Hacettepe University

Computer Engineering Department

BBM479 Project Proposal Report

Project Details

Title	**Campus Navigator: Campus-Oriented Smart Assistant for the Visually Impaired**
Short Description (max. 200 words)	Campus Navigator is a voice assistant application designed to assist visually impaired individuals on campus. The app allows users to ask real-time voice-based questions and receive immediate spoken responses, enhancing their navigation experience. It utilizes both voice inputs and live images from the user's surroundings to provide accurate and context-aware answers. By leveraging a Visual Question Answering (VQA) model, the system processes the user's query and the accompanying image, then generates a response based on both the visual and textual data.
Supervisor	**Assoc. Prof. Hacer Yalım Keleş**
Technical and Scientific Difficulty	( ) Easy ( ) Mediocre (X) Challenging
External Support	( ) Yes (X) No

If yes,	**Type**	**Details**
	( ) Company Funding / Support	Company name: Amount:
	( ) TÜBİTAK Project Fund	Type: Amount:
	( ) Other Fund	Source : Amount:

Group Members

	Full Name	**Student ID**
1	Ömer Faruk Horat	2200356086
2	Ecem Altınkeser	2210356019
3	Buğrahan Çıldır	2200356017
4	Doğukan Aytekin	2200356003

Project Summary (/ 20 Points)

Explain the project in summary, including your motivation to do the project, your solution plan in short and your expected outcome and impact. You have to summarize your project between 200-500 words.

Navigating university campuses can be a major challenge for visually impaired individuals, limiting their independence and ability to interact with their surroundings. Identifying objects, reading signs, finding landmarks, and avoiding obstacles can be especially hard in complex and changing campus settings. To address these challenges, we propose a voice-controlled smart assistant that combines real-time image recognition and Visual Question Answering (VQA) technology. The main goal of the system is to increase the independence, safety, and overall quality of life for visually impaired individuals, particularly within the university environment.

The project focuses on building a real-time assistance tool that uses both visual and voice inputs to answer questions about the user's surroundings. Recent progress in VQA, which combines computer vision and natural language processing (NLP), makes this possible. The system will process live images and user questions, using image captioning and VQA models to create spoken responses. This approach will help users navigate campus environments more effectively by providing critical information about objects, landmarks, and obstacles in real-time.

Our plan includes several key steps: conducting a review of the existing VQA models, selecting the most suitable model for our project, and developing a dataset of campus-specific images. We will also integrate NLP algorithms to process voice commands and convert them into text
queries for the VQA model. The system will capture images through a live camera inside a smart-glass device. Additionally, we will create a mobile app that connects to the smart glasses, sending the needed inputs to a cloud system where all processing and model functions are based. This will result in a fully integrated system that understands voice commands, analyzes images, and gives spoken responses to users.

The expected result is an accessible, practical tool that significantly improves campus navigation for visually impaired students. By helping them rely less on external assistance, the system will increase their independence and improve their educational experience. The project also aims to contribute more broadly to the field of assistive technology for visually impaired individuals.

Problem Definition and Literature Review (/ 20 Points)

Define your problem as clearly as possible. Explain your inputs, your context, your outputs and your limitations. Try to use a scientific language as much as possible. Where necessary use citations to existing literature to create context and clarify the problem. Equations, flow charts, etc. are welcome.

Navigating university campuses poses significant challenges for visually impaired individuals, limiting their ability to move independently and interact fully with their environment. These individuals encounter difficulties in identifying objects, reading signs, locating landmarks, and avoiding obstacles, which directly impacts their educational experience and overall quality of life. To address these challenges, the Campus-Oriented Smart Assistant for the Visually Impaired proposes a voice-controlled smart assistant that leverages real-time image recognition and question-answering capabilities. The primary objective of this system is to enhance the autonomy and safety of visually impaired individuals.

