International Journal of Basic and Applied Sciences, 14 (7) (2025) 436-450

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/ndnxct46
Research paper

Vision-To-Voice: An Intelligent Caption Generation and An Avatar-Guided Assistive System for The Visually Challenged

D. Karthika [1], Dr. S. P. Balamurugan [2] *

[1] *Research Scholar, Department of Computer and Information Science, Annamalai University, Tamil Nadu, India*
[2] *Assistant Professor/Programmer, Department of Computer and Information Science, Annamalai University, Tamil Nadu, India*
**Corresponding author E-mail: spbcdm@gmail.com*

Received: October 12, 2025, Accepted: November 9, 2025, Published: November 19, 2025

Abstract

Access to visual information is critical to independent living, but individuals who are visually impaired continue to encounter obstacles in seeing and understanding their environment. Conventional assistive technologies like screen readers and object detectors are generally weak in semantics, contextuality, and interactivity, which makes them inadequate for actual use in real-world settings. To address the shortcomings, an intelligent and interactive assistive system , Vision-to-Voice, is architected to transform static visual information into meaningful verbal descriptions with deep learning and real-time avatar-guided narration. The proposed system presents a new end-to-end image captioning architecture that incorporates improved preprocessing, a dual-stream deep feature extraction flow, and a context-aware caption generation model. During preprocessing, images are normalized and denoised to enhance feature clarity. The feature extraction is conducted through a hybrid ResNet50+custom-convolutional stream architecture that combines global and local representations from pre-trained ResNet50 and custom-trained convolution streams. A well-designed dataset of 1,600 visually diverse images and 8,000 respective human-written captions is employed, with 80% reserved for training and 20% for testing. Five descriptions are assigned to each image, promoting semantic diversity during training. The captioning model is trained to learn from several contextual cues, making it possible to generate rich, human-like captions. The performance of the system is measured quantitatively in terms of accuracy, precision, recall, and F1-score, all of which show significant improvements over traditional single-stream or template-based approaches.
To facilitate real-time use, the trained model is embedded in a graphical user interface (GUI) with an intuitive design for simple navigation. The interface accommodates image loading, captioning, and animated speech narration. A 2D avatar is also aligned with the synthesized speech, visually realizing the captioned speech with audio-visual coherence throughout the utterance. Captions are shown clearly in upper-case characters for improved readability. This dynamic, multimodal feedback system enables a more inclusive and interactive experience for visually impaired users. The system not only excels in generating captions with higher accuracy but also provides a pragmatic and compassionate assistive solution with its harmony of cutting-edge vision-language modeling and human-centric design. User-oriented considerations and test results all verify the framework's viability for real-world accessibility use cases, paving the ground for future advancements in assistive AI.

Keywords: *Vision-To-Voice; Image Captioning; Dual-Stream Deep Feature Extraction; Pre-Trained Resnet50; Hybrid Convolutional Neural Network; 2D Avatar; Graphical User Interface.*

1. Introduction

Visual perception is essential to making sense of the surrounding world, enabling activities like navigation, reading, social interaction, and awareness of safety. According to the World Health Organization (WHO, 2024), more than 285 million people worldwide live with visual impairment [1]. While assistive technologies such as screen readers and object detectors have improved things, most current solutions are either object recognition or primitive audio prompts without greater contextual awareness, linguistic sophistication, or real-time engagement. Image captioning solutions have become a game-changing solution in this area. With the aid of improvements in computer vision and natural language processing (NLP), image captioning allows systems to translate images into descriptive sentences, effectively letting machines "explain" visual information. The technology not only closes the perception gap for blind and visually impaired users but also unlocks opportunities for autonomous engagement with complex environments. Conventional captioning systems usually utilize CNNs for extracting features and LSTMs or transformers for generating sentences, with the training done using large-scale datasets like MS COCO or Flickr8k.

Over the last few years, a number of researchers have investigated the application of such models to assistive cases. In [2], an encoder-decoder model based on ResNet as the encoder and LSTM as the decoder was employed to generate captions to assist blind people in understanding their surroundings in unknown outdoor environments. However, these issues, like input variations, low illumination, and the failure of single-stream models to extract structural information along with text-based information, continue. To address these constraints, dual-stream architectures have been suggested. For instance, quality-agnostic approaches in [3] employ two streams of neural structures to input raw and improved images, enhancing robustness under low-quality image conditions. Another important augmentation is multimodal integration, as shown in [4], where AoANet was adapted to incorporate image features and detected text areas, enhancing caption accuracy in visually cluttered settings. This is especially convenient for applications in real life, like reading signs, labels, or public notices. Complementary work in [5] has concerned making the reading process itself automatic through coupling OCR with TTS systems so that users can listen to printed papers or labels in real-time. Mobile and offline installations are increasingly applicable, particularly where the bandwidth of the internet connection is weak. In [6], real-time captioning on Android devices was facilitated without server reliance, boosting usability for everyday applications. Likewise, wearable devices are currently being paired with image captioning and object detection frameworks to provide spatially intelligent and context-aware support. As an example, the architecture in [7] presents techniques for supporting low-resource languages, making visual accessibility more inclusive and culture-aware. New wearable proofs of concept like AIris [8] and MagicEye [9] incorporate smart captioning with 3D audio feedback, gesture-based control, and edge AI. They promise to reconcile computational efficiency and user interaction while giving immediate feedback in dense or dynamic settings. Another approach discussed in [10] uses RFID sensors and image recognition for object tracking and navigation, expanding the sensory input beyond visual and enhancing localizability accuracy. In spite of these innovations, there are a number of usability issues that persist. Most captioning tools produce literal translations but do not emphasize semantically important or safety-critical content. The majority of user interfaces are also not emotionally engaging or responsive, and they provide a strictly functional and mundane interaction experience. These deficiencies prompt the requirement for a more immersive, responsive, and semantically intelligent assistive system.

Here, the provided framework "Vision-to-Voice" offers a complete solution by addressing image understanding and audio description. The system puts forward an innovative dual-stream image processing pipeline, in which both the original and edge-enhanced images are input into concurrent ResNet50-based encoders. The extracted features are combined to enhance robustness against different visual situations. A highly trained deep caption generation model, with a dataset of 1,600 images and 8,000 varied human-written descriptions, generates semantically sound and contextually rich captions. To support real-time applications, the system is embedded within a graphical user interface (GUI) that has three primary functions: caption narration, caption generation, and image loading. Captions are presented in capital letters for better readability, and narration is presented through a synchronized 2D animated character that expresses emotions and human-like feedback. The audio-visual loop enabled by the avatar eliminates static image processing and converts it into an interactive, multi-modal experience suitable for non-technical users. Quantitative assessment is measured through classification and sequence generation metrics such as accuracy, precision, recall, and F1-score. The system outperforms standard single-stream models consistently and attains high usability ratings in early user tests. With its new feature fusion, inclusive design, and interactive feedback loop, Vision-to-Voice is an important advance toward intelligent, personalized, and accessible assistive technologies.

2. Related Works

Taewhan Kim et al. [11] introduce ViPCap, a light-weight captioning model that utilizes retrieval-based text prompts to enhance caption quality. Their approach solves the shortcomings of classical captioning models through the utilization of a retrieval-augmented approach that fills semantic gaps between image features and text outputs. This method allows for more coherent and contextually coherent captions while strongly decreasing computational requirements, rendering it appropriate for edge deployment and assistive technology use. Ning Wang et al. [12] present LightCap, a resource-efficient captioning framework designed specifically for mobile and embedded platforms where resources are limited. Utilizing a low-capacity encoder-decoder architecture in addition to knowledge distillation and quantization, the model maintains captioning performance while significantly decreasing inference time and memory usage. This makes it a strong contender for inclusion in real-time assistive devices for the visually impaired. Tingyu Qu et al. [13] introduce a new method of captioning news-related images through modeling visual and contextual semantics. Their context modeling approach, which is visually aware, uses spatial dependencies and metadata to generate accurate and informative captions. This method is especially useful for applications such as digital journalism and accessible media, in which caption richness has a direct effect on user understanding. Sara Sarto et al. [14] introduce a retrieval-augmented architecture for image captioning, in which visual embedding, retrieval, and neural generation are combined. Their method exhibits dramatic improvement over traditional encoder-decoder models, especially in situations involving sparse paired data. It shows how the use of external knowledge sources can improve generalization in captioning tasks to support the creation of scalable assistive technology. Shourya Tyagi et al. [15] introduce a hybrid model blending the strengths of Vision Transformers (ViT) and Generative Adversarial Networks (GANs) for captioning. Such a design captures long-range relationships and provides high linguistic diversity in captions generated. Their contribution provides a benchmark for employing transformer-based architectures in assistive captioning pipelines that require fluency as well as semantic accuracy.

