Memorial University of Newfoundland

Faculty of Engineering and Applied Science

Multilingual Captioning with Text-to-Speech

SUPERVISOR:
DR. WEIMIN HUANG

MEMBER:

SHUVAJIT BARUA - 202384127 (shuvajitb@mun.ca)

FORHAD AHMED KHAN - 202293581 (forhadak@mun.ca)

Outline

1. Abstract

With the rapid surge in digital media consumption, the challenge of understanding image-based content continues to grow for individuals facing linguistic and accessibility limitations. In response to this urgent issue, our research presents a powerful web-based system that automatically creating insightful captions for images, enhanced by in-depth facial analysis and displayed through multilingual text-to-speech (TTS) outputs. Users can use the cameras on their devices to upload or take pictures in real time. he system employs several state-of-the-art deep learning models — including BLIP, ViT-GPT2, and BLIP-2 — to produce precise and context-aware descriptions. These captions are further enhanced by facial attributes such as age, gender, and emotional expressions, then smoothly translated into Bengali, Hindi, Chinese, and English. High-quality audio is included with the translated texts to improve accessibility. For users with different linguistic and cultural backgrounds, the readability and inclusivity of visual information are improved by combining language and accessibility issues into a single pipeline.

2. Introduction

For visually challenged people, negotiating the physical world without sight creates unending and sometimes extraordinary problems. Daily activities—such as recognizing people in a room, grasping facial emotions during talks, or reading scenes in strange surroundings—demand help, usually from another person. His dependence might compromise liberty and lower self-assurance in social or professional encounters. Although necessary assistance gadgets, guiding dogs, and canes offer movement and spatial awareness, they cannot read visual information like who is present, what emotions they are displaying, or what activities are taking place in a particular scenario. According to WHO, around the world, at least 2.2 billion people have close or permanent vision impairment. Only 36% of people with refractive error-related distance vision impairment and 17% of those with cataract-related vision impairment worldwide are thought to have had access to a suitable remedy.[1].

Several of these challenges have crossed into the digital domain as technology develops. For social media, online education, digital journalism, and even public information systems, images and videos have taken center stage as the primary means of communication. For those who are blind or have limited eyesight, this shift has widened the accessibility divide. Although screen readers and alt-texts offer minimal support, they often miss deeper context and rarely convey emotional, positional, or multi-person details within an image. Most existing image captioning tools are monolingual and lack integration with real-time audio narration—further limiting their practical utility in diverse, multilingual communities.

Although automatic image captioning has seen notable progress, most existing systems remain limited in functionality and scope. Earlier models relied on handcrafted rules or template-based methods (Farhadi et al., 2010), while more recent approaches using CNN-RNN and Transformer-based architectures such as BLIP[5], ViT-GPT2 [4], and BLIP-2 [6] have improved descriptive accuracy. Nevertheless, these systems often fail to account for deeper contextual understanding—such as the age, emotion, or position of people in the image—and largely operate in a monolingual (primarily English-only) context. This limits their real-world usability across linguistic regions and among non-English speakers.

In response to these real-world and digital accessibility gaps, we propose a unified assistive system that automatically generates human-like image captions enriched with facial analysis, translates them into multiple languages, and delivers them through natural-sounding audio narration. Users can either upload images or capture them live using a webcam. Our system generates accurate and diverse descriptions using three advanced captioning models—BLIP, ViT-GPT2, and BLIP-2. These captions are enhanced using facial attribute detection (age, emotion, gender, position), translated into English, Bengali, Hindi, or Chinese, and synthesized into audio using Google TTS. Intelligent caching optimizes the whole process to increase speed and responsiveness. Our method offers a feasible, real-world answer for enabling visually impaired people to access and understand their environment independently and meaningfully by combining vision, language, emotional detection, multilingual support, and speech synthesis into a seamless workflow.

3. Literature Review

Automated image captioning and multilingual text-to-speech (TTS) systems have been radically altered by recent advancements in artificial intelligence, especially in computer vision, natural language processing (NLP), and voice synthesis. Modern inventions and relevant studies that have greatly supported photo captioning technologies, text-to-speech synthesis, and integration of these technologies into consistent systems are discussed in this paper.

3.1 Image Captioning

Farhadi et al. [2] suggested one of the first image captioning systems depending on manually created templates and rule-based methods. Although innovative, their approach lacked adaptability and did not produce thorough or contextually rich descriptions. Fixed rule systems and little visual-textual alignment caused most of the constraints. Vinyals et al. [3] proposed a CNN-RNN-based encoder-decoder architecture that indicated a significant shift in image captioning. Grammatically acceptable captions and contextually suitable ones were produced by the method through the combination of recurrent neural networks for language generation with convolutional neural networks for visual feature extraction. However, its sequential decoding nature limited training speed and constrained scalability.

Dosovitskiy et al. [4] introduced the Vision Transformer (ViT), adapting Transformer architectures—originally developed for NLP tasks—to vision-based data. ViT demonstrated superior performance on several image understanding benchmarks, proving that self-attention mechanisms could replace convolutional operations. Nonetheless, large-scale datasets were required for practical training by ViT, and high computational costs were suffered. Li et al. [5] proposed BLIP, a vision-language pre-training approach that combined image and text encoding for caption synthesis. Using cross-modal attention and contrastive learning, it successfully attained state-of-the-art performance on several benchmarks. Though accurate, BLIP was sensitive to domain-specific changes and needed significant pre-training efforts. Li et al. [6] expanded on their earlier work using BLIP-2, combining larger-scale pre-trained language and visual models to increase captioning accuracy. The des gn allowed smoother caption creation by removing the requirement for intermediate visual tokenization. For lightweight or real-time applications, nevertheless, the scale and complexity of the model presented difficulties. Rennie et al. [7] presented a reinforcement learning methodology termed Self-Critical Sequence Training (SCST), which enhanced caption production by utilizing signals like CIDEr. This approach enhanced caption fluency and alignment with human evaluations. Nevertheless, it caused instability during training and relied on meticulously selected reward measurements. To get captioning models to focus on important parts of images, Anderson et al. [8] created the Bottom-Up and Top-Down Attention model. This produces more detailed and visually supported descriptions. Unfortunately, the method's inference time and complexity increased because of the need for object detection as a pre-processing step.

3.2 Text-to-Speech (TTS) Systems

Tacotron was proposed by Wang et al. [9]. It was an end-to-end speech synthesis system. Character sequences were mapped directly to mel-spectrograms. The speech was natural and smooth. Concatenative and parametric methods were eliminated. But it was autoregressive. So misalignment and slower inference.

Tacotron 2 was introduced by Shen et al. [10]. The system integrated Tacotron with a WaveNet vocoder. The quality of speech was significantly improved. The results were clearer and more expressive. Despite the improvement, the model required high computational power. It also relied heavily on external alignment tools.

A non-autoregressive model was introduced by Yamamoto et al. [11]. The model was called Parallel WaveGAN. It was built for fast and high-quality speech synthesis. Real-time performance was achieved effectively. However, the model needed carefully paired training data to perform well. It allowed audio quality to be preserved while allowing real-time inference. Its performance in low-resource environments suffered, too, as it needed high-quality paired training data.

The gTTS API, a production-ready neural text-to-speech system supporting several languages, was created by Google AI [12]. Its accessibility features and smooth integration are appropriate for practical uses. Still, it is cloud-dependent and has no customization options for prosody or emotion. Amazon Web Services [13] and Microsoft Azure [14] launched commercial TTS systems, respectively Polly and Azure TTS, offering multilingual synthesis and expressive voice capabilities. Though they may limit fine-tuning and need online access, these tools provide scalability and simple implementation.

3.3 Multilingual Translation

By substituting recurrence with self-attention, Vaswani et al. [15] created the Transformer framework, hence transforming sequence-to-sequence translation. Over long sequences, this greatly enhanced training efficiency and translation quality. Though they are quite effective, Transformers need a lot of hyperparameter adjustment and computing power. Wu et al. [16] suggested Google's neural machine translation (GNMT) system, which used an attention-based encoder-decoder architecture to improve fluency and translation consistency. It had problems with uncommon words and needed a lot of inference time. By training a multilingual model on common vocabulary and sentence pairings, Johnson et al. [17] expanded the GNMT framework to allow zero-shot translation. Though it struggled with underrepresented language pairs, the method worked well across languages.

3.4 Integration of Image Captioning with TTS

Park et al. [18] introduced an integrated pipeline that generated image captions and synthesized them into spoken narratives. Designed to improve accessibility for visually impaired users, the system combined vision-language models with TTS synthesis. But multilingual support was missing. The system only used static image domains. A unified framework was proposed by Qiao et al. [19]. It used CNN-RNN based captioning with neural TTS models. Audio narration for assistive technologies was supported. But the system had latency. Flexibility across languages was also limited. A comprehensive system was proposed by Chuang et al. [20]. It included facial recognition, multilingual captioning and TTS synthesis. Descriptions were detailed and enhanced with facial analysis. The system was useful for accessibility and educational purposes. But deployment was more complex. Multiple pre-trained models were required. This gap is addressed in our project. We used captioning models like BLIP, ViT-GPT2 and BLIP-2. DeepFace was used for facial analysis. Google Translate was used for multilingual translation. gTTS was used for speech synthesis. All these components were combined into a single web-based system. The goal is to make digital accessible for diverse user groups.

4. System Overview

Many image captioning systems don’t support real-world accessibility. As a matter of fact, most of them support English only and the audio output is not provided. Human-centered features like age and emotion are not included in those systems as they mainly prioritize object recognition. As a result, users with visual impairments face significant barriers. Those who rely on audio input in multiple languages also face difficulties. Accessibility is not met in practical scenarios. To solve this, we have developed a web-based solution that is designed to improve visual accessibility and is also appropriate for visually impaired users and non-English speaking users. Users can upload a picture or capture a webcam photo through the browser. Based on user selection, one of the modern captioning models was used. Each caption is refined by facial analysis where key visual attributes are identified. These key attribues are age, emotion, and facial position. The description becomes more precise and user-focused. The translation is done only after the caption is finalized. Four languages are currently supported in the system, those are English, Hindi, Bengali, and Chinese. The translated caption is then converted into clear and natural-sounding audio in the user's selected language. Moreover, Caching is used to improve performance. A hashed version of the image is stored to prevent repeated processing. The whole platform brings vision, language, and speech together. It provides a simple and effective interface for real-world accessibility.

