

Subjective Scoring Framework for VQA Models in Autonomous Driving

anonymised for VLADR workshop

¹anonymised for VLADR workshop

²anonymised for VLADR workshop

Corresponding author: anonymised for VLADR workshop

ABSTRACT The development of vision and language transformer models has paved the way for Visual Question Answering (VQA) models and related research. There are metrics to assess the general accuracy of VQA models but subjective assessment of the answers generated by the models is necessary to gain an in-depth understanding and a framework for subjective assessment is required. This work develops a novel scoring system based on the subjectivity of the question and analyses the answers provided by the model using multiple types of natural language processing models (bert-base-uncased, nli-distilBERT-base, allmpnet-base-v2 and GPT-2) and sentence similarity benchmark metrics (Cosine Similarity). A case study detailing the use of the proposed subjective scoring framework on three prominent VQA models- ViLT, ViLBERT, and LXMERT using an automotive dataset is also presented. The framework proposed aids in analyzing the shortcomings of the discussed VQA models from a driving perspective and the results achieved help determine which model would work best when fine-tuned on a driving-specific VQA dataset.

INDEX TERMS Semantic Analysis, Scoring Framework, Subjective Assessment, VQA Models

I. INTRODUCTION

AUTONOMOUS driving has been a large area of development, both commercially and in academic research, over the last decade or more [1], and continues to be an area of deep interest in the research community. However, while perception and control tasks are still undergoing significant research [2], [3], there has been an increase in interest in incorporating elements of trustworthiness [4], [5] and explainability [6]–[8] into autonomous driving. Visual Questioning Answering (VQA) is proposed as a part of the vehicle autonomy trustworthiness and interpretability solution [9], [10]. VQA can be used, for instance, to explain the behaviour of an autonomous vehicle to the vehicle occupants [11] as shown in Figure 1.

VQA models are designed to comprehend and respond to questions based on images. They operate by integrating computer vision and natural language processing techniques. Typically, they take an image as input, generally processed through a convolutional neural network (CNN) to extract meaningful visual features. VQA models also receive a natural language question related to the image, which undergoes language processing (including tokenization and word embeddings). The models will then combine the extracted image features and processed question features, creating a joint rep-

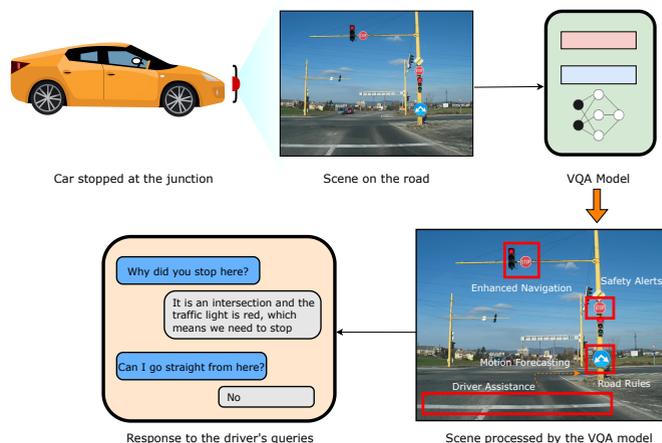


FIGURE 1. A demonstration of how VQA models work in a driving scenario

resentation that encapsulates the relationship between visual and textual information. This fused representation is further utilized to predict an appropriate answer to the given question, employing various machine learning approaches like neural networks and attention mechanisms. The goal is to provide accurate and relevant answers, demonstrating a comprehensive understanding of both the image content and the posed

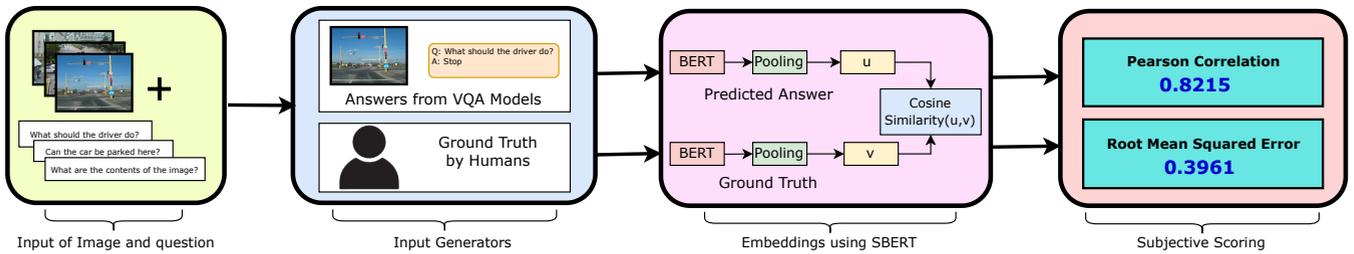


FIGURE 2. A Generalized Overview of the Subjective Scoring Framework proposed

question.

Diverse evaluation metrics exist in the literature for measuring the performance of VQA models. For instance, LXMERT [12] relies on accuracy for assessing its performance, while ViLBERT turns to test-dev accuracy for benchmarking [13]. ViLT, in contrast, juxtaposes test-dev accuracy with response time in comparison to other state-of-the-art models [14]. VAuLT uses a combination of Accuracy and Mac-F1 [15], and TVLT emphasizes latency as a pivotal metric [16].

In the VQA domain, there are some established evaluation methods and metrics that help measure the quality of responses to such questions:

- 1) **Human Evaluation:** Expert human evaluators assess the quality of responses based on predefined criteria. They may use rating scales or qualitative judgments that evaluate responses according to factors such as clarity, appropriateness, how informative an answer might be or overall quality. [17].
- 2) **Inter-Evaluator Agreement** measures how consistently different evaluators rate the same responses [18]. A high level of agreement indicates that the evaluation process is reliable.
- 3) Borrowed from machine translation evaluation, the **Bleu Score** assesses how closely the model's responses match good human responses [19]. It provides a measure of response similarity.
- 4) Similar to Bleu, the **METEOR score** evaluates response quality by considering synonyms and paraphrases [20]. It looks at how well the model's responses align with good human responses.
- 5) Initially designed for image captions, the **CIDEr score** can be adapted to VQA. It assesses response quality by considering consensus and diversity among human judgments [21].
- 6) Commonly used for text summarization, the **ROUGE Score** can also be applied to VQA to measure how well the model's responses match human responses [22].
- 7) **Response Length and Quality:** Evaluating the length of responses helps ensure they are neither too short nor overly long, depending on the nature of the question as done in [14].

In practice, the selection of appropriate metrics and evaluation methods should be tailored to the specific goals and character-

istics of the subjective questions under examination. Evaluating subjective questions requires a multifaceted approach that combines automated metrics with human assessments and user feedback to gain an understanding of response quality in the context of VQA.

A gap observed in the existing metrics is that there is an inability to compare the performance of VQA models to human answers. Evaluating the quality of responses to subjective questions in VQA models is a complex task. Subjective questions involve understanding context, and relevance, and generating human-like responses. A subjective question might be "What should the driver do?", in contrast to the more objective "What are the contents of the image?". The goal of this paper is to discuss a means by which the responses of a VQA model can be compared to human responses, given that humans can answer subjective questions in context. In this paper, we propose a subjective scoring framework tailored to the evaluation of visual question-answering models for autonomous driving. Figure 2 shows a high-level overview of our proposed method.

The principal contribution of this paper is to advance the development of a structured framework for assessing answers to subjective questions within VQA models in comparison to human answers, which in the authors' opinion is of great importance in the autonomous driving space. This undertaking substantively contributes to the academic and research community by propelling the frontiers of knowledge in natural language understanding and multimodal models.

The authors have presented very initial work as a conference submission [23]. However in this paper, we greatly expand on the previous work by using only a part of the results published in the conference for a case study validation. We have expanded the results of [23] by introducing a few more driving scenarios as further explained in the Section V-A. We use these results as an input in the framework proposed here.

Our paper is structured as follows. We start with an initial review of existing literature in this domain and highlight the deficiencies in the prior art we are overcoming with Section II. In Section III, we introduce the constituent components of the proposed framework and explain their significance in our context. To underscore the practical application of the framework, we present a case study in Section V-A of section V. Within this section, we leverage three distinct VQA models against a driving context dataset, showcasing how the frame-

work can be effectively utilized. We have shown a cursory set of results with just two questions, to demonstrate the utility of the proposal as can be seen in Section V. The findings of the case study are further discussed in the Section V-B. Some ideas on how we plan to further improve the framework has been written in Section VII.

The code and further details on the framework is available on our GitHub repository¹.

II. LITERATURE REVIEW

Vision Question Answering (VQA) models emerge as a fusion of both language and vision capabilities. Therefore, before proposing a novel method for VQA models, we survey the intricacies of how both language and vision models have been assessed independently. This holistic exploration ensures a well-rounded understanding of the challenges and methodologies associated with each domain, setting the stage for a nuanced and effective evaluation framework for VQA models.