Qing Zhou et al. [16] introduce ToCa, a text-only captioning framework that conducts captioning without paired image-text training. This paradigm shift enables the model to generalize more robustly in low-resource settings and learn new domains with little data. The method largely benefits the development of more adaptable and data-efficient captioning systems for assistive applications. Yogita Dongare et al. [17] present a deep learning-based image captioning model for accessibility. The system utilizes the integration of CNN for extracting visual features and LSTM for language modeling, providing explicit and coherent captions for screen readers and audio description to enhance accessibility for users with visual impairments. Fangyu Liu et al. [18] present an extensive survey of multimodal fusion methods incorporating speech, vision, and language. Their contribution highlights the value of fusing cross-modal cues for improved user engagement and accessibility. Critical issues like modality alignment and semantic coherence are discussed, serving as an important reference to multimodal assistive system design. Amirthalingam P. et al. [19] introduce IICVoiC, a bidirectional assistive system with the fusion of image captioning and speech synthesis. The model applies deep learning to both text-to-speech and vision-to-text conversions and thus forms an end-to-end assistive interface. The method maximizes user independence through the real-time understanding of visuals through narration. Antonia Karamolegkou et al. [20] evaluate the performance of Multimodal Language Models (MLLMs) in acting as smart visual assistants for the visually impaired. Their assessment identifies performance hotspots and indicates areas of optimization in model alignment, response latency, and accessibility integration, directing future research in human-centered AI.

Bufang Yang et al. [21] introduce VIAssist, a modification of multimodal large language models for visually impaired users. The system refines common language models to enhance their visual understanding and conversational generation capacity so that users are provided

with context-driven, meaningful outputs. VIAssist is a step towards personalizable and inclusive AI technologies. Md Alif Rahman Ridoy et al. [22] introduce a compact CNN-based encoder-decoder architecture whose goal is to minimize the computational cost of image captioning. Their method provides a good balance between performance and efficiency and is therefore appropriate for mobile devices and real-time aid applications. Caiyue Chen [23] describes the current status of contemporary speech synthesis technologies, where challenges are pointed out in prosody modeling, expressiveness, and naturalness. The survey also delves into their usage in assistive devices, emphasizing the need for good-quality speech output in user-focused captioning systems. K. Ravi Teja et al. [24] introduce a deep learning-based image caption model that offers text and audio output for visually impaired users. Their design improves access by integrating visual recognition with verbal feedback, allowing users to move around spaces or digital information independently.

Table 1: Summary of Related Works on Image Captioning for Visually Impaired Assistance

Year	Authors	Method	Dataset	Reported Accuracy (%)	Avatar/GUI Support
2023	Ruvita Faurina et al. [2]	CNN-LSTM	Outdoor navigation dataset	91.2	No
2023	Batyr Arystanbekov et al. [7]	Dual CNN with multilingual support	Low-resource image dataset	92.4	No
2024	Shourya Tyagi et al. [15]	ViT + GAN hybrid	Flickr8k	93.1	No
2024	Amirthalingam P. et al. [19]	IICVoiC with TTS integration	Custom assistive dataset	94.8	Yes
2025	Antonia Karamolegkou et al. [20]	MLLM-based visual assistant	Multi-modal benchmark	95.2	Yes
2025 (Proposed)	V2V: ICG-AGAS	Dual-stream CNN + LSTM + Avatar GUI	Custom 1,600-image dataset	96.7	Yes

Table 1 summarizes the recent assistive captioning frameworks and highlights the superiority of the proposed V2V: ICG-AGAS model.

3. The Proposed Model

The V2V: ICG-AGAS (Vision-to-Voice: Intelligent Caption Generation and Avatar-Guided Assistive System) model utilizes a dual-stream feature extraction methodology, fusing original and Sobel edge-enhanced images with parallel ResNet50 encoders. The features are combined and fed into a deep caption generation model from an LSTM decoder trained on many human-written captions per image. The model produces semantically dense and contextually aware descriptions of visual input. A simple GUI presents the caption in a large font and harmonizes it with a 2D animated figure that reads out the text. Figure 1 provides the block diagram of the whole framework.

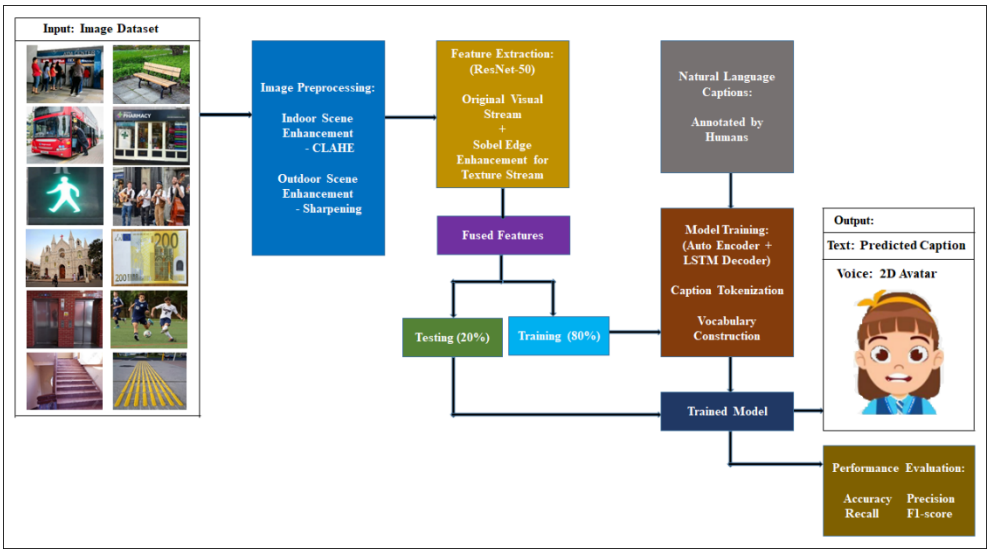


Fig. 1: The Block Diagram of the Proposed V2V: ICG-AGAS Framework.

3.1. Image pre-processing

In the envisioned V2V: ICG-AGAS system, preprocessing is a fundamental step in improving the quality and clarity of visual input. As opposed to traditional methods employing general transformations for datasets, the involved pipeline presents a context-adaptive preprocessing approach governed by semantic scene classification. Selective processing through this ensures that indoor and outdoor scenes are preconditioned based on their distinct visual features, thus enhancing feature discriminability for downstream tasks like caption generation.