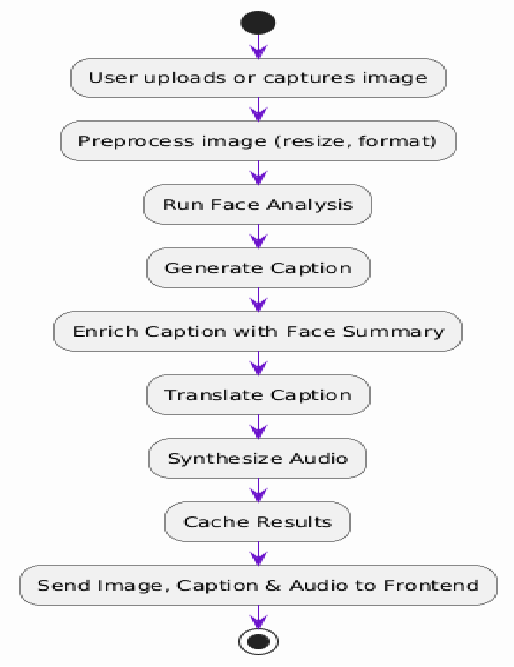

User uploads or captures image
Preprocess image (resize, format)
Run Face Analysis
Generate Caption
Enrich Caption with Face Summary
Translate Caption
Synthesize Audio
Cache Results
Send Image, Caption & Audio to Frontend

Figure: System Overview Diagram

5. Methodology

5.1 Workflow Steps

5.1.1 Image Input

We start the process by using our easy to use interface to take new photos with the user's webcam or upload existing digital images from the local drive. All photos undergo preparation processes to ensure uniformity and compatibility with the backend deep learning models, including resizing and format conversion.

5.1.2 Face Analysis and Semantic Context Enrichment

Facial attributes are analyzed in the next step. The DeepFace framework is used for this analysis, which makes the captions more meaningful and enriched as a result. Using deep convolutional neural networks, this component identifies and examines human faces in the picture, hence extracting characteristics such as age, gender, and emotional state. These human-centric elements provide an extra semantic layer that lets the machine include socially relevant signals into the last caption output and beyond simple object detection.

5.1.3 Image Caption Generation

The image—now enhanced with contextual signals—is sent to the chosen image captioning model for linguistic description following face analysis. Users have three state-of-the-art models to pick from: BLIP, ViT-GPT2, or BLIP-2. These models use transformer-based topologies and visual attention methods to encode the picture material and produce fluent, semantically correct captions. The algorithm may generate more informative and human-aware picture stories by combining facial characteristics with visual elements.

5.1.4 Caption Enrichment

The chosen model first produces a caption; then, through facial analysis findings, it is further developed. The original caption is augmented by this enrichment technique, adding structured human-centric information—such as the projected age, gender, emotional expression, and spatial position (e.g., left, center, right) of detected people. Embedding these semantic signals helps the system turn a simple description into a more prosperous, context-aware story that reflects what is in the picture and how the individuals inside it seem and interact spatially.

5.1.5 Translation

The enriched caption is then translated using the Google Translate API in this step. The system supports four significant languages. These include English, Bengali, Hindi, and Chinese. Transformer-based translation models are applied to improve accuracy and fluency.

5.1.6 Text-to-Speech

The translated caption is then spoken using Google's Text-to-Speech (gTTS) engine. Based on Tacotron and WaveNet architectures the speech is natural sounding and accessible for visually impaired users.

5.1.7 Output & Caching

The system caches results using SHA-256 hashing to be efficient and reduce redundant processing[21]. So repeated requests for the same image, language and model will be served quickly without processing again, making the user experience faster.

5.2 Captioning Models

The performance and richness of the captions in this system is dependent on the underlying models for visual understanding and human centric analysis. The system uses three state-of-the-art image captioning models—BLIP, ViT-GPT2 and BLIP-2—along with DeepFace for facial attribute extraction. Each model is chosen to fulfill a specific role in the pipeline to make it accurate, descriptive and inclusive.

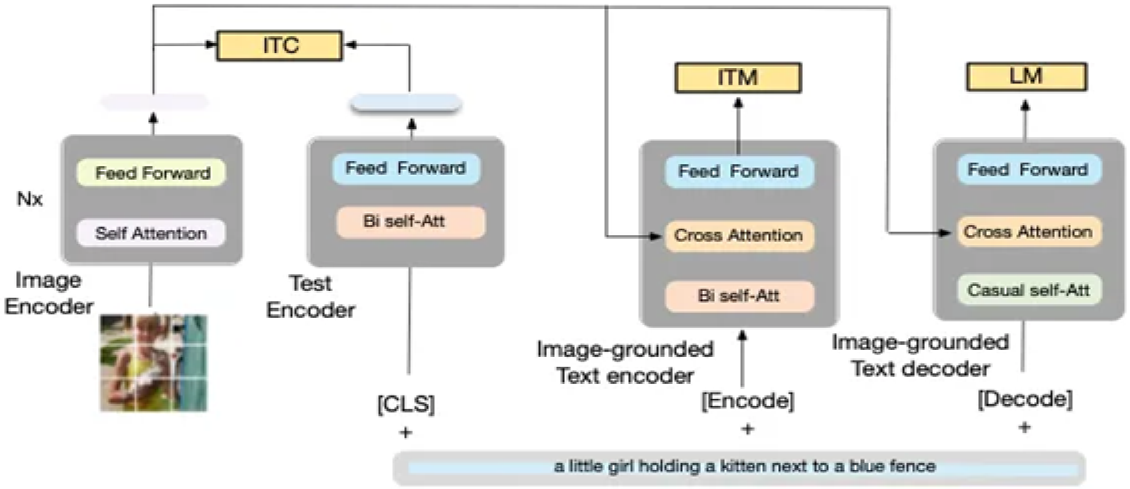
5.2.1 BLIP: Bootstrapping Language-Image Pretraining

BLIP (Bootstrapped Language-Image Pretraining) is a vision-language model that can do image captioning, picture-text retrieval and visual question answering in one go. It uses a Vision Transformer (ViT) as its visual backbone, breaking images into patches and processing them as tokens through self-attention layers to extract local and global visual features. On the language side, BLIP has a Transformer-based encoder-decoder structure that generates fluent and semantically grounded captions.

BLIP's bootstrapped training method allows it to learn from noisy internet data by generating weak captions and refining them as it goes. It can link words to image regions through cross-attention to coordinate visual and textual modalities. It is trained with a combination of objectives: contrastive learning for image-text pairs, matching to verify relevance and language modeling for generation.

BLIP generates accurate, coherent and contextually rich captions without fine-tuning. For real-world applications like our multilingual picture captioning system where accuracy and adaptability across languages and scenarios matter, its generalization and output are perfect.

BLIP is a solid baseline model in this system. It's versatile so it's good for people who need captions in various real-life scenarios. Its core message will be understood because it can maintain semantic fidelity between text and image.



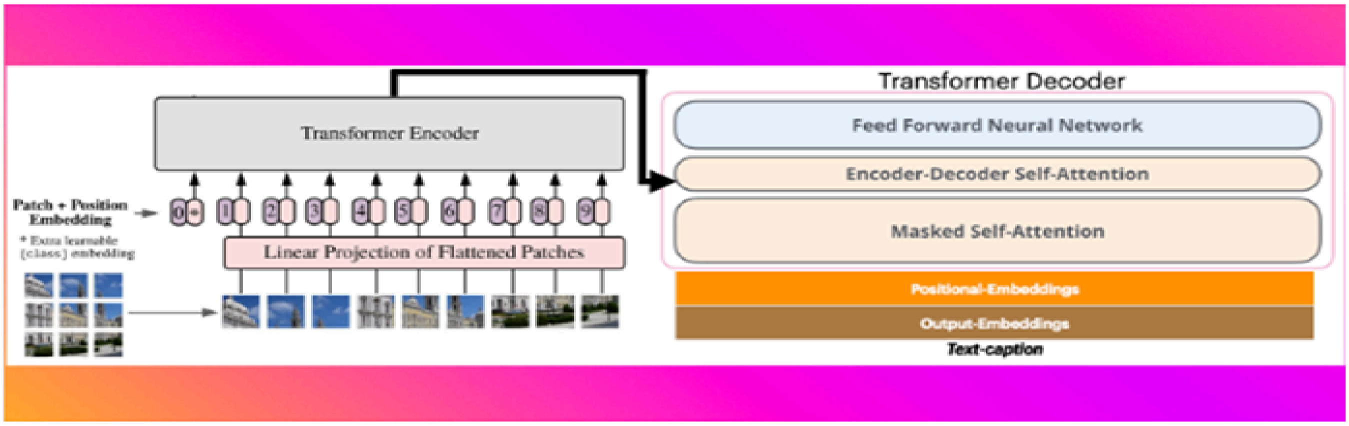
Figure[22]: BLIP

5.2.2 ViT-GPT2: Vision Transformer with GPT-2 Decoder

ViT-GPT2 is a hybrid image captioning system that uses Vision Transformer (ViT) for feature extraction and GPT-2, a pre-trained autoregressive language model, for text generation. ViT is the encoder here, it splits the image into fixed size patches and embeds them into a sequence of visual tokens. ViT captures local patterns and global structure through multi-head self-attention and summarizes the visual content into a set of rich contextual embeddings.

These visual embeddings are then projected into the same input space as GPT-2 and passed into the decoder. GPT-2, trained on a huge amount of text, uses its language modeling capability to generate text word by word. Because GPT-2 operates autoregressively, it generates one token at a time, using previously generated tokens and the visual context as input at each step.

ViT-GPT2 is particularly advantageous due to its modular design, allowing independent improvements or replacements of the visual or language components. Although its lightweight footprint and quicker inference make it a feasible alternative for edge devices, real-time applications, or when computational resources are limited, it may not always generate captions as nuanced as models trained end-to-end on multimodal tasks.