At the core of this problem is the need for real-time assistance that provides visually impaired users with immediate feedback about their surroundings. This feedback is crucial for identifying objects, landmarks, and obstacles in dynamic and continuously changing environments, thereby improving their campus navigation experience. In this context, our project integrates a Visual Question Answering (VQA) model, designed to process both voice and visual inputs to respond to user queries about their environment. Real-time image analysis is a critical component of this process. Recent studies on visual perception and image captioning indicate that generating image captions based on user queries significantly improves model performance, making VQA a vital feature of the proposed system.[5]

Visual question answering (VQA) is the task of answering questions based on an image mainly including two fields: computer vision and Natural Language Processing (NLP). Computer vision aims to teach a computer how to see while NLP aims to create interactions between machines and natural languages.[1] VQA models mainly have two sections: extracting features from the image using image captioning and generating an answer according to the extracted features using a large language model. There are multiple VQA models using different techniques to generate answers from an image. CogVLM replaces the shallow alignment in most VQA models with a trainable visual expert module in the attention and FFN layers.[2] BLIP-2 is a generic and compute-efficient method for vision-language pre-training that leverages frozen pretrained image encoders and LLMs.[3] xGen-MM (BLIP-3) is a new framework designed to scale up LMM training by utilizing an ensemble of multimodal interleaved datasets, curated caption datasets, and other publicly available datasets.[4]