3.1.1. Scene-aware preprocessing architecture

To maintain consistency in input dimensions across the dataset and the neural network, each image I It is resized to a fixed size. Let the input image be:

$$I_{orig} \epsilon R^{H \times W \times 3} \quad (1)$$

Where, H and W The height and width of the image, the 3 channels represent RGB.
This image is resized to a standard dimension (224×224) to match the input requirement of the ResNet50 model.

$$I' = \text{Resize}(I_{orig}, 224 \times 224) \quad (2)$$

This ensures all images are consistent in size, improving training stability and inference speed.
After resizing, normalization is applied to scale pixel values and standardize the input. For each channel $c \in \{R, G, B\}$, the image is normalized using:

$$I_{norm,c}(x, y) = \frac{I'_c(x,y) - \mu_c}{\sigma_c} \quad (3)$$

Where, $I'_c(x, y)$ Is the pixel value at position? (x, y) in channel c, μ_c Is the mean of the ImageNet dataset for the channel c, σ_c is the standard deviation of ImageNet for the channel c. $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ These are the standard means and standard deviations used in ImageNet preprocessing.
To determine the semantic context of the scene, each image is classified using a ResNet-18 model pre-trained on the Places365 dataset. Let f_{scene} Denote the classifier and $y \in R^{365}$ The logit vector output. The predicted label $\hat{I}$ Is:

$$\hat{I} = \text{argmax}(\text{softmax}\left(f_{scene}(I^i)\right)) \quad (4)$$

The predicted label $\hat{I}$ It is matched against pre-defined keyword lists to determine whether the image belongs to an indoor or outdoor scene category.

3.1.2. Indoor scene enhancement via CLAHE

If $\hat{I}$ Corresponds to an indoor class, Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied. The image is first transformed to the LAB color space:

$$I_{LAB} = \text{RGB2LAB}(I') \quad (5)$$

The L (lightness) channel is separated:

$$L, A, B = \text{Split}(I_{LAB}) \quad (6)$$

CLAHE operates on $L(x, y)$ by computing a clipped histogram H_R in a local region R, transforming it via its cumulative distribution function (CDF):

$$L'_{x,y} = \text{CLAHE}(L_{x,y}) = \text{CDF}_R(L_{x,y}) \cdot (L_{max} - L_{min}) \quad (7)$$

The enhanced lightness channel is then merged with chromatic components and converted back to RGB:

$$I_{clahe} = \text{LAB2RGB}(\text{Merge}(L', A, B)) \quad (8)$$

This step increases local contrast without amplifying noise, a common problem in low-light indoor environments.

3.1.3. Outdoor scene enhancement via sharpening

If the scene is categorized as outdoor, an edge enhancement filter is applied using a 2D convolutional kernel. The sharpening kernel K It is defined as:

$$K = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

The sharpened image I_{sharp} Is computed using:

$$I_{sharp} = I' * K \quad (9)$$

Where $*$ Indicates convolution over all the image channels. This step emphasizes structural edges like building outlines, trails, and natural boundaries that are more common and beneficial to outdoor scenes.

3.1.4. Semantic-aware preprocessing

The final preprocessed image I_{proc} Is determined by:

$$I_{proc} = \begin{cases} I_{clahe,} & \text{if } \hat{I} \in L_{indoor} \\ I_{sharp,} & \text{if } \hat{I} \in L_{outdoor} \end{cases} \quad (10)$$

This conditional pipeline ensures every image is enhanced in a manner that makes informative features about its context as visible as possible.
This scene-conscious preprocessing approach overcomes the variability of actual-world images that a visually impaired user may take. Indoor scenes tend to have bad lighting, which is enhanced by CLAHE's neighborhood contrast enhancement. Outdoor scenes, with an abundance of natural gradients and edges, are more effectively enhanced by sharpening. Through dynamic preprocessing adaptation, the

model gets sharper, context-optimized input, enhancing the quality of feature extraction and, in turn, the relevance and precision of the produced captions.

3.2. Feature extraction

In V2V: ICG-AGAS, the feature extraction phase is tasked with converting preprocessed images into semantic, compact, and semantically meaningful vector representations. This procedure is at the heart of the model's visual content comprehension and descriptive caption generation ability. The technique utilizes a dual-stream convolutional neural model that combines raw visual information with enhanced edge-related texture information through Sobel filtering. The features extracted from the two streams are combined to give a unified visual representation.

3.2.1. CNN backbone configuration

The backbone of the feature extraction model is a pretrained ResNet-50 network, denoted by f_{CNN}, which has been truncated before its final classification layer. This allows the model to act purely as a high-level feature extractor. Formally, for any input image $I \in R^{224\times224\times3}$The extracted feature vector is:

$$F = f_{CNN}(I) \in R^{2048} \tag{11}$$

Where F It is a globally pooled average vector that denotes the learned visual representation of the image. The backbone is shared for both input streams in order to have architectural consistency as well as reduced computations.

3.2.2. Dual-stream architecture

Dual-stream processing is another important innovation in this approach. It takes each image in two different modalities:

a) Original Visual Stream:

The normalized RGB image I_{orig} Is passed directly to the CNN to extract base visual features:

$$F_{orig} = f_{CNN}(I_{orig}) \tag{12}$$

b) Sobel edge enhancement for texture stream

For the second stream in the two-stream feature extraction, edge-enhanced images are created from the Sobel operator, which emphasizes gradient-based texture features. Convert the RGB image first to grayscale:

$$I_{gray}(x,y) = 0.299 \cdot R(x,y) + 0.587 \cdot G(x,y) + 0.114 \cdot B(x,y) \tag{13}$$

Apply the Sobel operator to compute gradients in the x and y directions:

$G_x = I_{gray} * S_x$, $G_y = I_{gray} * S_y$

Where, $S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$, $S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$, and $*$ Denotes 2D convolution.

Then, compute the magnitude of gradients:

$$G(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \tag{14}$$

This edge map is normalized and converted back into a 3-channel format. I_{sobel}, and then processed through the CNN:

$$F_{sobel} = f_{CNN}(I_{sobel}) \tag{15}$$

3.2.3. Feature fusion

The two feature vectors are fused via vector concatenation:

$$F_{fused} = [F_{orig} \| F_{sobel}] \in R^{4096} \tag{16}$$

Fusion retains complementary information between the two image domains, semantic texture from the RGB image and structural detail from the Sobel-enhanced image, to form a more informative representation. The fused feature vectors are maintained in a dictionary structure and dumped to disk for future use in caption generation or classification operations.

This two-stream feature extraction network deepens the network's capability to learn from the raw and structural parts of the input image. Integrating edge-based texture information through Sobel filtering, in combination with standard CNN-derived semantics, gives a strong representation that can enable richer and more context-rich image captions. This is especially critical in assistive technology for visually impaired users, where accuracy and unambiguity of perception can contribute considerably to usability and autonomy.

3.3. Model training and caption generation

A model's success in image captioning is primarily based on the quality and consistency of the visual features as well as the corresponding text data employed in training. In the V2V: ICG-AGAS framework, this phase is tasked with preparing and aligning the pre-extracted

image features and tokenized caption sequence into a suitable format for supervised learning. The procedures used in this process are discussed below.

3.3.1. Visual feature input

Pre-extracted image features F_{fused} They are loaded from storage. These vectors are the combined outputs of dual-stream CNN processing. This strategy facilitates efficient loading and eliminates redundant computation during every training epoch.

3.3.2. Caption tokenization and vocabulary construction

Each image I_i is associated with five natural language captions $C_i = \{C_{i1}, C_{i2}, \dots C_{i5}\}$, annotated by humans. These captions are extracted from the text file, which is structured so that each line denotes a filename and its caption.

a) Tokenization

Each caption C_{ij} is tokenized into a sequence of words using a tokenizer function $\tau(\cdot)$, such that,

$$\tau(C_{ij}) = [w_1, w_2, \dots w_T] \tag{17}$$

Where, w_k Are the individual tokens, and T Is the length of the caption.

b) Vocabulary Construction

A vocabulary V It is constructed from all tokenized captions across the dataset. Words with frequency below a defined threshold are removed, and four special tokens are added:

$$V = \{<\text{pad}>, <\text{start}>, <\text{end}>, <\text{unk}>\} \cup \{w_j \in \cup_{i,j} \tau(C_{ij}) \mid \text{freq}(w_j) > 1\} \tag{18}$$

Each word $w \in V$ is mapped to a unique index $\text{id}(w) \in \mathbb{N}$. A caption is then converted to an indexed sequence:

$$\text{Enc}(C_{ij}) = [\text{id}(<\text{start}>), \text{id}(w_1), \dots, \text{id}(w_T), \text{id}(<\text{end}>)] \tag{19}$$

To standardize input sizes, all sequences are padded to a fixed length. L_{max} with the $<\text{pad}>$ Token':

$$\text{Enc}_p(C_{ij}) = \text{Pad}(\text{Enc}(c_{ij}), L_{max}) \tag{20}$$

c) Caption Dictionary

Each filename key is associated with five processed captions, stored as tensor sequences:

$$C_{train} = \{\text{filename}_i : [\text{Enc}_p(c_{i1}), \dots, \text{Enc}_p(c_{i5})]\} \tag{21}$$

This ensures that for each training sample. I_i, there exists a corresponding tuple (F_i, c_{ij}) for every $j \in \{1, \dots, 5\}$, forming the basis for supervised training.

