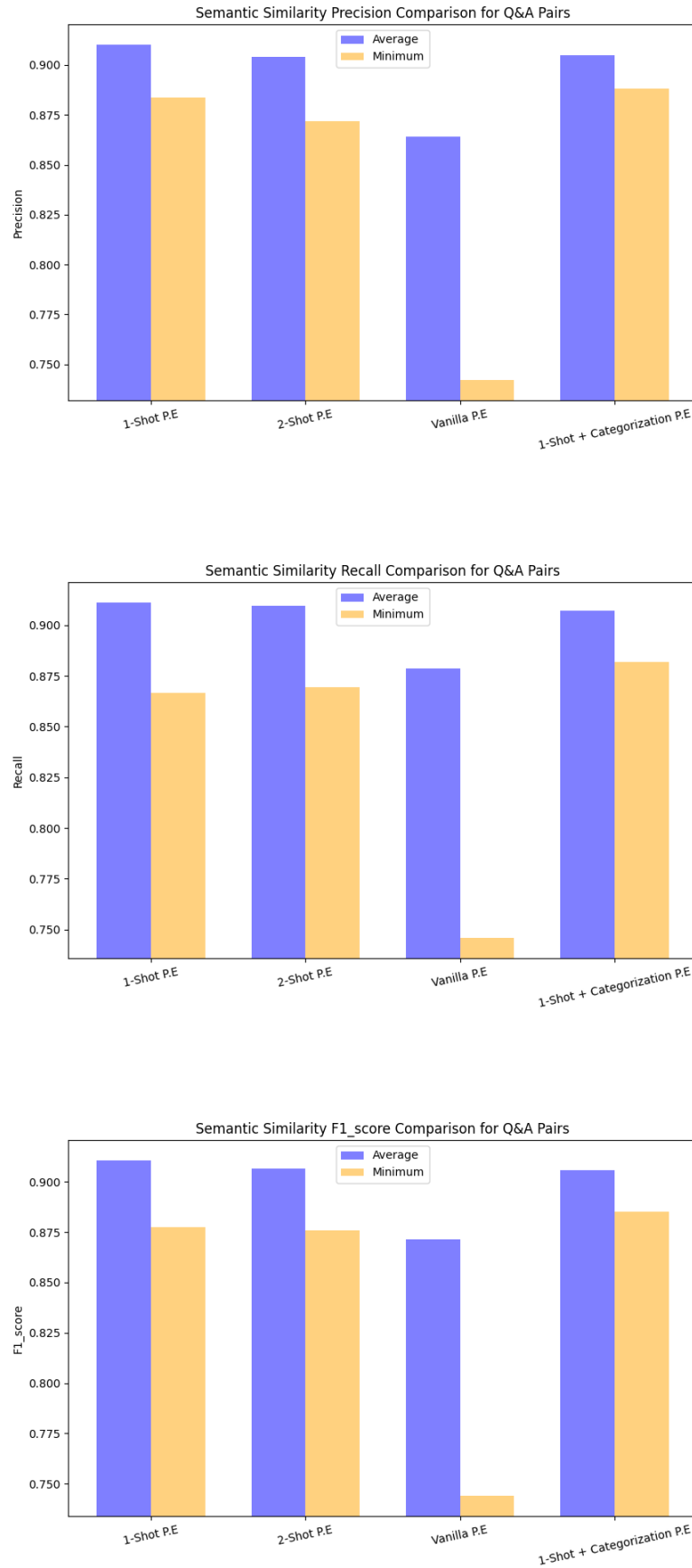


Figure 4.4: Semantic Similarity Scores for Q&amp;A Pairs





**METEOR Scores**

Table 4.6: Question

Method	Precision	Recall	F1-score
1-Shot	0.3730	0.3725	0.3719
2-Shot	0.3770	0.3915	0.3827
Vanilla	0.3412	0.3263	0.3328
1-Shot + Categorization	0.3741	0.3832	0.3776
Seq2seq w/SRL	0.2008	0.2417	0.2168

Table 4.7: Answers

Method	Precision	Recall	F1-score
1-Shot	0.2377	0.2379	0.2371
2-Shot	0.2291	0.2292	0.2278
Vanilla	0.1460	0.1463	0.1450
1-Shot + Categorization	0.2396	0.2340	0.2351

Table 4.8: Q&amp;A Pairs

Method	Precision	Recall	F1-score
1-Shot	0.2933	0.2890	0.2909
2-Shot	0.2888	0.2948	0.2908
Vanilla	0.2347	0.2254	0.2294
1-Shot + Categorization	0.2904	0.2929	0.2910

