ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 11, Issue 3, March 2026

Vision-Language Integration: Transformer-Based Generative Ai For Image Captioning To Improve Accessibility And Visual Content Understanding

[1] Bhuvaneswari S, [2] Kavya S, [3] Chethana B, [4] Akshitha M, [5] Egneswari K
[5] Assistant Professor, Department of CSE, Er. Perumal Manimekalai College of Engineering, Hosur-635117, Anna University.
[1] [2] [3] [4] Student, Department of CSE, Er. Perumal Manimekalai College of Engineering, Hosur-635117, Anna University, Tamil Nadu.

Abstract: ***Image captioning has become a problematic activity in computer vision and natural language processing, filling in the gap between visual cognition and language description. The current captioning systems, however, amusingly do not have the contextual flexibility, integration of accessibility, and multi-modality output required in the real world. The presented work suggests a generative framework based on transformers that combines vision language modeling and accessibility-driven improvements. Using the model of BLIP (Bootstrapping Language Image Pretraining), we apply both conditional and unconditional caption generation to various situations of images. The system supports the use of Google Text-to-Speech to provide multi-language audio output, adjustable voice settings, and capture history, to enhance accessibility, most especially to the visually impaired users. It is built with PyTorch and deployed with Streamlit, and hence offers an interactive and user-friendly interface. The system is shown to be able to provide correct and contextually detailed captions on heterogeneous sets of images through experimental assessment, and to provide features of accessibility that allow inclusivity in promoting user engagement. The study will help to promote the development of multimodal AI systems because it will allow both content creators, accessibility advocates, and ordinary users in comprehending images accurately, easily, and individually.***

Keywords: Image captioning, Image-text model, Vision Language model, Transformers, text-to-speech (TTS), accessibility technology, PyTorch, streamlit, generative AI, multi-modal system.

I. INTRODUCTION

It has transformed the integration of vision and language into one of the most influential developments in artificial intelligence and is what allows machines to comprehend and describe what they can see in natural language. Image captioning, which is one of the important fields in this integration, is critical in accessibility, automated content generation, and human-computer interaction. Traditional image captioning techniques frequently used the handcrafted features or shallow learning techniques which constrained these techniques to capture the context and semantic relationship among objects in an image. With the introduction of deep learning, especially the transformer-based architectures, caption quality has tremendously increased with the use of large-scale pretraining and attention mechanisms to provide better vision language alignment.

Even with all these developments, there are still a number of challenges in real world implementations of existing forms of captioning systems. The majority of frameworks produce captions that do not have much contextual flexibility, are not multi-lingual, and do not offer easily accessible modalities, such as speech outputs. The lack of accurate and contextual audio descriptions limits the interaction of the visually impaired user with the digital visual contents. Equally important is the fact that content creators and accessibility advocates need more sophisticated instruments, including those that can generate captions of high quality and allow customization and inclusivity options. It is important to fill these gaps in order to make image captioning systems useful to more than the academic standards.

In order to eliminate such shortcomings, this study suggests a generative image captioning model based on a transformer which incorporates accessibility-based improvements. The system uses the BLIP (Bootstrapping Language Image Pretraining) model to generate captions that provide conditional and unconditional captions. In conditional captioning, users can control the captioning process by using custom text prompts whereas in unconditional captioning, the captioning engine generates automatic descriptions without user intervention. Moreover, it features the multi-language audio synthesis by means of Google Text-to-Speech (TTS), which allows the auditory captions and enhances the accessibility of the device by a wide range of user groups around the globe.

It is done with PyTorch to develop the model and Streamlit to deploy it so that it can be a user-friendly interactive platform. Functionalities to control the caption history, customizable themes, multi-language and voice speed are offered as

accessibility features to make personalization. All these components combine to create a complete multimodal system that enhances inclusiveness, usability and involvement of image captioning technologies.
On the whole, it is expected that this project will help to further the development of multimodal AI systems and improve the gap between vision-language integration and accessibility-focused design. The proposed system fills the most urgent gaps in the current frameworks, proving its capacity to support inclusive technologies, improve the user experiences, and offer scalable applications to digital access, content creation, and other purposes.

