Towards Multilingual spoken Visual Question Answering System using Cross-Attention

Amartya Roy Chowdhury[⋆]
IIT Dharwad, India
amartya.chowdhury@iitdh.ac.in

Tonmoy Rajkhowa[⋆]
IIT (BHU) Varanasi, India
tonmoyrajkhowa.rs.ece24@itbhu.ac.in

Sanjeev Sharma
IIT (BHU) Varanasi, India
sanjeev.ece@iitbhu.ac.in

Abstract

Visual question answering (VQA) poses a multi-modal translation challenge that requires the analysis of both images and questions simultaneously to generate appropriate responses. Although VQA research has mainly focused on text-based questions in English, speech-based questions in English and other languages remain largely unexplored. Incorporating speech could significantly enhance the utility of VQA systems, as speech is the primary mode of human communication. To address this gap, this work implements a speech-based VQA system and introduces the textless multilingual visual question answering (TM-VQA) dataset, featuring speech-based questions in English, German, Spanish, and French. This TM-VQA dataset contains 658,111 pairs of speech-based questions and answers based on 123,287 images. Finally, a novel, cross-attention-based unified multi-modal framework is presented to evaluate the efficacy of the TM-VQA dataset. The experimental results indicate the effectiveness of the proposed unified approach over the cascaded framework for both text and speech-based VQA systems. Dataset can be accessed at https://github.com/Synaptic-Coder/TM-VQA.

1 Introduction

Recent focuses of the computer vision (CV) community have shifted its attention toward addressing challenges at the intersection of CV and natural language processing (NLP) (Voigt et al., 2021; Wiriyathammabhum et al., 2016). Visual question answering (VQA) (Antol et al., 2015) is a key domain that requires the participation of these fields. VQA system analyzes the visual content of an image, in conjunction with related queries, to generate relevant responses. It has a wide range of applications, including human-computer interaction (Li et al., 2022; Gao et al., 2022), content retrieval (Ding et al., 2024; Zhang et al., 2024), healthcare (Wu et al., 2022; Zhan et al., 2020), and surveillance (Toor et al., 2019). However, current VQA systems are often constrained by their reliance on text-based questions to produce answers. Given that speech is the fundamental and primary form of human communication, a speech-based VQA system may offer significant advantages over its conventional text-based counterpart. It may provide hands-free operation for users in situations where typing is impractical, such as while driving or in smart home settings. The ability to document responses as text also ensures that users retain the clarity and precision of visual information while benefiting from the convenience of spoken queries. Thus, this speech-based VQA system bridges the gap between natural human communication and machine understanding.

Implementing a spoken VQA system requires training using a dataset that comprises triplets of visual images, spoken questions based on those visuals, and textual answers. Due to the inexistence of datasets that meet these specifications, this work expands the VQA v2.0 dataset (Antol et al., 2015) due to its high-quality content and size which reflects real-world scenarios. Given that English-based VQA systems have garnered attention so far (Zhu et al., 2016; Krishna et al., 2017), this work takes a step forward to address this gap by translating the textual questions in three languages: German, French, and Spanish, using a machine translation (MT) model. Subsequently, these translations are synthesized into spoken form using a multilingual text-to-speech (TTS) system (Hayashi et al., 2021). This resulted in the creation of a textless multilingual visual question answering (TM-VQA) dataset containing 2.6M question-answer pairs (658k pairs for each language), derived from 123K images depicting real-life scenarios. To evaluate the efficacy of the TM-VQA dataset, for both text and speech-based VQA systems, the test set is further divided into binary "Yes/No", numerical answer type, and

[⋆]These authors have contributed equally to this work.

Proceedings of the 31st International Conference on Computational Linguistics, pages 9165–9175
January 19–24, 2025. ©2025 Association for Computational Linguistics

multi-class open-ended questions separately.

The cross-attention (CA) (Vaswani et al., 2017) mechanism in the Transformer (Vaswani et al., 2017) plays a crucial role in enabling effective interactions between two distinct modalities or input sequences. Unlike self-attention (SA), which models relationships within a single input sequence, CA attends and models the relationship between two distinct sequences. This approach is particularly useful when combining information from different sources, such as, from text or audio with images. This CA can focus on specific regions of the image relevant to the query, extracting pertinent details by aligning visual and textual representations. Thus, this work incorporates CA between the image and text or speech embeddings and provides a comparative analysis to demonstrate its effectiveness and relevance in VQA systems.

In summary, the primary contributions of this work are:

- Introduction of TM-VQA, a multilingual VQA dataset containing 2.6M question-answer pairs (658K for each language), derived from 123K images textual and spoken queries in four diverse languages: English (En), German (De), French (Fr) and Spanish (Es) to facilitate the development of text and speech-based VQA systems.

- A novel multi-modal unified framework employing cross-attention to analyze multilingual audio and image representations to generate responses.

- The performance of this unified system is investigated by incorporating various audio, image, and text features extracted from pre-trained state-of-the-art models and compared with cascaded ASR + VQA baselines. The results obtained claim superior performance in terms of accuracy for all categories of the test set over the cascaded system.