Figure 4.5: METEOR Scores for Questions

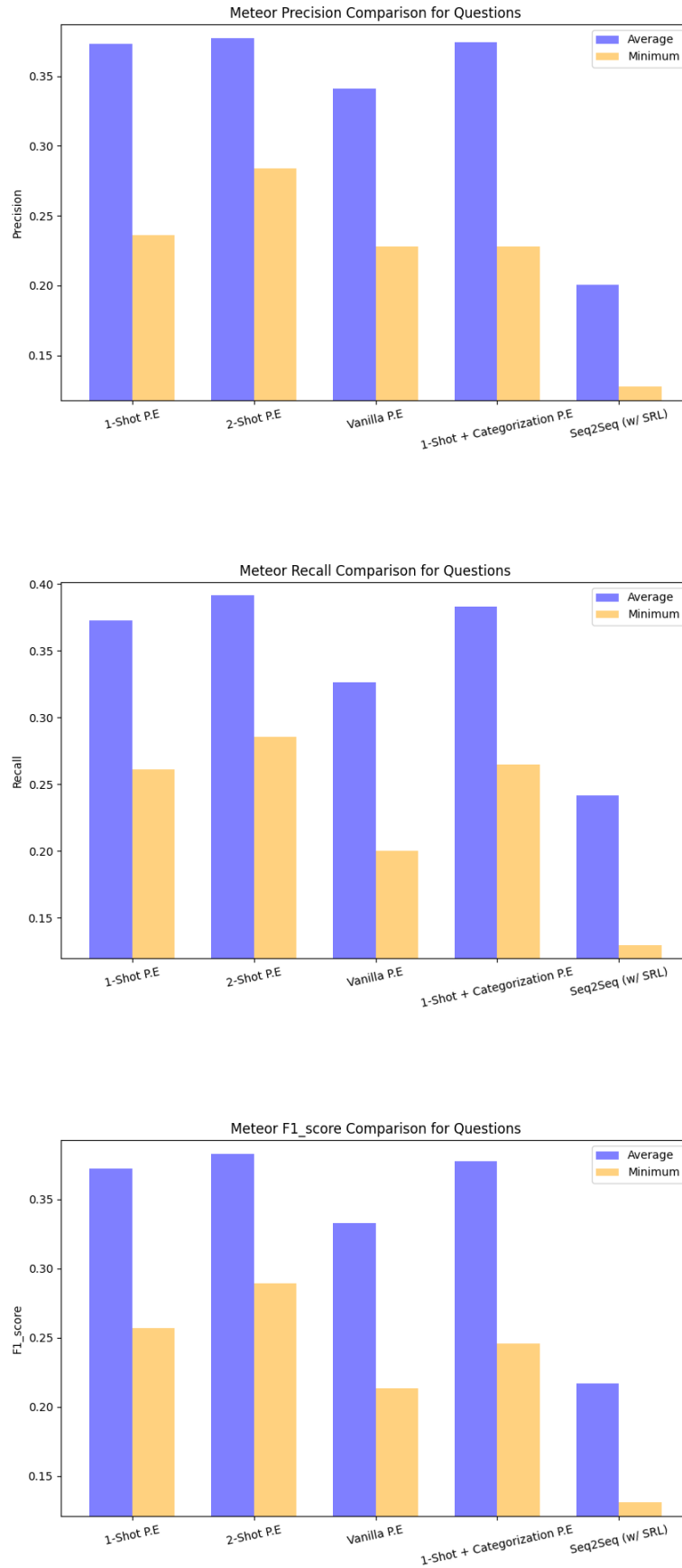


Figure 4.6: METEOR