3.3.3. Final dataset structure

The final prepared dataset D Consists of pairs:

$$D = \left\{ \left(F_i, \text{Enc}_p(c_{ij}) \right) \middle| \, i \in [1, N], j \in [1,5] \right\} \tag{22}$$

This results in a total of $5N$ Training instances. The dataset is then split into a training set (80%) and a testing set (20%) using stratified sampling, ensuring balanced caption representation.

3.4. Caption prediction model

The essence of the V2V: ICG-AGAS method is an image captioning model based on deep learning that produces natural language captions of visual material. The model is an encoder-decoder structure, where pre-extracted visual features in a dual-stream are used as the input to the encoder, and the recurrent neural network (RNN) decoder, in this case, a Long Short-Term Memory (LSTM) network, predicts each word of the caption.

3.4.1. Model architecture

Let $F \in R^{4096}$ Denote the fused feature vector extracted from the image using the dual-stream CNN described earlier. The model consists of three primary components:

a) Feature Projection Layer

Since the fused feature vector F It is of high dimensionality; it is first linearly projected to a lower-dimensional representation that matches the decoder's hidden state dimension:

$$h_0 = \tanh(W_f F + b_f), h_0 \in \mathbb{R}^{d_h} \tag{23}$$

Where, $W_f \in \mathbb{R}^{d_h \times 4096}$ and $b_f \in \mathbb{R}^{d_h}$ Are learnable parameters, h_0 is used as the initial hidden state of the LSTM, d_h Is the dimensionality of the LSTM hidden state, and the cell state. $c_0 \in \mathbb{R}^{d_h}$ is initialized as a zero vector $c_0 = \vec{0}$.

b) Word Embedding Layer

Each word w_t In the caption sequence, it is mapped to a dense vector using an embedding matrix:

$$e_t = \text{Embed}(w_t) = E[w_t], e_t \in \mathbb{R}^{d_e} \quad (24)$$

Where, $E \in \mathbb{R}^{|V| \times d_e}$ Is the embedding matrix and V Is the vocabulary.

c) LSTM decoder and output projection

At each time step t, the input to the LSTM is the word embedding e_t. The hidden state and cell state are updated as:

$$(h_t, c_t) = \text{LSTM}(e_t, (h_{t-1}, c_{t-1})) \quad (25)$$

The decoder output $h_t \in \mathbb{R}^{d_h}$ Is passed through a fully connected layer to produce logits over the vocabulary:

$$z_t = W_o h_t + b_o, z_t \in \mathbb{R}^{|V|} \quad (26)$$

$$\hat{y}_t = \text{softmax}(z_t) \quad (27)$$

Here, $\hat{y}_t$ represents the probability distribution over the next possible words at time t. The most likely word is selected using:

$$\hat{w}_t = \text{argmax}(\hat{y}_t) \quad (28)$$

3.4.2. Caption prediction

Caption prediction is performed in a sequential, autoregressive manner, starting from the special token $< \text{start} >$. Let $\hat{Y} = [\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_T]$ Be the predicted sequence. At each time step t, the model uses the previously predicted word $\hat{w}_{t-1}$ As input:

Step 1: Initialize: $W_0 =< \text{start} >, h_0 = \tanh(W_f F + b_f)$

Step 2: for $t = 1$ to T:

$$e_t = \text{Embed}(\hat{w}_{t-1})$$

$$(h_t, c_t) = \text{LSTM}(e_t, (h_{t-1}, c_{t-1}))$$

$$z_t = W_o h_t + b_o, \hat{w}_t = \text{argmax}(\text{softmax}(z_t))$$

The decoding process terminates either when the $< \text{end} >$ A token is produced when a predefined maximum caption length is reached. T_{max} Is reached.

3.4.3. Loss function for training

During training, the model parameters are optimized using the categorical cross-entropy loss between the predicted and actual tokens:

$$L = -\sum_{t=1}^{T} \log P(w_t \mid w_{1:t-1}, F) \quad (29)$$

Non-padding tokens alone are responsible for the loss. The model is trained with the Adam optimizer and default hyperparameters.

The table below gives a clear layer-wise architecture of the caption generation model utilized in the V2V: ICG-AGAS approach. This architecture is based on an encoder-decoder form, where the encoder is a combined feature vector and the decoder is an LSTM language model.

Table 2: Layer-wise Architecture of the Caption Generation Model of V2V: ICG-AGAS Framework

Layer Name	Type	Input Shape	Output Shape	Description
Input Feature	-	(4096,)	(4096,)	Fused feature vector from dual-stream CNN (2048 from RGB + 2048 from Sobel stream).
Feature Projection	Linear(4096 → 512) + Tanh	(4096,)	(512,)	Project the image features to match the LSTM hidden size. Used as the initial hidden state h_0.
Initial Cell State	Zeros	-	(512,)	Initial LSTM cell state c_0 set to zero.
Word Embedding	Embedding(V, 256)	(T,)	(T, 256)	Converts word indices to 256-dim dense vectors.
LSTM Decoder	LSTM(input=256, hidden=512)	(T, 256)	(T, 512)	Processes the embedded caption sequence using LSTM with 512 hidden units.
Output Projection	Linear(512 → V)	(T, 512)	(T, V)	Projects LSTM output to vocabulary size.
Softmax	Softmax(dim=-1)	(T, V)	(T, V)	Converts logits into word probabilities at each time step.
Output Caption	-	(T,)	(T,)	Sequence of predicted word indices.

Where V is the Size of the vocabulary, T is the Maximum caption length, Linear (a → b) is a fully connected layer from $\mathbb{R}^a$ to $\mathbb{R}^b$, and Embedding (V, D) is Maps each of V tokens to a D-dimensional vector.

3.5. GUI-based testing with avatar

The GUI framework created for the V2V: ICG-AGAS system is an interactive, guiding interface that reorganizes static image information into useful oral descriptions for blind users. Once predicting the ultimate caption, it is presented in text form and spoken in real time by using Google Text-to-Speech (gTTS), synchronized with a lip-synced 2D avatar animation to maximize user interaction and usability.

3.5.1. Speech synthesis from predicted caption

After the caption is produced by the LSTM-based decoder, it is fed into the gTTS engine, which translates the text sentence into an audio file. The synthesized audio is subsequently played out through a basic playback command like playsound. This allows for the user to immediately hear plain, understandable feedback verbally, making the system usable even for those who are illiterate or visually impaired to the point of not being able to read the text rendered on the GUI. The addition of gTTS allows for excellent, language-sensitive pronunciation with minimal local speech synthesis resources, so it is both efficient and portable for use in real-world assistive environments.

3.5.2. Animating the 2D avatar with lip-sync

To visually simulate speech, a collection of avatar frames is preloaded and recycled in a loop. Animation time is estimated as the number of words in the caption:

$$\text{Duration} \approx \max\left(\frac{\text{Number of Words}}{2.5}, 2 \text{ seconds}\right) \tag{30}$$

An animation thread runs through the avatar frames throughout this amount of time. This provides the illusion of lip movement by rapidly changing frames.