For large language models, several studies have paved the way for understanding the nuances of subjective evaluation. Prior works have explored diverse approaches, including human judgment and user feedback, to appraise the efficacy of language models in capturing contextual nuances, coherence, and overall linguistic quality. Similarly, vision-specific models have undergone rigorous evaluation, with researchers emphasizing the importance of subjective assessments in discerning the visual fidelity, interpretability, and overall perceptual quality of generated content. Building upon this foundation, our research endeavors to propose a subjective scoring framework that draws inspiration from the discussed works.

In [24], Aldahdooh et al. discuss the necessity for subjective quality assessment to validate the performance of objective measures on visual data, with video processing technologies as the selected visual data in their study. Language models used in recent tasks have been trained on large corpora, often collected from large groups of people where subjective judgments differ among different social groups. Disproportionate representation of opinions might create undesirable outcomes from such language models. It is imminent there is an increased need for a framework which can assess language model predictions from a subjective analysis perspective.

Durmus et. al. [25] developed a framework to assess the quality of output from a Large Language Model. The study has conducted three main experiments. The first experiment gave the result that participants from several European countries and America, Canada, and Australia are closer than opinions from other countries, but the trained LLM model will have a significantly different, yet biased output which will not cater to a wide range of countries. The second experiment was to prompt the model to imbibe cultural diversity by considering opinions from China, and Russia which have complex yet rich cultural values. The third experiment con-

cluded that translation of different languages will not necessarily cover the context, which requires deeper knowledge of social contexts. A decoder-only transformer fine-tuned with Reinforcement Learning from Human Feedback (RLHF) was proposed and a similarity metric was calculated from the probabilities of the predicted answers. Language models are prone to assumptions and biases due to the use of diverse human conversations in the training phase, hence more research to mitigate potential biases/discriminations and qualitative assessment of predictions is needed.

Wu et. al. [26] studied text summarization and how the models perform in the effective capture of nuances, interestingness, comprehensiveness, and such specific dimensions of a summary that are of particular interest to a human reader. The automatic evaluation metrics like BLEU [19] and ROGUE [22] were studied in this experiment to identify the factors missed by the metrics, yet relevant to the context. The authors experimented with Diverse Role Player Evaluation (DRPE) to identify the quality of expressions from the summarized text. Role player-based evaluation is via voting and a DRPE score was calculated as a joint probability of vote count and reasoning. Role player-based evaluation is best suited when the model predictions are more text-oriented and will be directed toward multiple end-users. Recent research led to development of vast number of VQA algorithms and automated assessment of predictions is necessary to encompass important attributes like semantic similarity and subjectivity of the generated text.

In [21], consensus-based evaluation of Image Description is studied. The automated metric proposed by the authors uses Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n -gram for encoding. TF-IDF is a traditional encoding method for sentence similarity but lacks semantic understanding of the tokens in the sentence [27]. The frequency of n -grams in the candidate sentence is assessed against the reference sentence. Consensus-based protocol evaluates how often the humans validate the candidate sentences as 'similar' to the reference sentence. The model's performance was evaluated for sentence similarity using various aspects like grammaticality, saliency, and accuracy, but semantic closeness is not evaluated in this study.

Bashir et. al. [28] attempted subjective evaluation of answers using Machine Learning and Natural Language Processing. The Natural Language Processing methods like tokenization, stemming, lemmatization, and case folding were used as pre-processing steps on the subjective input prior to word2vec embedding and followed by stop words removal before the machine learning step. The similarity score of this subjective text is evaluated using Word Mover's Distance (WMD) or Cosine similarity. Further in this experiment, the authors train a Multi-Class Classifier, Multinomial Naïve Bayes for classifying the text data into four categories, identified based on the score obtained from the prediction module. A final score is predicted based on the classification value obtained and the overall score generated. Although the authors conclude with experimental values that cosine

¹The repository will be made public on acceptance

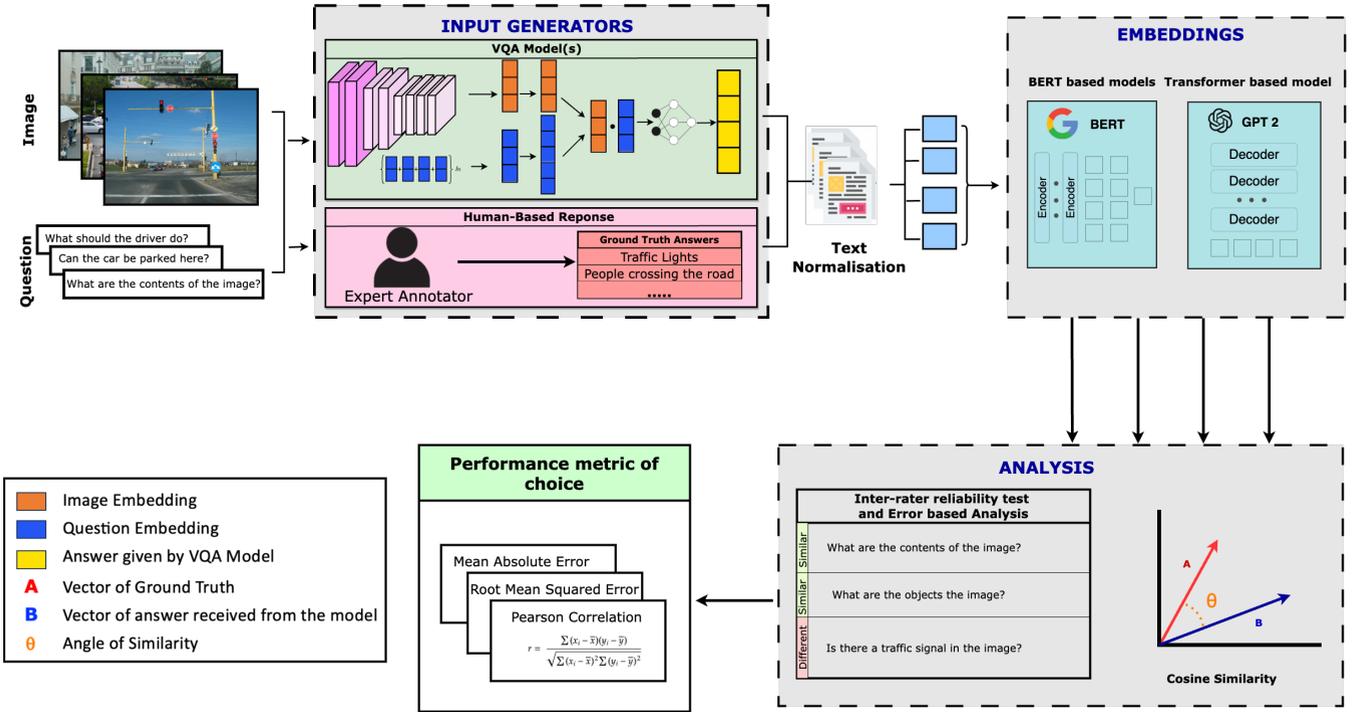


FIGURE 3. Architecture of the Subjective Scoring Assessment Framework

similarity do not perform well with semantic similarity as opposed to WMD and cosine similarity works better in cases where semantics are preserved well, their experimental steps did not use embedding methods that preserve semantic characteristics of the text during preprocessing.

The framework proposed emphasizes in retaining the semantic characteristics of the text during the embedding stage, to enable semantic closeness comparison using cosine similarity.

III. SUBJECTIVE SCORING EVALUATION ARCHITECTURE

In this section, we consider the architectural framework proposed for Subjective Scoring assessment, which is visually represented in Figure 3. We start by giving a general overview of the architecture before giving more details on each component.

In the initial phase of our architecture, we formulate specific questions about the subject matter depicted in an image. These questions aim to ascertain what’s happening in that image. To provide a basis for comparison, we ask human expert annotators to provide the correct answers to these questions as explained in Section III-A. The results of the poll and the questionnaire are available in Section V-A. Following this, we employ pre-trained VQA models to generate their own answers to these questions. A deeper discussion of the results observed from the poll can be found in [23].

All the questions, the answers obtained from humans (referred to as Ground Truth Answers), and the VQA model-generated answers are documented together. The Ground Truth answers serve a crucial role in our study. They help

us gauge whether the VQA model’s answers are contextually relevant to the image.

The pre-processing stage generally involves various text operations such as case normalization, tokenization, stopwords removal, lemmatization, and stemming. In our case, both the Ground Truth and Model-Generated Answers are often concise and limited to a few words, we opted for case normalization only, to streamline the data for all embedding models, ensuring consistency and manageability.

The embedding stage involves the application of pre-trained models, such as SBERT and GPT-2. SBERT (Sentence-BERT), a modified BERT framework designed for generating semantically related text embeddings for pairs of sentences [29]. Given our need for text embeddings for both Model-Generated Answers and Ground Truth Answers, the SBERT model streamlines this process. The utilization of SBERT greatly simplifies the simultaneous acquisition of text embeddings for these paired texts. Since the individual word embeddings for words in sentences are different, when we take the average or mean of the embeddings it will generate a different vector than the sentence vector. This is why we used SBERT which generates sentence embeddings and preserves the semantic context more efficiently.

