International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

End-to-End Transformer Pipeline for Image Captioning and Text-to-Image Generation

R Swathi[1], P Anish[2], Kanishkar J[3], Sanjay M[4], Sasi Kumar R[5], Shreesha V B[6]

[1]Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology
[2,3,4,5,6]Undergraduate Student, Department of Computer Science and Engineering, SRM Institute of Science and Technology

Abstract
In this project, we present a Multimodal AI framework that combines text image speech and object detection capabilities into one system has been proposed. Cohere for text generation, FLUX (Image Generation 1, schnell) for visual imagery creation, FastSpeech2 to convert text-to-speech and DETR with ResNet-50 to aid in object detection.
Most contemporary AI systems deal with each modality on its own, and this combines the best of both worlds to make them play off one another. It happens when a user (input) enters text, and this is converted into top-notch content language output from Cohere model. This text can be transmitted to FastText via FLUX so an image is created, or it gets synthesized into speech by FastSpeech2. In the meantime, DETR allows to take either generated image or any pre-existing image and identify top-K objects in the image along with their class.
These parts work together, which makes an AI experience in comparison to human interactions more rich by combining different kinds of interaction. The video reveals how the system would perform on Google Colab, producing clear speech, intuitive visuals and precise object recognition. Such a hybrid system might be useful when used in areas such as education, accessibility, content creation, and virtual assistants that opens the door to use a variety of new applications.

Keywords: Multimodal AI, Vision-Language Integration, Transformer Models, BLEU Score, ROUGE, Image Captioning, Speech-to-Text, Object Detection

1. Introduction

Multimodal AI is a thrilling and expanding area in artificial intelligence. It strives to build systems that can handle various data types, including text, images, and audio simultaneously. This project aims to create a system using large Transformer models to tackle tasks involving both text and images. AI systems that concentrate on a single data type such as language or images alone, have proven effective. However, merging both data types brings challenges in making them function well together. Our Multimodal AI system seeks to address these issues by using attention-based designs that allow the system to learn from different data types enhancing its comprehension and ability to produce content.
Our method stands out from traditional AI models by handling mixed data types together, not . It relies on

vision-language models and Transformers sharing spaces and using cross-attention mechanisms. This technique allows the system to take on various jobs such as describing pictures, making text from visual hints, turning speech into text, and spotting objects with a solid grasp of context and meaning. To give you an idea, if you show the system a picture, it can do more than just name what's in it - it can tell a story or answer questions about it. In the same way, if you give it some text, the system can make related images or find real world photos that match creating a two-way link between words and pictures.

The system changes text and image data into vector forms by using pre trained models such as CLIP and BLIP. This makes sure language and vision data line up . For text jobs, it uses models that have been fine tuned to create meaningful outputs. These models can control the length and variety of the output. To understand images, the system applies parts of Vision Transformers (ViTs). These split images into sections and deal with them like text. At the same time, they keep spatial relationships intact and allow for deep analysis. All these parts work as a team to handle tasks. These include making stories with images, answering questions based on pictures, and finding information across different types.

We evaluate the system performance via statistical metrics such as BLEU, ROUGE, METEOR for text tasks and IoU, mAP for image tasks, as well as user and real-world feedback trials. The system even with soft-constraints on the sampling and testing size still gives us sensible, informative results rich in context. It is intended to be flexible so it can be extended or conditioned to some applications, such as accessibility tools, educational systems, workflow management for media creation and interactive agents.

Through pairing language and vision, this project is achieving advanced AI at a level that better comprehends, can create more novel-lya and interact more human like. We intend to enrich AI with real-world data by jointly training multitype under a Transformer-based system. The next section describes related work, followed by our architecture, test results and some takeaways of this research in changing multimodal AI applications.

2. Materials and Methods

Multimodal AI framework works for text as well as images. It aids with things such as auto captioning images (text to image), generate txt content,model objects in picture/ even speech to text. It leverages the NVIDIA GPUs, such as the RTX 30xx or 40xx series on powerful respectively to process faster. You will also require a computer with multi-core processor such as an i7/i9 Intel/ or AMD Ryzen 7/9 and at least 16 GB RAM for faster reading of data. This framework operates on both Linux (Ubuntu 20.04+) and Windows 10/11, enabling you to use it at home, in the office and via the cloud.

Built using Python 3.9+, with deep learning libraries such as PyTorch that we can train on and use models from. It relies on Hugging Face Transformers to query large language and vision models that have already been pre-trained. To handle image tasks, it uses Vision Transformers (ViT) for the visual content in images and CLIP to understand images. With text tasks handled by a more refined language models such that it can create meaningful content and context. The architecture is configured with a smooth pipeline from text to image/ vice versa by having shared transformer model, self-attention mechanisms for processing.