Transformer Decoder
Feed Forward Neural Network
Encoder-Decoder Self-Attention
Masked Self-Attention
Positional-Embeddings
Output-Embeddings
Text-caption
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0* 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches

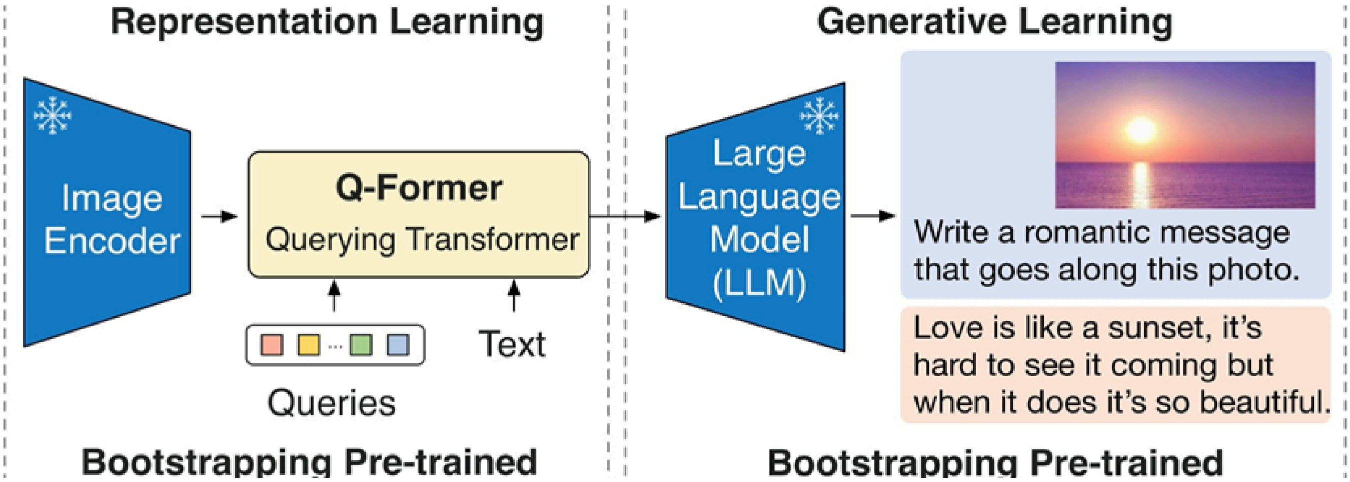
Figure[23]: ViT-GPT2

5.2.3 BLIP-2: Bootstrapped Language-Image Pretraining with Frozen Image Encoders

Designed to close the modality gap between vision and language, BLIP-2 (Bootstrapping Language-Image Pretraining 2) is among the most sophisticated image captioning systems, using a two-stage architecture. Unlike previous methods that closely link image and language processing, BLIP-2 separates the vision encoder from the language model to provide scalability and flexibility. In the first stage, a frozen vision encoder—typically a high-performance Vision Transformer-like ViT-g or EVA—extracts rich image features. These features are high-dimensional and retain detailed visual semantics from the input image.

In the second stage, a Querying Transformer (Q-Former) is introduced. This module learns a small set of trainable query tokens that interact with the frozen vision features via cross-attention. The output of the Q-Former is a compact, language-friendly representation of the image, which is then passed as a prompt to a pre-trained large language model (LLM) such as Flan-T5. The LLM generates fluent and highly contextual captions based on the prompt, leveraging its massive linguistic knowledge while remaining agnostic to the visual feature extraction process.

Among the many benefits of this decoupled architecture are that it dramatically lowers training costs by maintaining the vision and language models frozen or partially calibrated, allows zero-shot generalization to new tasks and domains, and fosters higher flexibility, allowing independent upgrades of components. BLIP-2 produces emotionally intelligent, nuanced captions, grasps complicated scenes, and recognizes abstract concepts. Among the three models in our system, it provides the best captioning quality, perfect when thorough interpretation and expressiveness take precedence.


Representation Learning
Generative Learning
Image Encoder
Q-Former
Querying Transformer
Queries
Text
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained
Bootstrapping Pre-trained

Figure[24]: BLIP-2

5.2.4 DeepFace

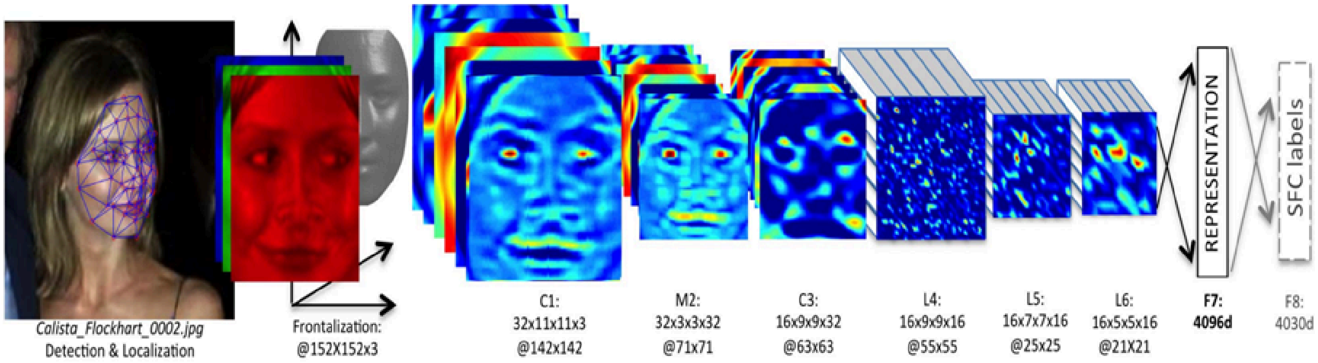
An open-source face analysis tool called DeepFace is part of the system that adds more meaning to the output captions by using data focused on people. DeepFace has deep convolutional neural networks that can find faces, align them, and pull out attributes in real-time. For people who have trouble seeing, this is an integral part of accessibility apps because it ensures that the labeling process includes the scene and the people in the picture, who they are, and how they're feeling.

First step is face recognition. DeepFace uses MTCNN (Multi-task Cascaded Convolutional Networks) or RetinaFace to detect the face in an image. These detectors work well in different lighting conditions and angles. Then DeepFace aligns the face using eyes, nose and mouth after detecting a face. This alignment stage ensures the face looks normal before examination and corrects any position variations.

A pre-trained embedding model like VGG-Face, Facenet, OpenFace or ArcFace is passed through them after faces are aligned. Feature vectors with many dimensions that stores face information are produced by these deep learning models. These vectors are then used to verify identities and analyze attributes. DeepFace has separate submodels for figuring out a person's age using CNNs that are based on regression, figuring out their gender using binary classification, and figuring out their emotions using multi-class classifiers that were trained on datasets like FER-2013. These categories can guess if someone is happy, sad, angry, neutral, or shocked.

In addition to essential facial traits, the system can also tell you where each recognized face is in the picture—whether it's in the middle, on the left, or right. It placed the bounding box's horizontal location with the image's width to do this. More detailed descriptions beyond object recognition are produced by structuring and appendixing the combined attributes—age, gender, emotion, and position—in the picture captions.

When DeepFace is added to the stream, the system makes captions that show what's happening, who's there, and how they look mentally. This combination makes the captions much more semantically prosperous and inclusive, primarily when used for assistive purposes in the real world. People with visual impairment can understand the environment clearly, making the experience more meaningful and informative. Also, because face analysis uses a lot of computing power, the system stores these results separately so that they don't have to be processed twice, which improves speed and responsiveness even more.



Figure[25]: Deepface

5.3 Caching

A key system optimization tool meant to reduce repeated processing, increase reaction times, and enhance user experience is caching. All computationally demanding activities are speech synthesis, translation, image captioning, and facial analysis. The system uses a thorough caching method to address this, saving and reusing past-produced outcomes whenever feasible.

The SHA-256 hashing method is key to this caching strategy. A commonly used cryptographic tool, SHA-256—or Secure Hash Algorithm 256-bit—produces a consistent 256-bit (32-byte) result from any input, including picture files. The unique digital fingerprint of the input data is this hash. Chosen for its properties, SHA-256 is deterministic, collision resistant and input sensitive. A different hash will be generated from even a small change to an image. This means that only the same images (irrespective of the file name) are considered the same during caching. The collision resistance ensures reliability since no two different inputs will generate the same hash in real world.

The system's caching of data, which stores results across three axes – the image captioning model used, the selected translation language and the face analysis output – is identified by this weird image hash. Previous results can be recalled quickly by the system without the processing pipeline being re-run for every user input. Moreover, face analysis results are stored separately, since language or captioning models have nothing to do with them. This design makes the system more efficient and scalable by eliminating recomputation.

The system becomes real-time and multilingual for captioning applications by combining SHA-256 hashing with a well organized cache structure, reducing duplicated processing and being super fast.

5.4 Language & Audio Handling

It's ability to offer multilingual support by combining latest translation and audio synthesis technology is what makes this solution. This will be super helpful for users with visual impairments or language limitations as it will ensure that both the visual description and face attribute analysis is displayed in the user's chosen language and via spoken audio.

5.4.1 Neural Machine Translation via Google Translate API

The translation module in our system uses Google Translate API which is built on top of Neural Machine Translation (NMT) powered by Transformer-based architectures as introduced by Vaswani et al. (2017). In essence this architecture changed the face of machine translation by allowing the network to look at the whole sequence at once and capture global dependencies between words regardless of their position in the sentence. Replacing traditional recurrent models with self-attention mechanisms achieved this.

The architecture follows an encoder-decoder model. The input sentence (usually in English) is converted into a sequence of high-dimensional vectors called contextual embeddings. Self-attention layers that look at every pair of words in the sentence enrich these embeddings. The meaning of each word and its role in the overall sentence structure is understood by the model through this.

The decoder produces the translated output in the target language by cross-attention to the encoded representations and one word-at-a-time prediction. Every decoding phase uses a cross-attention approach to highlight the most pertinent parts of the source sentence and combine this knowledge into the target-language generation process. This produces grammatically correct translations that are contextually relevant to the original meaning.

The fundamental benefit of this design is its capacity to predict complex linguistic events—including idiomatic phrases, nested clauses, and long-range relationships—with a better degree of accuracy than prior phrase-based or RNN-based models. Because of their parallelizable nature, transformers may also be trained and inferred quickly, making them suitable for real-time applications.

Our system makes this feature especially important. Translating the whole caption as a single semantic unit guarantees that the output is consistent, fluid, and contextually intact, as the input text to the translation module contains both descriptive captions and structured face summaries (such as emotion, age, and position). Translating parts in isolation might lead to fragmented or literal translations. Thus, it guarantees that the multilingual output is accessible but also natural-sounding and relevant.

5.4.2 Text-to-Speech Synthesis via gTTS (Google Text-to-Speech)

The system uses Google's Text-to-Speech (gTTS) service, which is based on a potent mix of Tacotron and WaveNet-inspired architectures, for turning translated text into spoken audio. Producing audio that is very natural, expressive, and comprehensible, these models indicate a significant advance in neural voice synthesis.