Scores for Answers

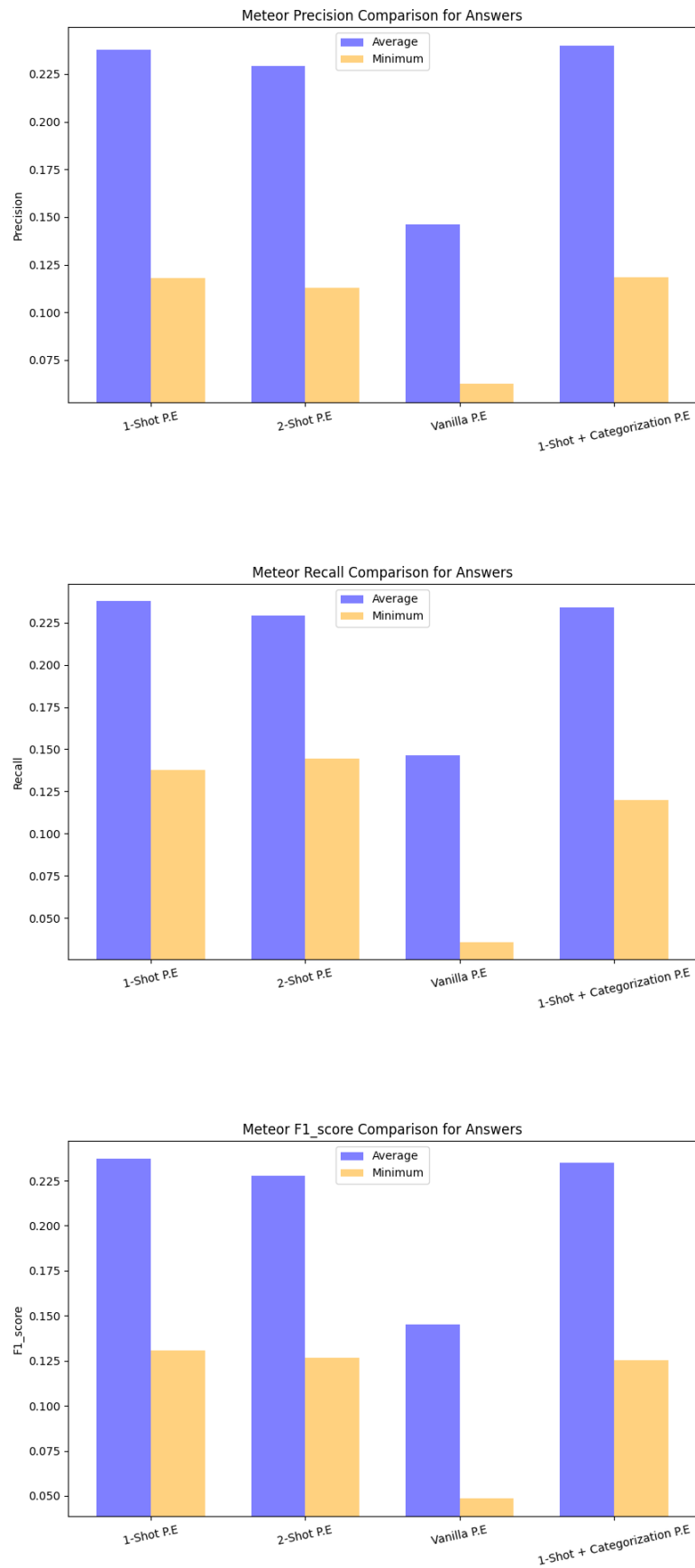
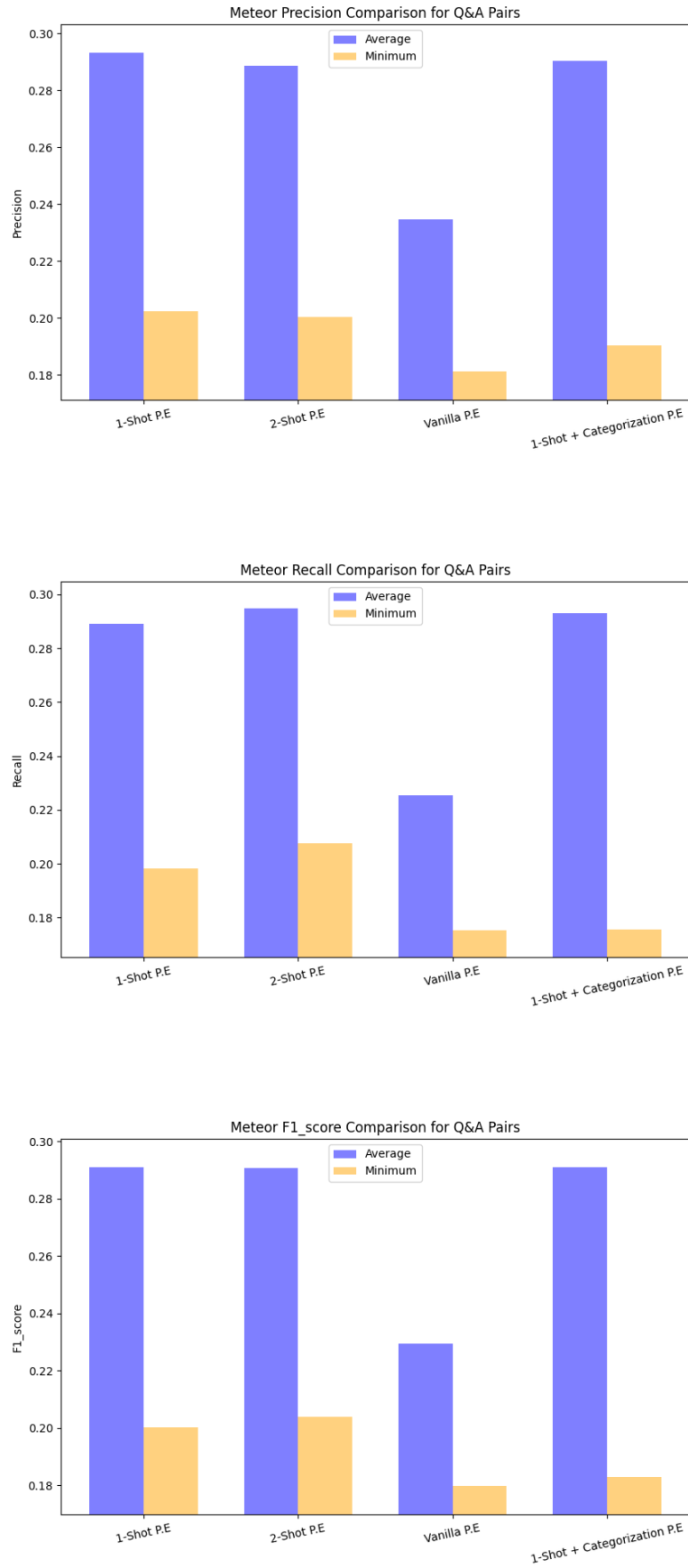