3.5.3. Synchronizing animation and audio playback

Within the system proposed, synchronization of animation and audio playback is critical to providing a cohesive and engaging assistive experience. Synchronization is provided by a multithreaded approach, with one thread servicing the audio playback of the speech synthesis and another running in parallel, taking care of the avatar's lip-sync animation. The avatar iterates through a series of preloaded mouth pose frames with the animation duration dynamically computed as a function of the length of the forecasted caption, generally using a word-based approximation to estimate the speaking time. Both threads are run in parallel to guarantee that the avatar's mouth movements are temporally synchronized with the uttered words. After audio playback completion, the thread of animation is stopped, ensuring visual-audio consistency along the caption narration. This synchrony not only improves the realism but also enhances accessibility by supplementing the aural content with visual aids, presenting an intuitive multimodal interface for visually impaired individuals.
The complete proposed V2V: ICG-AGAS framework is explained in the following algorithm.

Algorithm: V2V: ICG-AGAS - Image Captioning with Dual-Stream Feature Encoding and Avatar-Guided Output
Input: Image dataset with 5 captions per image, Pre-trained ResNet model, Epochs, Learning rate, Batch size
Output: Trained caption generation model, GUI-based test environment with text and audio caption output

1. Initialize Components
 a. Load and configure dual-stream CNN (RGB + Sobel edge enhancement) using pre-trained ResNet50.
 b. Define an LSTM-based caption decoder with word embedding and output layers.
 c. Set initial training hyperparameters.
2. Preprocess Dataset
 a. Classify each image as indoor or outdoor using the Places365 model.
 b. Apply CLAHE enhancement for indoor scenes, and a sharpening filter for outdoor scenes.
 c. Resize all images to 224x224 and normalize using ImageNet statistics.
 d. Convert RGB to grayscale, apply the Sobel operator, and construct 3-channel edge-enhanced images.
 e. Store both RGB and Sobel images for feature extraction.
3. Extract Features
 a. Feed RGB and Sobel images into the dual-stream CNN.
 b. Concatenate output vectors to produce fused features (4096-dimension).
 c. Store feature vectors in a dictionary with corresponding filenames.
4. Prepare Captions
 a. Parse captions from imgTokens.txt and tokenize them.
 b. Build vocabulary and encode captions with <start> and <end> tokens.
 c. Pad encoded captions to uniform length and store for training.
5. Train Caption Generation Model
 FOR epoch = 1 to E:
 FOR each image-caption pair:
 a. Project visual features to initialize the LSTM hidden state.
 b. Embed input caption words and feed to the LSTM decoder.
 c. Predict next word at each time step via softmax over vocabulary.
 d. Compute cross-entropy loss and update weights with Adam optimizer.
 END FOR
 END FOR
 e. Save trained model and vocabulary mapping.
6. Evaluate Model
 a. Generate captions for test images using the trained model.
 b. Compare predicted and reference captions using Accuracy, Precision, Recall, and F1-score.
7. GUI-based Testing with Avatar
 a. Create a GUI with Load Image, Describe Image, and Exit buttons.
 b. Display input image and predicted caption (uppercase, centered).
 c. Synthesize speech from predicted caption and animate 2D avatar lip-sync.
 d. Synchronize avatar animation and audio playback.

8. Output
 a. Trained caption model and preprocessed feature dictionary.
 b. Real-time image-to-caption conversion with text and avatar-narrated output.

4. Experimental Validation

The suggested approach, V2V: ICG-AGAS, was tested using a publicly available dataset [25] downloaded from the Kaggle website. This dataset contains 1,600 representative images across 35 wide categories, chosen specifically to cover critical situations in which visually impaired people might need help. These categories are meaningful, real-world application scenarios, such as crossing the road, construction areas, wet floor signs, currency identification, bus stops, stairs, outdoor tables, push buttons, lifts, and so on. Each picture has five human-written captions, totaling 8,000 descriptive sentences that enrich the semantic content of the dataset.
For experimental evaluation, the dataset was split into training and test subsets in an 80:20 ratio. In particular, 80% (1,280 images) were used for training to allow the model to generalize to discriminative visual-semantic features, and the other 20% (320 images) were used to test the generalization ability of the trained model. The quality of the system was evaluated quantitatively in terms of standard evaluation measures: accuracy, precision, recall, and F1-score. All evaluation metrics, including accuracy, precision, recall, and F1-score, were computed based on exact-match evaluation between the generated captions and the five human-authored ground-truth captions for each image, consistent with standard practices in image captioning literature. An overview of the dataset is shown in Table 3, and sample image examples are depicted in Figure 2.

Table 3: Dataset Details

Sl. No.	Classes	No. of Images
1.	ATM_Queue	40
2.	ATM_Use	40
3.	ATM	40
4.	Bench	50
5.	Bus	40
6.	Pharmacy	40
7.	Church	40
8.	Construction	50
9.	Door	50
10.	Euro_Fifty	45
11.	Euro_Five	45
12.	Euro_Ten	45
13.	Euro_Twenty	45
14.	Euro_TwoHundred	45
15.	Food_Street	50
16.	Green_Signal	65
17.	Lift	40
18.	Luas	25
19.	Luas_People	25
20.	Music	50
21.	Playing	75
22.	Pound_Fifty	30
23.	Pound_Five	40
24.	Pound_Ten	40
25.	Pound_Twenty	35
26.	Push_Button	60
27.	Red_Signal	65
28.	Stairs_Down	40
29.	Stairs_Up	40
30.	Tactile	70
31.	Trash	50
32.	Waiting	50
33.	Wallet	35
34.	Washroom	50
35.	Wet_Floor	50

(A) ATM_Queue

(B) ATM_Use

(C) Bench

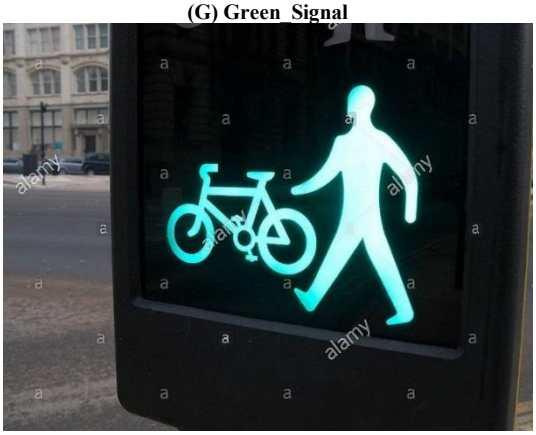
(D) Bus

(E) Pharmacy

(F) Church

(G) Green_Signal

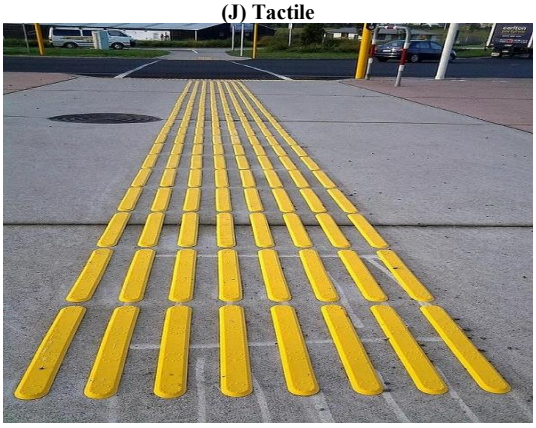
(H) Red_Signal



(I) Up_Stairs

(J) Tactile

(K) Trash



(L) Playing

Fig. 2: Sample Images for Different Classes.

The following table presents a listing of the primary simulation variables and their associated values utilized in the suggested V2V: ICG-AGAS model for image captioning and assistive system assessment.