2 Related Works

2.1 VQA datasets

The landscape of VQA systems evolved significantly with the introduction of various datasets to cater different aspects of VQA tasks (Gao et al., 2018; Li et al., 2018; Goyal et al., 2017; Marino et al., 2019; Liang et al., 2024; Singh et al., 2019; Desta et al., 2018; Gokhale et al., 2020; Gao et al., 2024; Goel et al., 2021; He et al., 2020; Rajkhowa et al., 2024b). VQA v1.0 (Antol et al., 2015), derived from the Microsoft COCO dataset, features a balanced question and free-form answer. DAQA (Fayek and Johnson, 2020) utilizes images collected from NYU-Depth V2 (Silberman et al., 2012) that focuses on indoor scenes with questions in natural language. CLEVR (Johnson et al., 2017) incorporates synthetic objects and complex questions to assess visual reasoning abilities. GQA (Hudson and Manning, 2019) emphasizes compositional reasoning and diverse relationships between objects. VizWiz (Bigham et al., 2010; Gurari et al., 2018), containing images and questions from blind users, uniquely addresses accessibility challenges. TDIUC (Kafle and Kanan, 2016, 2017) serves as a diagnostic dataset to highlight the limitations of existing VQA models. However, these datasets were domain-specific and relatively smaller in size. In contrast, VQA v2.0 (Antol et al., 2015) offers several advantages over its predecessors by introducing a more balanced distribution of question types. Additionally, this dataset maintains high-quality images covering real-world scenarios. Despite the improvements, these datasets facilitated the development of only text-based VQA systems.

The SBVQA dataset (Zhang et al., 2017) is the first to enable the development of a spoken VQA system. It consists of 200 hours of synthetically generated spoken questions and 1 hour of human-recorded speech in English. Similarly, the fact-based VQA (FVSQA) (Ramnath et al., 2021) dataset introduced a multilingual dimension consisting of 5 hours of synthetically generated spoken questions in English, Hindi, and Turkish. SBVQA 2.0 (Alasmary and Al-Ahmadi, 2023) incorporated speaker variability to better mimic real-world conditions. SBVQA and SBVQA 2.0 primarily focused on the English language, while FVSQA, despite its multilingual approach, is smaller in scale. Hence, existing datasets spanned limited domains that either lacked sufficient size for multilingual scenarios or catered predominantly to the English language. Moreover, these datasets are limited to open-ended questions, neglecting other types, such as binary "Yes/No" and numerical questions, which are crucial for real-world applications. This inspired the development of TM-VQA, an extension of the VQA v2.0 dataset, which offers greater size and coverage of a broader spectrum of domains, including various question types, thereby making it more suitable for real-world use cases.

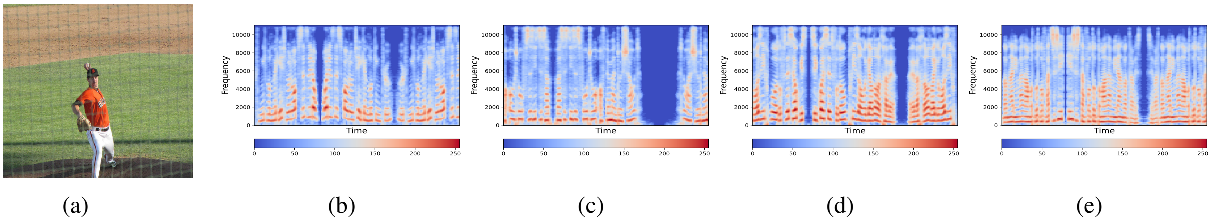


Figure 1: (a) Image and spectrogram visualization of corresponding speech-based question & answer in (b) En (**Q:** " What is this photo taken looking through?", **Ans:** Net), (c) De (**Q:** " Durch was ist dieses Foto entstanden, durch das man blickt?", **Ans:** Net), (d) Fr (**Q:** " Quelle est cette photo prise en regardant à travers ?", **Ans:** Net), (e) Es (**Q:** "Qué es esta foto tomada mirando?", **Ans:** Net).

Table 1: Statistical overview of TM-VQA dataset containing the number of images and question-answer pairs with the duration information of spoken questions (in hours) for the four languages.

Set	**# of Images**	**# Q & A pairs**	**Audio duration (in hours)**			
			ENGLISH	**GERMAN**	**FRENCH**	**SPANISH**
Train	82,783	443,757	220.4	258.17	220.82	201.82
Test	40,504	214,354	95.66	106.51	98.99	97.21
Total	**123,287**	**658,111**	**316.06**	**364.68**	**319.81**	**299.03**

2.2 VQA systems

Current VQA systems typically adopts a multimodal framework consisting of an image and question encoder, and a fusion mechanism (Khan et al., 2020; Lu et al., 2023). The image encoder extracts features from the visual input, while the question encoder processes the text-based or speech-based queries. The fusion component then concatenates these two streams of representations and passes to an architecture consisting of RNNs (Rumelhart et al., 1986; Jordan, 1997) or Transformers (Vaswani et al., 2017), and appropriate responses are generated at the output of a classification layer. State-of-the-art models, such as vision-transformer (Dosovitskiy et al., 2021), ResNet (He et al., 2016), VGG (Simonyan and Zisserman, 2014), Faster RCNN (Ren et al., 2016), etc were employed to extract image features, whereas LaBSE (Feng et al., 2020), Clip (Radford et al., 2021), Word2Vec (Church, 2017), etc were utilized for textual feature extraction. For audio representations, large acoustic models such as Wav2Vec2 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and Whisper (Radford et al., 2023) and statistical models such as Kaldi along with hand-crafted features such as mel filterbanks and mel frequency cepstral coefficients (MFCC) were employed. Fusion techniques, such as MCB (Fukui et al., 2016), MLB (Kim et al., 2016), and MUTAN (Ben-Younes et al., 2017), were proposed to capture the interaction between image and text or speech encoder. A double fusion network was proposed to extract coarse and fine-grained features from the images (Tian et al., 2022). However, most existing frameworks relied on individual models for each language-specific question-answer pair and lacked an unified framework.