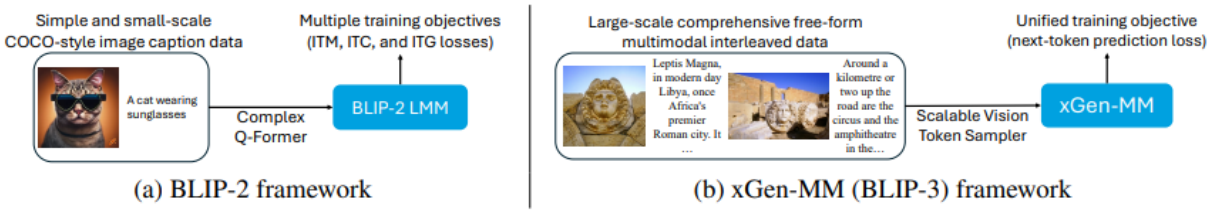

Fig 1. BLIP-2 and xGen-MM frameworks

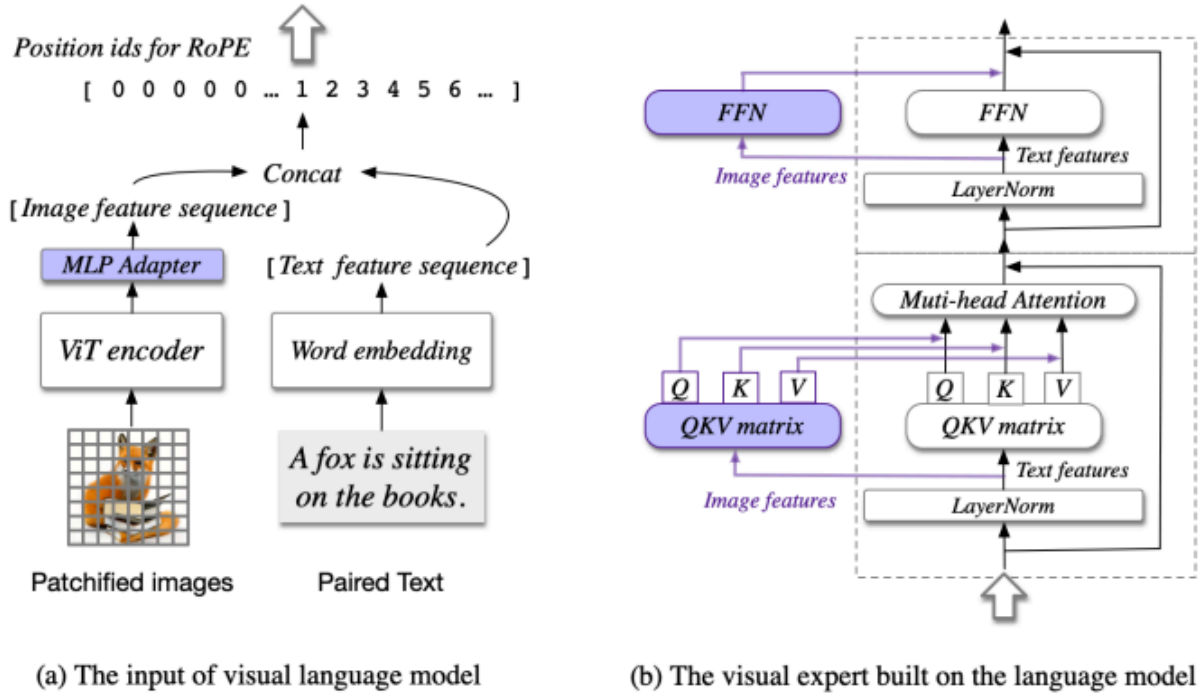

Fig 2. Architecture of CogVLM

Natural Language Processing (NLP) for Voice Assistants: NLP models are used to convert spoken questions from the user into structured data that can be interpreted by the system. This involves automatic speech recognition (ASR) to transcribe spoken words and natural language understanding (NLU) to derive meaning. NLP tasks often leverage transformer models, like BERT or GPT, for interpreting and generating language .

In our system, the VQA model analyzes visual data from the user's surroundings and produces spoken responses to voice-based queries. The inputs primarily consist of live images via smart glasses or, alternatively, through a smartphone and user's voice queries. The outputs are guiding verbal responses that provide context-aware feedback to the user.

Key challenges include ensuring real-time image processing across diverse lighting conditions, maintaining high system accuracy, guiding the user to ask relevant questions, and delivering timely, contextually appropriate responses. The project builds upon recent advancements in mobile computer vision, image captioning, and VQA models to improve the standard of living for visually impaired individuals.

By enhancing the accuracy and speed of real-time navigation assistance, this system aims to significantly reduce the dependence of visually impaired individuals on others, ultimately improving their autonomy and campus experience.