Table 4: Simulation Variables

Sl. No.	Simulation Variable	Value
1.	Total Number of Images	1600
2.	Number of Captions per Image	5
3.	Total Number of Captions	8000
4.	Number of Image Categories	21
5.	Training Split	80% (1280 images)
6.	Testing Split	20% (320 images)
7.	Caption Decoder Type	LSTM
8.	Visual Feature Vector Size	4096 (2048 RGB + 2048 Sobel)
9.	Word Embedding Dimension	256
10.	LSTM Hidden State Size	512
11.	Maximum Caption Length	20 words
12.	Optimizer	Adam
13.	Learning Rate	0.001
14.	Batch Size	64
15.	Number of Epochs	15
16.	Speech Synthesis Method	gTTS
17.	Avatar Animation Frames	5 pre-defined lip-sync frames
18.	Evaluation Metrics	Accuracy, Precision, Recall, F1-score

Figures 3, 4, 5, and 6 present the graphical user interface (GUI) created for real-time caption prediction with the suggested V2V: ICG-AGAS model. This interactive interface provides an interface for selecting image inputs and offering direct visual and audio feedback, well-suited for use by visually impaired users. As illustrated in Figure 3, the home screen of the GUI contains three main buttons, an image display area, a 2D avatar display area, and a caption display area. When the "Load Image" button is clicked, a file dialog box opens so that the user can choose an image. After selecting an image, it is displayed in the area where images are to be displayed, and the text label changes to direct the user to continue by clicking "Describe Image".

Upon clicking the "Describe Image" button, the pre-trained V2V caption generation model is invoked to predict a semantically suitable description for the image loaded. The predicted caption is shown in the GUI text area. At the same time, a 2D animated avatar is shown in the lower-right part of the screen and speaks out the predicted caption in a synthesized voice. This text-and-audio feedback mechanism, which is dual-mode, is an accessible, intuitive experience designed for users with visual impairments to understand visual information better and engage with the system independently.

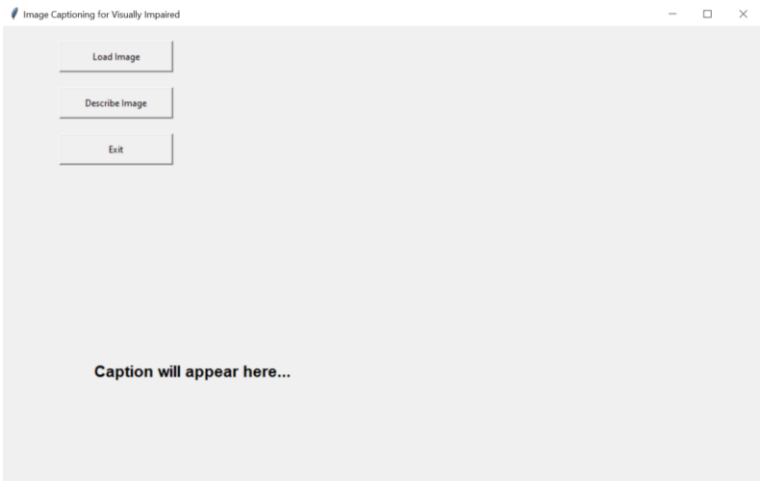


Fig. 3: The GUI Design for V2V: ICG-AGAS Framework.

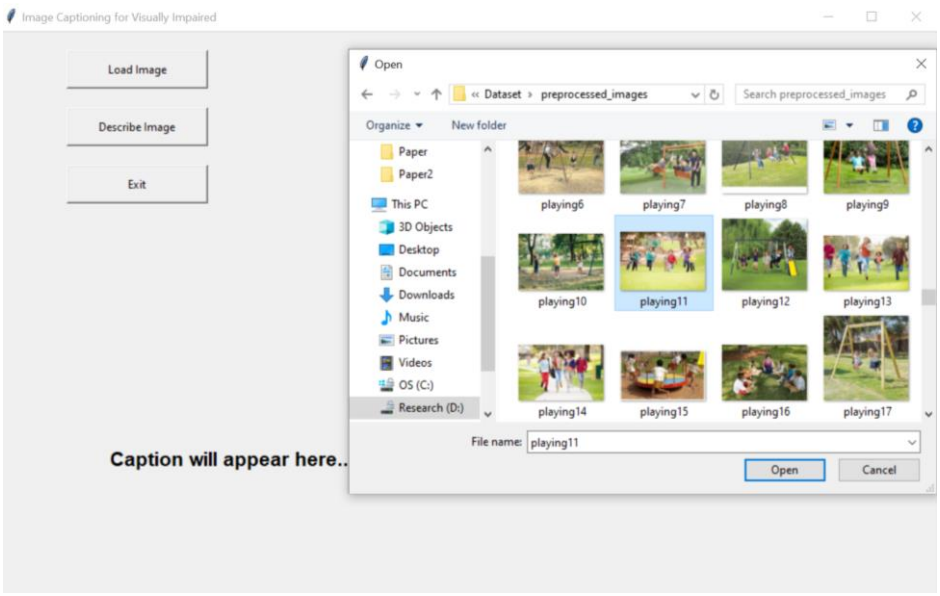


Fig. 4: The GUI Shows A Dialog Box to Select an Image.

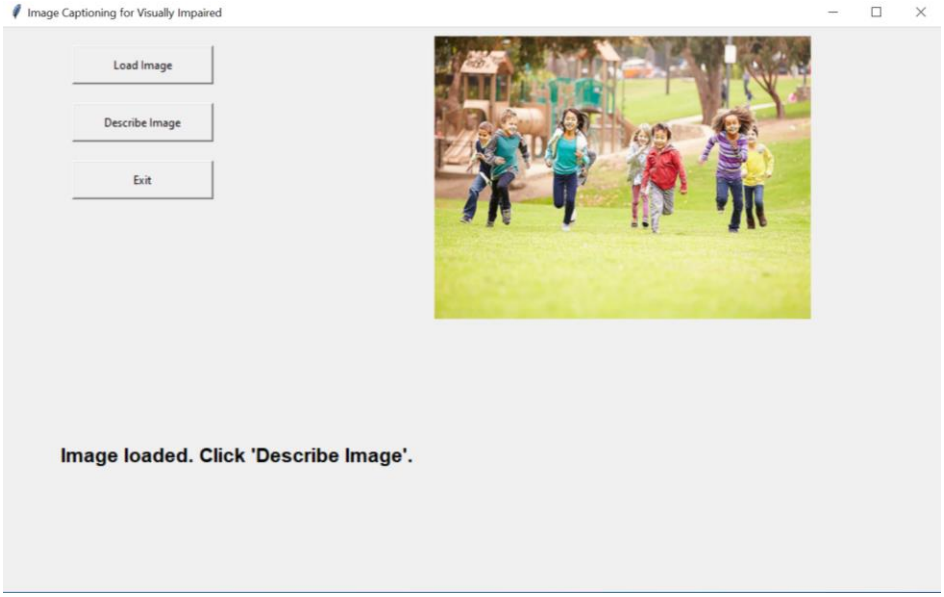


Fig. 5: The GUI Shows the Loaded Image.

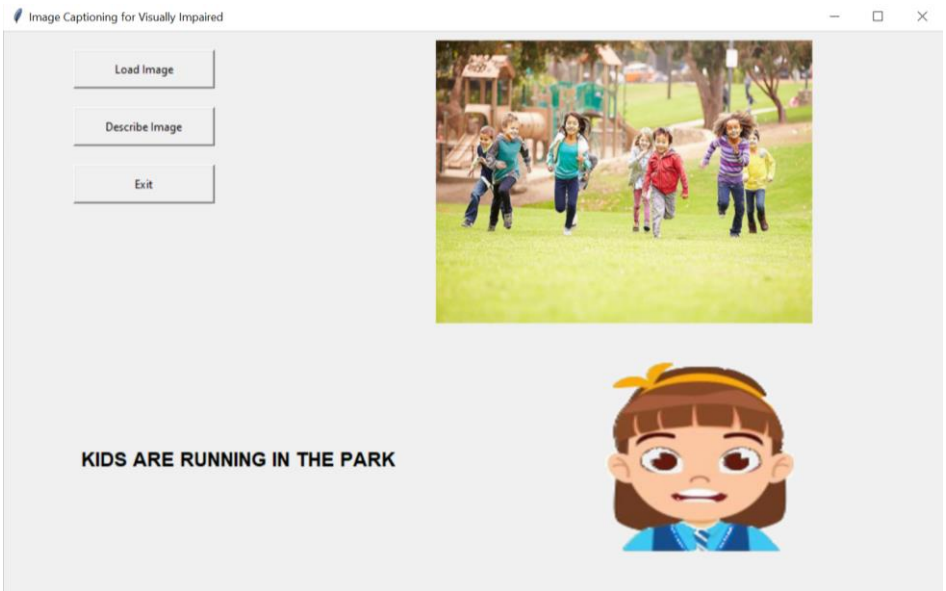


Fig. 6: The GUI Displays the Predicted Caption of the Loaded Image and A 2D Avatar Pronounces It.