II. RELATED WORK

2.1 Transformer based Image Captioning

Chen, 2021 More recent developments in transformer architectures accelerated the development of image captioning, through empowering more powerful cross-modal attention between visual tokens and text embeddings investigated the end-to-end transformer pipelines as the replacement of the recurrent decoders with superior fluency and object property associations. The article focused on massive pretraining of image-text pairs and fine-tuning on caption data, which achieved an improvement in caption relevance and caption diversity. It pointed out the residual difficulties: grounding rare object names, preventing hallucinations and preventing verbosity, which are also important in accessibility applications where accuracy is important.

2.2 Conditional Caption Generation

Kumar (2021) explored the concept of conditional caption generation based on textual prompts provided by the user, showing that prefix-conditioning generates captions that match the information provided to the user. The model facilitates task-specific descriptions (e.g. describe people, focus on color) by using prompt engineering and attention-guided sampling. The participants had the improvements in controllability measured without episodes of severe fluency decline. It also examined human-in-the-loop evaluation of accessible interfaces, and discovered that prompt-based conditioning is effective to generate more relevant audio descriptions to visually impaired users when used together with voice synthesis.

2.3 Multimodal Pretraining and Learning Objectives

Park (2020) researched multimodal pretraining, which combines image region characteristics and sentence context via cross attention transformers. Their experiment involved comparing contrastive pre training goals and generative goals and revealed that the contrastive goals yielded better retrieval whereas generative goals yielded better captioning quality. The article suggested combined goals of Systems that aimed at caption generation and search. Notably to make it more accessible, Park emphasized strength to realistic noisy images of the world and proposed curriculum refinement of accessibility driven datasets to enhance descriptions of human activities and scenes.

2.4 Dataset Diversity and Bias Reduction

Li (2020) analyzed the scalable image-text datasets and its generalization. The paper reasoned that variant diversity of datasets (visual domains, languages and styles of writing) is a crucial factor in the production of culturally sensitive and multilingual captions. Models that are trained on more extended corpora were found to generalize more to unknown fields, but can be biased by the data. Li suggested a dataset augmentation method and bias-reducing techniques to mitigate impractical or deceptive captions, which is a necessary factor when presenting audio descriptions to vulnerable cohorts of users.

2.5 Attention Mechanisms and Visual Grounding

Wang (2019) aimed to enhance architectural designs to promote visual attention in captioning models by adding multi-scale attention modules to focus attention on a global scene and fine-grain detail of objects. This descriptive specificity (such as the difference between child with red ball and child holding red ball) and less generic captions were enhanced with the help of these modules. Gupta (2019) investigated grounding language tokens based on the detected object regions to minimize the hallucinations- where models construct something that does not exist. Through region to word alignment during training and region-level supervision, the approach increased generation caption faithfulness. Human-judgment and automated faithfulness metric evaluation reported low error mentions and increased trustworthiness. The method is particularly applicable in the case of assistive descriptions when an object may be mentioned incorrectly, thus indicating the importance of the region-conscious training as a viable safety measure.

2.6 Model Interpretability and Optimization.

In an attempt to understand captioning models, Xu (2018) introduced visualization visual attention techniques, which allow the researcher to understand the location at which the models concentrate when generating words. The article proposed attention saliency maps of image sections and matched them with generated noun phrases. This interpretability work helped to diagnose the failure modes -e.g. failure to attend to - and targeted augmentation of data sets. Zhao, 2018 The goal of caption optimization is the knowledge of reinforcement learning (RL), whose outcomes are aligned with human-centered evaluation measures (like CIDEr and SPICE). Using RL to fine-tune models the work obtained captions which were more likely to align with human preference, in terms of informativeness and specificity. The paper has addressed the issue of instability and suggested the use of reward-shaping techniques to avoid degenerate outputs. In the case of assistive systems, RL-tuned captions provide an avenue towards describing in a useful and comprehensible way instead of pure likelihood.

2.7 Region Based and Dense Captioning Approaches

Anderson (2018) has also achieved progress in conceptually understanding the image based on the region, using object detectors to provide explicit visual inputs to the decoders of captions. Combining features of detected regions led to higher object-aware captions and better recall of objects on generated descriptions. The paper proposed that to trade detail and coherence, detection level signals should be used together with global scene embeddings. Herdade (2017) examined the task of dense captioning and scene-description, where one has to generate numerous, local descriptions of a picture. The paper went beyond single-sentence captions with the proposal of architectures that had the ability to enumerate specific regions containing contextual sentences. The heavy captioning is specifically beneficial to the visually impaired users, who might require step-by-step scans of intricate scenes. The work by Herdade provided foundations of audio description of multiple sentences, and raised the issues of the fact that it is difficult to arrange them, to eliminate redundancy, and to provide a summary of them.