3 TM-VQA dataset

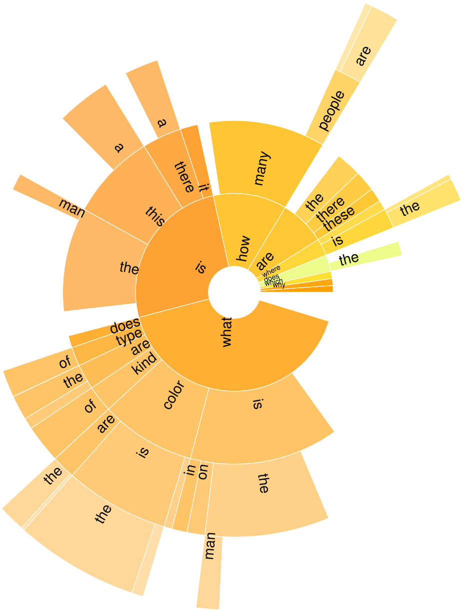


Figure 2: Pie diagram signifying the distribution of words for all question types.

The TM-VQA dataset extends the VQA v2.0 and incorporates textual and spoken questions, gener-

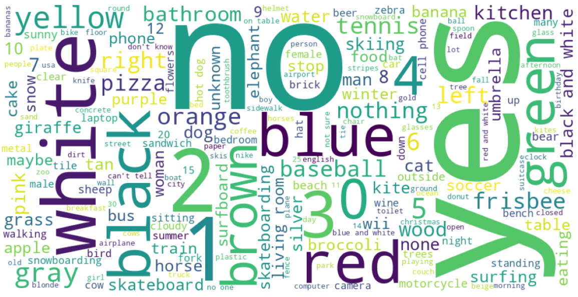


Figure 3: Word cloud representation for answers to different question types.

ated using the SeamlessM4T multimodal and multilingual AI translation model (Barrault et al., 2023), in English, German, French, and Spanish. This dataset comprises a total of 1,298.87 hours (320 hours for each language) of spoken queries. The dataset's statistics, including the number of images and related question-answer pairs along with the duration of the audio files, are presented in Table 1. Subsequently, this dataset is divided into non-overlapping train and test sets. To provide a deeper analysis, the test set is further categorized into "Yes / No", "open-ended", and numerical response-based questions, and the performance will be evaluated for these categories separately. Figure 1 illustrates an image of a baseball player overlooking a net and spectrogram visualization of queries posed in English, German, French and Spanish respectively. Figure 2 represents the pie distribution of first four words for the questions in English language, where the innermost ring represents the first word and outwardly radiating rings represent the subsequent words. The length of the arc is proportional to the number of questions containing a particular word. For clarity, words having a frequency of less than 35 are omitted. Figure 3 represents the word cloud for answers for all the categories combined. The most frequent answers are represented using larger fonts while the less frequent by smaller fonts.

4 Proposed methodology

4.1 Background

The transformer encoder mainly consists of multi-head attention (MHA) and multi-layer perceptron (MLP) blocks. Given input audio embeddings $\mathbf{X_a}$ and input image embeddings $\mathbf{X_i}$, the encoder applies MHA on the inputs and is represented as:

$$\mathbf{Y_a} = \mathbf{X_a} + \mathrm{MHA}(\mathbf{X_a}), \tag{1}$$

$$\mathbf{Y_i} = \mathbf{X_i} + \mathrm{MHA}(\mathbf{X_i}) \tag{2}$$

where $\mathbf{Y_a}$ and $\mathbf{Y_i}$ are the output from the MHA block. For conciseness, we skip the LayerNorm (LN) in both the MHA and MLP layers. From (1), MHA can be defined as:

$$\mathrm{MHA}(\mathbf{X_{a,i}}) = \mathrm{Softmax}(\frac{\mathbf{Q}\cdot\mathbf{K}^{\top}}{\sqrt{\mathrm{d_k}}}).\mathbf{V} \tag{3}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ denote linear transformation layers. Outputs $\mathbf{Y_a}$ and $\mathbf{Y_i}$ obtained from (1). and (2) are then fed to another MHA layer:

$$\mathbf{Z_a} = \mathbf{Y_a}' + \mathrm{MHA}(\mathbf{Y_a}'), \tag{4}$$

$$\mathbf{Z_i} = \mathbf{Y_i}' + \mathrm{MHA}(\mathbf{Y_i}') \tag{5}$$

Let $\mathbf{F_f} = \mathrm{CA}(\mathbf{X_a}, \mathbf{X_i})$ denote the fusion operation between the audio and image modalities, where CA denotes the cross-attention between the audio and image embeddings. In order to incorporate the fusion operation into the encoder block, the fusion module is applied before and after the MHA layer. Thus, (1) and (4) can be re-written as:

$$\mathbf{Y_a}' = \mathbf{Y_a} + \mathrm{CA}_\mathrm{a}^{(\mathrm{I})}(\mathbf{X_a}, \mathbf{X_i}), \tag{6}$$

$$\mathbf{Z_a}' = \mathbf{Z_a} + \mathrm{CA}_\mathrm{a}^{(\mathrm{II})}(\mathbf{Y_a}', \mathbf{Y_i}'), \tag{7}$$

where $\mathbf{Z_a}'$ is the output from one of the Transformer encoder. Similarly, if we want to fuse the image with audio representation then (2) and (5) can be re-written as:

$$\mathbf{Y_i}' = \mathbf{Y_i} + \mathrm{CA}_\mathrm{i}^{(\mathrm{I})}(\mathbf{X_i}, \mathbf{X_a}), \tag{8}$$

$$\mathbf{Z_i}' = \mathbf{Z_i} + \mathrm{CA}_\mathrm{i}^{(\mathrm{II})}(\mathbf{Y_i}', \mathbf{Y_a}'). \tag{9}$$