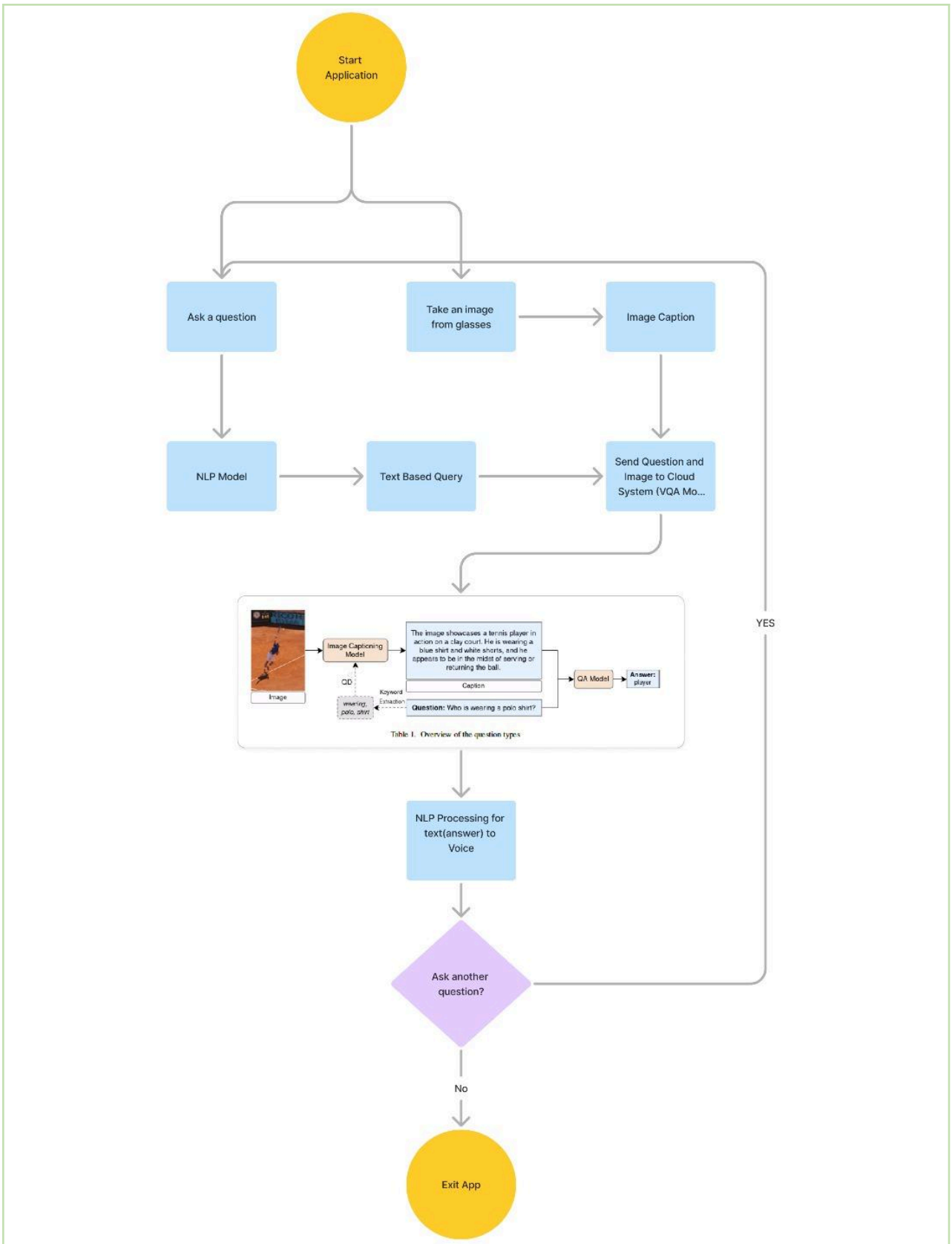

Fig 3. Campus Navigator flowchart

References

[1] Sharma, H., & Jalal, A. S. (2021). A survey of methods, datasets and evaluation metrics for visual question answering. Image and Vision Computing, 116, 104327.

[2] Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., & Tang, J. (2023, November 6). COGVLM: Visual Expert for Pretrained Language Models. arXiv.org. https://arxiv.org/abs/2311.03079

[3] Li, J., Li, D., Savarese, S., & Hoi, S. (2023, January 30). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv.org. https://arxiv.org/abs/2301.12597

[4] Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M. S., Kendre, S., Zhang, J., Qin, C., Zhang, S., Chen, C., Yu, N., Tan, J., Awalgaonkar, T. M., Heinecke, S., . . . Xu, R. (2024, August 16). XGen-MM (BLIP-3): a family of open large multimodal models. arXiv.org. https://arxiv.org/abs/2408.08872

[5] Özdemir, Ö., & Akagündüz, E. (2024, April 12). Enhancing Visual Question Answering through Question-Driven Image Captions as Prompts. arXiv.org. https://arxiv.org/abs/2404.08589

Solution Plan (/ 20 Points)

Explain the potential paths to solution. You should propose at least one solid plan to attack the problem. Dissect your plan into steps and clearly identify the inputs and outputs of each step. You are not expected to provide the technical details of each step. Provide a weekly timeline/Gantt chart displaying the relevant weeks for each step.

Project Planning (Weeks 1–2)

Task: Create the project schedule and allocate resources.
Inputs: Project scope, team members, timeline.
Outputs: A finalized project plan and schedule, as shown in the Gantt chart.

Technical Research and Development (Weeks 2–4)

Step 1: Literature Review for the VQA Model (Weeks 2–4)

Goal: Conduct research on existing VQA models and their performance, especially in visually impaired applications.
Inputs: Research papers, case studies, and technical documentation.
Outputs: An understanding of the latest advancements in VQA models and technologies applicable to the project.