We employed ‘bert-base-uncased,’ ‘nli-distilbert-base,’ and ‘all-mpnet-base-v2’ from the Sentence Transformer library (as discussed in Section III-D). These models generate text embeddings for both Ground Truth Answers and Model-Generated Answers. ‘bert-base-uncased’ is pre-trained on book corpora and English Wikipedia, ‘nli-distilbert-base’ on a substantial corpus of 570k human-written English sentence

pairs, and ‘all-mpnet-base-v2’ on 1 billion sentence pairs gathered from various domains [29]. These models are trained for diverse downstream tasks, including text similarity.

In our experiments, GPT-2 [30] was also employed to generate word embeddings for both Ground Truth Answers and Model-Generated Answers, facilitating the assessment of similarity between the two. There are many LLMs which are not free to use for comparative studies. We wanted to use models that are accessible to researchers without any additional costs. Another reason to use SBERT and GPT-2 is the computational cost, time and the specific tasks for which these models are trained on.

Upon acquiring the embeddings for both the Ground Truth answers and Model-Generated answers, we proceed to determine their similarity. To assess the similarity between the text embeddings, the cosine similarity measure was applied, yielding a score for each pair of Model-Generated Answers and Ground Truth Answers, quantified on a scale from 0 to 1.

The subsequent phase of this experiment entails a process known as subjective scoring (inter-rater reliability and error-based analysis) which involves the assignment of numerical scores for human-answers (ground truth), using a scale that ranges from 0 to 1 to evaluate all Model-Generated answers. These assigned scores serve as a benchmark for assessing the quality and appropriateness of the model-generated responses. Obtained cosine similarity scores are subsequently employed to establish the Pearson’s correlation between the human expert evaluations and the scores generated by the language models. Performance evaluation metrics, such as RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) are further employed to gauge the performance of the models in generating similarity scores for the textual data pairs as discussed in Section III-G.

To summarise, the proposed subjective scoring framework involves formulating questions about image content, obtaining human-annotated Ground Truth Answers, and using pre-trained models like SBERT and GPT-2 for generating text embeddings. The embeddings are assessed for similarity using cosine similarity, and subjective scoring assigns numerical scores (0 to 1) to human-answers and evaluates Model-Generated answers against them. Performance metrics like Pearson’s correlation, RMSE, and MAE are employed to assess the models’ performance in generating similarity scores for textual data pairs, enhancing the reliability of subjective scoring.

The major aspects used in the creation of the Subjective Scoring Framework have been discussed in the following subsections.

A. GROUND TRUTH HUMAN-BASED RESPONSES

We selected nine(9) VQA models after evaluating them for user interface quality, code replication ease, and compatibility with the selected pre-trained models. The initial experiment aimed to enhance the models’ performance using the driving related dataset, explicitly focusing on signboard interpretation [31]. However, the results revealed limited comprehen-

sion of driving-related matters by the models. This led us to conduct an additional experiment with 10 computer vision researchers, presenting them with contextually minimal images from our dataset, mirroring the approach used with the pre-trained models.

This approach allowed us to create a controlled environment for the evaluation, with a focus on the correlation between the visual content of the image, the posed question, and the most appropriate answer. The highest-rated answers, as determined through a poll conducted among the expert panel for each question posed, were designated as the ground truth for the dataset used which has been further discussed in Section V-A. It ensured that we introduced a significant human perspective into our research.

Humans are capable of nuanced understanding and reasoning, which might involve contextual, common sense, or background knowledge [32]. Humans also have the ability to generalize their understanding across various contexts and adapt to new scenarios [32]. Evaluating the performance of VQA models against human responses helped gauge the models’ ability to generalize and adapt to different visual scenes and questions. Additionally, by analyzing the discrepancies between human and model responses, at a later stage, we can identify areas where the models are underperforming and conduct targeted work on improving them.

B. VQA MODELS

In our proposed architecture, we examine three VQA models, namely ViLT, ViLBERT, and LXMERT in the Subjective Scoring Framework. These pre-trained models are used to generate predicted answers for a set of ten images from a driving-related dataset, which includes widely varying attributes in the scene. Here, we give a brief introduction of the VQA models used:

- **ViLT** (Vision and Language Transformer) commissions the transformer module to extract and process visual features in place of a separate deep visual embedder. This design leads to significant runtime and parameter efficiency. The authors fine-tuned ViLT-B/32 on the VQAv2 train and validation sets while reserving 1,000 validation images and their related questions for internal validation for the Visual Question Answering part of the model [14].
- **ViLBERT** (Vision-and-Language BERT) is a model developed for learning task-agnostic joint representations of image content and natural language [13]. BERT architecture is extended to a multi-modal two-stream model, processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers. To fine-tune ViLBERT on VQA, a two-layer MLP on top of the element-wise product of the image and text representations has been learned and mapped to 3,129 possible answers. The authors treat VQA as a multi-label classification task – assigning a soft target score to each answer based on its relevance to the 10 human answer responses. The VQA model is trained

with a binary cross-entropy loss on the soft target scores using a batch size of 256 over a maximum of 20 epochs.

- **LXMERT** (Learning Cross-Modality Encoder Representations from Transformers) is a large-scale Transformer model that consists of three encoders: an object relationship encoder, a language encoder, and a cross-modality encoder [12]. The model uses the Adam optimizer with a linear-decayed learning rate schedule and a peak learning rate at $1e - 4$. The model is trained for 20 epochs which is roughly 670K4 optimization steps with a batch size of 256. The pretraining of VQA tasks, however, is only for the last 10 epochs because this task converges faster and empirically needs a smaller learning rate.

C. TEXT NORMALISATION

Text normalization refers to the process of transforming text into a standardized or normalized form [33]. Text normalization helps ensure that different but semantically equivalent expressions are represented consistently. For example, converting all text to lowercase or applying stemming can help in capturing the same meaning across different forms of a word. By normalizing text, the dimensionality of the input space is reduced. This is particularly important when working with embeddings, as it helps in creating more compact and meaningful representations of text. In a typical VQA experiment, questions can be posed in various ways, and answers might have different forms. Normalizing text can help in handling these variations and making the model more robust to different input styles.

When applying SBERT or similar embedding models to VQA, it's common to preprocess both the questions and answers using text normalization techniques. These techniques may include lowercasing, stemming, lemmatization, removing stop words, and handling special characters. The goal is to create a standardized representation of the text data that captures the underlying semantic meaning and allows the embedding model to produce meaningful and consistent representations.

D. EMBEDDING MODELS

A word embedding is a learned representation for text where words that have the same meaning have a similar representation [34]. The framework employs four transformer-based models including three pre-trained Bert-based models and a GPT model to calculate embeddings, whose basic architectures are shown in Figure 4. BERT is a bidirectional transformer model pre-trained on extensive unlabelled text data from sources like Wikipedia and book corpora and is known for its robust performance in semantic textual similarity tasks, albeit with a higher computational cost [35]. However, there are various BERT-based models available, pre-trained on different corpora for various downstream tasks. The BERT-based models used in this study are bert-base-uncased, nli-distilbert-base, all-mpnet-base-v2, and another transformer model GPT-2.

The utilization of multiple embedding models presents several notable advantages. Firstly, diverse representations of textual data are facilitated by distinct embedding models, each capturing linguistic and semantic information through unique perspectives. This diversity enhances the representation of input text, encapsulating a broader spectrum of features and nuances within the data. Secondly, employing multiple models affords a comprehensive understanding of the input text, elucidating its intricacies from various analytical viewpoints. Each model emphasizes different aspects of the data, enriching the overall comprehension. Lastly, the amalgamation of embeddings from multiple models, known as ensemble learning, holds the potential to bolster performance and robustness. This amalgamation effectively mitigates biases or limitations inherent in individual models, resulting in an ensemble representation that leverages the strengths of each constituent model, thus enhancing the overall quality and reliability of the embedded data.