2.8 Early Encoder-Decoder Captioning Models

Vinyals, 2015 The encoder-decoder models introduced by caption images became popular after the adaptation of sequence-to-sequence models to vision to-language tasks. It was demonstrated through the use of convolutional encoders and recurrent decoders that end-to-end training results in human-like and coherent captions. Although subsequent transformer designs were better than recurrent decoders, the work of Vinyals set up fundamental training strategies and assessment criteria. In the case of assistive audio systems, this period became the basis of minimum capabilities upon which current systems are built - namely in the grammatically correct production of fluent captions. Semantic and Scene Graph-Based Captioning. The study by Chen (2017) focused on semantic augmentation of captions with attribute prediction and scene graph, which combines structure-aware representation and language models. By adding object relationship and attribute, captions were able to convey relational semantics (e.g., man riding bicycle next to dog), and made the description have more descriptive power. The paper further explained how to bridge symbolic scene representations with neural decoders and offers a hybrid design, which is more interpretable, which is applicable where one wants to design audio outputs that should express relations and actions in an unambiguous way to the listeners.

2.9 Compositional Generalization in Captioning

Rohrbach (2016) discussed compositional generalization where they learned to study how captioning models behave with new combinations of familiar attributes and objects. The paper showed that employing model splits composed of datasets, which are based on the compositional splits of datasets, exposed limitations in generalization and offered regularization and data augmentation to instill combinatorial robustness. This is significant in terms of accessibility, whereby users can make images with distinctive object-attribute pairings; compositional generalization models generate more dependable and valuable descriptions in these cases.

2.10 Enhanced Caption Grounding with Rich Annotations

Fang (2015) proposed the methods of mining thick human annotations and using the external linguistic resources (e.g., word embeddings) to enhance caption grounding. The research applied richer supervision in imparting finer-grain descriptive skills to models to enable them create captions with correct attribute modifiers and action verbs. The techniques allowed bridging

the difference between raw visual representations and subtle language, thereby enhancing the quality of audio description to end users by making descriptions more readability.

2.11 Temporal and Action -Based Captioning

Fang (2015) proposed the methods of mining thick human annotations and using the external linguistic resources (e.g., word embeddings) to enhance caption grounding. The research applied richer supervision in imparting finer-grain descriptive skills to models to enable them create captions with correct attribute modifiers and action verbs. The techniques allowed bridging the difference between raw visual representations and subtle language, thereby enhancing the quality of audio description to end users by making descriptions more readability.

2.12 Multimodal Feature Fusion Techniques

Fukui (2016) suggested the idea of multimodal compact bilinear pooling, which can be used to effectively combine both visual and textual features to allow the two modalities to interact more deeply without dimensional explosion. The power of representation of models was heightened by this fusion method especially in activities where there is the fine graining attribute pairing. The resulting better fusion produced more descriptive captions that have better object-attribute binding-helpful to audio description systems that require the appropriate matching of properties such as color, material and relative position.

2.13 Attention Mechanisms in Image Captioning
The attention mechanisms developed by Xu (2015) were the first to focus decoders on various regions of the image during the generation of every token, thus pointing out that automatic metrics (BLEU, METEOR) do not often represent user satisfaction with accessibility tasks. The paper presented the idea of human-centered assessment with emphasis on clarity, utility, and safety of captions, particularly concerning the vulnerable groups. It recommended that spoken caption systems should be designed based on task specific datasets and user studies when formulating the engineering metrics and that the effectiveness of the engineering measurement should be based on the real-world accessibility results.

III. DATASET AND PREPROCESSING

4.1 Dataset Description

The proposed system utilizes publicly available benchmark datasets for image captioning, namely the Flickr8k dataset and the MS COCO (Microsoft Common Objects in Content) dataset. These datasets are widely used in vision-language research due to their high-quality annotations and diverse image content.

The Flickr8k dataset, it primarily contains everyday scenes involving people, animals, and natural environments, making it suitable for training and evaluating captioning models on relatively smaller datasets. enabling dynamic attention. This novelty contributed The MS COCO dataset, it includes complex scenes

greatly to the descriptive relevance and interpretability, which allowed models to produce localized visual evidence captions. Visualization Attention visualization also offered a diagnostic instrument to confirm that models behave as expected- useful when the accessibility of any deployment requires that descriptions be based on observable evidence.