VQA Model Development (Weeks 4–14)

Step 1: Choosing the VQA Model (Weeks 4–6)
Goal: Select the most suitable VQA model based on research.
Inputs: Insights from literature review, performance metrics.
Outputs: A VQA model chosen for implementation.

Step 2: Creating the Image Captioning Model (Weeks 5–6)
Goal: Develop a model that generates captions for images.
Inputs: Publicly available image captioning models, datasets, and image processing techniques.
Outputs: A trained image captioning model.

Step 3: Researching Datasets (Weeks 5–7)
Goal: Investigate and evaluate datasets relevant to VQA and image captioning.
Inputs: Dataset collections, dataset comparison metrics.
Outputs: A set of relevant datasets identified for further use.

Step 4: Training and Testing the VQA Model (Weeks 7–11)
Goal: Train the selected VQA model and test its performance.
Inputs: Labeled dataset, model architecture, training algorithms.
Outputs: A trained VQA model with performance metrics.

Step 5: Creating and Labeling a Dataset (Weeks 10–14)
Goal: Collect and label a dataset of campus-specific images with questions and answers.
Inputs: Campus images, image metadata, labeling tools.
Outputs: A labeled dataset tailored to campus scenarios.

Step 6: Performance Optimization (Weeks 12–14)
Goal: Optimize the trained model for real-time performance, improving accuracy and reducing latency.
Inputs: Trained model, performance metrics, optimization algorithms.
Outputs: An optimized model ready for deployment.

NLP Module Development (Weeks 12–14)

Step 1: Developing a Model for Recognizing Voice Commands (Weeks 12–14)
Goal: Create a model that can recognize and process voice commands.

Inputs: Speech datasets, voice recognition algorithms.
Outputs: A functional voice recognition model.

Step 2: Integration of Natural Language Processing Algorithms (Weeks 12–14)
Goal: Integrate NLP algorithms to process user commands and convert them into text queries for the VQA model.
Inputs: NLP algorithms, speech-to-text conversion libraries.
Outputs: A fully integrated NLP system with voice command recognition.

Step 3: Sending the Generated Answer to the User via Voice (Weeks 12–14)
Goal: Implement a text-to-speech (TTS) system that conveys the VQA model's answer to the user.
Inputs: TTS libraries, text outputs from the VQA model.
Outputs: A system that delivers answers via voice output.

Weekly Timeline

Weeks 1–2: Creating project schedule
Weeks 2–4: Literature review for the VQA model
Weeks 4–6: Choosing VQA Model
Weeks 5–7: Creating Image Captioning Model
Weeks 5–7: Researching datasets
Weeks 7–11: Training and testing the VQA model
Weeks 10–14: Creating and labeling a dataset
Weeks 12–14 : Performance optimization
Weeks 12–14: Developing a model for recognizing voice commands
Weeks 12–14: Integration of natural language processing algorithms
Weeks 12–14: Sending the generated answer to the user via voice

WP No.	WP Name	WP Description	Assigned to	Weeks													
				1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Project Planing	Creating project schedule	Team	X	X												
2	Technical Research and Development	Literature review for the VQA model	Team		X	X	X										
3	VQA Model Development	Choosing VQA Model	Ecem, Buğra				X	X	X								
		Creating Image Captioning Model	Doğukan, Ömer Faruk					X	X	X							
		Researching datasets	Buğra					X	X	X							
		Training and testing the VQA model	Ecem, Buğra							X	X	X	X	X			
		Creating and labeling a dataset	Team										X	X	X	X	X
		Performance optimization	Team												X	X	X
4	NLP Module Development	Developing a model for recognizing voice commands	Buğra												X	X	X
		Integration of natural language processing algorithms	Ecem, Ömer Faruk												X	X	X
		Sending the generated answer to the user via voice	Doğukan												X	X	X

Methodology (/ 20 Points)

Explain the methodology you will use in each of the steps you have described under your solution plan. Here, you are expected to give more technical details about each solution step. Also explain how each member of the project will contribute by assigning members to steps. If you are assigning more than one member to a step, explain their specific role and how the work will be divided among them.