$\mathrm{CA}^{(\mathrm{I})}$ and $\mathrm{CA}^{(\mathrm{II})}$ are explained in the next section.

4.2 Fusion module

The proposed fusion module consists of two CA blocks for learning the interaction between the audio and image embeddings. The output from these CA blocks are passed through an MLP layer followed by a CNN layer. The two CA blocsk were employed to incorporate cross-modality learning. The first CA block identifies the agreement between the image and audio information. In a high level scenario, this CA block extracts important regions in the image based on the query. Mathematically, from the first CA block, the cross attention scores $\mathrm{CA}_\mathrm{a}^{(\mathrm{I})}$ for audio and $\mathrm{CA}_\mathrm{i}^{(\mathrm{I})}$ for the image is computed as:

$$\mathrm{CA}_\mathrm{a}^{(\mathrm{I})} = \mathrm{Softmax}\left(\frac{\mathbf{Q_a}\cdot\mathbf{K_i}^{\top}}{\sqrt{\mathrm{d_k}}}\right)\cdot\mathbf{V_a} \tag{10}$$

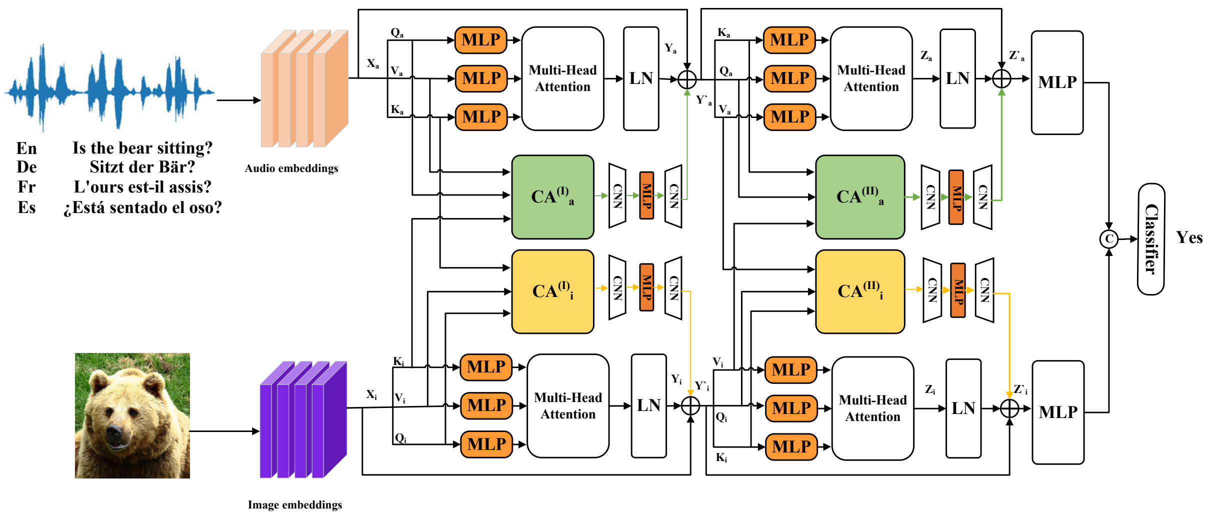


Figure 4: Architectural representation of the proposed unified architecture supporting audio input in four languages.

.

$$\mathrm{CA}_{\mathrm{i}}^{(\mathrm{I})} = \mathrm{Softmax}\left(\frac{\mathbf{Q_i} \cdot \mathbf{K_a}^{\top}}{\sqrt{\mathrm{d_k}}}\right) \cdot \mathbf{V_i} \qquad (11)$$

The second CA block tries to discover inner-relations from one modality to another as:

$$\mathrm{CA}_{\mathrm{a}}^{(\mathrm{II})} = \mathrm{Softmax}\left(\frac{\mathbf{Q_i} \cdot \mathbf{K_i}^{\top}}{\sqrt{\mathrm{d_k}}}\right) \cdot \mathbf{V_a} \qquad (12)$$

$$\mathrm{CA}_{\mathrm{i}}^{(\mathrm{II})} = \mathrm{Softmax}\left(\frac{\mathbf{Q_a} \cdot \mathbf{K_a}^{\top}}{\sqrt{\mathrm{d_k}}}\right) \cdot \mathbf{V_i} \qquad (13)$$