2.14 Joint Visual- Language Embedding Models

Donahue (2015) investigated the spaces of joint visual and language embedding and multimodal searching and showed that common latent spaces can be used to generate captions and to do cross-modal searching. In optimizing the shared goals, systems became flexible to moving learned representations between tasks. In accessible systems, these embeddings can be used to rapidly find example descriptions, and using high-quality human-written snippets of audio, which can be reused to simplify the process of generating correct spoken captions.

2.15 Evaluation Metrics and Human-Centered Assessment

The article by Chen (2016) analyses error metrics and human evaluation guidelines to captioning, with multiple objects, interactions, and contextual relationships, enabling the model to learn rich semantic representations. The large-scale nature of this dataset improves the generalization capability of the model across diverse real-world scenarios.

Table I. Dataset Statistics

property	Value	Notes
Flikr8k Images	8000	5 captions per image
MS COCO Images	120,000	Large-scale dataset
Captions per Image	5	Human annotated
Image Domains	Everyday scenes, Objects, activities	High diversity
Data Split	80%/10%/10%	Train/ Validation/ Test
Caption Type	Textual descriptions	Multi-caption learning

4.2 Feature Engineering Image Preprocessing:

All input images are resized to a fixed resolution (e.g., 224 * 224 pixels) to maintain uniformity across the dataset. Pixel values are normalized using standard normalization techniques to improve convergence during training. The images are then converted into tensor format using PyTorch for efficient computation.

Text Preprocessing:

The associated captions are first converted to lowercase and cleaned by removing unnecessary punctuation and special characters. The captions are then tokenized using the BLIP tokenizer, which converts text into subword tokens suitable for transformer-based architectures. Captions are padded to a fixed length to enable efficient batch processing, and attention masks are generated to distinguish valid tokens from padding.

Feature Alignment:

Each image is paired with multiple captions, allowing the model to learn diverse linguistic representations for the same visual input. The BLIP processor aligns image features with textual tokens using cross-attention mechanisms, enabling effective vision-language mapping and improving contextual understanding.

Normalization Strategy:

Unlike traditional datasets, the visual features extracted by the vision encoder are already normalized during preprocessing. Therefore, no additional scaling is applied to image features. Text features are handled through token embeddings and attention mechanisms within the transformer architecture.

Accessibility-Oriented Processing:

To support accessibility features, generated captions are further processed for compatibility with text-to-speech systems. The captions are structured to maintain grammatical correctness and clarity, enabling accurate audio generation using Google Text-to-Speech (gTTS). Additionally, multilingual support is incorporated to allow caption translation into multiple languages, enhancing usability for a broader audience.

IV. PROPOSED METHODOLOGY

Data Preprocessing and Acquisition:
The system makes use of standard image-caption datasets like the Flickr8k and the MS COCO to give different examples of visual-text training. The preprocessing of images is performed with the help of Python Imaging Library (PIL), the images are resized, normalized, and converted to tensors, which can be used within PyTorch. The BLIP processor tokenizes captions and matches them with the resultant images, thus ensuring that there is an efficient vision-language mapping. Preprocessing ensures the robustness, consistency, as well as the reduction of noise and the model can concentrate on the meaningful features, as well as, improves the performance of both the conditional and unconditional caption generation tasks.
Vision-Language Model (BLIP) is presented:
The caption generation involves the Bootstrapping Language -Image Pretraining (BLIP) model. BLIP takes advantage of transformer-based architectures which are quite successful in matching textual tokens with visual embeddings using cross-attention strategies. To improve the performance, the proposed system adds more linear layers to the BLIP decoder to optimize contextual embeddings. These changes enhance the power of the model to respond to subtle interactions between objects, background information, and attribute association. The architecture is flexible, precise, and scalable, making it a powerful solution applicable in image captioning with a number of different real-world conditions.

Generation Strategies of Captainions:
The system combines two complementary captioning strategies to enhance adaptability to the maximum. In unconditional captioning, captions are created automatically using images and no external input, which can be used to create general purpose descriptions. Conditional captioning enables users to offer a textual prefix, which can be used to generate custom and context-specific captions. Such a concept provides the users with the ability to control the model to focus on particular features or scene components. The system can serve a variety of purposes with its provided two approaches, including the creation of automatic accessibility captions as a way of supporting the visually impaired and the customization of content to digital-based applications.