Project Planning (Weeks 1–2)

During the Project Planning phase, our team will establish a structured workflow by defining clear roles and setting a timeline to ensure all tasks are completed on schedule. The first step is creating a detailed project timeline with key milestones like VQA model research, dataset preparation, and NLP integration. Weekly goals will be assigned for each deliverable, ensuring clear expectations.

We will define the core phases of the project—research, development, and testing—based on requirements. This includes determining necessary tools (e.g., programming environments, frameworks) and assigning roles based on expertise. Each member will focus on tasks where they can contribute most effectively.

By the end of this phase, we will have a concrete roadmap with clear deadlines, assigned tasks, and a well-coordinated workflow, ensuring smooth project execution.

Technical Research and Development (Weeks 2–4)

Step 1: Literature Review for the VQA Model (Weeks 2–4)

The team will conduct a comprehensive review of research papers on VQA models, focusing on both recent advancements in model architectures and their specific applications for visually impaired users. Possible technologies and open-source implementations (e.g., CogVLM, BLIP-2) will be studied to identify the most suitable frameworks for our project. In parallel, current solutions and real-world applications of VQA in assistive technologies will be researched to understand existing challenges and opportunities. Additionally, an in-depth analysis of relevant datasets (e.g., VQA 2.0, VizWiz) will be carried out to assess their applicability to accessibility-focused tasks.

Buğra: Researching datasets
Ecem: Researching VQA models
Ömer: Researching image captioning techniques
Doğukan: Researching keyword extraction technique

VQA Model Development (Weeks 4–14)

Step 1: Choosing the VQA Model (Weeks 4–6)

Available VQA models (CogVLM, BLIP-2, xGenMM) will be compared, analyzing their architectures and performance to determine which is the most resource-efficient and suitable for our project. Open-source codes of these models will be studied to evaluate their computational requirements, ease of integration, and adaptability for accessibility-focused tasks, ensuring the chosen model aligns with the project's needs for visually impaired users.

Ecem & Buğra: Compare VQA models and test their open-source codes

Step 2: Creating the Image Captioning Model (Weeks 5–6)

Image captions are obtained through the visual grounding and chat variations of the CogVLM, FuseCap, and BLIP-2 OPT2.7b models. When determining the image captioning method, we pay attention to both its alignment with our resource capacity and its high performance in image captioning benchmarks. We employ 4-bit quantization to CogVLM and use F16 precision for BLIP-2 OPT2.7b.

Ömer Faruk Horat and Doğukan Aytekin will work together on each part of the image captioning model development, contributing equally throughout the process. They will jointly evaluate CogVLM, FuseCap, and BLIP-2 OPT2.7b

models, applying 4-bit quantization to CogVLM and F16 precision to BLIP-2 OPT2.7b to optimize resources. Both will contribute to performance benchmarking, testing the models on campus-specific images, and reviewing metrics to ensure caption quality and contextual relevance. They will also collaborate on analyzing results to finalize the optimal model setup or model combination to best serve visually impaired use

Step 3: Researching Datasets (Weeks 5–7)
The focus of this step is to identify and select the most relevant datasets to train and evaluate the VQA model. The research will include the analysis of popular datasets such as VQA 2.0 and GQA, with special attention to their applicability for accessibility-focused tasks. Key factors such as image diversity, question relevance, explanation quality, and the presence of accessibility-related content will be evaluated. As a result of this step, suitable datasets for the project will be identified.

Step 4: Training and Testing the VQA Model (Weeks 7–11)
This step involves training the selected VQA model to generate accurate answers based on visual inputs. The process begins by dividing the data into **training, validation, and testing sets** to ensure robust performance evaluation. Hyperparameters such as **learning rate, batch size, and dropout rates** will be tuned to achieve optimal results. Continuous monitoring of **metrics like training loss and accuracy** will help track the model's progress, with adjustments made as needed to improve performance. After training, comprehensive testing will be conducted on **campus-specific images** to validate the model's accuracy, ensuring it aligns with the needs of visually impaired users.

