Hybrid Transformers with Multi-scale Feature Extraction for Vision-Language Tasks: CvT, MaxViT, and CoAtNet

Miranda Surya Prakash, *Member, IAENG*, Madhavi Mallam

***Abstract*—Multi-scale feature extraction remains a critical limitation in classification-focused image encoders like Vision Transformer (ViT) and ResNet, which struggle to capture localized features and global context essential for vision-language tasks. This study adapts hybrid transformers—Convolutional Vision Transformer (CvT), Multi-Axis Vision Transformer (MaxViT), and Convolution and Attention Network (CoAtNet)—originally developed for classification, to assess their effectiveness in Visual Question Answering (VQA) and Image Captioning, benchmarked against ViT. For VQA, nine models integrate these encoders with BERT, RoBERTa, or GPT-2 through a stacked cross-attention module, evaluated on DAQUAR, CLEVR, VizWiz, PathVQA, and SuperCLEVR3D datasets using accuracy, WUPS, and F1 scores. For Image Captioning, four models pair the encoders with a custom LLaMa-3 based decoder and a spaCy frequency tokenizer, assessed on Flickr8k with ROUGE scores. Results show hybrid transformers outperform ViT: CoAtNet excels in open-ended VQA tasks, MaxViT in reasoning-heavy scenarios, and CvT in efficiency-driven cases, while CoAtNet leads in captioning ROUGE scores, followed by MaxViT. This multi-task analysis indicates that hybrid transformers, by extracting hierarchical and spatial features at multiple scales, address weaknesses of ViT in localization and relational reasoning to varying degrees. CoAtNet shows strength across diverse tasks, MaxViT excels in complexity, and CvT offers efficiency. These findings highlight multi-scale approaches as a route to advanced vision-language models, with potential applications in medical imaging and 3D reasoning.**

***Index Terms*— Hybrid Transformers, Visual Question Answering, Image Captioning, Vision and Language**

I. Introduction

VISION and language integration attracts significant attention from researchers, with efforts concentrating on advanced pre-training and Large Language Models (LLMs) like BERT [1], RoBERTa [2] , GPT-2[3], and LLaMa-3[4] to enhance accuracy in tasks such as VQA and image captioning. These LLMs frequently refine textual outputs, yet image encoding in these systems receives less scrutiny. Most studies employ standard architectures like Faster R-CNN, ResNet, VGG, ViT [5], or Swin Transformer.

Manuscript received November 27, 2024; revised May 27, 2025.

Miranda Surya Prakash is a PhD candidate in the Department of Electronics and Communication Engineering, PES Institute of Technology and Management (Affiliated to Visvesvaraya Technological University, Belagavi-590018), Shimogha, Karnataka, India (Phone: +916361711795, e-mail: mirashc23@gmail.com).

Madhavi Mallam is a Professor and Head of Electronics and Communication Engineering, PES Institute of Technology and Management (Affiliated to Visvesvaraya Technological University, Belagavi-590018), Shimogha, Karnataka, India (Phone: +91 9182160989, e-mail: gurumadhu432@gmail.com).

Although effective, these encoders, originally designed for classification or detection, struggle to provide the flexible, multi-scale visual representations demanded by modern vision-language tasks. As datasets evolve to encompass real-world scenarios, medical imaging, and 3D reasoning, traditional architectures often fail to match the capabilities of advanced LLMs, underscoring the need for innovative encoder designs like hybrid transformers.

Traditional feature extraction methods include grid-based, region-based, and patch-based approaches, each offering distinct advantages. Grid-based methods like ResNet divide images into equal cells by preserving spatial structure but they lack deep contextual connections. Region-based techniques like Faster R-CNN prioritize object detection, capturing local details but overlook global scene context. Patch-based methods like ViT divide images into fixed patches and process them using transformers to capture long-range relationships while losing hierarchical structure and fine spatial details. Convolutional Neural Networks (CNNs)[6] like ResNet excel at extracting local features such as edges and textures but remain constrained by fixed receptive fields, limiting their ability to capture long-range relationships vital for vision-language reasoning. ViT enhances scalability and global context, yet its flat, patch-based design omits multi-scale hierarchies and local precision, reducing its effectiveness for tasks requiring detailed localization or relational reasoning.

Hybrid transformers address these limitations by combining convolutional and transformer mechanisms. CvT combines convolutional tokenization with transformer layers that merge local region-based features with global patch-based context in a hierarchical framework. MaxViT applies multi-axis attention which combines grid and block-based processing to capture broad spatial layouts and fine details across scales. CoAtNet integrates convolutional depth with attention-based breadth, producing detailed hierarchical and spatial representations. These designs offer a versatile approach to feature extraction, suited to vision-language demands. Although established in image classification, hybrid transformers remain untested in vision-language modeling, presenting an unexplored area.

This study pursues two objectives:

- To compare CvT, MaxViT and CoAtNet against ViT as image encoders pairing with LLMs as text encoders in VQA and image captioning across datasets including DAQUAR, CLEVR, VizWiz, PathVQA, SuperCLEVR3D, and Flickr8k.
- To evaluate their multi-scale feature extraction capabilities for open-ended and reasoning-based vision-language tasks and to assess how hierarchical and spatial representations improve cross-modal understanding.

To our knowledge, this work is among the first to apply hybrid transformers to vision-language modeling, extending their scope beyond classification. The study aims to improve vision-language integration and provide insights into optimizing image encoders for real-world and specialized applications.

II. Related work

We outline the shift from early convolutional and recurrent models to transformer-based systems, ending with an evaluation of hybrid transformer architectures.

A. Vision and Language Frameworks

Combining visual and textual data forms the basis of VQA and image captioning, leading to a range of architectural developments. Early VQA work, by [7], merges image and question features by concatenation, feeding them into RNNs. The simple concatenation approach is widely used in various studies [8]-[14]. This basic approach faces difficulties in fully blending multimodal features. At the same time, image captioning progresses with the NICG [15], which uses a CNN-RNN encoder-decoder setup to turn static visual features into captions. A similar setup is done by [16], [17].

Later designs shift to transformer-based methods. CNN+Transformer combinations [18], [16], [19], [83] follow by ViT+RNN setups [21] that use patch-based processing for wider context. Current systems, pairing ViT with LLMs [22], [13], [14] , are in vogue. The uniform patching of ViT reduces detail and spatial accuracy, resulting in weaker performance on tasks like PathVQA and image captioning, which require precise descriptions. Our research builds on this by introducing hybrid transformers, which mix convolutional focus with transformer scope, to improve outcomes in both VQA and captioning.

B. Embedding Techniques

Embedding strategies are vital for aligning visual and textual data into a common space, supporting VQA and captioning.

Visual embeddings

Visual embeddings turn images into numerical vectors, including region-based features from object detectors [23]-[25], grid-based CNN outputs [26]-[29] , and patch-based encodings of ViT [13], [22]. Region-based features do well in object-focused VQA but miss the full scene for captioning, while grid-based methods keep spatial order but lack wider connections. This conversion is essential for processing visual and textual information together. Grid, patch, and region-based techniques can overcome the challenges of representing complex visual features in a low-dimensional space. ViT embeddings work well for understanding the whole image but have trouble capturing the small details needed for writing captions or giving exact VQA answers. We use cross-modal alignment to enhance the fusion of visual and textual embeddings, improving coherence in joint representations. These issues point to a need for embeddings that balance local and global aspects, which our hybrid transformer method targets.

Textual Embeddings

Textual embeddings convert text into vectors that match visual features. Early methods like GloVe and Word2Vec [23], [30] gave basic meaning mappings, replaced by models like BERT and RoBERTa [13], [14], which produce detailed embeddings through pretraining. These embeddings improve question understanding in VQA and caption clarity in image description, but their success depends on strong visual encodings. Our study uses these textual embeddings with hybrid visual encoders to boost performance in both tasks.