The first stage of the synthesis pipeline is Tacotron. A time-frequency representation of audio that captures tone, pitch and rhythm, mel-spectrogram is a sequence-to-sequence model that converts input text. Tacotron converts characters or phonemes into hidden representations using an encoder, then uses an attention mechanism to map and align these representations to the desired acoustic space. By this alignment Tacotron can preserve prosody, stress patterns and pronunciation so the output speech has human-like expressiveness across languages and dialects.

Once the mel-spectrogram is generated it is passed into WaveNet which is a vocoder. Developed by DeepMind, WaveNet is a deep generative model that can produce high fidelity audio by modelling the waveform directly at the sample level. WaveNet generates the next audio sample based on all previous samples, unlike traditional vocoders that use simple rules or pre-recorded samples, and the input spectrogram, resulting in smooth and very realistic speech. It captures subtle intonations, breathing pauses and other natural characteristics of human speech.

Together Tacotron and WaveNet form a two stage synthesis pipeline: Tacotron does the linguistic-to-acoustic mapping and WaveNet refines that into waveform level audio with high temporal precision. The architecture can accommodate many languages and voices so the expression and intelligibility is consistent across dialects and speaking pace.

To help users with visual impairments or reading challenges our system uses this pipeline to convert enriched multilingual captions into speech. In addition to making image based digital content more accessible it helps auditory learners and non-native language users by giving them accurate, emotionally engaging narration in the language they prefer.

6. System Architecture

The developed application adopts a modular client-server architecture, facilitating scalability, maintainability, and cross-platform accessibility. The system supports real-time, multilingual image captioning, face attribute analysis, and audio narration generation, integrating multiple state-of-the-art AI models and web technologies within a cohesive pipeline.

6.1 Frontend Architecture

The frontend of the system uses conventional web technologies—HTML5, CSS3, and vanilla JavaScript—to provide a user interface that is responsive, lightweight, and platform-independent. The design of this product will enable users with different degrees of technical knowledge, especially those who depend on assistive technologies, to have a smooth and intuitive experience.

Images can be entered into the user interface in two distinct ways: either by uploading previously recorded photographs from the local file system or by capturing new images directly from the camera. The getUserMedia API drives the webcam functions by allowing users real-time access to the camera feed on their device. The video frame is drawn onto an off-screen HTML5 element when a camera captures a picture. It is then converted to a binary Blob object stored in JPEG format. Since it is compact, appropriately encoded, and appropriate for transmission, the binary data is eligible for transfer over HTTP.

After selecting or capturing the image, the frontend gathers further user inputs, including the image chosen captioning model (BLIP, ViT-GPT2, or BLIP-2) and the desired target language (English, Bengali, Hindi, or Chinese). The inputs and binary image data are encapsulated in a FormData object and submitted asynchronously to the backend using the Fetch API. This asynchronous communication preserves frontend responsiveness throughout the processing pipeline and eliminates the need for page reloads. This asynchronous communication maintains frontend responsiveness during the processing pipeline and removes the necessity for page reloads.

A loading animation is presented to indicate system activity during processing. Once captioning, translation, face analysis, and text-to-speech synthesis are finished, the backend sends back a structured JSON payload. This payload has:

- The caption text has been enriched and translated.
- The annotated picture URL shows face bounding boxes and labels found.
- A list of facial features, including age, gender, emotion, and relative location.
- Caption quality is shown by evaluation scores such as BLEU and METEOR.
- A link to the generated audio voiceover.

The frontend dynamically analyzes this response and modifies the interface to match. The caption is shown in a consistent style, and the annotated image is displayed on-screen. Users can play the narration straight in the browser using a <audio> tag that embeds the returned audio file. Visually challenged people and those with restricted reading abilities find this function essential.

An internal data structure is additionally run by the frontend to handle outcomes from several images. "Next" and "Previous" buttons let users go between previously uploaded or shot images, enabling side-by-side comparison of outcomes across several languages and models.

Accessibility is regarded as one of the main design priorities. The application is usable with screen readers and other assistive devices by ensuring that the user interface (UI) includes focusable elements, high-contrast text, keyboard navigation support, and semantic HTML markup. Screen sizes, from desktops to mobile devices, are fluidly adjusted by the layout without compromising functionality.

To sum up, the frontend design connects the user interaction layer with strong backend powers to create a real-time, interactive, and open platform for multilingual picture captioning and audio narration.

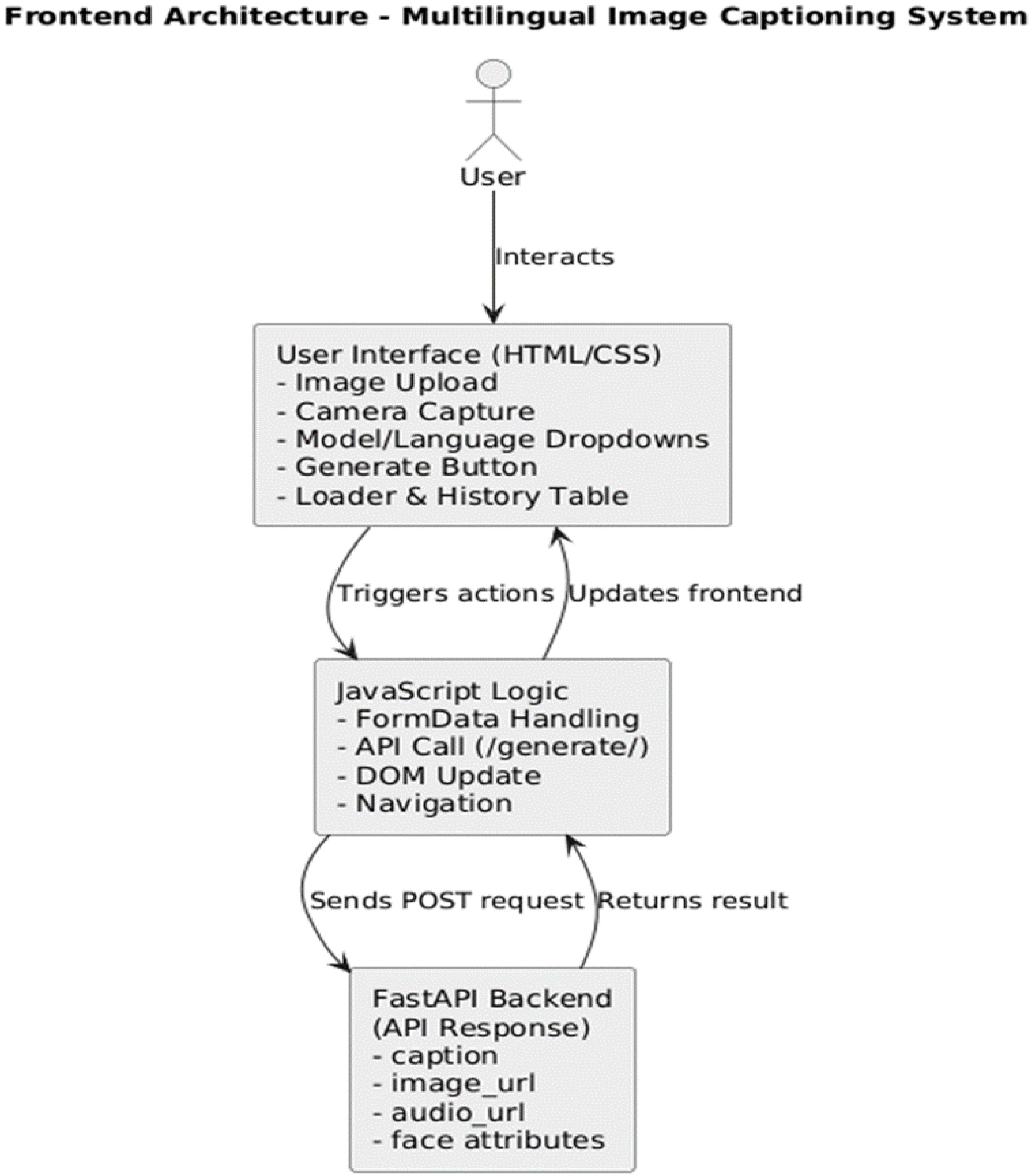
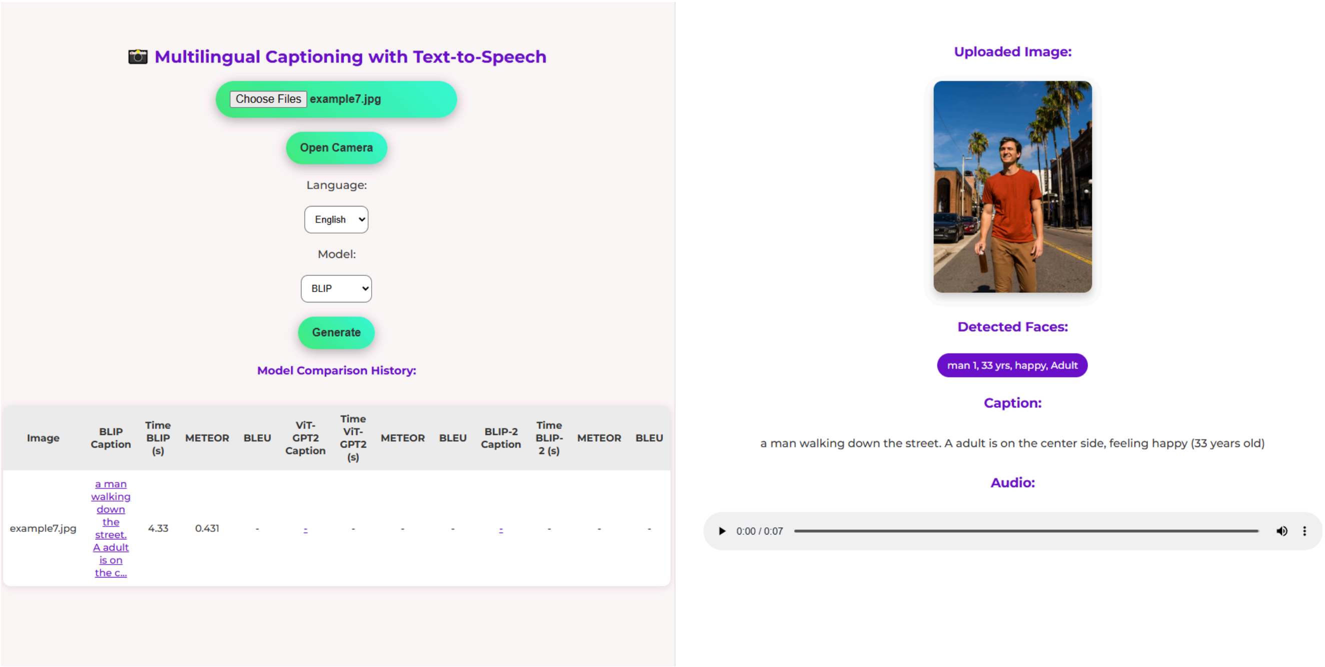


Figure: Frontend Architecture



📷 Multilingual Captioning with Text-to-Speech

example7.jpg

Language: English

Model: BLIP

Model Comparison History:

Image	BLIP Caption	Time BLIP (s)	METEOR	BLEU	ViT-GPT2 Caption	Time ViT-GPT2 (s)	METEOR	BLEU	BLIP-2 Caption	Time BLIP-2 (s)	METEOR	BLEU
example7.jpg	a man walking down the street. A adult is on the c...	4.33	0.431	-	-	-	-	-	-	-	-	-

Uploaded Image:

Detected Faces:

man 1, 33 yrs, happy, Adult

Caption:

a man walking down the street. A adult is on the center side, feeling happy (33 years old)

Audio:

Figure: Frontend view

6.2 Backend Architecture and Integration

FastAPI is a current Python framework made for high-performance asynchronous APIs used to build the backend system. This design lets the platform handle many requests for picture processing at the same time, keeping latency low even when a lot of requests are coming in at the same time. FastAPI uses asyncio and ThreadPoolExecutor to split up computationally intensive jobs like face detection, model inference, and audio generation into separate threads. This keeps the user interface responsive.