Step 5: Creating and Labeling a Dataset (Weeks 10–14)
The goal of this step is to collect and label a dataset of campus-specific images with relevant questions and answers to align the model with real-world use cases. Each image will be annotated with contextual questions and corresponding answers. The questions will focus on objects, locations, and interactions commonly encountered on campus, ensuring that the VQA model can generate useful responses. To maintain data quality, multiple quality checks will be performed to ensure annotation consistency and accuracy. This labeled dataset will become a crucial part of the training and testing process, improving the system's ability to provide meaningful answers in real-world settings.

Step 6: Performance Optimization (Weeks 12–14)
The team will work together to improve the system's efficiency and responsiveness. Focusing on reducing latency and resource load, team members will optimize model parameters, refine code, and adjust configurations. To ensure these improvements meet project requirements, the team will conduct iterative testing to evaluate processing speed, memory usage, and response

times, making adjustments as necessary. Throughout the process, the team will document their findings and finalize performance benchmarks to confirm the system meets desired standards before moving on to the next phase.

NLP Module Development (Weeks 12–14)

Step 1: Developing a Model for Recognizing Voice Commands (Weeks 12–14)
The focus of this step is to develop a model that can accurately recognize and process user voice commands. This will involve integrating speech-to-text (STT) libraries such as Google Speech-to-Text to convert spoken input to text. The output of the STT module will serve as a textual representation of the user's spoken commands, enabling seamless interaction with the VQA system.

Step 2: Integration of Natural Language Processing Algorithms (Weeks 12–14)
The goal is to integrate NLP algorithms with speech-to-text conversion libraries to process user commands and convert them into text queries for the VQA model. By leveraging these NLP algorithms, the system will recognize and interpret voice commands from users, ensuring accurate conversion into text for the VQA model to process. The output will be a fully integrated NLP system that enables seamless voice command recognition, allowing users to interact with the VQA model through natural speech.

Step 3: Sending the Generated Answer to the User via Voice (Weeks 12–14)
The text-based response generated by the VQA model will be converted into a speech format using a **Text-to-Speech (TTS)** engine. This will allow the system to communicate the answer verbally to the user. The speech output will be clear and contextually appropriate, ensuring that visually impaired users receive the information they need in real-time, making navigation more intuitive and accessible without relying on visual cues.

Doğukan: Converting textual outputs of the VQA model to speech format to send the user.

Outcome and Impact (/ 20 Points)

Explain the expected outcome of your project. If it is a software product, try to include example screen designs, if it is a hardware product, try to provide detailed technical specifications, if it is research output try to explain the outcome's contribution to the field. Also, explain the potential impacts of your results. These may be how the result will be used in real life, how it will change an existing process, or where it will be published, etc.

Campus Navigator is designed specifically for visually impaired individuals. The system will function entirely through voice commands and provide audio

responses based on real-time image analysis. There is no graphical interface, as the system is built to communicate solely via auditory interaction.

Users will interact through spoken commands (e.g., "What is this building?"), and the system will respond with audio, describing the surrounding environment based on real-time captured images. The assistant will take periodic photos from the user's smartphone or a wearable camera to gather visual data about the surroundings, which will be processed by the VQA model.

If a minimal user interface is required (for setup or configuration), it will be designed with large, high-contrast buttons and simple navigation to accommodate visually impaired users. However, once the application is running, all communication will be through natural language processing (NLP) and speech synthesis.

This project is based on a practical VQA-based assistant that does not require visual interaction, allowing visually impaired users to navigate real-world environments independently. It bridges the gap between AI advancements and accessibility, expanding the scope of how VQA systems can be applied in real-life scenarios.

Our project is designed to help visually impaired individuals navigate our university campus independently, without needing assistance from others. For example, a user can ask for directions to a specific building or classroom, and the system will provide audio instructions to guide them. It will also help them cross roads within the campus, avoid obstacles while walking, and locate important landmarks such as entrances, elevators, or stairs.