Figure 4.7: METEOR Scores for Q&amp;A Pairs

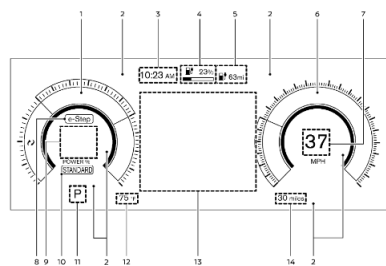


## 4.4 Findings

As shown in section 4.3, the METEOR and Semantic Similarity scores showed that the 2-shot prompt engineering configuration had the highest similarity scores when compared to the ground truth questions and answers.

When observing the METEOR precision, recall, and F1-measure scores for purely the questions, we see that the 2-Shot method is consistently the highest scoring method. This is followed by the 1-Shot + Categorization, and the 1-Shot prompt engineering methods respectively. This is consistent with widely accepted principles in the field of prompt engineering, where few-shot learning is thought to be the best method of improving the quality of results. On the opposite side of the scale however, lies the Seq2seq model that employed SRLs, which consistently had the worst metric scores. When analyzing the generated questions, we noticed that this was due to the Seq2seq model generating text based on the wording and/or word order of the sentence. Figure 4.8 illustrates an instance of this phenomenon, wherein the question posed to the Seq2Seq model is derived from the wording found within the generated caption (The caption contains: "On the right, there is another half-circle gauge..."). This style of captioning, introduced more opportunities for obvious, simplistic question that did not rely on the relationships between the functions of different objects in the images (Example: Figure 4.8, 2-Shot method).

Figure 4.8: Example of Generated Results (Image: WBA0027X)



Method	Question
Seq2seq w/ SRL	What is the second half-circle gauge?
Vanilla Prompt Engineering	What time is displayed in the center of the image?
1-Shot Prompt Engineering	What does the 'P' symbol inside a circle on the car dashboard indicate?
2-Shot Prompt Engineering	Is the vehicle in motion according to the image?
1-Shot + Categorization Prompt Engineering	Where is the Fuel Gauge located as per this diagram?

Surprisingly, when we take a look at the Semantic Similarity scores for the questions, we see that the 1-Shot prompt engineering method is higher in terms of F1-measure compared to the 1-Shot + Categorization prompt engineering method. This can likely be attributed to GPT-4V "overfitting" to match the examples according to the categorization, since the example image's questions all relate to the locations of certain components in the image (see section 3.5.2). As mentioned earlier, our ground truth was tailored towards asking questions that align with what a car owner would naturally inquire about, independent of the image. Thus, the prompt containing the categorization might have influenced GPT-4V to match with the example questions to a higher degree (compared to 1-Shot without categorization).