6.2.1 Image Captioning Module

Our system has three Transformer-based picture captioning models that users can choose from: BLIP, ViT-GPT2, and BLIP-2. These models are all built using the Hugging Face Transformers library.

- BLIP is loaded with its vision encoder and language decoder components and utilized for generating fluent, semantically aligned captions. It's effective for general-purpose captioning with balanced performance.

- ViT-GPT2 decouples image encoding (using Vision Transformer) and text decoding (using GPT-2). This model is used when the system requires faster, lightweight generation—especially useful in low-resource or real-time scenarios.

- BLIP-2 is the most advanced in our stack. It uses a frozen visual encoder and a Querying Transformer to format the output for large language models like FLAN-T5. This architecture enables zero-shot generalization, and we leverage it when the image context is complex or nuanced.

Each model is preloaded into memory at server startup, optionally offloaded to the GPU if available, allowing on-demand inference with minimal overhead.

6.2.2 Facial Attribute Analysis with DeepFace

The system combines DeepFace, which runs to enhance image captions with human-centric context.

- Face detection and alignment using backends like MTCNN or RetinaFace.

- Age estimation, gender classification, and emotion recognition using pre-trained CNNs such as VGG-Face or ArcFace.

Attributes are extracted and organized with spatial metadata for every identified face. The backend incorporates these descriptors in the final caption and uses image width to calculate each face's relative position. This enhances the output with a semantic richness that helps people with vision impairment.

6.2.3 Multilingual Translation Pipeline

The enhanced caption—including facial summary—is translated using the Google Translate API, which internally runs Neural Machine Translation (NMT) controlled by Transformer-based encoder-decoder topologies.

Our performance:

- Both the original caption and face summary are passed **as a unified text block**.

- The translated output is guaranteed to stay natural and context-aware by preserving grammatical and semantic consistency.

- Translation is available for English, Bengali, Hindi, and Chinese, hence supporting worldwide access.

6.2.4 Audio Narration with gTTS

Google Text-to-Speech (gTTS) then gets the converted caption for audio creation. Built on Tacotron, gTTS transforms text into mel-spectrograms; WaveNet creates realistic speech waveforms.

In our project:

- The translated caption (with numbers converted to words for TTS clarity) is synthesized into **MP3 format**.

- Audio generation is handled asynchronously to avoid blocking the main server thread.

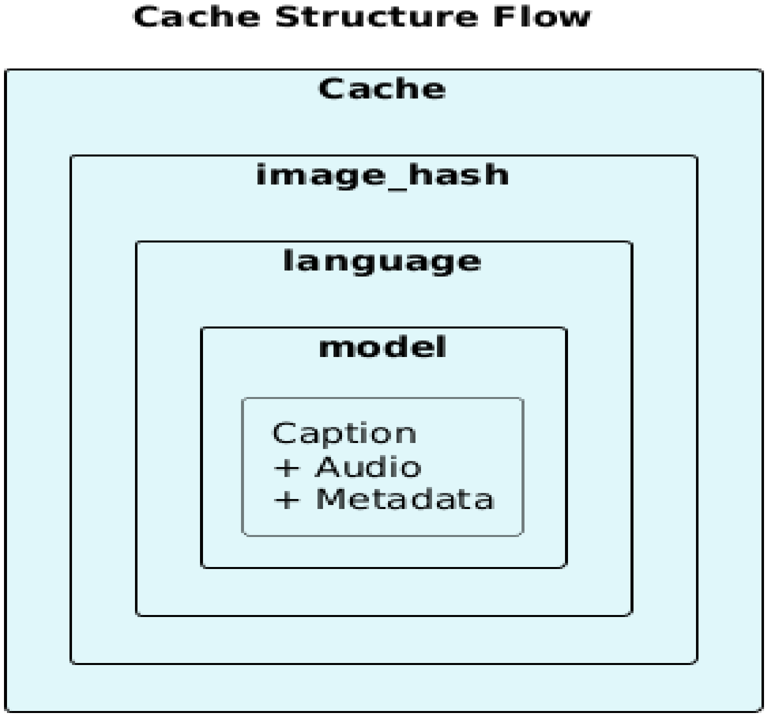
- This makes the system especially useful for **visually impaired users** who rely on auditory output.

6.2.5 Caching Architecture

We use a multi-level caching mechanism to maximize the whole workflow. This is how it functions:

- Every image file's unique identification is calculated using SHA-256 hashing. Caching uses this hash as the main key.

- The cache is structured hierarchically:



- We also store DeepFace results independently, so the same face data can be reused across languages or models without reprocessing.

This mechanism avoids repeated translation, TTS synthesis, or model inference for the same input—leading to faster response times and reduced GPU/CPU usage.

6.2.6 Robust Fallback Mechanisms

To keep the system reliable, both translation and TTS synthesis have fallback options. The system defaults to the original English caption if translation services fail—due to network issues, API restrictions or unsupported language inputs. Likewise, if gTTS can't produce audio for a certain language, like unsupported regions or synthesis interruptions, the system defaults to producing voice in English with the same caption. These fallbacks ensure the system continues to work and users get answers even under external service constraints.

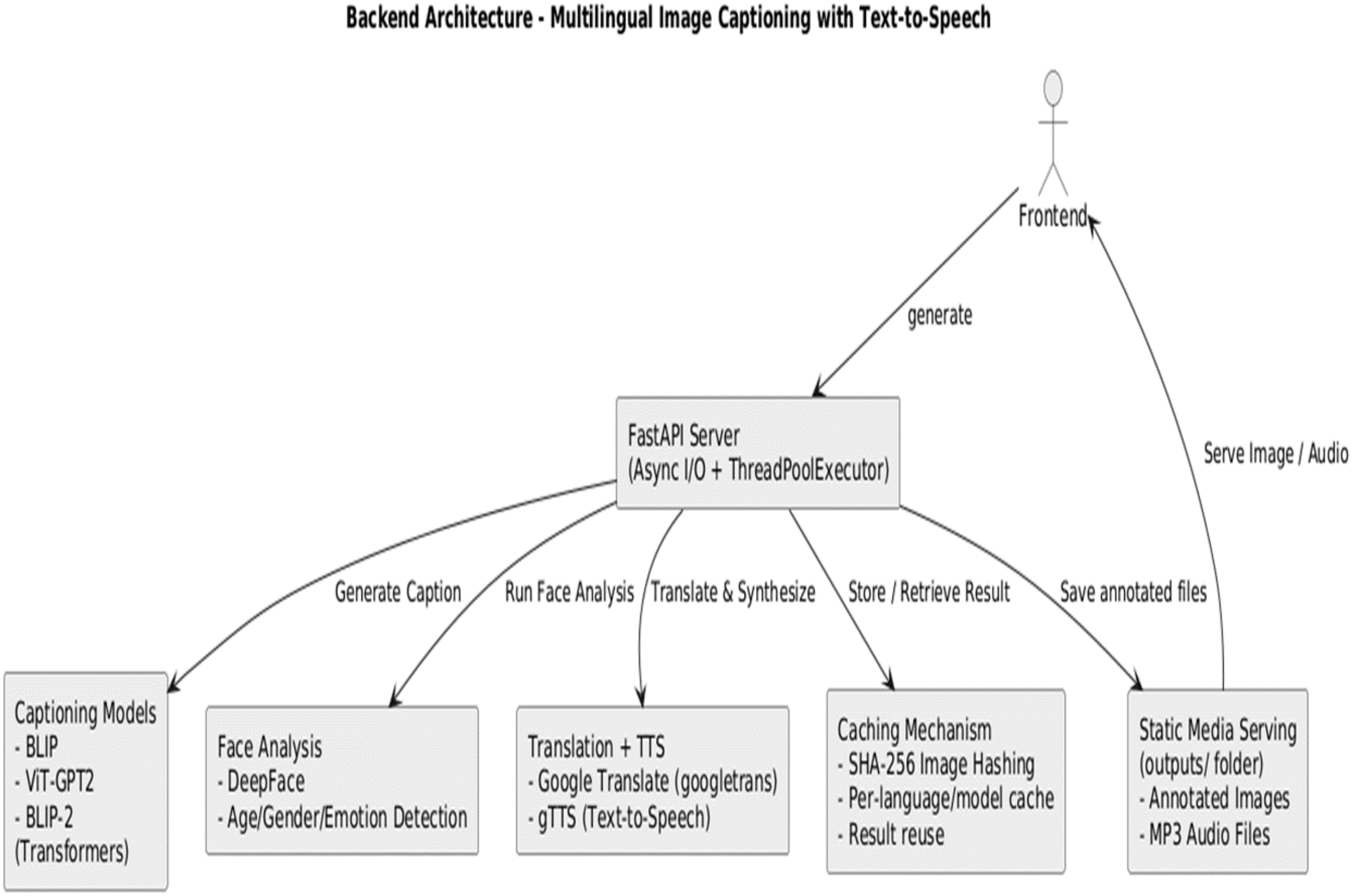


Figure: Backend Architecture

7. Technology Stack

Component	Technology Used
Frontend	HTML, CSS, JavaScript
Backend	Python, FastAPI
Captioning Models	BLIP, ViT-GPT2, BLIP-2
Face Analysis	DeepFace
Translation	Google Translate API (googletrans)
Text-to-Speech	gTTS
Additional Tools	OpenCV, PIL, Num2Words, PyTorch

8. Testing & Evaluation

8.1 Testing Setup

A dedicated testing setup was designed to test the effectiveness, accuracy and contextual understanding of the multilingual image captioning and text-to-speech (TTS) system. The system has multiple complex components—facial attribute analysis, caption generation, language translation and audio synthesis—all of which are critical to accessibility and descriptive richness. So a controlled and comprehensive testing approach was taken.