Following prior work (Houlsby et al., 2019), a similar bottleneck module is employed consisting of a down and up-projection layer made of CNN blocks, along with a non-linear activation function. The output $\mathbf{Z_a}'$ and $\mathbf{Z_i}'$ are passed through an MLP layer and then concatenated before the final classification layer having a softmax activation function. This produces the final output $\hat{\mathbf{y}} \in \mathbb{R}^C$, where C denotes the number of classes (unique words) in the textual answers from the training subset. Each of these classes were converted into an one-hot encoding vector having a dimension size equal to the number of unique words. This is done to compute binary cross-entropy loss ($\mathcal{L}_{\mathrm{BCE}}$) between the prediction and the ground-truth labels. Mathematically, $\mathcal{L}_{\mathrm{BCE}}$ is represented as:

$$\mathcal{L}_{\mathrm{BCE}} = \mathrm{BCE}(\hat{\mathbf{y}}, \mathbf{y}) \qquad (14)$$

where, $\mathbf{y}$ denotes the original label, $\hat{\mathbf{y}}$ denotes the prediction and BCE is the binary cross entropy function computed between $\mathbf{y}$ and $\hat{\mathbf{y}}$. In this way, the proposed model learns to align the audio and image representations for analysis and generate appropriate responses. Figure 4 presents architectural representation of the unified framework.

5 Experimental methodology

5.1 Visual representations

Image features were extracted using pre-trained CLIP-ViT (Radford et al., 2021), Swin Transformer (Liu et al., 2021), ResNet-152 (He et al., 2016), and Faster-RCNN (Ren et al., 2016) to provide a comparative analysis. The CLIP-ViT is employed as it is trained using an image-text pair and can differentiate the image and text in the embedding space. Swin Transformer computes representations using shifted windows and enhances the accuracy by confining the self-attention computation to non-overlapping local windows while simultaneously maintaining cross-window connections. It can model information at various scales while maintaining a linear computational complexity. ResNet-152, a deep convolutional neural network (CNN) and part of the ResNet (Residual Networks) is known for its use of skip connections to combat the vanishing gradient problem. It contains 152 layers and is widely used for feature extraction tasks due to its high accuracy and efficiency in handling complex visuals. Faster R-CNN is an advanced object detection model that integrates a region proposal network (RPN) with CNN to generate object proposals which are then passed to a classification layer for returning bounding boxes.

This model can efficiently identify object regions and classify them in a single, end-to-end framework.

5.2 Text representations

For textual feature extraction, LaBSE (Feng et al., 2020), mBERT (Devlin, 2018), and mRoBERTa (Liu, 2019) are utilized. LaBSE (Language-agnostic BERT Sentence Encoder), a pre-trained model, is designed for generating high-quality sentence embeddings that can work across multiple languages. This model enables cross-lingual tasks, such as multilingual retrieval and translation. It produces language-agnostic representations and is ideal for applications requiring semantic understanding in diverse languages. Multilingual BERT (mBERT), a pre-trained model, is designed to handle 104 languages and provides language-agnostic embeddings for cross-lingual tasks. It is useful for text embedding extraction as it generates contextualized representations across different languages. Multilingual RoBERTa (mRoBERTa), a variant of the RoBERTa model, is designed to handle multiple languages and is an improvement over multilingual BERT. It uses optimized training techniques, which makes it effective for cross-lingual tasks such as translation, sentence classification, and language understanding across various languages.

5.3 Audio representations

For the extraction of audio features, pre-trained large acoustic models, such as Whisper large-v3 (Radford et al., 2023) and Wav2Vec2 (Baevski et al., 2020) are employed. Whisper large-v3 is a multilingual acoustic model, known for its multitasking ability which can accurately transcribe audio and perform robust speech-to-text tasks across various languages. It is particularly useful in audio feature extraction for tasks like transcription, language identification, and speaker recognition, due to its high accuracy and robustness across noisy environments. The Wav2Vec2-based XLSR-128 (Cross-Lingual Speech Representations), a large-scale pre-trained model, designed for speech recognition tasks is employed and is highly useful for audio feature extraction as it can learn robust and language-agnostic speech representations making it effective for audio processing tasks.

5.4 Baseline and proposed model settings

A cascaded pipeline of an ASR followed by a text-based VQA system is taken as a baseline (Rajkhowa et al., 2023, 2024a). The ASR transcribes the queries spoken in a particular language into text. This text is then concatenated using CA and passed to the Transformer encoder to generate the responses in English text. This cascaded system is compared with the proposed unified architecture that can directly accept audio representations. This unified architecture can bypass the intermediate stage of ASR transcription and also reduce the error propagation from ASR to the VQA as seen in the cascaded approach. Furthermore, it also has the advantage of lower latency due to the involvement of a single module and can be effectively incorporated into edge devices.

5.5 Experimental Settings

The proposed framework is trained for 300 epochs having a batch size of 256. The model (best checkpoint) corresponding to the lowest validation loss is selected for evaluating the performance. Adam optimizer is used having a learning rate (LR) of 3×10^{-4}, along with the CosineAnnealing scheduler. All these experiments were conducted using 4 H100 GPU having 80 gigabytes (GB) of high-bandwidth memory (HBM2e) employing Ubuntu 20.04 LTS as the Operating System.

5.6 Evaluation Metrics

To effectively assess the performance of VQA systems, it's crucial to choose appropriate metrics for all the categories of question-answer scenarios. Metrics like Top-1 accuracy offer straightforward insights into how well the model can identify answers. Top-1 focuses on the accuracy of the highest-ranked prediction. This metric is commonly used for VQA system evaluations.