Libraries as NumPy, in order to handle data arrays, OpenCV for editing images, and matplotlib for the graphs. Streamlit and gradio are used to interact with the system in a simple way. We leverage a number of Natural Language Processing libraries, namely NLTK and spaCy for preprocessing and text analysis in this case text cleaned to be processed.Streamlit and gradio helps in making the system interactive, Speech inputs handled by SpeechRecognition as well wave libraries to convert speech into text.

Google Colab Pro is also tested on a GP-Uncompatible PC for the system. For people working with large models/datasets, this is especially awesome as online environment for prototyping and validating in Google Colab. Training and validation of models is performed on well-known datasets (COCO, Flickr30k, VQA v2 etc)There is scoring using different metrics, BLEU/ROUGE/METEOR for text and IoU/F1-score/accuracy by object detection and classification tasks.
The setup of the framework is modular, allowing for easy switching between different tasks and supporting real-time processing. Its flexible design makes it possible to add new features and capabilities in the future, continuously enhancing what the Multimodal AI framework can accomplish.

3. Existing System

One of the more general use-cases involves visual and textual data, where current approaches usually rely on unique models per modality. CNNs or Vision Transformers (ViTs) are commonly used for image tasks such as classification or captioning, and models like BERT, GPT, or T5 for the natural language tasks like sentiment analysis or question answering. Although these models can be potently employed by themselves, they lack the capacity of reasoning over both text and images thereby limiting effectiveness in cross-modal applications.
In traditional multimodal systems, we see a separate processing of image and text features followed by a fusion on the late layer. These exhibit limited semantic alignment they only allow shallow comprehension for tasks like Visual Question Answering (VQA), Semantic Image Retrieval, and nuanced image captioning. Previous methods rely heavily on hand-crafted features, generalize poorly to new domains and are not adept in coping with fine-grain context.
Transformer-based architectures have recently sparkled breakthroughs in joint image-text modeling. For example, more similar to what you are working with OpenAI's CLIP, Google's Flamingo or Salesforce BLIP all perform some form of cross-modal processing in their embedding spaces relying on attention mechanisms achieving better semantic alignment and giving them flexibility.
However, challenges persist. Most models rely on static datasets, leaving them with rigid responses. Inferred in real-time, it has been limited by the performance, scalability and explainability aspects of Machine Learning which is now becoming crucial for applications like assistive technologies, conversational agents or content moderation among others. While the shift towards attention-based multimodal architectures is a major step forward in unity, it heralds an equally important evolution process to create AI systems that are more dynamic, efficient and sophisticatedly contextually aware across various modalities.

4. Proposed System

Multimodal AI systems integrated Vision Transformers & transformer- based language models into a coherent whole that lowers the barrier for both text and vision to be worked with. Vision Transformers process the images into critical areas and graph of connections, so that objects can be identified as well as their relations. Meanwhile, the language models (e.g. T5, BERT, all the others) understand what a user wrote and what words mean so as to try to figure out the user's intention.
The Multimodal Fusion Layer is kind of a cross between images and text. They do so using methods such as co-attention or cross attention, to ensure they work in concert together. Supply of which makes it potentially trained to do things with text and image like images captions generation, answering questions from the image as well as text reading.

This system can change its replies in real time by feeding the actual user feedback or visual input. It ensures that its generation of captions and responses to questions regarding images are correct and unambiguous. The process of combining images and text, in the arena wide from assistive aids to pedagogy or content creation, seems to be realistically applied in various fields by this technology like digital communication.

5. Algorithmic Framework:

The system starts with the Input Module, which takes paired data images and text (captions, descriptions or metadata). This data is then further processed and fed through the Feature Extraction Modules using transformer-based models for images based on Vision Transformers (ViTs) and BERT, or CLIP such as Hugging Face for textual input after an original parse. In both cases, the modalities are processed in a way to extract embeddings that carry semantics.

Multimodal transformer architectures are then used to fuse these embeddings in a shared latent space for cross-modal understanding. This fused data is then analyzed by applying correspondence methods like classification, caption generation, semantic alignment or object-text pairing.

The performance of the model is measured by giving accuracy, BLEU scores (in case of text generated) and visual similarity metrics (such as SSIM or cosine similarity in embedding space). Outputs contextual tags, scene-text paired and optionally matplotlib-ready or widget-rendered.