C. Hybrid Transformers

Vanilla Transformers struggle with vision due to limited spatial bias and difficult optimization. A simple improvement involves introducing a hybrid architecture consisting of a transformer and convolution. Although there is no comprehensive research on the application of hybrid transformers in multimodal tasks, these models show exceptional performance in areas such as object detection and image classification[31].

CvT shows its versatility in Computer Vision tasks. Researchers have used CvT for image retrieval reranking [32] and for object detection[33], [34] achieving improved accuracy and efficiency. It used 3D convolutions to capture spatiotemporal features in video analysis, which helps in lip reading [35]. It is employed to effectively suppress the additive white Gaussian noise (AWGN) in images through a residual learning approach[36]. The integration of CvT with Spiking-CvT (SCvT) improves classification tasks by utilizing the combined capabilities of transformers and spiking neural networks [37].

CvT is appropriate for classifying high-resolution remote sensing images because it can capture details during end-to-end training[38]. It uses a multi-head attention mechanism within the CvTSRR framework to address social relationship detection beyond images[39]. It is used in the medical field to automatically generate chest X-ray reports and predict captions for medical images [40], [41]. Handwritten digit recognition has proven to be a strong suit for CvT [42].

CoAtNet, a hybrid model showed promise in various applications: Paddy crop disease detection [43], classification of musical notes[44], brain tumor[45] , cotton leaf segmentation[46], breast cancer analysis[47], identification of nematodes [48], recognition of parasite eggs [49]and classification of space objects [50].

MaxViT excelled in several applications. MaxViT-UNet [51] focused on medical image segmentation. MaxSR [52] used MaxViT for single-image super-resolution of individual images. MaxViT is used for feature extraction[53], representation learning, and classification of Constant-Q Transform (CQT) spectrograms[54] to enhance speech emotion recognition. The effectiveness of MaxViT in categorizing tomato leaf diseases is shown by [55].

From our search of the literature, it is clear that hybrid transformers show superiority in capturing visual relationships. To achieve a deeper understanding and improve performance on vision and language tasks, we extend this study by combining hybrid transformers with language models.

III. Model Selection

Inspired by hybrid transformers, we choose pre-trained CvT, CoAtNet, and MaxViT as image encoders for vision-language tasks, combining Transformer and CNN strengths for multiscale feature extraction, evaluated against baseline ViT. For VQA, we apply BERT, RoBERTa, and GPT-2 to extract text features, while selecting LLaMa-3-inspired model as decoder for captioning.The key features of the CvT, CoAtNet, and MaxViT architectures are compared in Table 1. This table highlights their unique methodologies and design choices, such as their attention mechanisms, hierarchical structures, local and global feature extraction strategies, and how they handle positional encoding and attention.

A. *CvT* [56]

CvT improves computer vision tasks by integrating convolutions into ViT, blending CNN local context with transformer global attention. It uses a hierarchical structure with Convolutional Token Embeddings (CTE) for spatial down sampling and feature expansion, capturing local information through strided convolutions. Convolutional Transformer Blocks (CTBs) replace standard blocks by applying depth-wise separable convolutions in projections to model local, spatial contexts efficiently. Combining CNN subsampling and local fields with transformer global attention, CvT removes explicit positional embeddings, using inherent spatial structure for efficiency across visual tasks.

B. *CoAtNet* [57]

CoAtNet combines CNN and self-attention to improve generalization and capacity for vision tasks. It uses MBConv blocks with depth-wise convolutions in early stages for local feature extraction and translation equivariance, followed by transformer blocks with relative attention in later stages for global context. Convolutions always precede attention in a sequential multi-stage hierarchy, capturing features at multiple abstraction levels. This design merges convolution efficiency with attention adaptability and deliver strong performance even with limited training data, avoiding ad-hoc hybrid solutions.

C. *MaxViT* [58]

MaxViT boosts efficiency by combining MBConv and multi-axis self-attention (Max-SA) within blocks, stacked vertically in a hierarchy. MBConv with depth-wise convolutions comes before Max-SA to capture local features, followed by block attention for local interactions and grid attention for sparse global mixing with fixed windows and grids. This cuts self-attention complexity from quadratic to linear, advancing beyond Swin Transformer masking requirements. Max-SA and MBConv blocks support local and global interactions across all stages, with MBConv serving as conditional positional encoding by removing separate positional layers, and excelling in reasoning tasks.

IV. Proposed Architecture

A. *Visual Question Answering*

The VQA model integrates a text encoder, an image encoder, fusion module, and a classifier to process image-question-answer triples across multiple configurations. The text encoder uses a pretrained transformer which tokenizes questions and generates a fixed-dimensional embedding from the initial token of the last hidden state.

The fusion module applies a two-layer cross-attention mechanism. In each layer, the image embedding acts as the query, attending to the text embedding as key and value with four attention heads. Mathematically, this stacked fusion mechanism is formalized as $F = \{A_1, A_2\}$with iterative refinement defined as $v_I^{(i)} = A_i\left(v_I^{(i-1)}, v_T\right)$ for $i \epsilon\{1,2\}$. Within each layer, text embeddings are projected via $k_T = W_p v_T + b_p$ followed by multi-head attention $a_i = MHA\left(v_I^{(i-1)}, k_T, k_T\right)$with $H + 4$ heads. The attention output is processed through a residual connection and layer normalization as $v_I^{(i')} = LN\left(v_I^{(i-1)} + a_i\right)$, then refined by a feed-forward network $f_i = W_2 GELU\left(W_1 v_I^{(i')} + b_1\right) + b_2$ with a $4x$ dimension expansion, followed by another residual connection and normalization $v_I^{(i)} = LN\left(v_I^{(i')} + f_i\right)$. Trainable setups employ single-vector fusion by averaging embeddings into one vector, while frozen setups use sequence-based fusion with mean pooling or text embedding repetition.

Nine VQA models are evaluated, grouped by text encoder.. The first configuration uses trainable GPT-2 with 768D paired with CvT with 384D by applying single-vector fusion and a 384D classifier. The second configuration uses frozen GPT-2 with MaxViT producing a 49×1152 sequence, applying sequence-based fusion with mean pooling and a 512D classifier.

TABLE I
COMPARISON OF KEY FEATURES IN CVT, COATNET, AND MAXVIT ARCHITECTURES

Feature	CvT	CoAtNet	MaxViT
Architecture	Hybrid CNNs and ViT	Hybrid CNNs and ViT	Hybrid CNNs and ViT
Local Feature Extraction	Convolutional Token Embedding (CTE)	Depth-wise convolutions in MBConv	MBConv with depth-wise convolutions
Global Context Extraction	Transformer blocks for global attention	Transformer blocks for global attention	Max-SA for global interaction
Hierarchical Structure	CTE layers for down sampling	Sequential stacking of CNN and Transformer	Vertical stacking of Max-SA and MBConv
Positional Encoding	None	Relative attention as CPE	Relative self-attention with Conditional Position Encoding (CPE)
Attention Mechanism	Convolutional Transformer Blocks (CTBs)	Transformer attention with CPE	Grid Attention (sparse global)
Multi-Scale Feature Extraction	Local edges and global region hierarchies	Fine textures and broad spatial layouts	Grid-based details and scene-wide patterns
Strengths	Combines local and global features	Effective with limited data, better generalization	Balanced local and global processing