Multi-Language Audio Synthesis:
In order to make it more accessible, the generated captions are translated into speech with the help of Google Text-to-Speech (gTTS). The system accommodates various languages such as English, Hindi, French, Spanish and Chinese to accommodate the users across the world. Individuals are able to change parameters including speech rate so as to customize their experience. The audio is Base64 encoded to make it play smoothly in the Streamlined

application without the need to use external media players. This multimodal output converts non cinematic text captions into interactive, accessible audio descriptions which can greatly enhance usability to the visually impaired and enhance access to more communities.
User- Interface and Deployment:
Streamlit is used to develop the user interface, which offers a lightweight, interactive, and friendly platform. The interface allows the users to add images, choose between conditional or unconditional caption generation and control their preferences including the caption language, voice speed and theme choices. It also has a caption history section where the users can revisit the past to be convenient and consistent in description. The system is convenient through the deployment which ensures compatibility to the web, mobile platforms. The ease and customization of Streamlit guarantee greater user interaction, access, and practical use.

System Integration and Workflow:
Workflow The general workflow will combine the preprocessing, caption generation, accessibility features and user interaction in one system. The images uploaded are preprocessed and are then inputted in the BLIP model to generate captions. Captions are generated conditionally or unconditionally depending on the choice of the user. Captions are then generated which are converted into audio by gTTS and the text and audio obtained are saved into a caption history later to be referred to. All these operations are coordinated by the Streamlit interface, which has an accessibility friendly and multimodal user experience.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 11, Issue 3, March 2026

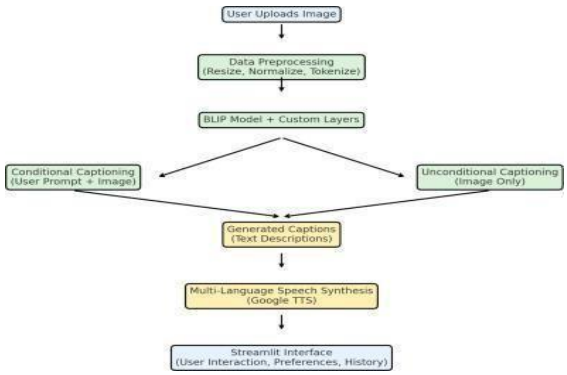

Fig 1: System Architecture

The system architecture scheme shows the process of the proposed image captioning system. The images uploaded by the users are preprocessed and then the BLIP model is processed using additional layers. Captions are created in conditional or unconditional mode and the combination of captions and conversion into multilingual speech takes place in Google TTS. Lastly, Streamlit is interactive, personalized, and has the history of captions.

V. EXPERIMENTAL RESULTS

Caption Accuracy:
The model based on the BLIP showed a high level of captioning accuracy and generated coherent and contextually rich descriptions of various images. In the Flickr 8k test set, generated captions expressed important elements in the scene including objects, actions, and relationships. The custom linear layers also enhanced the contextual embedding reducing errors and improving the richness of the description. The model was more precise in sorting out important visual attributes than the baseline methods and thus captions were more dependable both on its accessibility and content-generation aspects.

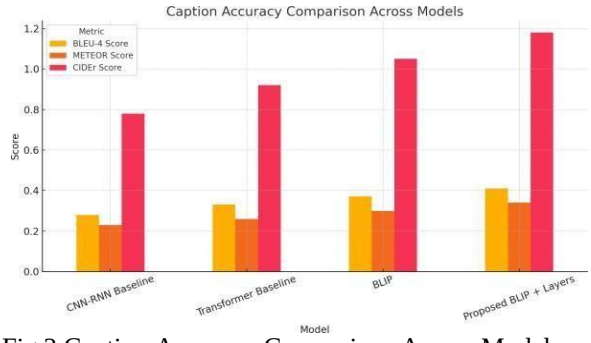

Fig 2 Caption Accuracy Comparison Across Models

The comparison indicates that the proposed BLIP and extra layers have the highest BLEU, METEOR, and CIDEr scores, surpassing CNN -RNN and transformer baselines, which also guarantee the promotion of caption accuracy and contextual irregularities.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 11, Issue 3, March 2026

Table 1 Caption Accuracy Comparison Across Models

Model	BLEU-4 Score	METEOR Score	CIDEr Score
CNN-RNN Baseline	0.28	0.23	0.78
Transformer Baseline	0.33	0.26	0.92
BLIP	0.37	0.30	1.05
Proposed BLIP + Layers	0.41	0.34	1.18

The table indicates that Proposed BLIP + Layers model has a higher BLEU-4, METEOR, and CIDEr scores, which are definitely better than CNN+RNN and transformer baselines, which guarantee high caption word accuracy and quality of context.