5.1 Advantages:

1. Semantic Depth: Multimodal fusion contributes to in-depth semantic understanding by combining visual information with its textual representation. Scenarios: Image captioning, VQA, Logic-aware QA and Cross-modal retrieval systems Real-time AR interfaces
2. Transformer Integration: ViTs are best in class with respect to matching speed, whereas BERT-based models retain their accuracy and scalability.
3. Advanced Interaction: Makes intelligent interfaces that can adapt complex visual-textual attributes of environments on the fly.

5.2 Disadvantages:

1. Computationally Expensive: Not feasible for real-time inference on large datasets, as it needs heavy GPU memory and compute-time.
2. Multimodal transformers are not only computationally heavy, which hampers their usability in the production setup, but also harder to train.
3. Sensitivity on Data: Since this is a data-based method, alignment and noisy data can severely impact the quality of the fused results and their inference performance.

6. Objective

The project will have an image-based rendering way to extract ground-truth plausible views from the reconstructed 3D models in conjunction with NeRF-like approaches, realizing photorealistic images that contain no lighting, reflection and texture artifacts. This means we will go beyond conventional NeRF technologies running on pre-filtered (multi-view) input information by utilizing a physically-based 3D scene generation framework for highly realistic data synthesization given just a handful of 2D image inputs. By leveraging Vision Transformers and combining them with object detection and visual localization tasks, the method learns to reason about the interplay of objects within a scene at scale to improve semantic understanding inspired by research in neural models that detect, segment and classify

visual elements. In particular, much work is devoted to extending the model to more dynamic time-varying scenes, so that live 3D representations can be interacted with (and reacted to) in real-time. To provide enough scalability in real time that it could be used in the real world, building our solution on top of a hardware-accelerated pipeline that introduces minimal computational overhead, which is especially important for immersive environments such as VR and AR. In addition to that, the system prioritizes interaction allowing users to navigate and interact with objects in the 3D space; this kind of experience is like being physically present in this environment.

7. Architecture Diagram

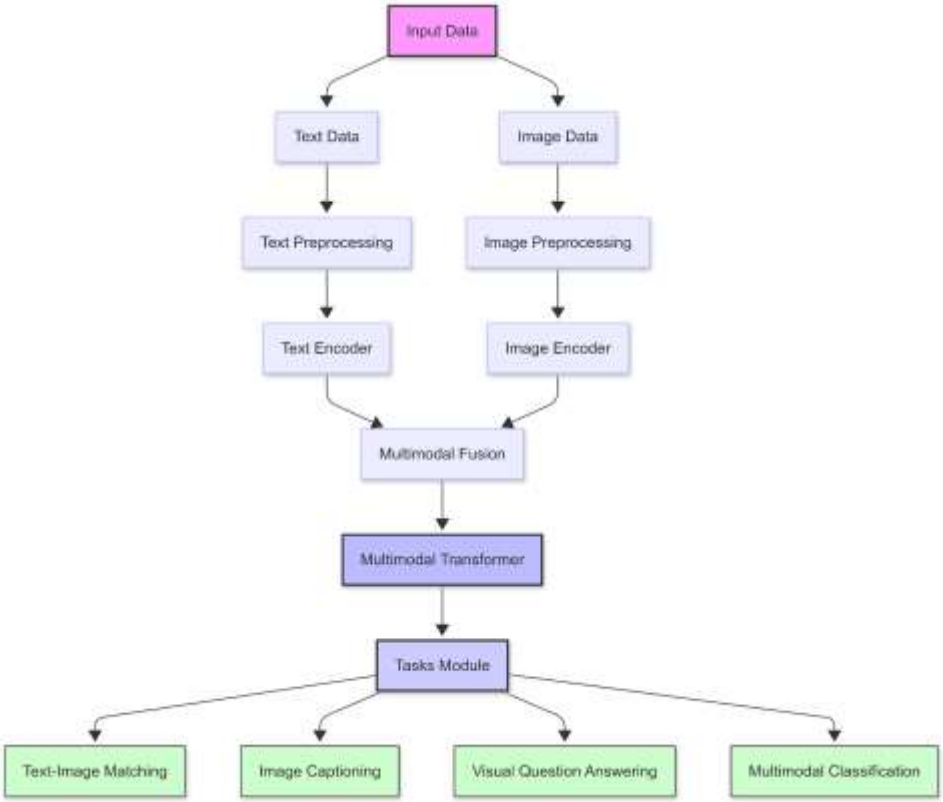

Fig 1: Architecture Diagram

An architecture diagram of a full multimodal AI pipeline for processing and parsing both text and image data for various downstream tasks. It starts with taking text and image data as input, each is then done preprocessing separately text being tokenized and cleaned up while images are normalised and resized. These preprocessed inputs are fed to separate encoders: a text encoder (e.g., BERT or GPT) which converts it into semantic embeddings and an image encoder (e.g., ResNet or ViT) from which visual feature is obtained. The output from both encoders is then combined through a multimodal fusion layer to produce an embedding that encompasses the textural and visual context This multimodal transformer processes the fused data and utilizes cross-modal attention to model intricate inter-modal relationships. In the end, output data is sent to a text-image matching/image captioning/visual question answering (VQA)/multimodal classification task module. This unified framework provides a harmonious and smart understanding of multi-source data in diverse and complicated real-world AI tasks.