To support this, a custom evaluation dataset was created with 50 manually curated images, each with 5 human written reference captions. These captions describe the visual scenes, actions and emotional or contextual cues and are the benchmark for model performance. For focused demo and analysis, 3 images were selected from the dataset. These images are solo and group subjects, varied emotional states and diverse scene composition making them ideal for caption quality and face-aware enhancements.

Each image was run through the 3 integrated image captioning models—BLIP, ViT-GPT2 and BLIP-2—and their outputs were evaluated in English using 2 widely used automatic evaluation metrics—BLEU and METEOR.

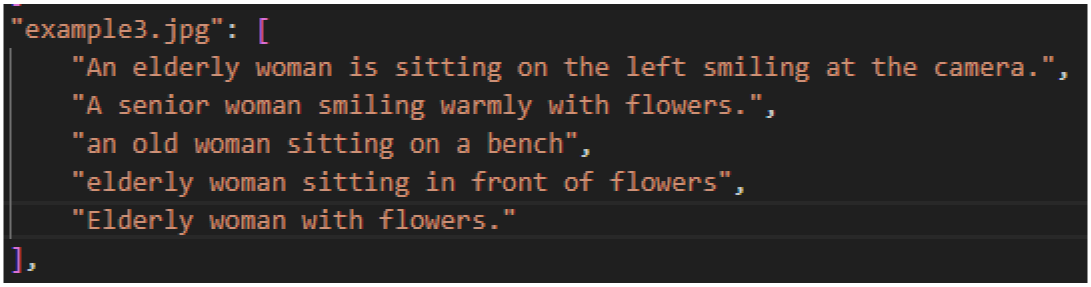
BLEU (Bilingual Evaluation Understudy) is a precision-based metric that measures how many n-gram overlaps are there between the generated caption and reference captions. It mainly focuses on surface level similarity, gives higher scores to captions that match the exact wording and structure of the reference texts. But BLEU can sometimes undervalue semantically correct but syntactically different outputs.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) on the other hand, considers both precision and recall and includes stemming, synonym matching and paraphrasing. It aligns generated captions to reference captions based on meaning rather than exact word matches, so it's more sensitive to semantic correctness and linguistic diversity. This makes METEOR better capture the contextual quality of captions especially when there are multiple valid phrasings. Both together gives a more balanced view—BLEU for surface fluency and syntactic match and METEOR for semantic correctness and descriptive.

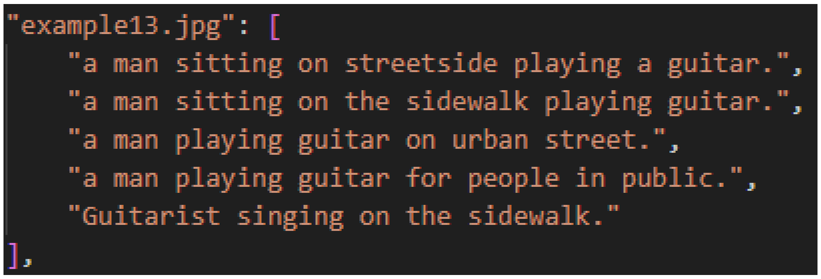
8.2 Reference Captions

We carefully crafted the reference captions to ensure consistency in tone, detail, and structure across all 50 images. These captions avoid overly simplistic phrasing and aim to provide realistic, human-style descriptions. Each image includes five corresponding captions that vary in sentence structure and lexical choice while preserving semantic meaning.

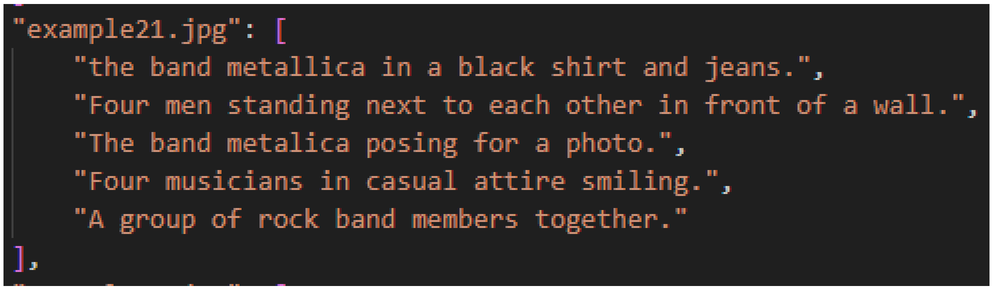
The snippets of the reference captions for the three tested images are given below:



```
"example3.jpg": [
    "An elderly woman is sitting on the left smiling at the camera.",
    "A senior woman smiling warmly with flowers.",
    "an old woman sitting on a bench",
    "elderly woman sitting in front of flowers",
    "Elderly woman with flowers."
],
```



```
"example13.jpg": [
    "a man sitting on streetside playing a guitar.",
    "a man sitting on the sidewalk playing guitar.",
    "a man playing guitar on urban street.",
    "a man playing guitar for people in public.",
    "Guitarist singing on the sidewalk."
],
```



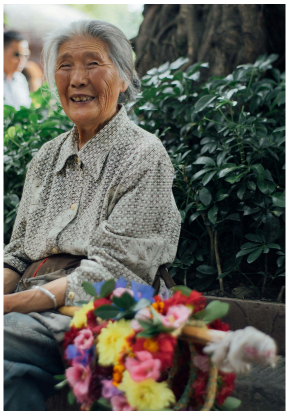
```
"example21.jpg": [
    "the band metallica in a black shirt and jeans.",
    "Four men standing next to each other in front of a wall.",
    "The band metalica posing for a photo.",
    "Four musicians in casual attire smiling.",
    "A group of rock band members together."
],
```

The ability of the models to capture not just basic objects or people, but also context, mood, and relationships within the image is helped to be evaluated by these references..

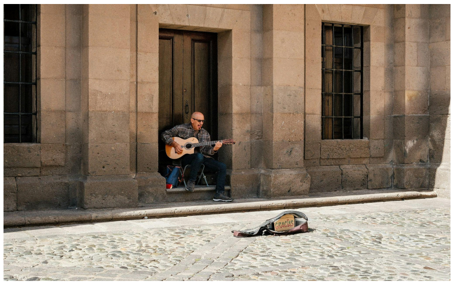
8.3 Results and Evaluation

As part of the evaluation process, our multilingual image captioning system was tested using three representative images selected from a larger set of 50 manually annotated samples. Each image is accompanied by five human-written reference captions, allowing BLEU and METEOR scores to be computed for captions generated by the three models: BLIP, ViT-GPT2, and BLIP-2.

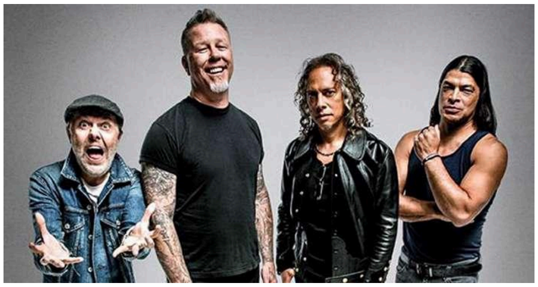
The test images used in the evaluation are given below:



example_3.jpg



example_13.jpg



example_21.jpg

The performance scores for each model across these three test images are presented in the table below:

Image Name	Model	Generated Caption	BLEU	METEOR
example_3.jpg	BLIP	an old woman sitting on a bench with flowers.	0.8034	0.2505
	ViT-GPT2	a woman sitting in front of a bunch of flowers.	0.4518	0.0787
	BLIP-2	a woman sitting on a bench.	0.7598	0.2078
example_13.jpg	BLIP	a man sitting on a bench playing a guitar.	0.4317	0.7472
	ViT-GPT2	a man sitting on the sidewalk with a stuffed animal.	0.5373	0.5319
	BLIP-2	a man playing a guitar.	–	0.5628
example_21.jpg	BLIP	the band metallica in a black shirt and jeans.	0.6606	0.8018

	ViT-GPT2	three men standing next to each other in front of a wall.	0.8314	0.0980
	BLIP-2	Metallica.	–	0.0543

Here are the results for each model across different image contexts. BLIP produces the most well rounded captions with good syntax and semantics. BLIP-2 while sometimes producing shorter outputs captures the essence with high contextual relevance. ViT-GPT2 is efficient and fluent but produces more general or surface level descriptions. Having both BLEU and METEOR scores gives us a complete view of each model's precision and linguistic quality and we can support multiple models based on different user priorities.

8.4 Multilingual Caption Generation

8.4.1 Captions Generated with Uploaded Image

To show the system's multilingual capabilities we generated captions for 3 test images in 4 languages: English, Bengali, Hindi and Chinese. These translations were generated using Google Translate API after facial enrichment was added to the original English captions. The results show the system can generate contextually relevant, emotionally aware and linguistically diverse descriptions.