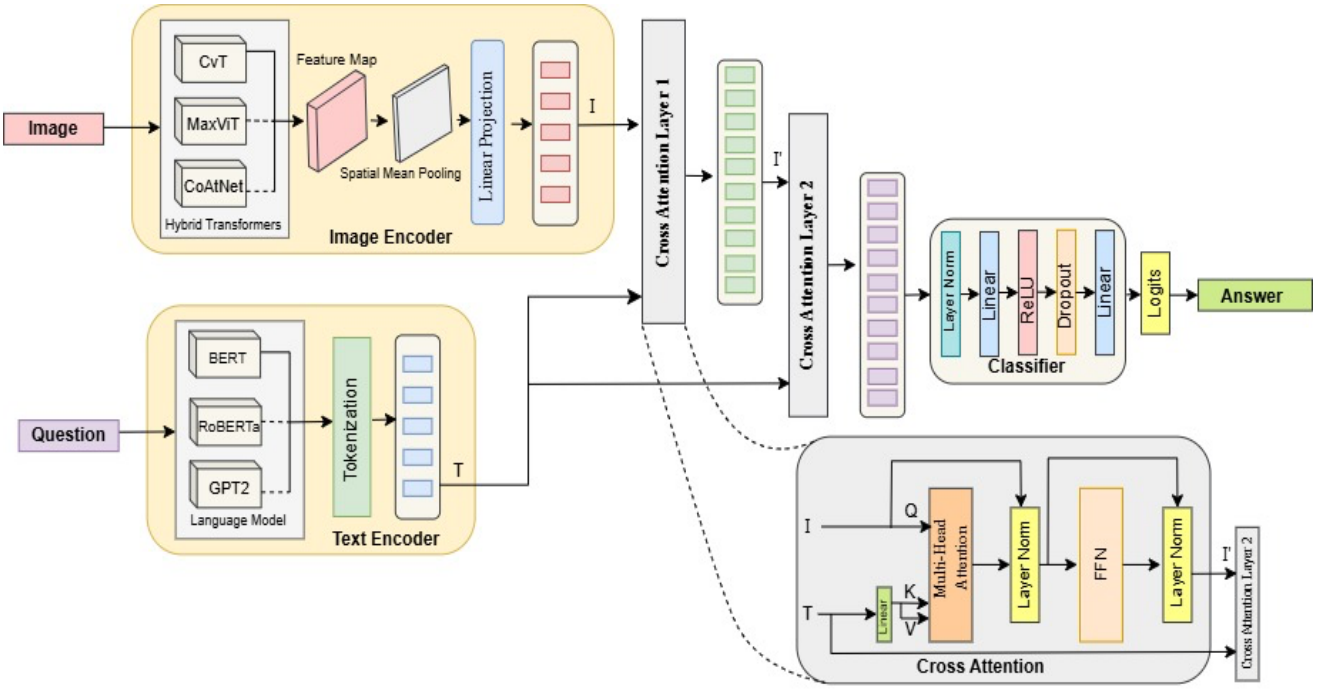

Fig. 1. VQA architecture with pretrained hybrid image and text encoders fused through stacked cross-attention

The third configuration uses frozen GPT-2 with CoAtNet producing a 49×1536 sequence, applying sequence-based fusion with text embedding repeated 49 times and a 512D classifier.

The VQA model fuses image embeddings and text embeddings using two stacked cross-attention layers, where image features iteratively attend to text features. A linear projection aligns text embeddings to the image space and refine multimodal representations. This enhances reasoning for answer prediction by capturing complex inter-modal dependencies through two-layer iterative attention. The classifier maps fused embeddings to the answer space. Fig.1 depicts proposed VQA architecture unified by a stacked cross-attention mechanism.

B. Image Captioning

An image captioning framework as shown in Fig. 2 pairs the same hybrid transformer-based image encoder with a LLaMa-3-inspired Transformer decoder to generate text descriptions from images on the Flickr8k dataset. The image encoder processes 224×224-pixel inputs, with the classification head removed to extract raw feature embeddings. These high-dimensional outputs are projected to a 768D space to align with the input of the decoder.

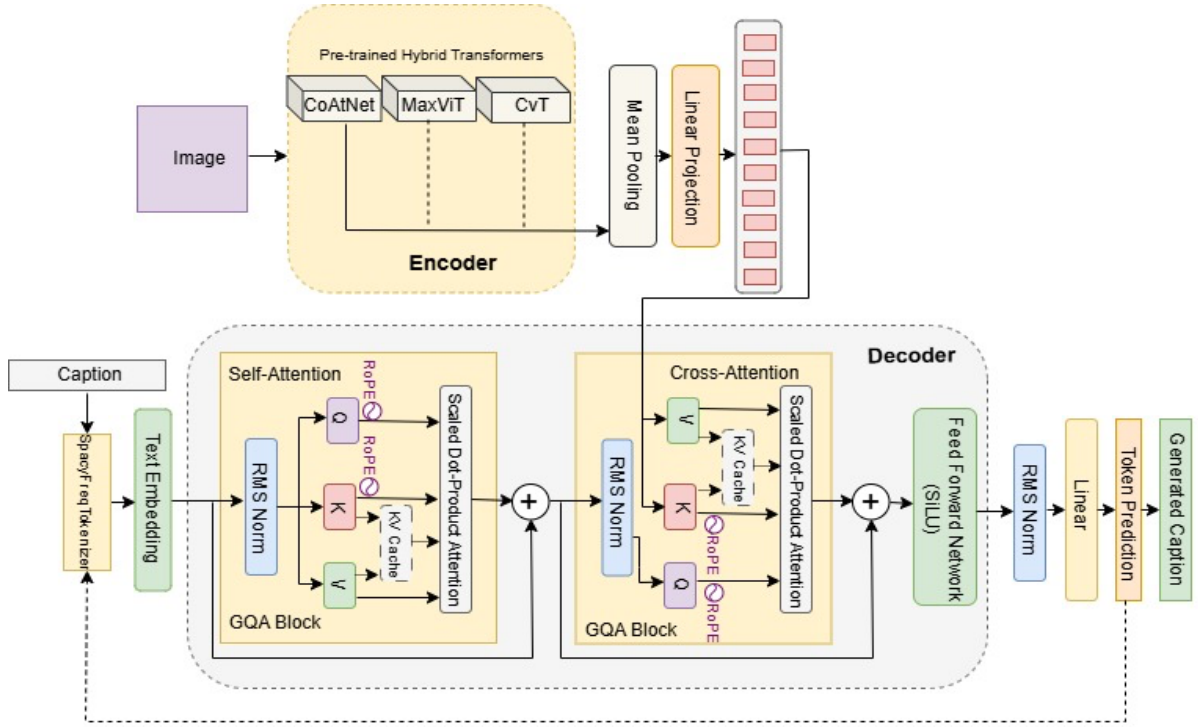

Fig. 2. Proposed Image Captioning framework integrating a pretrained hybrid transformer-based image encoder with a LLaMa-3-inspired Transformer decoder

Encoder weights remain frozen, and training optimizes the projection layer and decoder. The decoder, a single-layer Transformer, maps vocabulary indices to 768D using a token embedding layer. It applies two causal self-attention mechanisms: one processes the token sequence incrementally, and the other attends to image features. A feed-forward network with three linear projections and SiLU activation follows. Grouped Query Attention (GQA) uses 32 query heads and 8 key/value heads (24D per head), with Rotary Positional Embeddings (RoPE) and Root Mean Square Normalization. During inference, a key-value cache improves efficiency. A spaCy-based tokenizer builds a vocabulary with a frequency threshold of 1, includes special tokens, and later pads sequences to 30 tokens.

V. Dataset, Metrics, Settings, Results

A. Datasets

We use three generic datasets: DAQUAR, CLEVR, and VizWiz, and two domain-specific datasets: PathVQA and SuperCLEVR3D for VQA. For captioning we employ a single generic dataset Flickr 8k. Table 2 explains dataset characteristics, including total images, QA pairs, captions, subset sizes, train/test splits, and ratios for VQA and captioning in our comparative study.

B. Evaluation Metrics

To evaluate the performance of our VQA approach, we use three metrics Accuracy, F1 and Wu-Palmer similarity (WUPS)[24] score as they align well with task-specific evaluation standards. They capture different aspects of model effectiveness that a single metric might miss.

Accuracy measures how well the model provides correct answers based on the visual input. Accuracy in VQA systems is calculated as in Equation 1.

$$Accuracy = min(\frac{x}{3}, 1) \quad (1)$$

When evaluating, full credit is awarded if the predicted answer matches at least three of ten annotator responses, with x representing the count of such matches.