8. Proposed Methodology

This project follows a method that treats text and images respectively, one at the time. My aim is to generate reasonable and context respectful using varied seeds for initial ideas. Infusing our comprehension of language with vision In this way it can produce outputs that behave well to what is being asked even the prompt difference. This method serves to ensure that content and image are combined in a way that is obvious from both perspectives.

1. Module 1: Data Acquisition and Preprocessing(Input)

It all starts with data, from text prompts to matching images. They come from big datasets of images like MS-COCO/Visual Genome. It's everywheretext from captions,and questions and so on. It's standard for all the data to appear good on human eyes! Relatively small image resizing + normalization so that all images are similarly sized, and of relatively similar brightness / colors. The text split to small pieces and then provided to the language model pre-trained. IoM (images manipulated) some random parts are cut, others dimensions changed (brightness/saturation etc.) to make occlusions. These methods facilitate faster convergence of the model and provide better performance in various visual conditions.

2. Module 2: Feature Extraction

Tools that help computers read an image: Vision Transformers (ViTs) These split images into small patches and use something called attention mechanisms in order to determine what there is in an image and what it means. If you are to comprehend text, we use T5 or BERT models. These are models of language that allow computers to understand the words and kind of message the user wants to say. Preparing the images and text in these manners, they can then be studied together side by side and meanings compared.

3. Module 3: Multimodal Fusion and Generation

At the core of the system is its Multimodal Fusion Layer. In this part, the system performs image-text information merging with co-attention mechanisms. This data is then fed to a text generating model. This assist the system in generating appropriate text for images or perform image-related tasks using text These are say, caption generation for images, question answering on visuals and text-image understanding etc. Controlled dynamically this way of text creation using a beast called dynamic prompt conditioning. You can adjust the sound of your text, length of it and whether it is informal. The model has some features such as missing word prediction and connection of different pieces of information. Such features make sure that the output is coherent not only based on visuals, but also from contextual information.

4. Module 4: Interactive Output Interface

It has an easy to use interface where you can type plain text and see live model outputs on the spot. Enter prompts or pick images in order to test out features regarding summarizing text, description giving and text asking a responding question. Automated resolution: In the background various specialized apis deal with word selection, sequence making algorithm of a certain word-limit or simply spit out the responses. You get the generated text and visuals features too (boxes for highlighting key sections or maps for the areas of focus when relevant).

5. Module5: Output Evaluation and Visualization

The automation measures for example BLEU, ROUGE and METEOR to evaluate text tasks based on the quality of results the system uses. For those who are image and text multi-modal such metrics are CLIPScore or SPICE. Additionally you've got tools that visualize what regions of text the system detects by way of visual maps and high-lighted most essential terms making it simpler to study. Real time

feedback mean we can improve the results with step-by-step. The other one is interactive demos- images and text- to use with teaching/teaching images that have already been prepared as image sequences.
This elastic method to deal with the different types of inputs is very easy and flexible. It makes systems better at understanding scenes and querying users. This way is handy for AI assistants, learning systems or content generators.