A short description of all the embeddings used is as follows:

- **BERT base uncased embedding:** BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based deep learning model that learns contextualized word representations. BERT-base-uncased refers to a specific variant of BERT that is the base model and is trained on uncased (lowercase) text. The 'uncased' aspect implies that the model ignores capitalization differences, making it appropriate for tasks where capitalization is not critical. BERT-base-uncased generates high-dimensional vector representations (embeddings) for words or tokens in the input text, capturing rich semantic and contextual information.
- **NLI-distilBERT base embedding:** DistilBERT is a distilled, smaller, and faster version of BERT while retaining a similar level of performance. NLI-distilBERT refers to a variant of DistilBERT that is specifically trained for Natural Language Inference (NLI) tasks. NLI involves determining the relationship between a given hypothesis and premise (usually entailment, contradiction, or neutral). The 'base' in 'NLI-distilbert-base' denotes the base architecture and size of the model.
- **mT5 (all-mpnet-base-v2) embedding:** 'all-mpnet-base-v2' is not a standard term related to mT5 (multilingual Translation Transformer), which is a transformer-based model designed for multilingual translation tasks. However, all-mpnet-base-v2 could refer to a particular variant or version of a model based on the MPNet (Multilingual Pre-trained language model) architecture. MPNet is a transformer model similar to BERT but designed for multilingual tasks.
- **GPT-2 embedding:** GPT-2 (Generative Pre-trained Transformer 2) is a transformer-based language model developed by OpenAI [30]. GPT-2 has received generative pre-training on a massive corpus of web text and it generates high-quality, coherent text based on the

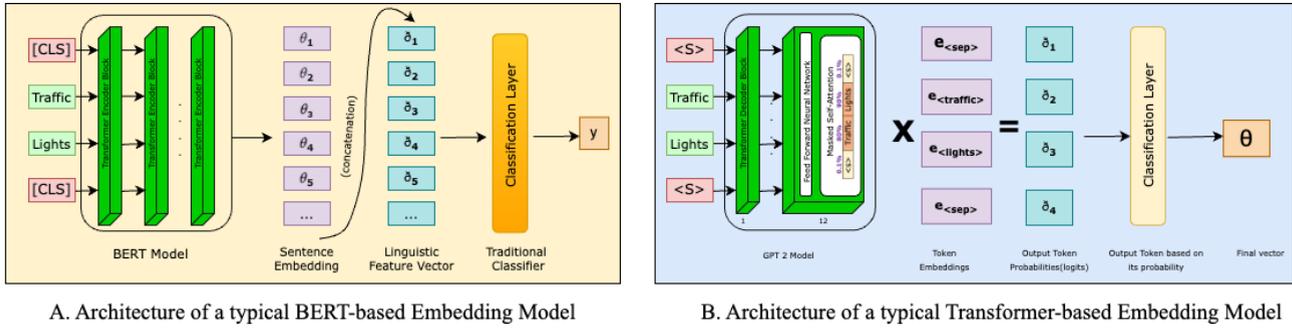


FIGURE 4. Basic Architecture of the embedding models used in the Framework

context provided in the input. GPT-2 embeddings refer to the vector representations generated by passing input text through the GPT-2 model. These embeddings capture contextual and semantic information and are widely used in various natural language generation tasks, creative writing, and more.

When employing vectors derived from multiple embeddings, a preservation of semantic nuances within textual content is observed. Consequently, the utilization of these vectors in the computation of cosine similarity leads to a more accurate evaluation of semantic affinity or proximity, surpassing the conventional word2vec embeddings and sentence similarity metrics.

The choice of specific natural language processing models in the framework is driven by their proven effectiveness in semantic tasks, i.e., their ability to capture semantic nuances and similarities between textual data pairs and availability for researchers without additional costs for reproducibility and further fine-tuning if needed for their respective projects. BERT, including bert-base-uncased, nli-distilBERT-base, and all-mpnet-base-v2, is selected for its bidirectional transformer architecture, pre-trained on diverse corpora, making them effective in understanding the context and semantics of subjective information. This is crucial for evaluating the subjectivity of Model-Generated answers in comparison to human-annotated Ground Truth Answers.

GPT-2 is included due to its capability to generate word embeddings, enabling the assessment of similarity between Ground Truth and Model-Generated answers. Its language generation capabilities make it relevant for assessing the appropriateness and quality of Model-Generated responses in a subjective scoring framework. These models collectively offer a comprehensive analysis of textual data pairs, balancing performance and accessibility in the subjective scoring framework.

E. COSINE SIMILARITY

Cosine similarity is a metric used to measure the similarity between two non-zero vectors of an inner product space. It measures the cosine of the angle between the vectors, indicating how closely they are related in terms of orientation [36]. Cosine similarity is often used in Natural Language

Processing and Information Retrieval tasks to determine the similarity between documents, words, or other text representations in a high-dimensional space [37].

In the context of VQA, cosine similarity is specifically valuable for evaluating the similarity or closeness between a model-generated answer and the ground truth. The formula used to calculate cosine similarity in our experiment is:

$$\begin{aligned} \text{similarity}(A, B) &= \cos(\theta) = \left(\frac{A \cdot B}{\|A\| \|B\|} \right) \\ &= \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2 \sum_{i=1}^n B_i^2}} \end{aligned}$$

where $A = \text{Ground Truth}$ $B = \text{Generated Answer}$

The reasons why cosine similarity is important in this scenario are:

- i. **Evaluation of Model Performance:** Cosine similarity helps evaluate how close the generated answer is to the correct or expected answer. A high cosine similarity indicates that the generated answer is similar to the reference answer, which is desirable in VQA tasks.
- ii. **Numerical Comparison of Vector Representation:** Cosine similarity provides a numerical measure of similarity between two vectors (representing answers in this case). The Transformer based models convert the input text data into high-dimensional vectors called embeddings. Cosine similarity measure is used to find the similarity between those vectors, allowing for a quantitative assessment of how well the generated answer aligns with the correct answer which is further discussed in section V-A.
- iii. **Optimization and Fine-Tuning:** VQA models can be fine-tuned using cosine similarity as a loss function. The goal here would be to optimize the model to generate answers that have high cosine similarity with the ground truth answers during training.

F. INTER-RATER RELIABILITY TEST AND ERROR-BASED ANALYSIS

In subjective scoring, inter-rater reliability serves as the statistical measure to assess the consistency between two or more raters or graders while evaluating the same data. It is the degree of similar or dissimilar judgements of different raters indicating the reliability of the measurement process across multiple graders. Inter-rater reliability not only enhances the validity and credibility of assessments but also guarantees quality, fosters fairness, and supports effective decision-making. Error-based analysis, such as RMSE, focuses on quantifying discrepancies between predicted and actual scores. It helps in continuous improvement, optimizing model performance, and ensuring fairness by identifying and rectifying biases.

Both approaches contribute to the reliability and precision of subjective assessments, with inter-rater reliability emphasizing human agreement and error-based analysis focusing on the accuracy and optimization of scoring models.

G. METRICS

The performance of VQA models can be tested using any metrics but we recommend using root mean squared error and mean absolute error. Using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to evaluate the performance of VQA models provides insights into the models' prediction accuracy by quantifying the discrepancy between predicted scores and ground truth scores. RMSE and MAE also help to understand the magnitude of errors in these predictions. A lower RMSE or MAE suggests that the model's predictions are closer to the true scores, indicating a more accurate VQA system. RMSE and MAE also allow for comparison of the VQA model's performance with other models or benchmarks. They provide a standardized metric that enables researchers to assess how well the model performs relative to established baselines or other state-of-the-art approaches. We also used Pearson's correlation measure in order to determine the degree of correlation between the two texts. A short description of the metrics is given below:

- Pearson Correlation is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is used to assess how closely two variables are related, and it provides a numerical value indicating the degree of correlation. In the context of VQA models, it helps to assess how closely the VQA model's predictions align with the actual answers or scores provided in the dataset.

$$r = \frac{\sum(P_i - \bar{P})(O_i - \bar{O})}{\sqrt{\sum(P_i - \bar{P})^2 \sum(O_i - \bar{O})^2}}$$

where r is the Pearson Correlation co-efficient, P_i is the VQA model's predicted answer scores, O_i is the Ground Truth answer scores, \bar{P} is the mean of the predicted answers' scores, and \bar{O} is the mean of the Ground Truth answers' scores.

- Root Mean Squared Error (RMSE) is a widely used statistical metric that measures the average magnitude of the errors or residuals between predicted values and actual (observed or ground truth) values in a dataset.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (P_i - O_i)^2}{N}}$$

where P_i is VQA models's prediction value, O_i is ground truth value and N is the total number of data points. RMSE gives higher weight to larger errors because of the squaring. It is sensitive to outliers and penalizes them more severely. RMSE is suitable when you want to account for the magnitude of errors and favor models that minimize larger errors.

- Mean Absolute Error (MAE) is a statistical metric used to measure the average magnitude of errors between predicted values and actual (observed or ground truth) values in a dataset.

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |P_i - O_i|$$

where P_i is VQA models's prediction value, O_i is ground truth value and N is the total number of data points. MAE treats all errors equally and does not give higher weight to larger errors. It is less sensitive to outliers and provides a more straightforward measure of average prediction error. MAE is often used when you want to assess the overall accuracy and reliability of predictions without emphasizing large errors.

The Subjective Scoring Framework designed using the discussed embedding models and cosine similarity enables researchers to compare and choose the most effective model based on the accuracy of embeddings in capturing semantic similarity. Additionally, it serves as a performance benchmark, quantifying the models' ability to represent the relationship between questions and images accurately. This approach also enriches the overall understanding of input features by leveraging diverse representations from different models. Analyzing differences in cosine similarity scores provides insights into model behavior and interpretation of questions and images.

IV. COMPUTATIONAL RESOURCES USED

The subjective framework has been developed using publicly available pre-trained models in our framework which is open access and can even be run on 2 CPU cores and 2.30 GHz of CPU Frequency. While the framework is available on our GitHub page, the VQA models used for the case study are not. The three VQA models used are open-access pre-trained models that have been inferenced on one Nvidia GeForce RTX 3080 GPU with 10240 CUDA cores, 12GB memory and 1.67 GHz CPU Frequency. The computational time required to run an inference for one question in ViLBERT was 27

seconds, ViLT was 26 seconds and for LXMERT, it was 23 seconds.