WUPS metric evaluates the semantic difference between predicted answers and ground truth answers.

It is computed as shown in Equation 2.

$$WUPS(A,T) = \frac{1}{N_q}\sum_{i=1}^{N_q} \min\left\{\prod_{a\in A_i}\max_{t\in T_i} WUP(a,t), \prod_{t\in T_i}\max_{a\in A_i} WUP(a,t)\right\} \quad (2)$$

Here, A signifies the set of predicted answers, N_q denotes the total number of questions, T indicates the set of target answers, A_i and T_i correspond to the predicted and target sets for the ith question, and $WUP(a,t)$ measures the semantic similarity between predicted answer a and ground truth answer t.

The F1 score is a harmonic mean of precision and recall which evaluates answer accuracy by comparing predicted and ground truth answers. Precision is the fraction of correct elements in the prediction, while recall is the fraction of correct elements relative to the ground truth, typically based on exact or semantic matches. For evaluating caption quality, we employ ROUGE [25] metric which systematically compares generated captions against human-written references.

C. Experimental Settings

For VQA, input data consisting of image-question-answer triples are pre-processed for model compatibility. Questions are tokenized with encoder-specific tokenizers, padded to 24 tokens, and truncated if needed: GPT-2 uses an end-of-sequence token, RoBERTa adds padding dynamically, and BERT includes an explicit padding token. Answers are cleaned by removing spaces and selecting the first entry if multiple are provided, then mapped to numerical labels via a vocabulary index. Images for CvT are resized to 224×224 pixels using bilinear interpolation, normalized to a range of 0 to 1, and transposed to channel-first format; MaxViT and CoAtNet use pretrained library transformations. Pytorch is used and training employs the Hugging Face Trainer API in two setups. The high-throughput setup applies to 7 models with a batch size of 32, 8 data loader workers, AdamW with a learning rate of 1×10^{-4}, a linear warmup over 10% of steps, fp16 precision, and early stopping with a patience of 3 and threshold of 0.001. The low-throughput setup applies to 2 frozen-encoder models with a batch size of 16, 4 workers, and the same optimization settings. Experiments run on a single CUDA-enabled GPU.

For Captioning, the framework operates in PyTorch Lightning with a GPU accelerator and a fixed random seed. The decoder uses a dropout rate of 0.27 and two worker threads with persistent workers. Training runs for only four epochs with a batch size of 64, AdamW with a learning rate of 3.1×10^{-4}, a MultiStepLR scheduler with a decay factor of 0.69 at epoch square root milestones, and gradient clipping at 3.77. ROUGE score assesses caption quality.

TABLE II
OVERVIEW OF DATASETS FOR VQA AND IMAGE CAPTIONING COMPARATIVE ANALYSIS

VQA Datasets								
Dataset	Total Images	Total QA Pairs	Images in Subset	QA Pairs in Subset	Images (Train/Test)	QA Pairs (Train/Test)	QA-to-Image Ratio	Mean QA per Image (Train/Test)
DAQUAR [11]	1,447	12,468	1,447	12,468	1,012 / 435	8,714 / 3,754	8.62	8.61 / 8.63
VizWiz [59]	39,704	32,842	13,816	11,447	9,671 / 4,145	8,033 / 3,414	0.83	0.83 / 0.82
CLEVR [60]	100,000	999,968	2,800	27,999	1,959 / 841	19,589 / 8,410	10.00	10.00 / 10.00
SuperCLEVR3D[61]	30,000	993,864	3,176	28,560	2,223 / 953	19,989 / 8,571	33.13	8.99 / 8.99
PathVQA [62]	4,288	32,632	3,754	28,573	2,627 / 1,127	20,046 / 8,527	7.61	7.63 / 7.57
Image Captioning Datasets								
Dataset	Total Images		Total Captions		Images (Train/Test)	Captions (Train/Test)	Captions to Image Ratio	Mean Captions per Image (Train/Test)
Flickr8k [63]	8091		40,460		6000/1000	30,000/5000	5.00	5.00/5.00

TABLE III
EVALUATION OF HYBRID TRANSFORMERS FOR VQA ON DAQUAR, VIZWIZ, CLEVR, PATHVQA AND SUPERCLEVR3D

Model		Dataset															
Image	Text	DAQUAR			CLEVR			VizWiz			PathVQA			SuperCLEVR3D			Param-eters
		Acc %	WUPS @9	F1	Acc %	WUPS @9	F1	Acc %	WUPS @9	F1	Acc %	WUPS @9	F1	Acc %	WUPS @9	F1	
ViT		17.7	23.2	7.79	43.2	46.5	29.6	65.6	66.5	65.6	45.1	46.6	42.9	29.9	33.8	21.2	210.15M
CvT		23.1	28.0	14.4	42.9	46.3	30.3	63.1	63.9	63.1	46.7	47.2	44.5	29.8	34.3	17.1	133.73M
MaxViT		27.8	32.6	23.2	43.8	47.1	31.8	67.6	68.5	67.6	53.8	54.6	50.3	30.6	34.9	22.6	244.42M
CoAtNet		29.6	34.4	25.8	44.3	47.5	34.2	67.2	67.9	66.8	56.1	57.0	52.0	29.9	34.3	17.7	290.34M
ViT		16.5	22.4	7.33	41.6	45.1	34.5	57.4	58.4	56.7	44.9	45.4	40.3	28.4	33.0	16.3	225.31M
CvT		21.1	26.1	12.6	42.2	45.5	34.5	54.4	55.4	38.8	39.7	40.3	37.9	27.8	31.6	18.8	148.89M
MaxViT		26.2	31.3	20.9	43.5	46.7	34.3	64.1	64.9	63.5	44.1	44.8	39.7	34.1	38.2	32.5	259.57M
CoAtNet		29.5	36.5	27.6	43.5	46.7	32.3	60.0	60.9	58.8	45.2	45.8	40.7	29.4	33.6	20.4	305.51M
ViT		8.13	13.5	2.97	28.6	31.5	21.6	53.8	37.6	54.9	29.6	30.3	19.3	14.8	18.3	4.17	225.10M
CvT		10.7	14.8	2.95	27.7	30.4	20.7	53.8	54.9	37.6	29.6	30.3	19.4	14.5	19.1	5.12	148.68M
MaxViT❄		12.7	17.9	6.70	28.3	31.2	21.9	63.5	64.5	60.4	36.0	37.5	24.0	16.5	20.0	6.11	16.050M
CoAtNet❄		12.6	18.1	6.48	30.4	33.5	15.7	53.8	54.9	37.6	36.9	38.4	23.7	16.3	20.0	6.03	64.047M

VI. EXPERIMENTAL RESULTS AND ANALYSIS

A. *Visual Question Answering*

The evaluation of ViT, CvT, CoAtNet, and MaxViT of Table 3 highlights their distinct architectural designs and performance across multiple datasets. ViT, built entirely on a transformer structure, extracts feature globally by dividing images into 16×16 patches. This approach lacks sensitivity to local details and spatial relationships, as reflected in its lower scores: 17.7% accuracy on DAQUAR and 43.17% on CLEVR. CvT combines convolutional layers with transformers, creating a hierarchical, region-focused feature extraction process. With a 384D embedding and 133.73M parameters when paired with BERT, it balances efficiency and capability, achieving 23.13% accuracy on DAQUAR. MaxViT uses a multi-axis attention mechanism, processing features in a layered grid and block structure, with a 1152D (frozen) or 768D (trainable) embeddings. This design supports strong spatial and relational understanding, shown by 27.83% accuracy on DAQUAR and 53.85% on PathVQA. CoAtNet integrates deep convolutional and attention layers, operating with a 1536D (frozen) or 768D (trainable) embeddings. It delivers the highest performance of 29.65% accuracy on DAQUAR and 56.15% on PathVQA due to its ability to handle both broad and specific tasks effectively. CvT improves on ViT by incorporating regional focus, aiding object detection with moderate success, though its lighter structure restricts it in complex scenarios.