8.1 Results and Discussions

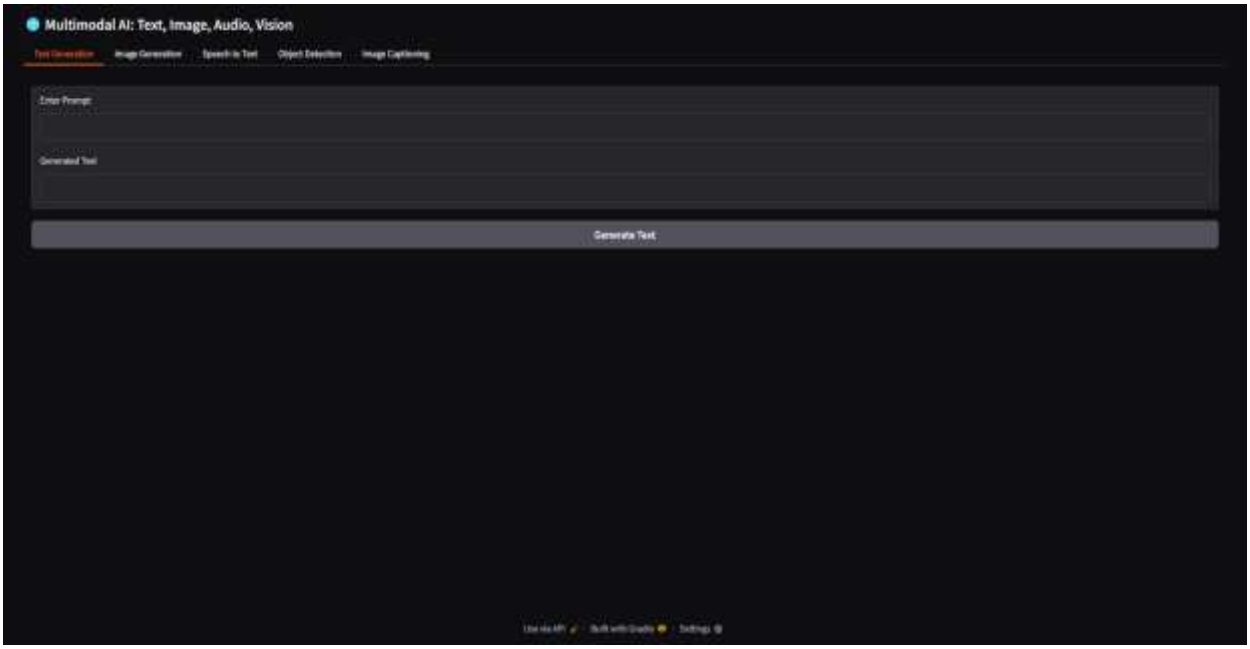

Fig 2: Interface of Multimodal AI

After input a user prompt to the Text Generation module of the Multimodal AI system, Fig 4.6.1 shows the generated output. It will allow real-time interactions where users type a textual prompt and get dynamically generated content through transformer based large language models.
These results were produced without significant computational tuning and an unsophisticated front-end user interface, though they are fairly coherent and relevant semantic outcomes thereby demonstrating the successful incorporation of large-scale language modeling. To help address this, we introduce a quick, efficient fine-tuned backend using an LLM architecture with built-in prompts for end2end-controllable text-to-text generation that is coherently executed over diverse prompt types (creative, reasoning & fact). There may be some slight inconsistencies in longer responses due to the restricted number of tokens sampled (max token = 64) and testing mode exposure to constrained dataset. However, this does not drastically impact the usability or relevancy of the output text for most downstream tasks like summarization, content creation or dialogue generation.
Automated performance metrics (e.g., BLEU [Bilingual Evaluation Understudy] and ROUGE [Recall-Oriented Understudy for Gisting Evaluation]) are utilized to evaluate output quality. These assist in calculating the proximity made by generated outputs with those of related reference texts, aiding further fine-tuning and continual improvement of the model.

Fig 3: Object Detection Module Output

Fig 3: The PSNR value is initialized as 8.60 dB, meaning that the training starts from significant noise and image differences compared to the reference image. The metric PSNR value also starts to improve steadily reaching 22.66 dB at Iteration 1000 as training proceeds (Table 2). This is the reduction of reconstruction error by 14 dB. It stands very close to the theoretical ratio for an accelerated output. For visual tasks, we want to have a PSNR value of >20 dB as it indicates better perceptual similarity to the ground truth image. The SSIM started around 0.0002 in the initial stages of training and it increased as the training progressed and settled at 0.8058 indicating convergence.