V. RESULTS AND DISCUSSION

A. CASE STUDY

1) Images Used

The Table 1 is the comprehensive list of the images, questions and answers used in the case study discussed. These images were selected from the MS COCO dataset [38]. It can be seen from the category column that we tried to keep diverse driving scenarios in mind while designing the questions. The rationale behind each of them is listed below:

- **Dark Setting:** Tests the model's ability to recognize and answer questions about objects, signs, or situations in low-light conditions, which may require understanding context or relying on limited visual cues.
- **Light Setting:** Evaluates the model's performance in scenarios with different lighting conditions, assessing whether it can adapt to varying levels of brightness and handle challenges like glare or reflections effectively.
- **Parking:** Tests the model's comprehension of parking-related questions, including identifying parking signs, understanding parking regulations, and recognizing parking lot layouts or parking manoeuvres.
- **Railway Crossing Signboards:** Assesses the model's capability to recognize and interpret railway crossing signs accurately, which is critical for understanding potential hazards and ensuring safety near railway tracks.
- **Pedestrian Crossing:** Evaluates the model's understanding of pedestrian crossings and its ability to answer questions related to pedestrian safety and traffic regulations in areas with pedestrian activity.
- **Traffic Scene:** Tests the model's comprehension of complex traffic scenes, including identifying vehicles, traffic signs, signals, and understanding traffic flow dynamics in different driving environments.
- **Accident:** Assesses the model's capability to recognize and interpret accident scenes, which may involve identifying damaged vehicles, assessing the severity of the situation, and understanding relevant signs or signals.
- **Road Signboards:** Evaluates the model's ability to recognize and interpret various road signs accurately, including speed limits, directions, warnings, and regulatory signs, across different scenarios.
- **Roadworks:** Tests the model's understanding of roadwork-related signs and situations, assessing its ability to recognize temporary changes to road conditions and navigate through work zones safely.
- **Men at Work in the Middle of the Road:** Assesses the model's comprehension of roadwork zones and its ability to recognize hazards posed by road workers or maintenance crews, requiring cautious navigation.
- **Fallen Signboard:** Tests the model's ability to identify and respond to unexpected obstacles or hazards on the road, such as fallen signboards, which may require adapting to changes in a driving scenario.

By testing VQA models with questions across these diverse categories, researchers can evaluate the models' generalization capabilities, robustness, and understanding of a wide range of driving scenarios, ultimately informing improvements in their performance and reliability for real-world applications.

2) Dataset collection

The authors conducted a survey consisting of two specific questions, namely "What are the contents of the image?" and "What should the driver do?", targeting the chosen set of images all pertaining to driving scenarios as described in Section V-A1.

The survey was distributed among a cohort of ten Computer Vision researchers who provided responses to the questions based on the available options and the accompanying images. The answer that received the most votes was selected as the ground truth. By establishing a consensus answer based on the majority vote for every question, a reference point has been created against which the accuracy of the model can be measured. This allows for quantitative assessment and comparison of different models.

In real-world scenarios, questions about images often have multiple valid interpretations or perspectives. For example, when asking "What should the driver do?" in the context of a driving scenario, different experts or people might provide varying but valid responses based on their individual viewpoints. Having diverse answers in the ground truth aligns with the complexity and diversity of human understanding and decision-making. However, to reduce the complexity of the experiment, the scope has been limited to only one Ground Truth answer per question. The framework will be able to handle multiple answers to one question when taken in as a superset instead of a subset as we are doing in the case study. To handle multiple answers, we need to aggregate the embeddings either by averaging or concatenating the embeddings of multiple ground truth answers to create a composite embedding representation for the question-image pair.

The rationale behind asking both subjective and objective questions, namely "What are the contents of the image?" and "What should the driver do?", is to assess the model's ability to comprehend and respond to different types of questions in the context of visual information. Questions like "What are the contents of the image?", require the model to understand and interpret the visual content and provide a descriptive answer. These questions evaluate the model's capability to recognize objects, scenes, and other relevant visual elements depicted in the image. A question like "What should the driver do?", require the model to provide a specific action or response based on the given visual information. These questions assess the model's understanding of driving scenarios and ability to reason about the appropriate course of action.

By including both subjective and objective questions, the experiment aims to evaluate different aspects of the model's performance. Subjective questions focus on the model's visual comprehension and scene understanding abilities, while

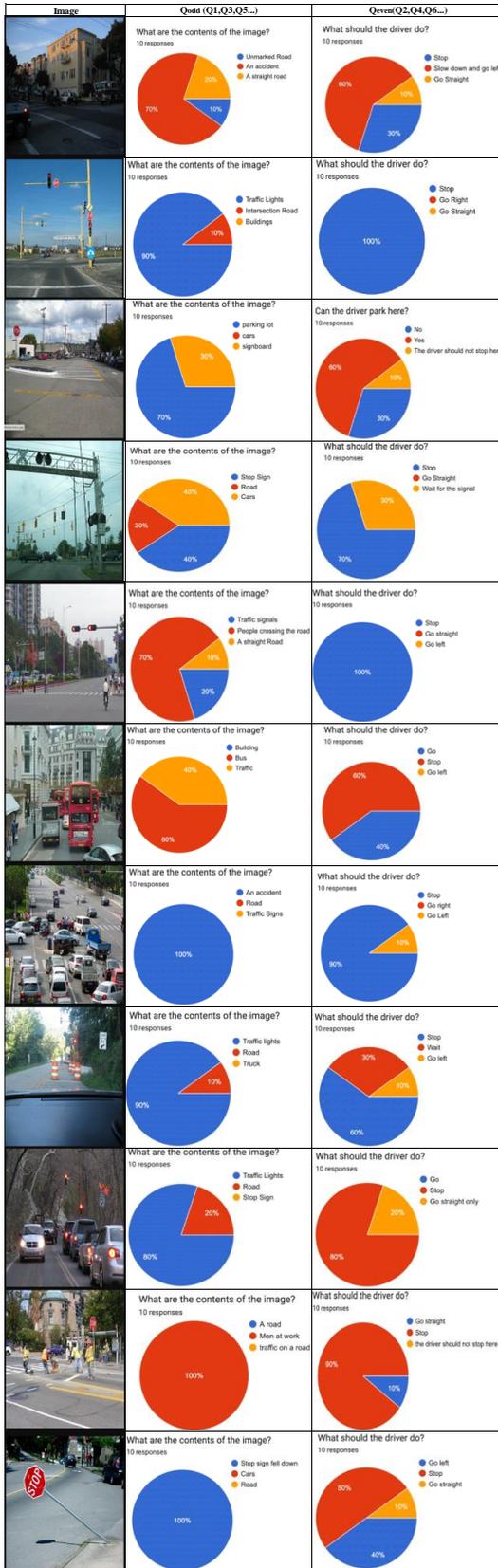


FIGURE 5. Ground Truth Poll Results

objective questions assess its ability to provide contextually appropriate and practical responses in a driving context. This comprehensive evaluation helps to gauge the model’s overall proficiency in visual question answering and its potential utility in real-world applications such as self-driving cars.

The answers received from the three VQA models discussed have also been listed. It provides a visual representation of the model’s performance in addressing the posed questions, allowing for an assessment of their effectiveness based on the ground truth. The rationale behind comparing the answers of three VQA models with human answers and using colour coding (green for correct, orange for wrong, yellow for partially correct) is twofold. Firstly, this approach visually highlights performance and discrepancies between the models and human responses. This visual representation allows for a quick and intuitive understanding of the accuracy and effectiveness of the models in comparison to human performance. Secondly, such an approach allows engineers to set upper and lower bounds on acceptable performance which in turn informs the classification stage of the modelling process. Further discussion of this rationale is considered in the paper ‘Towards a performance analysis on pre-trained Visual Question Answering models for autonomous driving’ [23].

As explained in the Section III-A, the chosen Ground Truth answers have been collected using a poll among Computer Vision researchers in the D2ICE Group ². The results of the poll have been shown in the Figure 5. As can be seen in the Table 1, every image was posed with 2 questions. The questions and the answers proposed have been shown in columns corresponding to the images in Figure 5. The ground truth for the respective questions was evaluated against the answers generated by the pre-trained models.

3) Data Preprocessing- Text Normalisation

Text normalization refers to the process of standardizing and simplifying text data to ensure that it is consistent and more easily processed by models [33]. Converting all text to lowercase in our experiment to ensure uniformity in capitalization, which is particularly useful for models like BERT-based and NLI-distilBERT-base that are case-insensitive. This can be observed in the Table 2 where the models are performing better after normalisation. It also ensures that the text input to the models is consistent, helping to avoid issues where the same word with different capitalization or punctuation is treated as different tokens. The models BERT, DistilBERT, and GPT-2 have predefined vocabularies. Text normalization helps ensure that words in the text match the tokens in the model’s vocabulary, preventing out-of-vocabulary issues.