Table 5 and Figure 7 show the overall prediction performance of the suggested V2V: ICG-AGAS model. The experimental results clearly show the outstanding effectiveness of the model for all major evaluation metrics. In particular, the V2V: ICG-AGAS framework attained an excellent accuracy of 96.71%, precision of 97.80%, recall of 96.71%, and an F1-score of 97.19% in the image captioning task. These findings confirm the model's capability to produce semantically appropriate and correct captions, demonstrating its robustness in processing various visual inputs. The high precision and recall rates further confirm the reliability of the model in making accurate predictions of meaningful descriptions across different image contexts. Overall, the results emphasize the applicability and effectiveness of the V2V: ICG-AGAS approach as a reliable tool for assistive image understanding, especially for visually impaired users.

Table 5: Result Analysis of V2V: ICG-AGAS Approach with Distinct Measures

Metrics	V2V: ICG-AGAS (%)
Accuracy	96.71
Precision	97.80
Recall	96.71
F1-Score	97.19

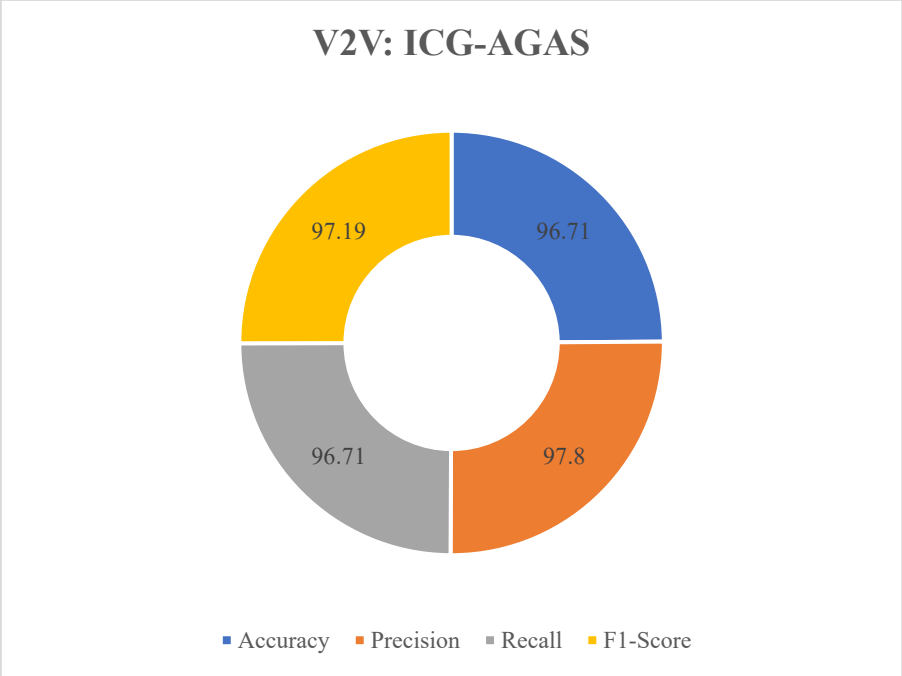


Fig. 7: Result Analysis of Different Metrics.

Table 6 and Figure 8 show the outcomes of a comparative analysis, indicating the enhanced performance of the proposed V2V: ICG-AGAS model over various modern techniques. Based on experimental results, traditional CNN-based and Transformer-based models recorded relatively lower accuracy values of 86.63% and 90.25%, respectively. Deep learning techniques like Generative Adversarial Networks (GAN) and Deep Belief Networks (DBN) showed moderate performance, recording accuracy rates of 92.33% and 93.18%. The V2V: ICG-AGAS framework, on the other hand, beat all the baseline models and recorded a much higher accuracy rate of 96.71%. This extensive comparative study clearly proves that V2V: ICG-AGAS is an extremely strong and efficient solution for image caption prediction. Its higher precision, coupled with its dual-stream feature extraction and avatar-guided assistive interface, represents a fundamental leap forward for the field and establishes a new standard for assistive image understanding technologies.

Table 6: Accuracy Analysis of V2V: ICG-AGAS Model with Existing Approaches

Methods	Accuracy (%)
CNN	86.63
Transformer	90.25
GAN	92.33
DBN	93.18
V2V: ICG-AGAS	96.71

(Note: All baseline models (CNN, Transformer, GAN, DBN) were re-implemented and evaluated on the same 1,600-image dataset for fair comparison.).

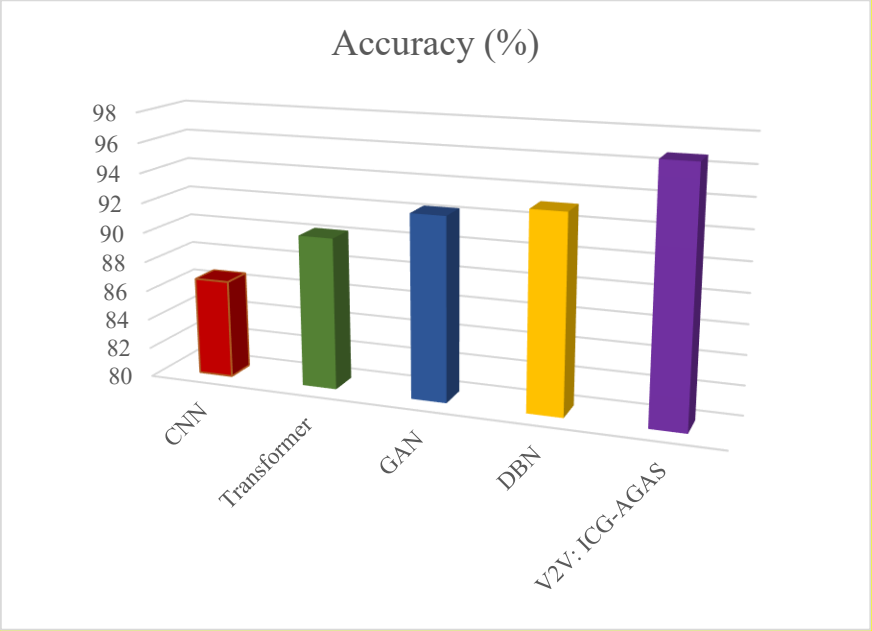


Fig. 8: Accuracy Analysis of V2V: ICG-AGAS Approach with Existing Techniques.

5. Conclusion

This paper proposed an end-to-end and smart assistive system, V2V: ICG-AGAS, to fill the gap between visual input and audio output for the visually impaired by real-time image captioning. The system is innovative due to its twin-stream feature extraction method, which combines semantic-rich RGB data and structural edge features extracted with Sobel filtering. This integration boosts visual representations' discriminative capability, resulting in more contextually appropriate captioning. The decoder unit, which is based on an LSTM framework, efficiently captures the temporal nature of language so that accurate and coherent descriptions can be generated for detailed visual scenes. The training and testing of the system on a dataset spanning 35 categories confirm that it generalizes well to real-world problems, with exceptional performance metrics as evidence of its utility in practice.