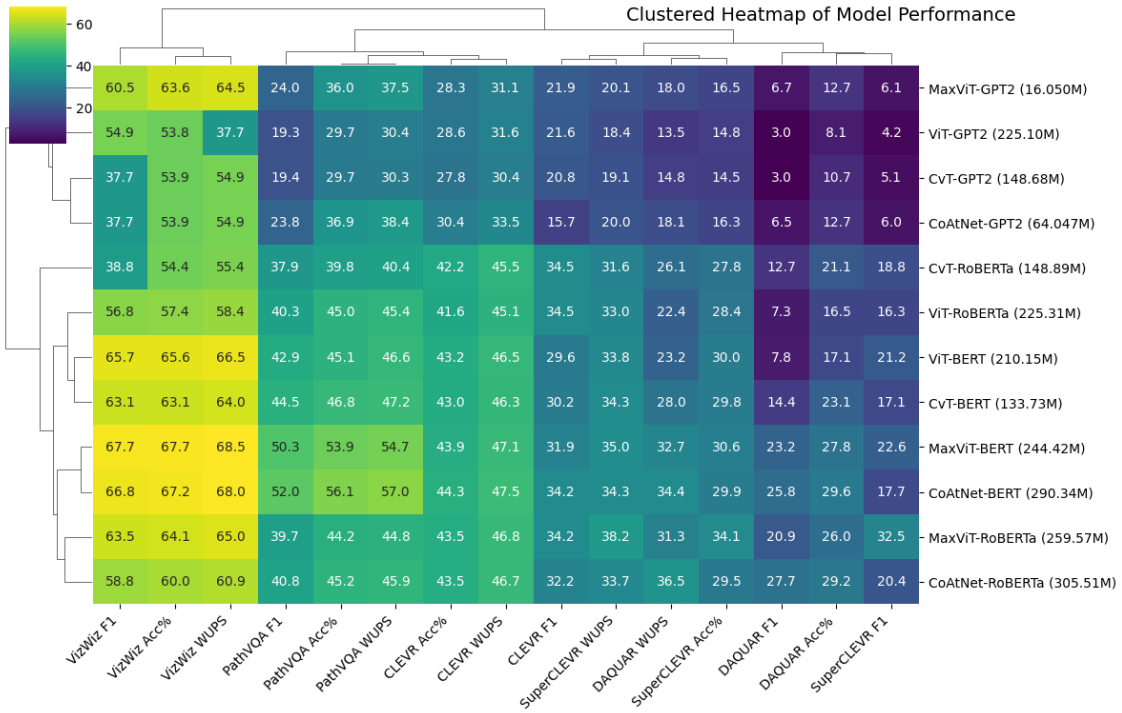

Fig. 3. Heatmap of Model Performance Across Five Datasets with Clustering

MaxViT performs well in spatial and relational tasks, especially in 3D contexts like SuperCLEVR3D achieving 34.10% accuracy with RoBERTa but requires more computational resources. CoAtNet stands out, excelling in crowded scenes in DAQUAR proving its effectiveness on limited data, precise reasoning on CLEVR, and specialized domains like PathVQA, though it demands significant resources and is less tailored for 3D tasks. Performance data show CoAtNet and MaxViT leading, with CoAtNet slightly ahead while CvT offers a practical middle ground and ViT trails due to its lack of layered structure. Figure 3 displays a heatmap of twelve model scores across five datasets with dendrograms indicating clustering of similar models and metrics. This heatmap shows how twelve VQA models perform on all five datasets. Colors show the scores: yellow is high with 68.45, purple is low value with 2.97. We clustered similar models and metrics to group them The rows and columns near each other are alike. The dendrograms on the top and left show how we grouped them: short branches mean high similarity. For example, MaxViT-BERT shines in VizWiz (yellow), while ViT-GPT2 struggles in DAQUAR (purple), and clustering helps us see these trends

Our top-performing models, CoAtNet+BERT and CoAtNet+RoBERTa, achieve an impressive accuracy of 29.65% and a WUPS of 36.55%, outshining state-of-the-art (SOTA) methods like SAN and Att-LSTM. These results highlight their superior precision and semantic grasp, rivalling even Bayesian and DPPnet approaches. They also come remarkably close to MTAN. Table 4 details accuracy, WUPS, and F1 scores for hybrid transformers and benchmarks on DAQUAR, showing top performance.

The CoAtNet + BERT and MaxViT + BERT models, achieving CLEVR accuracies of 44.33% and 43.85%, underperformed compared to SOTA methods such as NS-VQA (99.8%) due to training on only 2.8% of the dataset of about 2,800 images. The limited CLEVR subset restricted the ability of model to capture the complexity of dataset, while the single GPU constrained training efficiency. Their WUPS scores of 47.52% and 47.10% indicate robust semantic comprehension despite these constraints. With access to the full dataset and enhanced computational resources, these models are expected to attain SOTA-level performance. Table 5 provides performance scores of hybrid transformers compared to SOTA on CLEVR, with training limits.

On VizWiz, the proposed MaxViT+BERT model achieves 67.69% accuracy and CoAtNet+BERT reaches 67.18%, yet they exhibit reduced performance relative to BERT-RG, which attains 79.85% accuracy, potentially due to its integration of dual image encoders, ResNet and VGG, for enriched feature extraction.

Our end-to-end trained approach outperforms pretrained VLMs like CLIP, achieving stronger predictive capability and WUPS. This highlights the advantage of our unified, task-adaptive design over static pretrained representations. Table 6 lists the scores of hybrid transformers and SOTA VQA models on VizWiz.

TABLE IV
EVALUATION RESULTS OF HYBRID TRANSFORMERS AND EXISTING STATE OF THE ART VQA MODELS ON DAQUAR

Model	Visual Representation	Feature Relationship	Fusion	Language Representation	Acc.	WUPS	F1
Ask YourNeurons [30]	ResNet-152	Grid based Global	Summation	LSTM,GloVe embedding	21.67	27.99	-
Bayesian Answer [65]	ResNet	Spatial and Global	Bayesian network	Skip-thought Vectors	28.96	34.74	-
DPPnet [28]	VGG-16	Global	Dynamic parameter	Skip-thought vector, GRU	28.98	34.80	-
SAN [66]	VGGNet	Local and spatial	Stacked attention	CNN/LSTM	29.30	35.1	-
Multi-objective [23]	Faster R-CNN	Spatial and Semantic	Element multiplication	Word2Vec, LSTM	29.20	35.34	-
Att-LSTM [9]	Attributes	Global	Concatenation	Doc2Vec, LSTM	29.23	35.37	-
MTAN [67]	ResNet	Local and Spatial	Multi-tier attention	Supervised Embedding GRU	33.53	40.28	-
Deep Walk [12]	VGG ConvNet	Semantic and Spatial	Concatenation	Deep Walk Embedding	40.56	50.09	-
Cross-Attention [13]	ViT	Global	Concatenation	BERT	28.0	33.0	-
Transformer [14]	ViT/Swin	Global	Concatenation	BERT/RoBERTa	-	35.1	-
CoAtNet-RoBERTa	CoAtNet	Hierarchical and Spatial	Stacked Cross Attention	RoBERTa	29.24	36.55	27.67