6 Results

Table 2 presents a comparative analysis between cascaded and unified text-based VQA systems. Here, Faster-RCNN and mRoBERTa are used as image and text encoders. From this table, it can be observed that the unified system outperformed its cascaded counterpart. The cascaded approach is known to suffer error propagation from the ASR transcription to other text-based systems and this effect can be observed in the case of text-based VQA. A similar trend can also be observed in

Table 2: Comparative analysis between text-based cascaded and unified VQA systems employing Faster-RCNN and mRoBERTa as image and text encoder. Performances computed using Top-1 accuracy metric are expressed in %.

Image Encoder	Text Encoder	Source language	Model Type	All	Other	Yes/No	Num
Faster-RCNN	mRoBERTa	ENGLISH	Cascaded	64.57	55.29	83.32	44.67
			Unified	**68.49**	**58.67**	**86.77**	**47.42**
		GERMAN	Cascaded	54.46	53.37	76.54	44.42
			Unified	**56.49**	**54.54**	**78.89**	**45.42**
		FRENCH	Cascaded	52.45	51.45	73.27	41.5
			Unified	**54.58**	**53.55**	**75.79**	**43.35**
		SPANISH	Cascaded	52.57	47.39	72.31	40.58
			Unified	**54.69**	**49.87**	**73.82**	**41.82**

Table 3: Performance ablation (expressed in % using Top-1 accuracy metric) of the proposed unified text-based system incorporating distinct state-of-the-art models in the image and text encoder for English source language.

Image Encoder	Text Encoder	All	Other	Yes/No	Num
CLIP-ViT-B	mBERT	64.76	54.63	81.54	45.21
	mRoBERTa	64.57	**55.5**	82.31	45.49
	LaBSE	**64.98**	54.65	**82.43**	**46.5**
Swin Transformer	mBERT	63.89	**54.67**	83.45	45.43
	mRoBERTa	**64.56**	54.23	**83.79**	**45.56**
	LaBSE	64.28	53.78	83.76	45.38
ResNet	mBERT	60.33	50.43	79.4	43.6
	mRoBERTa	**60.86**	**51.2**	78.54	43.27
	LaBSE	60.59	50.39	**79.72**	**43.57**
Faster-RCNN	mBERT	66.54	57.12	86.05	47.2
	mRoBERTa	**68.49**	**58.67**	86.77	**47.42**
	LaBSE	67.87	58.32	**87.55**	46.79

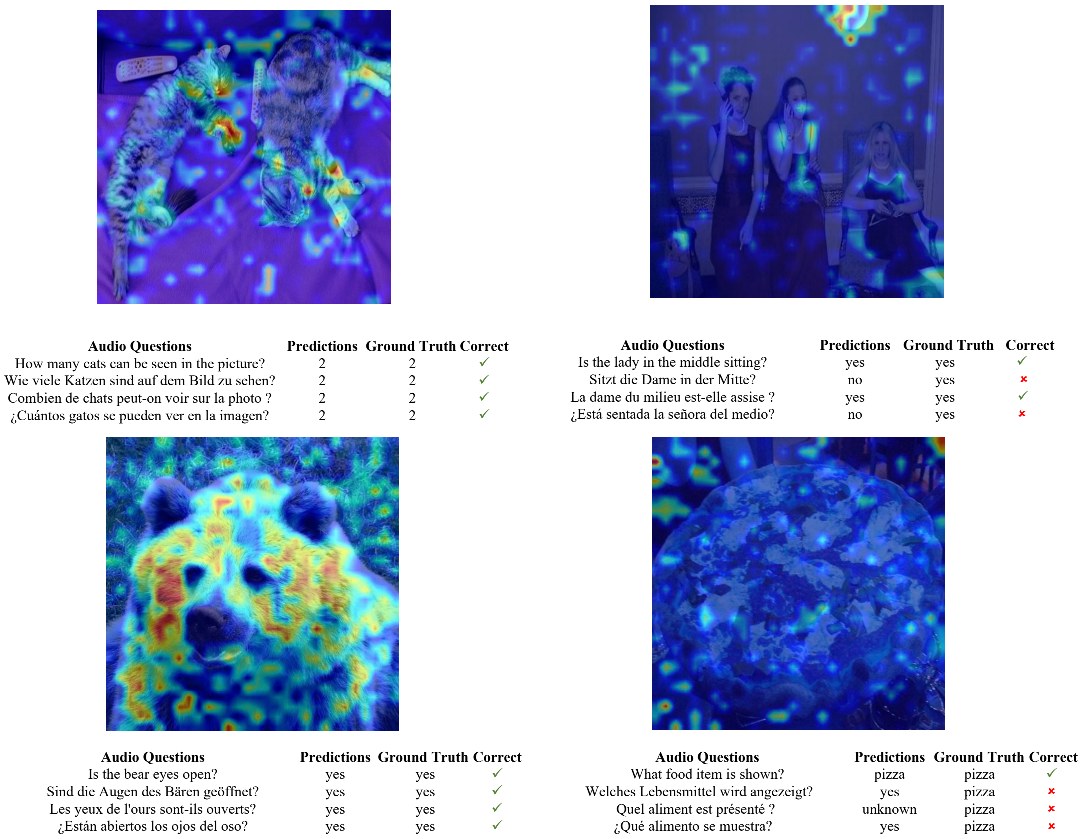
Table 4: Comparative analysis between speech-based cascaded and unified VQA systems employing Faster-RCNN and Whisper as image and audio encoder. Performances computed using Top-1 accuracy metric are expressed in %.