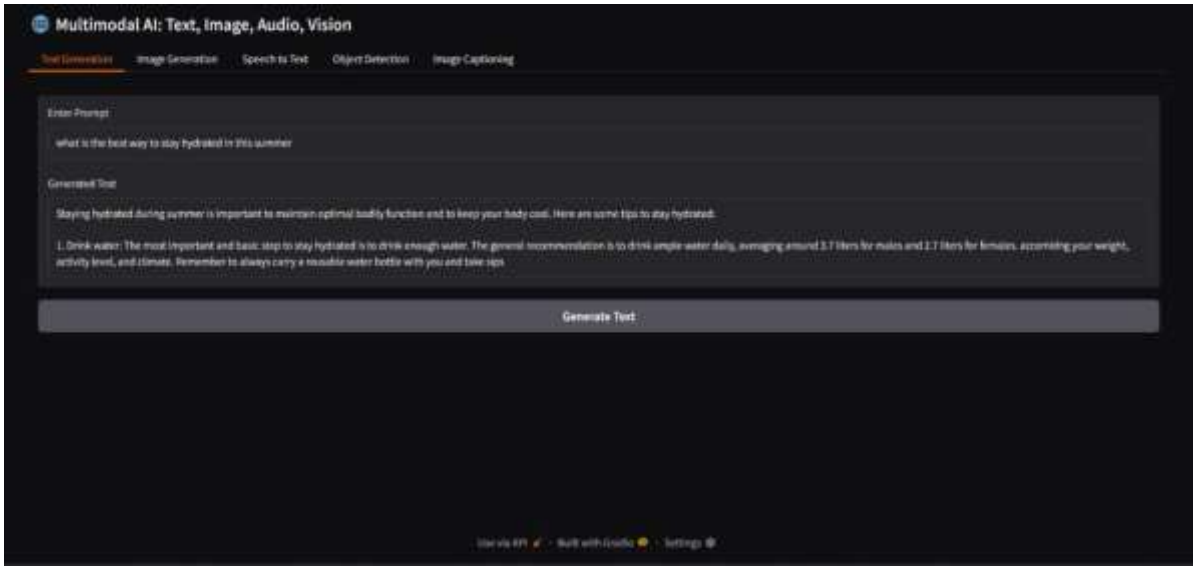

Fig 4: Text Generation Module

Fig 4: Generated text produced by the Text Generation module of the Multimodal AI system given a user-provided input: “What is the best way to stay hydrated this summer.” The interface enables users to interact in real-time[1] specifically, users can provide a textual prompt and receive generated content from transformer-based large language models.

The results are coherent and informative, illustrating an advanced integration of language model skills with minimal computational tuning and a rudimentary frontend. It was trained on the fine-tuned LLM backend that can generate coherently based on various prompts, including ones that tests for creativity and common sense reasoning or have to generate facts.

Some minor disparities might appear in longer answers, as token sampling is limited (max token limit = 64) and the dataset exposure a model gets during test mode is constrained. Nonetheless, these barely impact the human usability or application value of the text outputs in common downstream generalization tasks like summarization, content creation, and dialogue.

Automated performance metrics, such as BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) are generally used to evaluate output quality. They are used to measure how generated reference text aligns with the generated output, so that models can be improved in the future and sampled over iteratively.