TABLE 1. Comparison of Responses: Human Answers versus Selected Models

Index	Category	Image	Questions	Human Answers	ViLBERT	ViLT	LXMERT
Q1	Dark setting		What are the contents of the image?	An Accident	Trucks	Cars	Cars
Q2			What should the driver do?	Slow down and go left	Run	Stop	Go
Q3	Light Setting		What are the contents of the image?	Traffic Lights	Clouds	Traffic Lights	Power Lines
Q4			What should the driver do?	Stop	Sleep	Stop	Stop
Q5	Parking		What are the contents of the image?	Parking Lot	Clouds	Cars	Cars
Q6			Can the driver park here?	Yes	No	No	Yes
Q7	Signboard		What are the contents of the image?	Road	Paint	Traffic Lights	Cars
Q8			What should the driver do?	Stop	Sleep	Stop	Go
Q9	Pedestrian Crossing		What are the contents of the image?	People crossing the road	Clouds	Buildings	People
Q10			What should the driver do?	Stop	Run	Stop	Stop
Q11	Traffic		What are the contents of the image?	Traffic	Trucks	Buses	Buses
Q12			What should the driver do?	Go	Run	Stop	Go
Q13	Accident		What are the contents of the image?	An Accident	Clouds	Cars	Cars
Q14			What should the driver do?	Stop	Stop	Stop	Stop
Q15	Signboard		What are the contents of the image?	Road	Windows	Trees	Concrete
Q16			What should the driver do?	Wait	Sleep	Stop	Stop
Q17	Roadworks		What are the contents of the image?	Traffic Lights	Trucks	Cars	Cars
Q18			What should the driver do?	Stop	Stop	Stop	Go
Q19	Men at work		What are the contents of the image?	Men at work	Trucks	People	Cars
Q20			What should the driver do?	Stop	Sleep	Stop	Stop
Q21	Fallen Signboard		What are the contents of the image?	Stop sign fell down	Trucks	Stop sign	Cars
Q22			What should the driver do?	Go left	Eat	Stop	Stop

TABLE 2. Results of performance for the three VQA models- ViLT, ViLBERT, and LXMERT by using the metrics- Mean Absolute Error, Root Mean Squared Error, and Pearson Correlation before and after text normalisation of the dataset used in the Case Study

Before Normalisation:

	Mean Absolute Error			Root Mean Squared Error			Pearson Correlation		
	ViLBERT	ViLT	LXMERT	ViLBERT	ViLT	LXMERT	ViLBERT	ViLT	LXMERT
NLI-distilbert-base	0.5581	0.4901	0.5241	0.7698	0.7743	0.7427	0.7686	0.6495	0.6588
all-mpnet-base-v2	0.2814	0.2955	0.3961	0.5694	0.6241	0.5697	0.8286	0.7447	0.8215
BERT-base-uncased	0.6464	0.4937	0.5124	0.8264	0.7774	0.7777	0.7077	0.6109	0.5183
GPT2	0.8999	0.6632	0.6633	0.9715	0.8937	0.8929	0.2522	0.3903	0.1952

After Normalisation:

	Mean Absolute Error			Root Mean Squared Error			Pearson Correlation		
	ViLBERT	ViLT	LXMERT	ViLBERT	ViLT	LXMERT	ViLBERT	ViLT	LXMERT
NLI-distilbert-base	0.5581	0.4901	0.5241	0.7698	0.7743	0.7427	0.7686	0.6495	0.6588
all-mpnet-base-v2	0.2814	0.2955	0.3961	0.5694	0.6241	0.5697	0.8286	0.7447	0.8215
BERT-base-uncased	0.6636	0.4976	0.5259	0.8375	0.7822	0.7841	0.6657	0.6042	0.5321
GPT2	0.9008	0.6617	0.6612	0.9720	0.8926	0.8914	0.3027	0.5497	0.3913

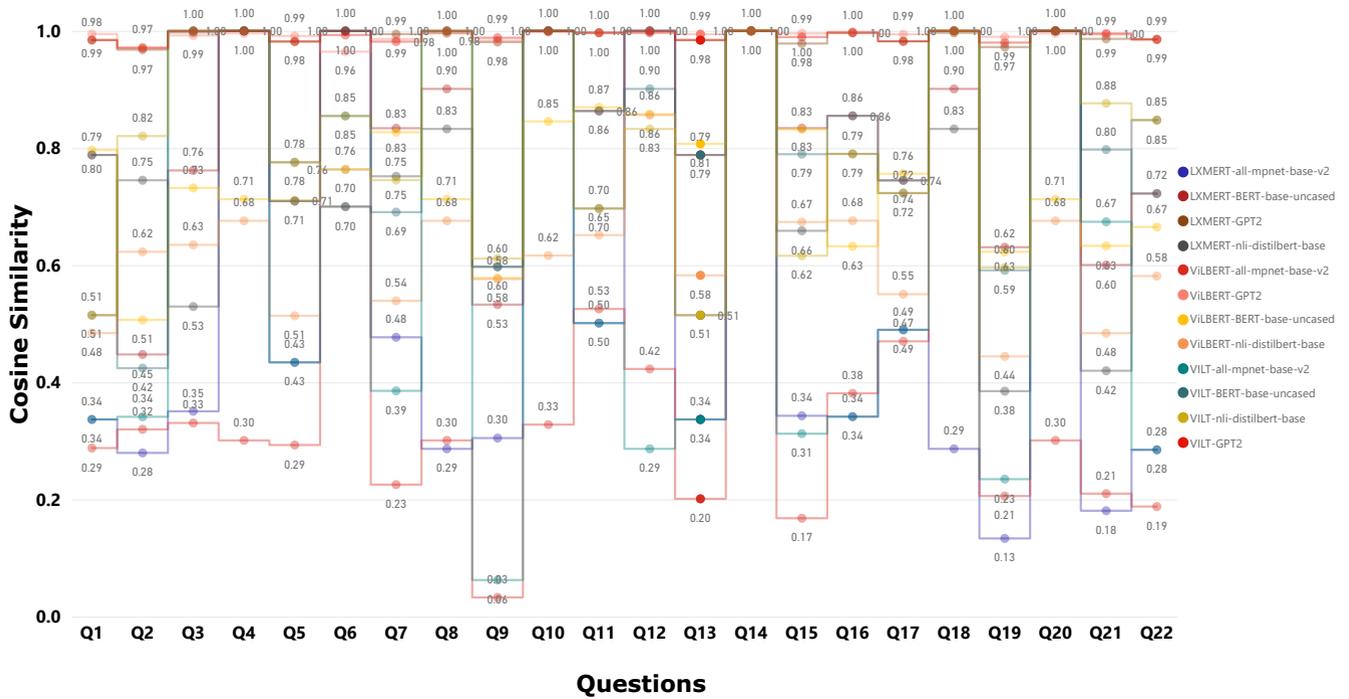


FIGURE 6. Analyzing Semantic Consistency: A graph of Cosine Similarity Values of predicted answers by the chosen VQA Models- ViLBERT, ViLT, and LXMERT with Ground Truth Across Questions in the Case Study

B. DISCUSSION

The experiment has been divided into two stages, generating a similarity score for the predicted and Ground Truth Responses using Cosine Similarity for the chosen VQA models as shown in Figure 6 and then in the further step, the subjective score given by the humans is evaluated against the obtained similarity score using the error-based methods (RMSE and MAE). Pearson Correlation is used to assess how closely the predicted answers and ground truth answers are related as shown in Figure 7.

It must be noted that all the values of cosine similarity lie in the range [-1,1]. A higher cosine value implies that the predicted answer by the model is closer to the ground truth and vice versa. In Figure 6, it can be observed that GPT-2 has given the most unreliable results of all the embedding models used. The reason for this is that GPT-2 has not been trained for text similarity or contextuality tasks. We fine-tuned all the embedding models used in the framework for text-similarity, it is noted that GPT-2 still performed poorly. It is primarily a generative model trained on large text corpora for tasks like text completion and generation. While it can capture some level of context, it does not perform well in tasks that require understanding and representing intricate relationships between multiple pieces of text while comparing them. Additionally, text similarity and contextuality tasks involve

consideration of linguistic ambiguity, understanding subtle nuances, disambiguating complex sentences and assessment of clustering that may occur in relation to expert responses. GPT-2, like many language models, may struggle with such challenges. This has also been highlighted in the values shown in Tables 3 and 4. However, it can be clearly seen from the results presented in Figure 7 that if embedding models are giving bad values (like GPT-2), our framework is able to identify those anomalies (like false-positives) as shown in the low Pearson Correlation values of all the three models; ViLT (0.549661), ViLBERT (0.302661) and LXMERT (0.391322).