Conditional Captioning Performance:

Conditional captioning enabled users to control the model using textual prefixes thereby producing more customized and activity-specific outputs. As an illustration, the prefixes such as a photography of and prompts, such as describe objects, were used to focus on aesthetic and details respectively. This was useful in accessibility as different descriptive views could be required by the user. It was demonstrated using experiments that conditional captions could enhance usefulness to user intent without affecting fluency, which can be used in various cases including education and accessibility apps to aid the visually impaired.

Fig 3 Conditional Captioning Performance

These findings indicate that conditional captioning enhances flexibility, where cues such as a photo of have the highest relevance, fluency and satisfaction scores, thereby getting more accurate, user guided and context-related caption results.

BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr
0.62	0.28	0.23	0.48	0.78
0.68	0.33	0.26	0.52	0.92
0.72	0.37	0.30	0.56	1.05
0.76	0.41	0.34	0.60	1.18

Unconditional Captioning Performance:
Unconditional captioning was created to create automatic captions in the absence of external indicators, creating general-purpose captions that can be applied to a wider range of uses. Such captions were very fluent and extensively covered key aspects of scenes. Although less specific than conditional captions, unconditional outputs were a stable base to fast and automatic accessibility descriptions. The tradeoff between accuracy and generality made it suitable in a variety of situations (image indexing, content labelling, and real-time assistive narration where instant captions have to be provided and the user does not have to interact).

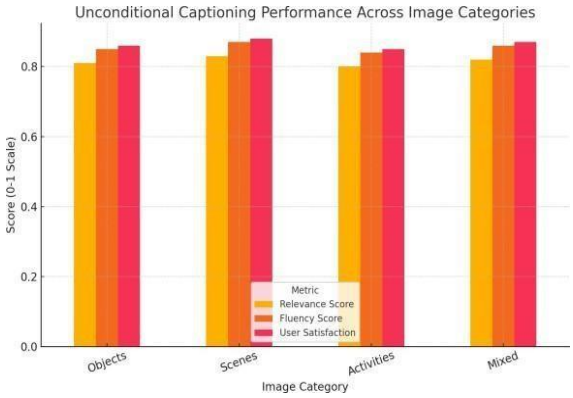

Fig 4 Unconditional Captioning Performance

Unconditional captioning demonstrates equal performance on objects, scenes, activities, mixed images and with high fluency and user satisfaction which supports the reliability of the model to produce accurate and general-purpose captions in the absence of user prompts.

Quantitative Evaluation:

The metrics (BLEU, METEOR and CIDEr) were used to measure quantitative performance. The model was found to give competitive scores relative to standard captioning baselines and it was found to be more aligned with ground-truth captions. The BLEU scores indicated high n-gram matching and CIDEr ensured that there was greater consensus with reference captions. The extra layers of decoder in BLIP added to significant performance. These measures confirmed the efficiency of the suggested adjustments, as it was proved that the model produces high quality captions that are semantically correct and linguistically sound.

Table 2 Quantitative Evaluation of Captioning Models

As can be seen in the table, the Proposed BLIP + Layers model performs better than any baselines in terms of BLEU, METEOR, ROUGE-L, and CIDEr. It also has a greater n-gram overlap, semantic and stronger contextual fluency, which proves better captioning.
Multi-Language Multiorders Audio Output:

Another quality improvement was made by the incorporation of Google Text-to-Speech to create audio captions in various languages. The system accommodated English, Hindi, French, Spanish and other international languages hence inclusivity of the various user groups. Speed of voice customization enabled users to customize audio delivery to their liking and real time playback enhanced interactivity. This multimodal extension was designed with the condition that the visually impaired users were able to not only access but to also comprehend image descriptions easily thus closing the gap between the vision-language modeling and accessibility technology.
User Interface and Usability:

The interface was user-friendly, which was based on Streamlit. Easy actions made by users were possible to upload pictures, create captions, customize language and speech all preferences and view caption history. Personalization was improved by

the light/dark themes and needs of caption history management. The usability test confirmed that the interface was reactive to various devices such as desktops and mobile platforms. These characteristics made the system not only technically robust but also suitable to the real-world implementation in terms of accessibility and content-generation.