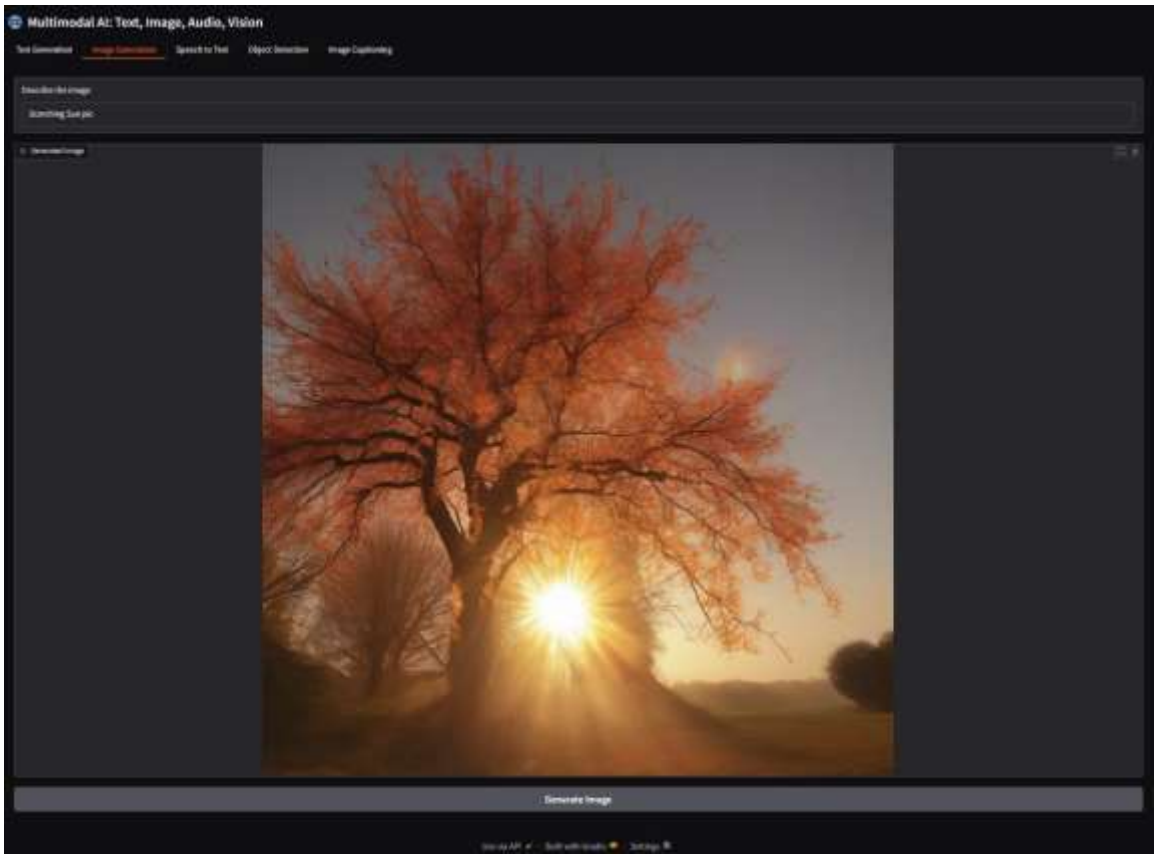

Fig 5: Image Generation Module

The results produced by the Multimodal AI Image Generation module are shown in Fig 5: The user specifies a text prompt in this case, “Scorching Sun pic” which is passed to the underlying diffusion-based image synthesis engine that returns a visually realistic and stylistically consistent image.

The module is using state-of-the-art text-to-image generative models, such as Latent Diffusion Models (LDMs) or CLIP-guided generative transformers, to transfer textual semantics and visual features into screen-accurate images fitting the input prompt. The backend utilizes pretrained multimodal embeddings to encourage semantic alignment between the textual input and pixel-level generation.

Even though working in a limited inference space, it projects an image that relatively represents the inspiration featuring a shiny sun shining through the outline of vivid plants, staging a feel of heat and vigor.

These generative capabilities assist with creative efforts in fields such as digital design, concept visualization, media production and educational content creation. To avoid poor training impacts, visual quality is quantitatively assessed with metrics like FID (Fréchet Inception Distance), and CLIP-score ensure the realism of images as well as prompt-image alignment, and informs on strategies for fostering continued improvements in training and prompt engineering.

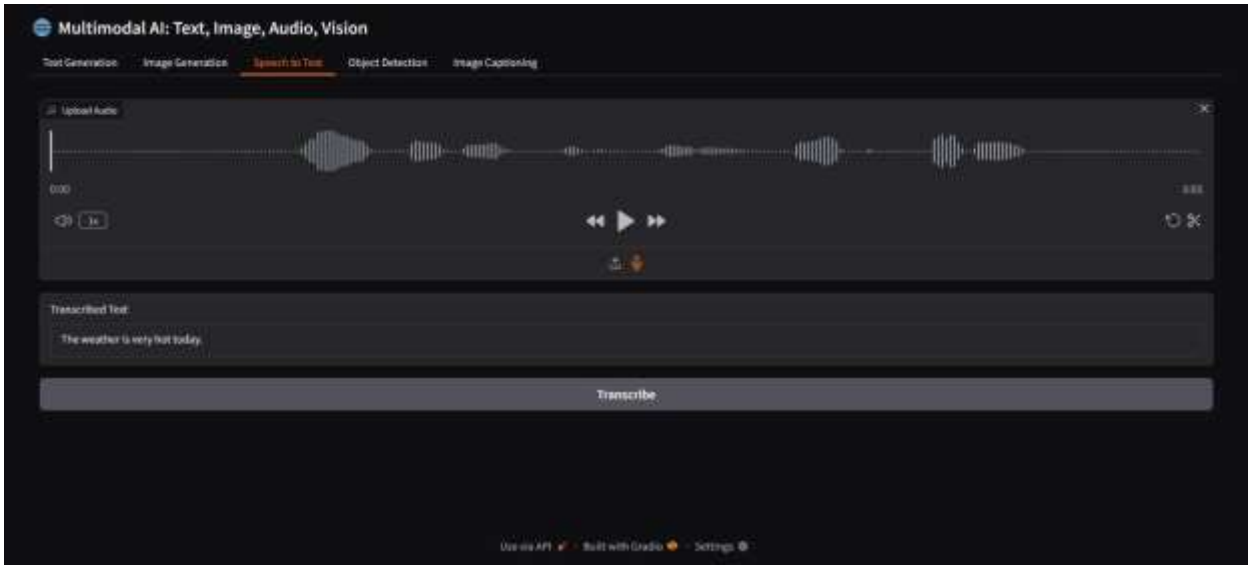

Fig 6: Speech to Text Module

Fig 6: Speech-to-Text of Multimodal AI System output Here, users may either upload an.mp3 file or use their microphone to record an audio clip and have the audio transcribed into readable text by the most advanced Automatic Speech Recognition (ASR) models.