## 4.5 Summary

In this chapter, we evaluated our proposed methods/models on the task of Knowledge-based Visual Question Generation. For each method, we calculated the precision, recall, and F1-Measure per metric. These results prove the utility and efficiency of providing examples (few-shot learning) to GPT-4V in order to boost the quality of results generated, while also emphasizing the importance of word order and sequence in sentences given as input to Seq2seq models.

# Chapter 5

## Conclusion

In conclusion, our study addresses challenges that are encountered in the development and application of Knowledge-Based Visual Question Generation (K-VQG). Our study lays the groundwork for further research within this field, and develops methods that can be used as a starting point for this task. Our work addresses obstacles such as the scarcity of existing research in K-VQG, and the difficulty in obtaining datasets tailored towards knowledge-based questions sourced from images. Our contributions aim to assist other researchers by providing a functional foundational analysis of K-VQG approaches as well as insights into the employment of large multimodal datasets like GPT-4V through prompt engineering. Alongside, we introduce the use of a sequence-to-sequence (Seq2Seq) model with Semantic Role Labels for this task, which was traditionally used for textual question generation. With this we showcase its efficacy in generating questions based on image descriptions provided by GPT-4V. Our investigation delves into various distinct methods, each analyzed for its impact on the resulting questions and answers. After an exhaustive analysis of the resulting generated questions, we found that in the case of prompt engineering for K-VQG, contextual support and providing examples of questions and answers significantly improve the quality of results received. Further implications of the applications, potential, and reliability of these methods and more, will need to be

explored in future work.

## 5.1 Thesis Contribution Highlights

This main contributions of this thesis can be outlined as follows:

- First we develop five distinct methods of generating knowledge-based questions from images
- Alongside this, we develop a prompt engineering method to generate captions from images, which served as input to one of our K-VQG methods
- Then, we curated a set of 132 Knowledge-based questions and answers from 17 images to serve as ground truth for our study. This dataset then underwent a rigorous validation process to assess and improve its content.
- We then developed a categorization technique to automatically determine the role of each image in the dataset.
- Finally, We explored different approaches to analyze the quality of the generated questions and answers, including using automated evaluation methods like semantic similarity scores.

## 5.2 Limitations

The proposed models in this research demonstrated great results on automatic evaluation. However, there are some limitations to this work that would need to be considered.

- While this work’s main objective is the development of K-VQG methods, the proposed methods were not given a diverse combination of datasets for a variety contexts and applications. Although, we see promising correlations between the



methods' questions and our ground truth, we would require a more diverse set of scenarios/images in order to fully understand the capabilities of these methods.

- Additionally, there is a necessity to significantly expand the volume of available images, questions, and answers, as our dataset was quite limited (451 image dataset, 132 knowledge-based questions and answers from 17 images).
- Lastly, it's important to consider a wider range of contextual approaches. This could involve methods with more few-shot learning examples or those that make better use of textual context from the manual. In terms of categorization, a more customized prompt could also help GPT-4V classify image context more accurately.

## 5.3 Future Work

Given the novelty of this study, there is significant potential for future research to further develop its findings. Considering the limitations, it's imperative to contemplate the expansion of the data involved. This would consist of doing some of the following:

- Expand the number of ground truth questions, answers, and images to allow for more variations in the question and answer styles.
- Considering the diversity of vehicle types, incorporating owner's manuals from various models would enable a more extensive array of images to be included in the dataset.
- Incorporating images from a broader array of sources, like educational materials and real-world contexts in addition to owner's manuals, would enrich the dataset with a more extensive range of visual content. This dataset would then provide a more nuanced understanding of contextual information in a multitude of fields and help facilitate a more exhaustive analysis of the data.

# Bibliography

- [1] Muxin Chen, Xiyu Huang, Zijun Xiao, Jiasheng Wei, and Wenlung Hwang. Visual question generation for class acquisition of unknown objects, 2020.
- [2] Alireza Naeiji, Aijun An, Heidar Davoudi, Marjan Delpisheh, and Muath Alzghool. Question generation using sequence-to-sequence model with semantic role labels. In *Paper to be presented at the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)*, Dubrovnik, Croatia, May 2023.
- [3] Weijie Su, Xiaodan Zhu, Yu Cao, Bin Li, Lei Lu, Fan Wei, and Jing Dai. Generating natural questions about an image. <https://paperswithcode.com/paper/generating-natural-questions-about-an-image>, 2019. Papers with Code.