Image encoder	Audio encoder	Source language	Model Type	All	Other	Yes/No	Num
Faster-RCNN	Whisper	ENGLISH	Cascaded	55.22	48.45	68.12	41.37
			Unified	**58.53**	**50.67**	**69.19**	**43.89**
		GERMAN	Cascaded	51.67	42.78	64.45	39.38
			Unified	**54.87**	**44.88**	**67.47**	**41.47**
		FRENCH	Cascaded	48.56	42.24	66.57	38.39
			Unified	**51.75**	**44.47**	**68.52**	**40.41**
		SPANISH	Cascaded	47.69	39.68	63.44	34.47
			Unified	**52.5**	**42.17**	**66.3**	**38.84**

Table 5: Performance ablation (expressed in % using Top-1 accuracy metric) of the proposed unified speech-based system incorporating distinct state-of-the-art models in the image and audio encoder for English source language.

Image Encoder	Audio Encoder	All	Other	Yes/No	Num
CLIP-ViT-B	Whisper	54.24	48.2	68.66	**40.84**
	Wav2Vec2	**54.33**	**48.38**	**68.7**	40.47
Swin Transformer	Whisper	**54.37**	**48.6**	**68.54**	**40.79**
	Wav2Vec2	54.28	47.53	67.85	41.21
ResNet	Whisper	52.39	46.5	65.49	38.69
	Wac2Vec2	**53.67**	**46.63**	**65.87**	**38.9**
Faster-RCNN	Whisper	58.53	50.67	69.19	**43.89**
	Wav2Vec2	**58.69**	**50.73**	**69.68**	43.79

Table 4 where the unified approach consistently outperformed its cascaded counterpart for speech-based VQA systems. From these two tables, it can be inferred that a unified system can demonstrate better performance while being computationally efficient. Table 3 presents an ablation study for



Audio Questions	Predictions	Ground Truth	Correct
How many cats can be seen in the picture?	2	2	✓
Wie viele Katzen sind auf dem Bild zu sehen?	2	2	✓
Combien de chats peut-on voir sur la photo ?	2	2	✓
¿Cuántos gatos se pueden ver en la imagen?	2	2	✓

Audio Questions	Predictions	Ground Truth	Correct
Is the lady in the middle sitting?	yes	yes	✓
Sitzt die Dame in der Mitte?	no	yes	×
La dame du milieu est-elle assise ?	yes	yes	✓
¿Está sentada la señora del medio?	no	yes	×

Audio Questions	Predictions	Ground Truth	Correct
Is the bear eyes open?	yes	yes	✓
Sind die Augen des Bären geöffnet?	yes	yes	✓
Les yeux de l'ours sont-ils ouverts?	yes	yes	✓
¿Están abiertos los ojos del oso?	yes	yes	✓

Audio Questions	Predictions	Ground Truth	Correct
What food item is shown?	pizza	pizza	✓
Welches Lebensmittel wird angezeigt?	yes	pizza	×
Quel aliment est présenté ?	unknown	pizza	×
¿Qué alimento se muestra?	yes	pizza	×

Figure 5: Examples of different questions on each image show which region contributed the most to answering the question. We see that the bright spots are the positions that get the most attention when asking a particular question.

unified text-based VQA systems that incorporate distinct image and text representations extracted using various state-of-the-art pre-trained models for the English language. From this table, it can be observed that the Faster-RCNN and mRoBERTa combination provided the best performance. Faster-RCNN, through its bounding box identification, can effectively identify the object regions from the image. mRoBERTa has demonstrated superior cross-lingual context understanding capability among other pre-trained models. Table 5 presents an ablation study for unified speech-based VQA systems for the English language. Here, the comparison is made using various combinations of distinct image and audio representations. From this table, the Faster-RCNN and Wav2Vec2 combination outperformed the remaining combinations. The performance of Wav2Vec2-based systems is marginally better than Whisper. Overall, the accuracy for VQA systems incorporating English as the source language in input is higher than the rest of the languages with Spanish being the lowest. However, the English language has the advantage as the target language is in English and it is easier to map the attention between the source and target language pairs. In summary, it can be inferred that Faster-RCNN and mRoBERTa combination is better for text-based VQA systems and Faster-RCNN and Whisper combination is better for speech-based VQA systems.

Fig 5. denotes the attention maps of the cross-attention module in the proposed unified architecture. It can be observed that the model can extract the important parts of an image corresponding to a particular question. The bright spots denote the region that has received the most attention.

7 Conclusion

This work introduced a VQA dataset to facilitate speech-based VQA research and proposed a novel unified VQA framework that employs cross-attention. Experimental analysis indicates the superiority of this unified architecture over cascaded systems with Faster-RCNN, mRoBERTa and Wav2Vec2 as better models for image, text and audio representations.

8 Limitations

Cross-attention blocks increase the model's complexities thereby making it computationally expensive. Moreover, the TM-VQA dataset is skewed towards "Yes / No" type questions. Additionally, the model is not tested using real speech. These studies will be included in the future works.

Acknowledgments

The authors highly acknowledge the computing facilities provided by the Deep Learning and Communications laboratory at the Department of Electronics Engineering, Indian Institute of Technology (Banaras Hindu University), Varanasi.

References

Faris Alasmary and Saad Al-Ahmadi. 2023. SBVQA 2.0: Robust end-to-end speech-based visual question answering for open-ended questions. *IEEE Access.*

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. SeamlessM4T-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596.*

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620.