Objective questions often require a more factual understanding and may benefit from embeddings that capture detailed semantic information. On the other hand, subjective questions may involve more nuanced understanding and context, where different embeddings might excel. The mean of all the cosine values to the objective questions (Table 3) helps us determine that ViLT has performed well with BERT-base-uncased embeddings (0.76) whereas for subjective questions (Table 4), it has been observed that ViLT performed better with NLI-distilbert-base embeddings (0.9140). DistilBERT models, including NLI-distilbert, are often designed to be more computationally efficient while retaining much of the performance of the larger BERT models. If the task involves natural language inference as is the case for subjective questions, the embeddings from NLI-distilbert might be particularly well-suited for capturing the relationships between premises and hypotheses. BERT embeddings are better suited for capturing the nuances and details in the context of objec-

²The D²iCE (Data-Driven Computer Engineering) research group is a specialized hub that focuses on the convergence of academia and industry to address real-world challenges using AI and ML. For more info: <https://www.d2ice.ie/>

TABLE 3. Analyzing Semantic Consistency of objective questions: Cosine Similarity Values of predicted answers by the chosen VQA Models- ViLBERT, ViLT, and LXMERT with Ground Truth Across Questions in the Case Study

Index	ViLBERT				ViLT				LXMERT			
	NLI-distilbert-base	all-mpnet-base-v2	BERT-base-uncased	GPT2	NLI-distilbert-base	all-mpnet-base-v2	BERT-base-uncased	GPT2	NLI-distilbert-base	all-mpnet-base-v2	BERT-base-uncased	GPT2
Q1	0.4840	0.2878	0.7966	0.9948	0.5147	0.3366	0.7882	0.9845	0.5147	0.3366	0.7882	0.9845
Q3	0.6348	0.3306	0.7319	0.9922	1.0000	1.0000	1.0000	1.0000	0.5294	0.3505	0.7618	0.9972
Q5	0.5136	0.2929	0.7111	0.9914	0.7757	0.4340	0.7094	0.9819	0.7757	0.4340	0.7094	0.9819
Q7	0.5391	0.2253	0.8271	0.9862	0.7453	0.3854	0.6906	0.9817	0.7518	0.4769	0.8338	0.9946
Q9	0.5763	0.0326	0.5781	0.9832	0.6114	0.0624	0.5973	0.9883	0.5972	0.3048	0.5327	0.9810
Q11	0.6514	0.5252	0.8696	0.9963	0.6968	0.5011	0.8633	0.9970	0.6968	0.5011	0.8633	0.9970
Q13	0.5827	0.2011	0.8071	0.9944	0.5147	0.3366	0.7882	0.9845	0.5147	0.3366	0.7882	0.9845
Q15	0.6738	0.1679	0.8318	0.9958	0.6161	0.3123	0.7896	0.9895	0.6587	0.3428	0.8340	0.9785
Q17	0.5504	0.4700	0.7562	0.9944	0.7231	0.4897	0.7448	0.9823	0.7231	0.4897	0.7448	0.9823
Q19	0.4442	0.2059	0.6228	0.9899	0.5959	0.2346	0.5912	0.9801	0.3849	0.1334	0.6307	0.9725
Q21	0.4839	0.2099	0.6330	0.9939	0.8764	0.6741	0.7974	0.9956	0.4198	0.1807	0.6003	0.9864
Mean	0.5577	0.2681	0.7423	0.9920	0.6973	0.4333	0.7600	0.9878	0.5970	0.3534	0.7352	0.9855

tive questions, resulting in a higher mean cosine similarity for ViLT performance. It's understandable for different embeddings to excel in different aspects of language understanding, and the optimal choice may vary depending on the specific context and goals of the task at hand.

Models like BERT and its variants have been fine-tuned for tasks like semantic textual similarity and context-based question answering which can be seen in our results as well. Considering the current dataset of images and questions in the case study, it is observed that LXMERT with all-mpnet-base-v2 embeddings is the best model for the experiment which has been shown in the graph in Figure 6.

This assertion gains additional credence when we consider the comprehensive evaluation of the models using key metrics- Pearson Correlation, RMSE, and MAE, as visualized in Figure 7. LXMERT equipped with all-mpnet-base-v2 embeddings stands out as the top-performing model for the conducted experiment. A high positive value of the Pearson correlation coefficient (0.82150878) in VQA indicates a strong positive linear relationship between the model's predicted answers and the ground truth answers. This suggests that the model's answers agree with the ground truth, and the model is performing well in terms of providing accurate responses to the questions posed about images. Consequently, a low MAE (0.39607018) and RMSE (0.56971988) indicate that the model's responses are accurate and exhibit minimal deviation from the ground truth answers. Finer precision values of all our results can be referred at our GitHub page.

VI. LIMITATIONS AND BIASES

The proposed subjective scoring framework and its application in VQA models have some limitations and potential biases that have been considered and mitigated:

- **Human Annotation Bias:** The reliance on human-annotated Ground Truth Answers introduces the potential for bias based on individual annotators' perspectives. Variability in human interpretation and subjectivity may influence the benchmark against which Model-Generated answers are evaluated. To overcome this lim-

itation, we used multiple human annotators and selected the most voted answer to mitigate the impact of individual subjectivity, aiming for a more robust and reliable benchmark as discussed in Figure 5 .

- **Subjective Nature of Scoring:** Subjective scoring itself introduces a level of ambiguity and subjectivity. The assignment of numerical scores by human annotators may vary, leading to potential inconsistencies. Inter-rater reliability measures can mitigate this, but some subjectivity is inherent in the scoring process.
- **Text Embedding Techniques:** Different embedding models may capture semantic information differently, and the chosen models may not perfectly represent the complexity of subjective language, impacting the overall effectiveness of similarity assessment. However, we attempt to offer a comprehensive analysis of textual data pairs, balancing performance and accessibility in the subjective scoring framework by using both GPT-2 and BERT.
- **Computational Cost:** The use of complex transformer models like BERT and GPT-2 comes with high computational costs, limiting scalability for large datasets or real-time applications. This could hinder the practicality of the framework in certain contexts. To mitigate this, we have used publicly available pre-trained models in our framework which is open access and can even be run on 2 CPU cores and 2.30 GHz of CPU Frequency. The framework is accessible in the form of a Google Colab notebook on our GitHub page.

VII. CONCLUSION AND FUTURE WORK

In this work, a framework that can analyse the predictions generated by VQA models based on subjective and semantic attributes of the answers has been developed. The subjective analysis is aggregated using four(4) types of embeddings from natural language processing models and sentence similarity benchmark metrics.

Subjective questions typically involve nuanced aspects including contextual dependencies, the existence of multiple

TABLE 4. Analyzing Semantic Consistency of subjective questions: Cosine Similarity Values of predicted answers by the chosen VQA Models- ViLBERT, ViLT, and LXMERT with Ground Truth Across Questions in the Case Study

Index	ViLBERT				ViLT				LXMERT			
	NLI-distilbert-base	all-mpnet-base-v2	BERT-base-uncased	GPT2	NLI-distilbert-base	all-mpnet-base-v2	BERT-base-uncased	GPT2	NLI-distilbert-base	all-mpnet-base-v2	BERT-base-uncased	GPT2
Q2	0.6229	0.3196	0.5066	0.9719	0.8205	0.3411	0.4242	0.9709	0.7450	0.2795	0.4475	0.9679
Q4	0.6758	0.3008	0.7127	0.9958	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Q6	0.7635	0.7000	0.8548	0.9646	0.7635	0.7000	0.8548	0.9931	1.0000	1.0000	1.0000	1.0000
Q8	0.6758	0.3008	0.7127	0.9958	1.0000	1.0000	1.0000	1.0000	0.8327	0.2864	0.9010	0.9966
Q10	0.6164	0.3277	0.8455	0.9977	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Q12	0.8564	0.4228	0.8579	0.9990	0.8327	0.2864	0.9010	0.9966	1.0000	1.0000	1.0000	1.0000
Q14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Q16	0.6760	0.3812	0.6322	0.9986	0.7900	0.3414	0.8551	0.9968	0.7900	0.3414	0.8551	0.9968
Q18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8327	0.2864	0.9010	0.9966
Q20	0.6758	0.3008	0.7127	0.9958	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Q22	0.5814	0.1879	0.6652	0.9851	0.8477	0.2848	0.7223	0.9856	0.8477	0.2848	0.7223	0.9856
Mean	0.7404	0.4765	0.7727	0.9913	0.9140	0.7231	0.8870	0.9948	0.9135	0.6799	0.8934	0.9949

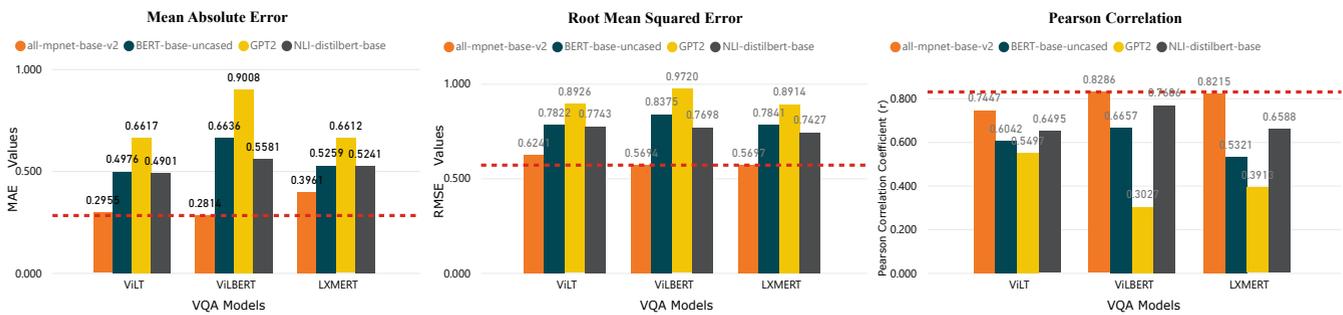


FIGURE 7. Comprehensive overview of performance for the three VQA models- ViLT, ViLBERT and LXMERT by using the metrics- Mean Absolute Error, Root Mean Squared Error and Pearson Correlation. The red dotted line is used to show the most significant values obtained.