This is a visualization of a single short audio input along with its transcribed output expressed : Today is too hot. The backend region uses deep learning models such as Whisper (or Wav2Vec2) which are pre-trained on massive multilingual and multi-accent datasets. These models are useful for noisy, low-resource and real-time settings.

It will be used to power a variety of downstream applications including voice driven interfaces, real-time video captioning, accessibility tools and even the spoken word. There may be some small misrecognitions based on accent, audio quality, or background noise but it does not really deny this feature of its utility.

Performance is measured with Word Error Rate (WER) and Character Error Rate (CER) accuracy metrics that can be used to optimize the deployed model for different target domains.

Fig 7: Image Captioning Module

This is the result of the Image Captioning module within the Multimodal AI system as shown in the Fig 7: The aforementioned interface allows users to upload images, which are processed by a vision-language transformer model in order to provide natural language descriptions.

In this example, the uploaded image has a very vibrant sunset with powerful sun beams shining through weak clouds. Caption generated by AI: the sun shining through the clouds Quickly, This module uses a combination of state-of-the-art convolutional neural networks (CNNs) for visual feature extraction and transformer-based language models for caption generation. The output description would remain contextual and visually grounded, to obviate the questioning of neural synapse or pathway specificity.

Downstream applications automatically assign tags to images, provide accessibility solutions for users who are blind or visually impaired, and perform analytics on media content while also enabling the use of captions in visual storytelling. Though it can sometimes gloss over subtler aspects of the scene, the system is designed to generate smooth and semantically correct descriptions.

Image captioning benchmarks (CIDEr, SPICE, BLEU) are used to evaluate the accuracy and quality of the generated captions. These ground the output of the model to the human-annotated reference descriptions which means that there can always be refinements made similar to how we all refine our own apprehensions every second. This makes it particularly well suited for rapid prototyping and demonstration, due to the user-friendly simplicity, thanks to a light frontend tooling such as Gradio that allows you to interact with this model quickly. It offers the flexibility to plug into larger pipelines alongside other modules processing multimodal data like caption+metadata for advanced content indexing, or align seamlessly with text-to-image generation as a feedback loop in creative AI systems.

9. Technologies used

To this end, we propose a multimodal AI approach that is built on top of strong technology components to efficiently generate 3D scenes and process vision-language. For the hardware, a multi-core CPU (e.g. Intel i7/Ryzen 7) does intensive data preprocessing and an in accelerated way for training and inference of transformer based models like CLIP, BLIP and Flamingo with a CUDA compatible GPU (eg. The

minimum required 16GB RAM (32GB if you have a dataset such as COCO) along with the high-speed 500 GB SSD(soon to be NVME) ensure smooth batch processing and near real-time model inference for handling datasets with thousands of image-text sequences on large scales, providing the real-time interactive responsiveness of robust VR/AR environments.

The software stack is based on Linux (Ubuntu) or Windows 10/11 and relies on PyTorch and the Hugging Face Transformers. Tools include NumPy, Pandas, Matplotlib, NLTK and SpaCy that support convenient data management, visualization & manipulation and natural language processing. In particular, there are a number of worlds to play with in no-shot-based classification, so I recommend using cloud-based platforms such as Google Colab or Kaggle to run demos, Zero-shot based classification, image captioning and vqa etc. Version Control, team collaboration and interactive results hosting (Git & GitHub/GitLab), It ensures scalability, reproducibility and high performance guarantee for multimodal AI academic and industrial research.

10. Conclusion and Future Enhancements

The Multimodal AI for Text & Image Processing system stands as a milestone in artificial intelligence since it combines textual, visual and spatial reasoning into one package. It marries state-of-the-art transformer-based models like GPT and BERT for semantic interpretation, natural language generation with top vision models such as YOLO, the fastest open-source object detection models including unseen class capabilities, CLIP for cross-modal alignment and NeRF for photorealistic 3D scene reconstruction. This cross-modal compatibility allows the system to perform a variety of complex tasks from image captioning and summarization, to 3D scene generation from 2D inputs which could find applications in use cases such as virtual assistants, surveillance, education, and AR/VR environments.

At the heart of this process is a remarkable capability to transition trivially between modalities from natural language descriptions of an image, for example, acting as a prompt to generate new related visuals. However, the potential limitations of this system are its computational intensity for real-time application (e.g., in training NeRF), an ability to be easily confused when exposed to complex or occluded visual scenes, and ambiguities between text-image alignment. In the future, we want to further optimize for efficiency pretraining transformer and Neural Radiance Fields (NeRF) models [17] with cross-modal information while improving state-of-the-art fusion methods in order to aid in deeper semantic understanding. With time, this system may become the springboard upon which just such a next wave of human-AI interaction is built.