Image	Model	Caption			
		English	Bengali	Hindi	Chinese
example_3.jpg	BLIP	an old woman sitting on a bench with flowers. A senior is on the left side, feeling happy (52 years old)	ফুলের সাথে বেঞ্চে বসে এক বৃদ্ধ মহিলা. একজন প্রবীণ বাম দিকে আছেন, খুশি বোধ করছেন (৫২ বছর বয়সী)	एक बूढ़ी औरत फूलों के साथ एक बेंच पर बैठी. एक वरिष्ठ बाईं ओर है, खुश महसूस कर रहा है (५२ वर्ष)	一个老太太坐在长凳上的鲜花. 大四学生在左侧, 感到快乐（五二岁）

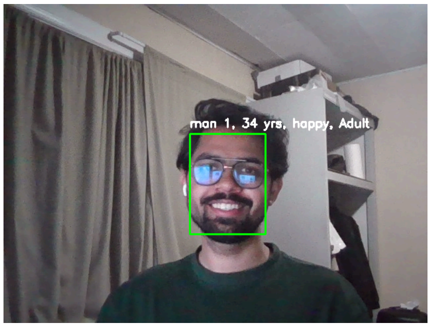
	ViT-GPT2	a woman sitting in front of a bunch of flowers. A senior is on the left side, feeling happy (52 years old)	এক মহিলা ফুলের সামনে বসে এক মহিলা. একজন প্রবীণ বাম দিকে আছেন, খুশি বোধ করছেন (৫২ বছর বয়সী)	फूलों के एक झुंड के सामने बैठी एक महिला. एक वरिष्ठ बाईं ओर है, खुश महसूस कर रहा है (५२ वर्ष)	一个女人坐在一束鲜花前. 大四学生在左侧, 感到快乐（五二岁）
	BLIP-2	a woman sitting on a bench. A senior is on the left side, feeling happy (52 years old)	একজন মহিলা বেঞ্চে বসে আছেন. একজন প্রবীণ বাম দিকে আছেন, খুশি বোধ করছেন (৫২ বছর বয়সী)	एक महिला एक बेंच पर बैठी. एक वरिष्ठ बाईं ओर है, खुश महसूस कर रहा है (५२ वर्ष)	一个坐在长凳上的女人. 大四学生在左侧, 感到快乐（五二岁）
example_13.jpg	BLIP	a man sitting on a bench playing a guitar. A adult is on the center side, feeling angry (43 years old)	গিটার বাজানো বেঞ্চে বসে থাকা এক ব্যক্তি. একজন প্রাপ্তবয়স্ক কেন্দ্রের পাশে আছেন, রাগ অনুভব করছেন (৪৩ বছর বয়সী)	गिटार बजाने वाली बेंच पर बैठा एक आदमी. एक वयस्क केंद्र की तरफ है, गुस्सा महसूस कर रहा है (४३ वर्ष)	一个坐在长凳上弹吉他的男人. 成年人在中央, 感到生气（四三岁）
	ViT-GPT2	a man sitting on the sidewalk with a stuffed animal. A adult is on the center side, feeling angry (43 years old)	গিটার বাজানো বেঞ্চে বসে থাকা এক ব্যক্তি. একজন প্রাপ্তবয়স্ক কেন্দ্রের পাশে আছেন, রাগ অনুভব করছেন (৪৩ বছর বয়সী)	भरवां जानवर के साथ फुटपाथ पर बैठा एक आदमी. एक वयस्क केंद्र की तरफ है, गुस्सा महसूस कर रहा है (४३ वर्ष)	一个坐在长凳上弹吉他的男人. 成年人在中央, 感到生气（四三岁）
	BLIP-2	a man playing a guitar. A	একজন মানুষ গিটার বাজান. একজন প্রাপ্তবয়স্ক কেন্দ্রের	एक गिटार बजाने वाला एक आदमी.	一个弹吉他的男人. 成年人在中央, 感到生气（四三岁）

		adult is on the center side, feeling angry (43 years old)	পাশে আছেন, রাগ অনুভব করছেন (৪৩ বছর বয়সী)	एक वयस्क केंद्र की तरफ है, गुस्सा महसूस कर रहा है (४३ वर्ष)	
example_21.jpg	BLIP	the band metallic in a black shirt and jeans. A adult is on the left side, feeling angry (40 years old), A adult is on the left side, feeling happy (35 years old), A adult is on the center side, feeling neutral (33 years old), A adult is on the right side, feeling neutral (33 years old)	একটি কালো শার্ট এবং জিন্সে ব্যান্ড ধাতব. একজন প্রাপ্তবয়স্ক বাম দিকে রয়েছেন, রাগ অনুভব করছেন (৪০ বছর বয়সী), একজন প্রাপ্তবয়স্ক বাম দিকে আছেন, খুশি বোধ করছেন (৩৫ বছর বয়সী), একজন প্রাপ্তবয়স্ক কেন্দ্রের পাশে রয়েছেন, নিরপেক্ষ (৩৩ বছর বয়সী) বোধ করছেন, একজন প্রাপ্তবয়স্ক ডানদিকে রয়েছেন, নিরপেক্ষ (৩৩ বছর বয়সী) বোধ করছেন)	एक काली शर्ट और जींस में बैंड मेटालिक. एक वयस्क बाईं ओर है, गुस्सा महसूस कर रहा है (४० वर्ष), एक वयस्क बाईं ओर है, खुश महसूस कर रहा है (३५ वर्ष का), एक वयस्क केंद्र की तरफ है, तटस्थ महसूस कर रहा है (३३ वर्ष का), एक वयस्क दाईं ओर है, तटस्थ महसूस कर रहा है (३३ वर्ष का)	乐队金属穿着黑色衬衫和牛仔裤. 一个成年人在左侧，感到生气（四〇岁），一个成年人在左侧，感到快乐（三五岁），一个成年人处于中央，感到中立（三三岁），成年人在右侧，感到中立（三三岁）
	ViT-GPT2	three men standing next to each other in front of a wall. A adult is on the left side, feeling angry (40 years old), A adult is on the	একটি প্রাচীরের সামনে একে অপরের পাশে দাঁড়িয়ে তিনজন লোক. একজন প্রাপ্তবয়স্ক বাম দিকে রয়েছেন, রাগ অনুভব করছেন (৪০ বছর বয়সী), একজন প্রাপ্তবয়স্ক বাম দিকে আছেন, খুশি বোধ করছেন (৩৫ বছর বয়সী), একজন	एक दीवार के सामने एक दूसरे के बगल में खड़े तीन आदमी. एक वयस्क बाईं ओर है, गुस्सा महसूस कर रहा है (४० वर्ष), एक वयस्क बाईं ओर है, खुश	三个人站在墙前. 一个成年人在左侧，感到生气（四〇岁），一个成年人在左侧，感到快乐（三五岁），一个成年人处于中央，感到中立（三三岁），成年人在右侧，感到中立（三三岁）

		left side, feeling happy (35 years old), A adult is on the center side, feeling neutral (33 years old), A adult is on the right side, feeling neutral (33 years old)	প্রাপ্তবয়স্ক কেন্দ্রের পাশে রয়েছেন, নিরপেক্ষ (৩৩ বছর বয়সী) বোধ করছেন, একজন প্রাপ্তবয়স্ক ডানদিকে রয়েছেন, নিরপেক্ষ (৩৩ বছর বয়সী) বোধ করছেন)	महसूस कर रहा है (३५ वर्ष का), एक वयस्क केंद्र की तरफ है, तटस्थ महसूस कर रहा है (३३ वर्ष का), एक वयस्क दाईं ओर है, तटस्थ महसूस कर रहा है (३३ वर्ष का)	
	BLIP-2	metallica -. A adult is on the left side, feeling angry (40 years old), A adult is on the left side, feeling happy (35 years old), A adult is on the center side, feeling neutral (33 years old), A adult is on the right side, feeling neutral (33 years old)	ধাতবিকা -. একজন প্রাপ্তবয়স্ক বাম দিকে রয়েছেন, রাগ অনুভব করছেন (৪০ বছর বয়সী), একজন প্রাপ্তবয়স্ক বাম দিকে আছেন, খুশি বোধ করছেন (৩৫ বছর বয়সী), একজন প্রাপ্তবয়স্ক কেন্দ্রের পাশে রয়েছেন, নিরপেক্ষ (৩৩ বছর বয়সী) বোধ করছেন, একজন প্রাপ্তবয়স্ক ডানদিকে রয়েছেন, নিরপেক্ষ (৩৩ বছর বয়সী) বোধ করছেন)	मेटालिका -. एक वयस्क बाईं ओर है, गुस्सा महसूस कर रहा है (४० वर्ष), एक वयस्क बाईं ओर है, खुश महसूस कर रहा है (३५ वर्ष का), एक वयस्क केंद्र की तरफ है, तटस्थ महसूस कर रहा है (३३ वर्ष का), एक वयस्क दाईं ओर है, तटस्थ महसूस कर रहा है (३३ वर्ष का)	Metallica-. 一个成年人在左侧，感到生气（四〇岁），一个成年人在左侧，感到快乐（三五岁），一个成年人处于中央，感到中立（三三岁），成年人在右侧，感到中立（三三岁）

8.4.2 Capotions Generated with Captured Image

The test image used in the evaluation for captured image is given below:



captured.png

The table shows the system can generate contextually relevant, emotionally aware, and linguistically diverse descriptions.

Image	Model	Caption				Processing time
Captured.png	BLIP	English	Bengali	Hindi	Chinese	1.57
		a man with glasses and a beard is smiling. A adult is on the center side, feeling happy (34 years old)	চশমা এবং দাড়িওয়ালা একজন লোক হাসছে. একজন প্রাপ্তবয়স্ক কেন্দ্রের পাশে আছেন, খুশি বোধ করছেন (৩৪ বছর বয়সী)	चश्मा और दाढ़ी वाला एक आदमी मुस्कुरा रहा है. एक वयस्क केंद्र की तरफ है, खुश महसूस कर रहा है (३४ साल पुराना)	一个戴着眼镜和胡须的男人在微笑. 成年人在中央, 感到快乐 (三四岁）	

	ViT-GPT2	a man with glasses in a room. A adult is on the center side, feeling happy (34 years old)	একটি ঘরে চশমাযুক্ত এক ব্যক্তি. একজন প্রাপ্তবয়স্ক কেন্দ্রের পাশে আছেন, খুশি বোধ করছেন (৩৪ বছর বয়সী)	एक कमरे में चश्मा वाला एक आदमी. एक वयस्क केंद्र की तरफ है, खुश महसूस कर रहा है (३४ साल पुराना)	一个房间里戴着眼镜的男人. 成年人在中央, 感到快乐（三四岁）	3.96
	BLIP-2	a man wearing glasses and a green shirt. A adult is on the center side, feeling happy (34 years old)	চশমা এবং সবুজ শার্ট পরা একজন লোক. একজন প্রাপ্তবয়স্ক কেন্দ্রের পাশে আছেন, খুশি বোধ করছেন (৩৪ বছর বয়সী)	चश्मा पहने एक आदमी और हरी शर्ट. एक वयस्क केंद्र की तरफ है, खुश महसूस कर रहा है (३४ साल पुराना)	一个戴着眼镜和绿色衬衫的男人. 成年人在中央, 感到快乐(三四岁）	28.68

8.5 Performance Benchmarking

Benchmarks focused on inference time and system responsiveness across modules.