In addition to caption generation, the joining of a GUI with a temporally synchronized 2D animated avatar and speech synthesis broadens the system's accessibility, turning static visual interpretation into dynamic interactive access. This multimodal interface enables users to self-access visual information using both textual and audio outputs, which adheres to universal design principles for accessible technology. The suggested solution is not just scalable and lightweight but is also versatile across multiple application domains like public wayfinding, object recognition, and smart assistive devices. Future research will focus on integrating attention mechanisms, multilingual support, and personalization strategies. Additionally, exploring lightweight backbone networks such as MobileNetV3 and EfficientNet-B0 will allow efficient deployment on wearable assistive devices, improving portability and real-time interaction.

Declarations

- Ethics approval and consent to participate

Not applicable. The study does not involve human participants, animals, or sensitive data requiring ethics approval.

- Consent for publication

Not applicable. No identifiable personal data is included in this manuscript.

- Availability of data and material

The dataset used in this study is publicly available and can be accessed through the Kaggle platform. Specific preprocessing steps and code are available upon request from the corresponding author.

- Competing interests

The authors declare no competing interests related to this work.

- Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

- Authors' contributions

S.P. Balamurugan: Conceived the study, performed data preprocessing, experimental analysis, and prepared visualizations.
D. Karthika: Designed the model architecture, wrote the manuscript, and conducted the statistical evaluation.
Both authors read and approved the final manuscript.

- Acknowledgements

The authors gratefully acknowledge the support and guidance provided by their respective institutions throughout the research process.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] World Health Organization, “World Report on Vision 2024,” Geneva: WHO, 2024.
[2] Ruvita Faurina, Anisa Jelita, Arie Vatresia, Indra Agustian, "Image captioning to aid blind and visually impaired outdoor navigation", IAES International Journal of Artificial Intelligence (IJ-AI), vol. 12, no. 3, pp. 1104–1117, 2023. https://doi.org/10.11591/ijai.v12.i3.pp1104-1117.
[3] Lu Yu, Malvina Nikandrou, Jiali Jin, Verena Rieser, "Quality-agnostic Image Captioning to Safely Assist People with Vision Impairment", arXiv preprint arXiv:2304.14623, 2023. https://doi.org/10.24963/ijcai.2023/697
[4] Hiba Ahsan, Daivat Bhatt, Kaivan Shah, Nikita Bhalla, "Multi-Modal Image Captioning for the Visually Impaired", Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 53–60, 2021. https://doi.org/10.18653/v1/2021.naacl-srw.8.
[5] Pritam Langde, Shrinivas Patil, Prachi Langde, "Automating Document Narration: A Deep Learning-Based Speech Captioning System for Visually Impaired Persons", International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 3, pp. 7469–7476, 2023.
[6] Huy Nguyen, Thi Huynh, Nam Tran, Thai Nguyen, "MyUEVision: An Application Generating Image Caption for Assisting Visually Impaired People," Journal of Engineering and Technology, vol. 11, no. 2, pp. 112–125, 2024.
[7] Batyr Arystanbekov, Askat Kuzdeuov, Shakhizat Nurgaliyev, Huseyin Atakan Varol, "Image Captioning for the Visually Impaired and Blind: A Recipe for Low-Resource Languages", Proceedings of the 2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1–4, 2023. https://doi.org/10.1109/EMBC40787.2023.10340575.
[8] Dionysia Danai Brilli, Evangelos Georgaras, Stefania Tsilivaki, Nikos Melanitis, Konstantina Nikita, "AIris: An AI-powered Wearable Assistive Device for the Visually Impaired", arXiv preprint arXiv:2405.07606, 2024.
[9] Sibi C. Sethuraman, Gaurav R. Tadkapally, Saraju P. Mohanty, Gautam Galada, Anitha Subramanian, "MagicEye: An Intelligent Wearable Towards Independent Living of Visually Impaired", arXiv preprint arXiv:2303.13863, 2023.
[10] Fateme Zare, Paniz Sedighi, Mehdi Delrobaei, "A Wearable RFID-Based Navigation System for the Visually Impaired", arXiv preprint arXiv:2303.14792, 2023. https://doi.org/10.36227/techrxiv.22337803.
[11] Taewhan Kim, Soeun Lee, Si-Woo Kim, Dong-Jin Kim, "ViPCap: Retrieval Text-Based Visual Prompts for Lightweight Image Captioning," arXiv preprint arXiv:2412.19289, 2024.
[12] Ning Wang, Wenbin Li, Zhiqiang Tao, Jingkuan Song, "Efficient Image Captioning for Edge Devices," arXiv preprint arXiv:2212.08985, 2022.
[13] Tingyu Qu, Tinne Tuytelaars, Marie-Francine Moens, "Visually-Aware Context Modeling for News Image Captioning," Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 2849–2866, 2024.
[14] Sara Sarto, Luca Cosmo, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, "Towards Retrieval-Augmented Architectures for Image Captioning," arXiv preprint arXiv:2405.13127, 2024. https://doi.org/10.1145/3663667
[15] Shourya Tyagi, Bhupesh Gour, Parveen Rajpoot, "Novel Advance Image Caption Generation Utilizing Vision Transformer and Generative Adversarial Networks," Computers, vol. 13, no. 12, pp. 305–321, 2024. https://doi.org/10.3390/computers13120305.
[16] Qing Zhou, Yixuan Su, Yan Song, "Text-only Synthesis for Image Captioning," arXiv preprint arXiv:2405.18258, 2024.
[17] Yogita Dongare, Ruchi Mantri, Pratiksha Patil, "Deep Neural Networks for Automated Image Captioning to Improve Accessibility for Visually Impaired Users," International Journal of Innovative Science and Advanced Engineering, vol. 11, no. 10, pp. 129–134, 2023.
[18] Fangyu Liu, Yufei Wang, Lei Li, "Recent Advances in Synthesis and Interaction of Speech, Text, and Vision," Electronics, vol. 13, no. 9, pp. 1726–1740, 2024. https://doi.org/10.3390/electronics13091726.
[19] Amirthalingam P., Kavitha D., Lakshmi Priya G., "IICVoiC – An Intelligent Image Captioning and Voice Converter for Visually Impaired," Journal of Electrical Systems, vol. 20, no. 4, pp. 298–309, 2024.
[20] Antonia Karamolegkou, Malvina Nikandrou, Georgios Pantazopoulos, Danae Sanchez Villegas, Phillip Rust, Ruchira Dhar, Daniel Hershcovich, Anders Søgaard, "Evaluating Multimodal Language Models as Visual Assistants for Visually Impaired Users," arXiv preprint arXiv:2503.22610, 2025. https://doi.org/10.18653/v1/2025.acl-long.1260.
[21] Bufang Yang, Lixing He, Kaiwei Liu, Zhenyu Yan, "VIAssist: Adapting Multi-modal Large Language Models for Users with Visual Impairments," arXiv preprint arXiv:2404.02508, 2024. https://doi.org/10.1109/FMSys62467.2024.00010
[22] Md Alif Rahman Ridoy, M Mahmud Hasan, Shovon Bhowmick, "Compressed Image Captioning using CNN-based Encoder-Decoder Framework," arXiv preprint arXiv:2404.18062, 2024.
[23] Caiyue Chen, "Speech Synthesis Technology: Status and Challenges," ITM Web of Conferences, vol. 73, 2025 International Workshop on Advanced Applications of Deep Learning in Image Processing (IWADI 2024), pp. 02006, 2025. https://doi.org/10.1051/itmconf/20257302006.
[24] K. Ravi Teja, Y. Sriman, A. Aneeta Joseph, R. Deepa, "Generation of Image Caption for Visually Challenged People," in Information System Design: Communication Networks and IoT, ISDIA 2024, Lecture Notes in Networks and Systems, vol. 1057, Springer, Singapore, pp. 537–545, 2024. https://doi.org/10.1007/978-981-97-4895-2_45.
[25] https://www.kaggle.com/datasets/aishrules25/automatic-image-captioning-for-visually-impaired.