TABLE V
EVALUATION RESULTS OF HYBRID TRANSFORMERS AND EXISTING STATE OF THE ART VQA MODELS ON CLEVR

Model	Visual Representation	Feature Relationship	Fusion	Language Representation	Acc.	WUPS	F1
LG-Capsule [21]	ResNet-101	Global and Local	Low-rank bilinear pooling	GRU	97.9	-	-
RAMEN [22]	Bottom-up- attention	Spatial	Early and late fusion	GRU	96.52	-	-
Formal logic [23]	Scene graph to facts	Logical reasoning	Rule-based logic inference	BART	99.16	-	-
NS-VQA [24]	Mask R-CNN	Hierarchical	Hierarchical program	LSTM	99.8	-	-
CoAtNet-BERT	CoAtNet	Hierarchical and Spatial	Stacked Cross Attention	BERT	44.33	47.52	34.19
MaxViT-BERT	MaxViT	Hierarchical, grid-based	Stacked Cross Attention	BERT	43.85	47.10	31.88

TABLE VI
EVALUATION RESULTS OF HYBRID TRANSFORMERS AND EXISTING STATE OF THE ART VQA MODELS ON VIZWIZ

Model	Visual Representation	Feature Relationship	Fusion	Language Representation	Acc.	WUPS	F1
BERT-RG[25]	ResNet + VGG	Local and Spatial	SAN	BERT	79.85	-	-
MultiCLIPQA [26]	CLIP	Global	Concatenation	CLIP	61.49	-	-
MaxViT-BERT	MaxViT	Hierarchical, Grid-based	Stacked Cross Attention	BERT	67.69	68.45	67.66
CoAtNet-BERT	CoAtNet	Hierarchical and Spatial	Stacked Cross Attention	BERT	67.18	67.95	66.82

TABLE VII
EVALUATION RESULTS OF HYBRID TRANSFORMERS AND EXISTING STATE OF THE ART VQA MODELS ON PATHVQA

Model	Visual Representation	Feature Relationship	Fusion	Language Representation	Acc.	WUPS	F1
AMAM [27]	2Q and W2Q Attention	Spatial and Semantic	Bilinear Attention Networks	Glove, GRU	50.4	-	-
MMQ [28]	Meta Model	Local and Spatial	BAN, SAN	LSTM	48.8	-	-
Open-ended VQA [29]	ViT	Global	Causal Language Model	GPT2-XL, BioMedLM, BioGPT	63.6	-	-
Medical VQA[30]	ViT	Global	Concatenation	BERT	67.05	-	-
MaxViT-BERT	MaxViT	Hierarchical, grid-based	Stacked Cross Attention	BERT	53.85	54.69	50.29
CoAtNet-BERT	CoAtNet	Hierarchical and Spatial	Stacked Cross Attention	BERT	56.15	57.05	52.02

TABLE VIII
EVALUATION RESULTS OF HYBRID TRANSFORMERS AND EXISTING STATE OF THE ART VQA MODELS ON SUPERCLEVR3D

Model	Visual Representation	Feature Relationship	Fusion	Language Representation	Acc.	WUPS	F1
PO3D-VQA[61]	CNN + Render-and-Compare (Neural Meshes)	Probabilistic	Probabilistic Scene Parsing	LSTM (Seq-to-Seq Program)	75.64	-	-
MaxViT-RoBERTa	MaxViT	Hierarchical, grid-based	Stacked Cross Attention	RoBERTa	34.10	38.21	32.48
MaxViT-BERT	MaxViT	Hierarchical, grid-based	Stacked Cross Attention	BERT	30.61	34.99	22.65

Our research, constrained by limited GPU memory and runtime, permitted only partial training by achieving a maximum accuracy of 56.15% with CoAtNet+BERT as in Table 7. The results are below the SOTA benchmark of 67.05% on Medical VQA due to these computational limitations. Notably, this performance surpasses AMAM with 50.4% and MMQ with 48.8%, demonstrating competitive results despite reduced resources. Enhanced computational resources could further align our outcomes with leading SOTA approaches.

The evaluation of our models trained on the SuperCLEVR-3D dataset is shown in Table 8 reveals a significant performance gap compared to the SOTA PO3D-VQA, which achieves an accuracy of 75.64%, while our top-performing model, MaxViT+RoBERTa, records only 34.10% accuracy, with a WUPS score of 38.21 and an F1 score of 32.48. This study is significant as it examines the effectiveness of transformer-based architectures for 3D-aware VQA which provide critical insights into their applicability to complex spatial reasoning tasks.

Our models are limited by the lack of explicit 3D scene parsing and symbolic reasoning capabilities, essential for addressing the questions in datasets on object parts, 3D poses, and occlusions. Trained on a restricted subset of 2,223 images, our approaches exhibit reduced proficiency in capturing 3D spatial comprehension, depending instead on two-dimensional feature integration. This highlights the need for incorporating 3D-specific methodologies and expanding dataset size to improve performance. Our findings lay a groundwork for future advancements in robust 3D-aware VQA systems.

B. Image Captioning

Our framework, pairing hybrid transformer encoders with the spaCy-LLaMa Tokenizer and a decoder inspired by LLaMa-3, achieves ROUGE scores of 40.13, 40.76, 44.08, and 47.70 on Flickr8k. CoAtNet-LLaMa, scoring 47.70, outperforms many SOTA approaches. Its hybrid convolution-attention encoder extracts richer spatial features, while the LLaMa-inspired causal attention in the decoder ensures coherent captions.

IICG, scoring 61.35, surpasses all our models, likely because the multi-layer decoder outpaces our single-layer design. Simpler tokenization and limited depth of the decoder may restrain performance against such optimized SOTA approaches. Results reflect partial training for only 4 epochs, due to computational constraints, limiting direct comparison to fully optimized SOTA models. Table 9 summarizes the ROUGE scores achieved by the hybrid transformer-based image captioning framework on Flickr8k, detailing the impact of the frozen image encoder and LLaMa-3-inspired decoder configuration. Table 10 shows ROUGE scores of SOTA image captioning models on Flickr8k.

VII. CONCLUSION

This study establishes that hybrid transformers CvT, MaxViT and CoAtNet outperform the baseline ViT in vision-language tasks. Across all datasets hybrid models capitalize hierarchical and spatial feature extraction to exceed uniform patch-based approach of ViT. Our findings emphasize the power of multiscale representations: CoAtNet performs best in open-ended and domain-specific tasks on PathVQA using deep convolutional-attention fusion, MaxViT excels in reasoning scenarios on VizWiz and SuperCLEVR3D with multi-axis attention, and CvT offers efficiency on VizWiz with fewer parameters. In captioning on Flickr8k, hybrid encoders with a LLaMa-3-inspired decoder achieve reliable performance, with CoAtNet.

TABLE IX
EVALUATION OF HYBRID TRANSFORMERS FOR IMAGE CAPTIONING ON FLICKR8K

Model	Encoder	Decoder	ROUGE	Parameter
ViT-LLaMa	ViT	LLaMa-3 Inspired	40.13	103M
CvT-LLaMa	CvT		40.76	42.1 M
MaxViT-LLaMa	MaxViT		44.08	135M
CoAtNet-LLaMa	CoAtNet		47.70	180M

TABLE X
PERFORMANCE EVALUATION OF STATE-OF-THE-ART IMAGE CAPTIONING MODELS USING ROUGE ON THE FLICKR8K

Model	Encoder	Decoder	ROUGE
CNN+LSTM[78]	VGG16 Hybrid Places 1365+LSTM	LSTM	30.76
Hybrid Approach[24]	VGG16+ResNet50 + YOLO	BiGRU and LSTM	31.0
CNN+RNN[79]	ResNet	GRU	36.05
CNN+ SA- Bi-LSTM[80]	ResNet-101	SA-Bi-LSTM	46.1
Attention Based LSTM[81]	Inception V3	Attention Based LSTM	46.41
DenseAtt[25]	Faster R-CNN	Multilayer top-down attention LSTM	46.9
SS-ENSEMBLE[26]	ResNet-152	Attention-Based Multi-Modal LSTM	47.0
CNN + LSTM +ATTN[82]	ResNet-101	LSTM + Attention	47.74
CNN+LSTM [84]	DenseNet-201	LSTM + Attention	51.17
TT-LSTM[27]	XceptionNet	Bidirectional LSTM	51.9
ResNet-Transformer[19]	ResNet-101	Transformer layers	42.44
IICG[20]	DenseNet-121	Transformer	61.35
CNN+Transformer[18]	EfficientNetB0	WordPiece Tokenization, MPNetv2	32.76
ViT- BERT/GPT-2[22]	ViT	BERT/GPT-2	41.3/13.0
CoAtNet-LLaMa3	CoAtNet	LLaMa-3 Inspired	47.70

The success of these models stems from their architectural innovation. This multiscale paradigm not only enhances performance across tasks but also reveals a critical insight: effective vision-language modeling hinges on capturing the full spectrum of visual information, from fine-grained details to scene-wide patterns. Given the partial training and limited dataset scope due to computational constraints, access to better computational resources would likely have enabled our hybrid models to surpass SOTA approaches, fully realizing their potential in both accuracy and generalizability.