- **Captioning Speed:** Inference time varied by model. ViT-GPT2 took 2-3 seconds, BLIP 5-6 seconds, and BLIP-2 25-30 seconds on CPU.
- **TTS Latency and Quality:** gTTS module took under 3 seconds for typical sentence length. Audio was evaluated for clarity and pronunciation, consistent with neural TTS like Tacotron and WaveNet.
- **Caching Performance:** SHA-256 caching mechanism gave a big boost in response time. Repeated requests for the same image, model and language combination skipped redundant computation and took ~70% less time.

9. Challenges Faced

Several technical and architectural challenges were posed by the multilingual image captioning system with facial analysis and text-to-speech synthesis. Coordinating multiple AI models, managing external APIs, and ensuring performance scalability across a real-time, user-interactive web interface posed the main challenges.

9.1 Integration Complexities Across Diverse AI Models

One of the biggest challenges was integrating multiple state-of-the-art AI models, each with different architectural requirements, input formats and hardware demands. The system has three captioning models—BLIP, ViT-GPT2 and BLIP-2—built on different Transformer-based configurations and pretraining pipelines. Managing their concurrent use required careful memory management, device specific model loading (GPU vs CPU) and asynchronous request handling via FastAPI. Ensuring all models can run in the same runtime environment without resource contention required detailed optimization and load balancing strategies.

9.2 Handling Inconsistent API Responses from Translation and Speech Synthesis

The system relies on external APIs for translation (Google Translate) and speech synthesis (gTTS) which added more complexity. These services sometimes returned incomplete, delayed or locale incompatible responses especially for low resource languages or complex sentence structures. Some languages supported in Google Translate were not available in gTTS, resulting to mismatch between caption translation and speech output. To address this, robust fallback mechanisms were implemented to default to English when necessary to ensure system continuity even during partial service failures.

9.3 Efficient Caching and Performance Optimization

Another big challenge was caching and performance management especially for computationally intensive models like BLIP-2. BLIP-2's integration with large language models increased caption generation latency and memory usage especially on CPU only systems. To mitigate this, the backend implemented a SHA-256 based caching mechanism that stores results per image, per model and per language including facial analysis outputs independently. This reduced redundant computation and improved system responsiveness but required precise handling of cache invalidation and consistency to prevent stale or mismatched results.

In summary, these challenges showed the complexity of harmonizing cutting-edge AI models and third-party services into a real-time multilingual accessible web system. Overcoming them required iterative testing, architectural refactoring and modular design to ensure robustness, flexibility and user-centric performance.

10. Major Achievements and Experiences

The multilingual image captioning system with face analysis and text-to-speech synthesis is the culmination of bringing together multiple AI technologies into one application. Throughout the project we achieved several key milestones that made the system robust.

10.1 Integration of Diverse AI Technologies

One of the biggest achievements was integrating multiple advanced AI models from different domains—vision, language and speech. The system combines image captioning models (BLIP, ViT-GPT2, BLIP-2), facial attribute recognition (DeepFace), neural machine translation (Google Translate), and text-to-speech synthesis (gTTS). Integrating these heterogeneous models—each with different architecture, resource requirements and output format—required a modular and scalable backend architecture. Making these models work together is a testament to the technical depth and interdisciplinary nature of the project.

10.2 Development of a Robust Caching Strategy

To address the latency and performance issues from repeated model inferences and external API calls, the system uses a robust caching strategy using SHA-256 hashing. This allows us to precisely identify and reuse previously processed image-language-model combinations. We also cache face analysis results independently to avoid redundant facial attribute extraction. This caching mechanism made the application more responsive especially during repeated interactions and multi-model comparisons and is a practical example of performance optimization in real-time AI services.

10.3 Enhancement of Multilingual Accessibility

The system provides multilingual accessibility by offering real-time translation and audio narration in English, Bengali, Hindi and Chinese. The unified pipeline ensures that enriched captions with facial context are translated and vocalized across languages. This not only makes the system usable across diverse global audience but also aligns with the broader goal of inclusivity and digital accessibility. Translating visual content into spoken language helps users with visual impairment and non-native speakers to understand the content better.

These achievements show the effectiveness of our design methodologies, collaboration and iterative testing. We also gained practical experience in AI integration, performance tuning and user centric system design which will be useful for future research and application development in accessible multimedia systems.

11. Questions and Answers

During the development of the multilingual image captioning and text-to-speech system, we encountered several technical and design-related questions that needed to be investigated and solved. This section documents some of the most common questions, along with the answers and solutions. It reflects the iterative development process and shows the practical challenges of building an AI-powered multimedia system.

Q1: How to load multiple deep learning models without using too much memory?

Loading multiple Transformer-based models (e.g., BLIP, BLIP-2, ViT-GPT2) at the same time can consume a lot of system memory, especially on machines without GPU. To address this, models were preloaded with explicit device allocation—placing resource-intensive models such as BLIP-2 on the CPU, while assigning smaller models to the GPU if available. Lazy loading was avoided to reduce request latency, and memory reuse was managed using PyTorch's no_grad() and torch.cuda.empty_cache() for efficient inference.

Q2: What strategy was used to prevent redundant analysis and improve performance?

To eliminate redundant computation, a SHA-256 hashing mechanism was implemented to generate a unique identifier for each image. Results (captions, face analysis, translations, and audio) were cached based on this hash along with the selected model and language. For example, if an image was already processed with BLIP in English, selecting the same image again would bypass caption generation and face analysis, fetching results directly from cache. This dramatically reduced processing time and improved responsiveness.

Q3: Why were some languages not supported for TTS, and how was this handled?

Google's Text-to-Speech (gTTS) API does not support all languages supported by Google Translate. For example, while translation into Bengali or Hindi may succeed, gTTS may return errors for unsupported TTS locales. To handle this, a fallback mechanism was implemented that defaulted to English for audio generation if TTS synthesis failed for the target language, ensuring the system still produced usable output without breaking the user experience.

Q4: How were model-specific caption inconsistencies managed?

Different models generate captions with varying sentence structures and detail levels. To support consistent downstream processing (e.g., translation and TTS), the system preserved each caption separately in the cache, indexed by the image hash, selected model, and language. The frontend interface allowed users to view and compare outputs by model, with clear labels to indicate the source of each result.

Q5: How was the position of faces (left, center, right) determined in the image?

Facial positions were calculated by sorting detected faces based on the x-coordinate of their bounding box centers. The face with the smallest x-coordinate was considered "left," the one in the middle was "center," and the largest was “right.” This enabled more context-aware caption enrichment, such as: “An adult on the left appears surprised (age 34).”

Q6: How was language consistency ensured in both caption and face summary?

Initially, the base caption and face summary were translated separately, leading to language mismatches in TTS output. This was resolved by concatenating the enriched caption with the face summary into a single string and then translating the entire combined text. This ensured both content and structure were preserved uniformly across all supported languages during translation and audio synthesis.

Q7: What caused face analysis to repeat unnecessarily for the same image?

Face analysis was initially triggered on every new caption generation, even for cached images. This was fixed by decoupling the face analysis logic from the captioning and translation routines. Now, once facial attributes are computed and cached per image, they are reused across all model and language selections without reprocessing.

These questions and their respective resolutions played a vital role in refining the system, improving performance, and ensuring stability across user scenarios. By documenting them, we support reproducibility and provide valuable reference points for future development efforts in similar multimodal AI systems.

12. Future Enhancements

Future plans include making the system more robust, accessible and global. One of the big ones is to add offline translation and speech synthesis so we don’t have to rely on external APIs and can work in low connectivity or secure environments. The system will also support more languages so we can include more native languages and be more inclusive for international users. To make the output more descriptive we will add advanced scene classification models so the system can tell the difference between indoor, outdoor, natural or urban and provide more context relevant captions. Accessibility features will be improved including better support for screen readers and assistive technologies so we can be more usable for visually impaired users. We will also add voice command so users can control the interface with voice commands for a more intuitive and hands free experience.

13. Conclusion

This shows how multiple advanced AI components can be combined into one unified, accessible and multilingual image captioning system. By combining image captioning models, facial attribute analysis, neural machine translation and text to speech synthesis the system generates enriched visual descriptions that are language adaptable and audio accessible. Using state of the art models like BLIP, ViT-GPT2 and BLIP-2 the generated captions are contextually accurate and semantically meaningful and the DeepFace framework adds an extra layer of human centric detail by identifying age, gender and emotion.

Multilingual translation and speech synthesis allows the system to reach a global audience, addressing the accessibility challenges faced by visually impaired and language barrier users. Also the use of caching mechanism and asynchronous processing makes the system responsive and scalable. Overall, the potential of AI driven multimodal systems to enhance digital content accessibility is shown by this project, and the foundation for future research and development in inclusive human computer interaction is laid by this project.

14. References

1. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment
2. Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P. N., Rashtchian, C., Hockenmaier, J., & Forsyth, D. A. (2010). Every picture tells a story: Generating sentences from images. In *European conference on computer vision* (pp. 15–29). Springer.

3. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156–3164).

4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

5. Li, J., Selvaraju, R. R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 12888–12900).

6. Li, J., Gotmare, A., Selvaraju, R. R., Joty, S., Xiong, C., & Hoi, S. C. (2023). BLIP-2: Bootstrapped language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

7. Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7008–7024).

8. Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6077–6086).

9. Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In *Proceedings of Interspeech 2017* (pp. 4006–4010).

10. Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4779–4783).

11. Yamamoto, R., Song, E., & Kim, J. M. (2020). Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6199–6203).

12. Google AI. (2020). Cloud Text-to-Speech.

13. Amazon Web Services. (2021). Amazon Polly. https://aws.amazon.com/polly/

14. Microsoft Azure. (2022). Azure Cognitive Services – Text to Speech. https://azure.microsoft.com/en-us/products/cognitive-services/text-to-speech/

15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, 30.

16. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

17. Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics, 5*, 339–351.

18. Park, K., Hong, J., Kim, H., & Kim, J. (2019). Multimodal spoken content description for the visually impaired. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–12).

19. Qiao, J., Wang, Y., Sun, J., Li, Y., & Wang, Y. (2020). Assistive image captioning with neural TTS for visually impaired users. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 3034–3042).

20. Chuang, W. Y., Hung, Y. T., & Lin, Y. T. (2021). Accessible image captioning enhanced by face recognition and multilingual speech synthesis. *IEEE Access, 9*, 89564–89575.
21. Eastlake 3rd, D., & Hansen, T. (2011). *US secure hash algorithms (SHA and SHA-based HMAC and HKDF)* (No. rfc6234).
22. https://ahmed-sabir.medium.com/paper-summary-blip-bootstrapping-language-image-pre-training-for-unified-vision-language-c1df6f6c9166
23. https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transformers/
24. Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.
25. Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1701-1708).