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

Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

Mikyas T Desta, Larry Chen, and Tomasz Kornuta. 2018. Object-based reasoning in VQA. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1814–1823. IEEE.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024. MVQA: A dataset for multimodal information retrieval in PDF-based visual question answering. *arXiv preprint arXiv:2404.12720.*

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR.*

Haytham M Fayek and Justin Johnson. 2020. Temporal reasoning via audio question answering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2283–2294.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852.*

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847.*

Jingying Gao, Qi Wu, Alan Blair, and Maurice Pagnucco. 2024. Lora: A logical reasoning augmented dataset for visual question answering. *Advances in Neural Information Processing Systems*, 36.

Lianli Gao, Pengpeng Zeng, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. 2018. Examine before you answer: Multi-task learning with adaptive-attentions for multiple-choice VQA. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1742–1750.

Panpan Gao, Hanxu Sun, Gang Chen, Ruiquan Wang, and Minggang Li. 2022. Visual question answering for intelligent interaction. *Mobile Information Systems*, 2022(1):4232968.

Vatsal Goel, Mohit Chandak, Ashish Anand, and Prithwijit Guha. 2021. IQ-VQA: Intelligent visual question answering. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pages 357–370. Springer.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Vqa-lol: Visual question answering under the lens of logic. In *European Conference on Computer Vision*, pages 379–396. Springer.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Tomoki Hayashi, Ryuichi Yamamoto, Takenori Yoshimura, Peter Wu, Jiatong Shi, Takaaki Saeki, Yooncheol Ju, Yusuke Yasuda, Shinnosuke Takamichi, and Shinji Watanabe. 2021. Espnet2-TTS: Extending the edge of TTS research. *arXiv preprint arXiv:2110.07840*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Michael I Jordan. 1997. Serial order: A parallel distributed processing approach. In *Advances in psychology*, volume 121, pages 471–495. Elsevier.

Kushal Kafle and Christopher Kanan. 2016. Answer-type prediction for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4976–4984.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*, pages 1965–1973.

Aisha Urooj Khan, Amir Mazaheri, Niels Da Vitoria Lobo, and Mubarak Shah. 2020. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. *arXiv preprint arXiv:2010.14095*.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Rengang Li, Cong Xu, Zhenhua Guo, Baoyu Fan, Runze Zhang, Wei Liu, Yaqian Zhao, Weifeng Gong, and Endong Wang. 2022. AI-VQA: visual question answering based on agent interaction with interpretability. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5274–5282.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. 2018. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6116–6124.

Mingfu Liang, Ying Wu, et al. 2024. TOA: task-oriented active VQA. *Advances in Neural Information Processing Systems*, 36.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Siyu Lu, Yueming Ding, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. 2023. Multiscale feature extraction and fusion of image and text in VQA. *International Journal of Computational Intelligence Systems*, 16(1):54.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Tonmoy Rajkhowa, Amartya Roy Chowdhury, Prashant Bannulmath, KT Deepak, and SR Mahadeva Prasanna. 2023. Optimizing direct speech-to-text translation for un-orthographic low-resource tribal languages using source transliterations. In *(O-COCOSDA)*, pages 1–6. IEEE.

Tonmoy Rajkhowa, Amartya Roy Chowdhury, Hrishikesh Ravindra Karande, and SR Mahadeva Prasanna. 2024a. Evaluating the efficacy of large acoustic model for documenting non-orthographic tribal languages in india. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6475–6483.

Tonmoy Rajkhowa, Amartya Roy Chowdhury, Sankalp Nagaonkar, Achyut Mani Tripathi, and Mahadeva Prasanna. 2024b. TM-PATHVQA: 90000+ textless multilingual questions for medical visual question answering. In *Interspeech 2024*, pages 4034–4038.

Kiran Ramnath, Leda Sari, Mark Hasegawa-Johnson, and Chang Yoo. 2021. Worldly wise (wow)-cross-lingual knowledge fusion for fact-based visual spoken-question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1908–1919.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986. *Biometrika*, 71(599-607):6.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Weidong Tian, Haodong Li, and Zhong-Qiu Zhao. 2022. Dual capsule attention mask network with mutual learning for visual question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5678–5688.

Andeep S Toor, Harry Wechsler, and Michele Nappi. 2019. Biometric surveillance using visual question answering. *Pattern Recognition Letters*, 126:111–118.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Henrik Voigt, Monique Meuschke, Kai Lawonn, and Sina Zarrieß. 2021. Challenges in designing natural language interfaces for complex visual models. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 66–73.

Peratham Wiriyathammabhum, Douglas Summers-Stay, Cornelia Fermüller, and Yiannis Aloimonos. 2016. Computer vision and natural language processing: recent approaches in multimedia and robotics. *ACM Computing Surveys (CSUR)*, 49(4):1–44.

Qi Wu, Peng Wang, Xin Wang, Xiaodong He, and Wenwu Zhu. 2022. Medical VQA. In *Visual Question Answering: From Theory to Application*, pages 165–176. Springer.

Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. 2020. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354.

Jinxu Zhang, Yongqi Yu, and Yu Zhang. 2024. CREAM: Coarse-to-fine retrieval and multi-modal efficient tuning for document VQA. In *ACM Multimedia 2024*.

Ted Zhang, Dengxin Dai, Tinne Tuytelaars, Marie-Francine Moens, and Luc Van Gool. 2017. Speech-based visual question answering. *arXiv preprint arXiv:1705.00464*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.