Looking ahead, our research opens pathways to further refine vision-language systems by developing novel hybrid models that integrate CapsNet or ResNet with ViT. Such designs could amplify multiscale capabilities which include local textures from ResNet convolutions, hierarchical capsule from Capsule Network, and global context from the ViT. By emphasizing multiscale feature extraction, this study lays a foundation for advancing vision-language integration, with potential impacts in domains like medical imaging, 3D reasoning, and real-world scene understanding, where diverse scales of visual information are paramount.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Proceedings of NAACL-HLT 2019, pages 4171–4186, 2019.

[2] Y. Liu, N.Goyal, J. Du, M. Joshi, D, Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners." [Online]. Available: https://github.com/codelucas/newspaper

[4] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey "The LLaMa 3 Herd of Models," 2024, [Online]. Available: http://arxiv.org/abs/2407.21783

[5] A. Dosovitski, L.Beyer, A. Kolesnikov, D. Weissenborn "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[6] Y. X. Geng, L. Wang, Z. Y. Wang, and Y. G. Wang, "Central Attention Mechanism for Convolutional Neural Networks", IAENG International Journal of Computer Science, vol. 51, pp. 1642-1648, 2024.

[7] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. Zitnick, D.Bhatra and D, Parikh, "VQA: Visual Question Answering," 2015, [Online]. Available: http://arxiv.org/abs/1505.00468

[8] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask Your Neurons: A Neural-Based Approach to Answering Questions about Images," in *2015. IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2015, pp. 1–9. doi: 10.1109/ICCV.2015.9.

[9] Q. Wu, C. Shen, A. van den Hengel, P. Wang, and A. Dick, "Image Captioning and Visual Question Answering Based on Attributes and External Knowledge," 2016, [Online]. Available: http://arxiv.org/abs/1603.02814

[10] S. K. Raman, S. Venkatraman, A. Arbor -Mi, and D. Agnihotri, "Deep Learning Approach to Visual Question Answering",Tech.Rep.

[11] M. Ren, R. Kiros, and R. S. Zemel, "Exploring Models and Data for Image Question Answering.", 2015, [Online]. Available: arXiv:1505.02074

[12] Q. Li, F. Xiao, L. An, X. Long, and X. Sun, "Semantic concept network and deep walk-based visual question answering," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 15, no. 2s, 2019, doi: 10.1145/3300938.

[13] M. Rezapour, "Cross-Attention Based Text-Image Transformer for Visual Question Answering," *Recent Advances in Computer Science and Communications*, vol. 17, 2024, doi: 10.2174/0126662558291150240102111855.

[14] H. Tang, "Vision Question Answering System Based on Roberta and Vit Model," in *2022 International Conference on Image Processing, Computer Vision and Machine Learning, ICICML 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 258–261. doi: 10.1109/ICICML57342.2022.10009711.

[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," 2014, [Online]. Available: http://arxiv.org/abs/1411.4555

[16] Z. Yang, Z.Zhou, C.Wang and L.xu, "Image Guidance Encoder-Decoder Model in Image Captioning and its application", IAENG International Journal of Computer Science, vol 51, pp. 1385-1392, 2024.

[17] M. Li and Z. Zhou, "Vision-Text Bidirectional Collaborative Image Captioning Algorithm", IAENG International Journal of Computer Science, vol 52, pp. 515-523, 2025.

[18] A. Shetty, Y. Kale, Y. Patil, R. Patil, and S. Sharma, "Optimal transformers-based image captioning using beam search," *Multimed Tools Appl*, 2023, doi: 10.1007/s11042-023-17359-6.

[19] A. Patel and A. Varier, "Hyperparameter Analysis for Image Captioning," 2020, [Online]. Available: http://arxiv.org/abs/2006.10923

[20] J. Jansi Rani and B. Kirubagari, "An Intelligent Image Captioning Generator using Multi-Head Attention Transformer," *International Journal of Engineering Trends and Technology*, vol. 69, no. 12, pp. 267–279, 2021, doi: 10.14445/22315381/IJETT-V69I12P232.

[21] S. Elbedwehy, T. Medhat, T. Hamza, and M. F. Alrahmawy, "Efficient Image Captioning Based on Vision Transformer Models," *Computers, Materials and Continua*, vol. 73, no. 1, pp. 1483–1500, 2022, doi: 10.32604/cmc.2022.029313.

[22] W. Man Casca Kwok and W. Man Casca, "Image Captioning ViT/BERT, ViT/GPT." [Online]. Available: https://www.researchgate.net/publication/369655809

[23] Y. Xi, Y. Zhang, S. Ding, and S. Wan, "Visual question answering model based on visual relationship detection," *Signal Process Image Commun*, vol. 80, 2020, doi: 10.1016/j.image.2019.115648.

[24] M. Kaur and H. Kaur, "An Efficient Deep Learning based Hybrid Model for Image Caption Generation." [Online]. Available: www.ijacsa.thesai.org

[25] K. Wang, X. Zhang, F. Wang, T. Y. Wu, and C. M. Chen, "Multilayer Dense Attention Model for Image Caption," *IEEE Access*, vol. 7, pp. 66358–66368, 2019, doi: 10.1109/ACCESS.2019.2917771.

[26] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning Where to See and What to Tell: Image Captioning with Region-Based Attention and Scene-Specific Contexts," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 12, pp. 2321–2334, 2017, doi: 10.1109/TPAMI.2016.2642953.

[27] P. P. Khaing and M. T. Yu, "Two-Tier LSTM Model for Image Caption Generation," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 4, pp. 22–34, 2021, doi: 10.22266/ijies2021.0831.03.

[28] H. Noh, P. Seo and B. Han, "Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction.", 2015, [Online]. Available: http://arxiv.org/abs/1511.05756.

[29] H. Xu and K. Saenko, "Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering," 2015, [Online]. Available: http://arxiv.org/abs/1511.05234

[30] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask Your Neurons: A Deep Learning Approach to Visual Question Answering," 2016, [Online]. Available: http://arxiv.org/abs/1605.02697

[31] Z. Hu, H. Wu and L.He," A Neural Network for EEG Emotion Recognition that Combines CNN and Transformer for Multi-scale Spatial-temporal Feature Extraction", IAENG International Journal of Computer Science, vol 51, pp. 1094-1104, 2024.

[32] C. Zhang, S. Liwicki, and R. Cipolla, "Beyond the CLS Token: Image Reranking using Pretrained Vision Transformers," 2022.

[33] W. Jin, H. Yu, and H. Yu, "CvT-ASSD: Convolutional vision-Transformer Based Attentive Single Shot MultiBox Detector," 2021, [Online]. Available: http://arxiv.org/abs/2110.12364

[34] W. Jin, H. Yu, and X. Luo, "CvT-ASSD: Convolutional vision-Transformer Based Attentive Single Shot MultiBox Detector," in *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, IEEE Computer Society, 2021, pp. 736–744. doi: 10.1109/ICTAI52525.2021.00117.