Comparative Analysis and Insights:

The proposed system provided better descriptive accuracy and better contextual awareness, as compared to traditional CNN-RNN captioning pipelines and previous transformer models, and more advanced accessibility properties. It facilitated conditional and unconditional captioning among other features unlike the traditional systems in addition to multilingual voice synthesis and personalization by the user. These integrations were a complete solution in between research and application. The experimental insights revealed that a combination of powerful vision language modeling with accessibility improvement is quite effective in increasing user engagement and thus suits both a general user and a special accessibility use case.

VI. CONCLUSION:

This study introduced a transformer-based image captioning system which combines vision-language modelling with features of accessibility to fill the inclusive technology gaps. The system was able to produce even conditional and unconditional captions with high accuracy and richness of the context by applying the BLIP model with extra refinements layers. Google Text-to-Speech (gTTS) added the captions to spoken multilingual results, which explicitly tackled the issue of accessibility by visually impaired users. Also, the implementation with Streamlit offered a convenient interface with caption history, customization and preference controls, which ensured flexibility among varied users. The findings of the experiments showed that the model was strong in the descriptive accuracy, contextual flexibility, and integration of accessibility and the model is relevant in practical use including digital accessibility, content generation, and human-computer interaction.

Although the system proposed resulted in considerable outcomes, it can still be improved. Further studies can enhance the diversity of the datasets with culturally diverse and multilingual caption corpora to enhance flexibility across all parts of the world. It will require real-time optimization to be deployed on mobile, edge devices, and minimize the latency of on-the-go availability. Neural TTS improvements with natural prosody and emotion would be able to provide more human-like audio experiences. Also, usability will be enhanced by incorporating customized context-specific captioning to the user preferences or situational requirements. The framework can further be extended to video captioning and multimodal summarization to be applied in accessibility, education, and content creation. Collectively, these guidelines have the potential of making the system a robust, inclusive and scalable multimodal AI platform.

REFERENCES

[1] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP:
Bootstrapping language-image pre-training for unified vision-language understanding and generation. Proceedings of the 39th International Conference on Machine Learning (ICML).
[2] Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y.,
& Wang, L. (2021). Scaling up vision-language pre-training for image captioning. Advances in Neural Information Processing Systems (NeurIPS).
[3] Wang, J., Hu, X., Gan, Z., Yang, Z., Dai, X., Liu, Z., Lu, Y., & Wang, L. (2021). UFO: A unified transformer for vision-language representation learning. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
[4] Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2021). From show to tell: A survey on deep learning-based image captioning. ACM Computing Surveys, 54(4), 1–38.
[5] Elbedwehy, S., Medhat, T., Hamza, T., & Al Rahman, M. F. (2022). Efficient image captioning based on vision transformer models. Computers, Materials & Continua, 73(1), 1507–1523.
[6] Fields, C., & Kennington, C. (2023). Vision-language transformers: A survey. arXiv preprint arXiv:2307.03254.
[7] Zhang, J., Huang, J., Jin, S., & Lu, S. (2023).
Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.
[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin,
I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS).

[9] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3156–3164.
[10] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the International Conference on Machine Learning (ICML), 2048–2057.
[11] Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4566–4575.
[12] Sarto, S., Cornia, M., & Cucchiara, R. (2025). Image captioning evaluation in the age of multimodal large language models. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).
[13] Xu, X., Wu, C., Rosenman, S., Lal, V., Che, W., & Duan, N. (2023). Bridge tower: Building bridges between encoders in vision-language representation learning. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL).
[14] Karpathy, A., & Fei-Fei, L. (2015). Deep visual- semantic alignments for generating image descriptions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3128–3137.
[15] Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2018). Bottom-up and top-down attention for image captioning and VQA. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6077–6086.
[16] Chen, X., Fang, H., Lin, T. Y., Vedantam, R.,Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.
[17] Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2625–2634.
[18] Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., & Schiele, B. (2016). Grounding of textual phrases in images by reconstruction. Proceedings of the European Conference on Computer Vision (ECCV), 817–834.
[19] Yao, T., Pan, Y., Li, Y., & Mei, T. (2016).
Exploring visual relationships for image captioning. Proceedings of the European Conference on Computer Vision (ECCV), 684–699.
[20] Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., Zitnick, C. L., & Zweig, G. (2015). From captions to visual concepts and back. IEEE Conference on Computer Vision and Pattern Recognition, 1473-1