'correct' answers, the clustering of responses around different correct answers and a consideration of individual preferences in terms of autonomous driving. The framework proposed is instrumental in discerning the areas where the models might fail, guiding subsequent refinement efforts to improve their performance. In real-world applications, users frequently pose subjective questions that necessitate a good comprehension of context and consequently warrant nuanced responses.

It has been noted that even a question like 'what are the contents of the image?' has a great deal of subjectivity, which could become much more of an issue when case study responses are taken from a larger, more general subset of the population. As humans, we tend to say what the contents are in the context of the scenario. For example, in a driving scene, we will look at the road contents, rather than saying 'buildings and windows' in their background. However, this is not the same for a VQA model which depends on its' object detection algorithm for the contents of the image presented. The authors intend to conduct a simulated study that will have human researchers drive in the same environment as the pictures provided to the VQA model and observe what the humans perceive when they are in the driving seat and how similar or different it is to the observations of a VQA model. This will help us assess on training the object detection part of

a VQA model exactly to the needs of a driver when presented with a driving scenario. The authors also intend to consider how this ambiguity can be formally addressed by assessing the optimisation question that arises when a vector of cosine similarities is computed with reference to a range of expert responses.

The design of VQA systems for driving should invariably prioritize the end-users and their expectations of obtaining coherent and contextually relevant responses. The framework proposed contributes to meeting these different and sometimes diverse expectations, leading to a methodology that can support a diverse range of meaningful VQA interactions.

Furthermore, standardizing the evaluation process facilitates benchmarking and comparative analyses across different VQA models. Researchers can leverage this framework to scrutinize model enhancements, propelling advances within the evolving landscape of VQA models.

For future work, we are actively applying the proposed framework to assess the performance of various VQA models, including both pre-trained and finetuned models using a driving dataset (Nuscenes). The incorporation of the subjective scoring framework enhances our ability to conduct a nuanced and thorough analysis of a VQA model's effectiveness in

capturing the details of visual information³ related to driving scenarios.

We are also exploring different applications. For example, we are also working on using the framework to assess a VQA model that can detect defects on mobile screens, for instance, scratches, cracks, etc. We are also working on using the framework to assess a VQA model that can detect defects on mobile screens, for instance, scratches, cracks, etc. The objective of the project is to finetune the training and validation process for the specific question of screen defect detection so that there is a rigorous parameter-based data collection procedure for handling false positives or negatives that may be identified by a model in a real-time setting.

REFERENCES

- [1] L. Chen, Y. Li, C. Huang, B. Li, Y. Xing, D. Tian, L. Li, Z. Hu, X. Na, Z. Li, S. Teng, C. Lv, J. Wang, D. Cao, N. Zheng, and F.-Y. Wang, "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1046–1056, 2023.
- [2] L. Chen, Y. Li, C. Huang, Y. Xing, D. Tian, L. Li, Z. Hu, S. Teng, C. Lv, J. Wang, D. Cao, N. Zheng, and F.-Y. Wang, "Milestones in autonomous driving and intelligent vehicles—part i: Control, computing system design, communication, hd map, testing, and human behaviors," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 9, pp. 5831–5847, 2023.
- [3] L. Chen, S. Teng, B. Li, X. Na, Y. Li, Z. Li, J. Wang, D. Cao, N. Zheng, and F.-Y. Wang, "Milestones in autonomous driving and intelligent vehicles—part ii: Perception and planning," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 10, pp. 6401–6415, 2023.
- [4] D. Fernandez Llorca and E. Gomez Gutierrez, "Trustworthy autonomous vehicles," Publications Office of the European Union, Luxembourg, Tech. Rep. EUR 30942 EN - JRC127051, 2021.
- [5] X. He, W. Huang, and C. Lv, "Toward trustworthy decision-making for autonomous vehicles: A robust reinforcement learning approach with safety guarantees," *Engineering*, 2023.
- [6] Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *International Journal of Computer Vision*, vol. 130, pp. 2425–2452, 2022.
- [7] J. Chen, S. E. Li, and M. Tomizuka, "Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5068–5078, 2022.
- [8] L. Cultrera, F. Becattini, L. Seidenari, P. Pala, and A. D. Bimbo, "Explaining autonomous driving with visual attention and end-to-end trainable region proposals," *Journal of Ambient Intelligence and Humanized Computing*, 2023.
- [9] T. Deruyttere, S. Vandenhende, D. Grujicic, L. Van Gool, and M. F. Moens, "Talk2car: Taking control of your self-driving car," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 2088–2098.
- [10] S. Kumar, B. N. Patro, and V. P. Namboodiri, "Auto qa: The question is not only what, but also where," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2022, pp. 272–281.
- [11] S. Atakshiyev, M. Salameh, H. Babiker, and R. Goebel, "Explaining autonomous driving actions with visual question answering," in *IEEE Conference on Intelligent Transportation Systems*, 2023.
- [12] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019.
- [13] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.
- [14] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.
- [15] G. Chochlakis, T. Srinivasan, J. Thomason, and S. Narayanan, "Vault: Augmenting the vision-and-language transformer with the propagation of deep language representations," *arXiv preprint arXiv:2208.09021*, 2022.
- [16] Z. Tang, J. Cho, Y. Nie, and M. Bansal, "Tvl: Textless vision-language transformer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9617–9632, 2022.
- [17] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" *Computer Vision and Image Understanding*, vol. 163, pp. 90–100, 2017.
- [18] D. T. Beckler, Z. C. Thumser, J. S. Schofield, and P. D. Marasco, "Reliability in evaluator-based tests: using simulation-constructed models to determine contextually relevant agreement thresholds," *BMC medical research methodology*, vol. 18, pp. 1–12, 2018.
- [19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [20] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.
- [21] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
- [22] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
- [23] K. Rekanar, C. Eising, G. Sistu, and M. Hayes, "Towards a performance analysis on pre-trained visual question answering models for autonomous driving," in *Proceedings of the Irish Machine Vision and Image Processing Conference*, 2023.
- [24] A. Aldahdooh, E. Masala, G. Van Wallendael, P. Lambert, and M. Barkowsky, "Improving relevant subjective testing for validation: Comparing machine learning algorithms for finding similarities in vqa datasets using objective measures," *Signal Processing: Image Communication*, vol. 74, pp. 32–41, 2019.
- [25] D. Esin, N. Karina, L. T. I. S. Nicholas, A. Amanda, B. Anton, C. Carol, H.-D. Zac, H. Danny, and J. N. et al., "Towards measuring the representation of subjective global opinions in language models," *arXiv preprint arXiv:2306.16388*, 2023.
- [26] W. Ning, G. Ming, S. Linjun, L. Shining, and J. Daxin, "Large language models are diverse role-players for summarization evaluation," *arXiv preprint arXiv:2303.15078*, 2023.
- [27] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1. Citeseer, 2003, pp. 29–48.
- [28] B. M. Farrukh, A. Hamza, J. A. Rehman, K. Natalia, and B. S. S., "Subjective answers evaluation using machine learning and natural language processing," *IEEE Access*, vol. 9, pp. 158 972–158 983, 2021.
- [29] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019.
- [30] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, and J. Wang, "Release strategies and the social impacts of language models," 2019.
- [31] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The german traffic sign recognition benchmark: a multi-class classification competition," in *The 2011 international joint conference on neural networks*. IEEE, 2011, pp. 1453–1460.

³In the context of VQA, "visual information" refers to the images or visual content that the model is designed to understand and interpret. It includes details, features, and characteristics present in images related to driving scenarios, such as objects, road conditions, or any visual elements that are relevant to the task at hand. The VQA model processes this visual information to provide answers or responses based on the questions posed to it.

- [32] H. Mercier and D. Sperber, "Why do humans reason? arguments for an argumentative theory," *Behavioral and brain sciences*, vol. 34, no. 2, pp. 57–74, 2011.
- [33] D. Jurafsky and J. H. Martin, "Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing," *Upper Saddle River, NJ: Prentice Hall*, 2008.
- [34] Y. Goldberg and G. Hirst, "Neural network methods in natural language processing. morgan & claypool publishers (2017)," *zitiert auf*, p. 69, 2017.
- [35] D. Jacob, L. Kenton, C. Ming-Wei, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.
- [36] D. Gunawan, C. Sembiring, and M. A. Budiman, "The implementation of cosine similarity to calculate text relevance between two documents," in *Journal of physics: conference series*. IOP Publishing, 2018, p. 012120.
- [37] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *2016 4th International Conference on Cyber and IT Service Management*. IEEE, 2016, pp. 1–6.
- [38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.