[35] H. Wang, G. Pu, and T. Chen, "A Lip Reading Method Based on 3D Convolutional Vision Transformer," *IEEE Access*, vol. 10, pp. 77205–77212, 2022, doi: 10.1109/ACCESS.2022.3193231.

[36] A. W. H. Prayuda, H. Prasetyo, and J. M. Guo, "AWGN-based image denoiser using convolutional vision transformer," in *Proceeding - 2021 International Symposium on Electronics and Smart Devices: Intelligent Systems for Present and Future Challenges, ISESD 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ISESD53023.2021.9501567.

[37] S. Talafha, B. Rekabdar, C. Mousas, and C. Ekenna, "Spiking Convolutional Vision Transformer," in *Proceedings - 17th IEEE International Conference on Semantic Computing, ICSC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 225–226. doi: 10.1109/ICSC56153.2023.00044.

[38] A. Chirtu, N. C. Ristea, and A. Radoi, "Convolutional Transformers for Aerial Image Classification: a General to Specific Learning Curve," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 659–662. doi: 10.1109/IGARSS46834.2022.9884876.

[39] S. Shultana, L. Li, and N. Li, "CvTSRR: A Convolutional Vision Transformer Based Method for Social Relation Recognition," in *Proceedings - 2023 International Conference on Communications, Computing and Artificial Intelligence, CCCAI 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 64–68. doi: 10.1109/CCCAI59026.2023.00020.

[40] A. Nicolson, J. Dowling, and B. Koopman, "Improving chest X-ray report generation by leveraging warm starting," *Artif Intell Med*, vol. 144, 2023, doi: 10.1016/j.artmed.2023.102633.

[41] L. Lebrat, A. Nicolson, R. S. Cruz, G. Belous, B. Koopman, and J. Dowling, "CSIRO at ImageCLEFmedical Caption 2022," 2022. [Online]. Available: http://ceur-ws.org

[42] V. Agrawal and J. Jagtap, "Convolutional Vision Transformer for Handwritten Digit Recognition," 2022, doi: 10.21203/rs.3.rs-1984839/v1.

[43] A. Karunanithi, A. S. Singh, and T. Kannapiran, "Enhanced Hybrid Neural Networks (CoAtNet) for Paddy Crops Disease Detection and Classification," *Revue d'Intelligence Artificielle*, vol. 36, no. 5, pp. 671–679, 2022, doi: 10.18280/ria.360503.

[44] C. Setyo and G. P. Kusuma, "Recognition of music symbol notation using convolutional neural network," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 2, pp. 2055–2067, 2024, doi: 10.11591/ijece.v14i2.pp2055-2067.

[45] E. A. Mattar, "Classification of Brain Tumor Using CoAtNet Model," *Am J Biomed Sci Res*, vol. 20, no. 2, pp. 164–171, 2023, doi: 10.34297/AJBSR.2023.20.002687.

[46] J. Yan, T. Yan, W. Ye, X. Lv, P. Gao, and W. Xu, "Cotton leaf segmentation with composite backbone architecture combining convolution and attention," *Front Plant Sci*, vol. 14, 2023, doi: 10.3389/fpls.2023.1111175.

[47] K. Saednia, W. T. Tran, and A. Sadeghi-Naini, "A hierarchical self-attention-guided deep learning framework to predict breast cancer response to chemotherapy using pre-treatment tumor biopsies," *Med Phys*, vol. 50, no. 12, pp. 7852–7864, 2023, doi: 10.1002/mp.16574.

[48] R. A. Lika, N. H. Shabrina, S. Indarti, and R. Maharani, "Transfer Learning using Hybrid Convolution and Attention Model for Nematode Identification in Soil Ecology," *Revue d'Intelligence Artificielle*, vol. 37, no. 4, pp. 945–953, 2023, doi: 10.18280/ria.370415.

[49] N. AlDahoul, H. A. Karim, M. A. Momo, F. I. F. Escobar, V. A. Magallanes, and M. J. T. Tan, "Parasitic egg recognition using convolution and attention network," *Sci Rep*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-41711-3.

[50] N. Aldahoul, H. A. Karim, M. A. Momo, F. I. F. Escobara, and M. J. T. Tan, "Space Object Recognition With Stacking of CoAtNets Using Fusion of RGB and Depth Images," *IEEE Access*, vol. 11, pp. 5089–5109, 2023, doi: 10.1109/ACCESS.2023.3235965.

[51] A. R. Khan and A. Khan, "MaxViT-UNet: Multi-Axis Attention for Medical Image Segmentation," 2023, [Online]. Available: http://arxiv.org/abs/2305.08396

[52] B. Yang and G. Wu, "MaxSR: Image Super-Resolution Using Improved MaxViT," 2023, [Online]. Available: http://arxiv.org/abs/2307.07240

[53] K. L. Ong, C. P. Lee, H. S. Lim, K. M. Lim, and A. Alqahtani, "MaxMViT-MLP: Multiaxis and Multiscale Vision Transformers Fusion Network for Speech Emotion Recognition," *IEEE Access*, vol. 12, pp. 18237–18250, 2024, doi: 10.1109/ACCESS.2024.3360483.

[54] K. L. Ong, C. P. Lee, H. S. Lim, K. M. Lim, and T. Mukaida, "SCQT-MaxViT: Speech Emotion Recognition With Constant-Q Transform and Multi-Axis Vision Transformer," *IEEE Access*, vol. 11, pp. 63081–63091, 2023, doi: 10.1109/ACCESS.2023.3288526.

[55] S. Hossain, M. Tanzim Reza, A. Chakrabarty, and Y. J. Jung, "Aggregating Different Scales of Attention on Feature Variants for Tomato Leaf Disease Diagnosis from Image Data: A Transformer Driven Study," *Sensors*, vol. 23, no. 7, 2023, doi: 10.3390/s23073751.

[56] H. Wu *et al.*, "CvT: Introducing Convolutions to Vision Transformers," 2021, [Online]. Available: http://arxiv.org/abs/2103.15808

[57] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying Convolution and Attention for All Data Sizes," 2021, [Online]. Available: http://arxiv.org/abs/2106.04803

[58] Z. Tu *et al.*, "MaxViT: Multi-Axis Vision Transformer," 2022, [Online]. Available: http://arxiv.org/abs/2204.01697

[59] D. Gurari *et al.*, "VizWiz Grand Challenge: Answering Visual Questions from Blind People," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2018, pp. 3608–3617. doi: 10.1109/CVPR.2018.00380.

[60] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning," 2016, [Online]. Available: http://arxiv.org/abs/1612.06890

[61] X. Wang, W. Ma, Z. Li, A. Kortylewski, and A. Yuille, "3D-Aware Visual Question Answering about Parts, Poses and Occlusions," 2023, [Online]. Available: http://arxiv.org/abs/2310.17914

[62] X. He *et al.*, "Towards Visual Question Answering on Pathology Images." [Online]. Available: https://github.com/UCSD-AI4H/PathVQA

[63] M. Hodosh, P. Young, and J. Hockenmaier, "Framing Image Description as a Ranking Task Data, Models and Evaluation Metrics Extended Abstract.", 2013, doi: 10.1613/jair.3994.

[64] M. Malinowski and M. Fritz, "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input", Advances in neural information processing systems, 2014.

[65] K. Kafle and C. Kanan, "Answer-type prediction for visual question answering," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2016, pp. 4976–4984. doi: 10.1109/CVPR.2016.538.

[66] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked Attention Networks for Image Question Answering," 2015, [Online]. Available: http://arxiv.org/abs/1511.02274

[67] S. Manmadhan and B. C. Kovoor, "Multi-Tier Attention Network using Term-weighted Question Features for Visual Question Answering," *Image Vis Comput*, vol. 115, 2021, doi: 10.1016/j.imavis.2021.104291.

[68] Q. Cao, X. Liang, K. Wang, and L. Lin, "Linguistically Driven Graph Capsule Network for Visual Question Reasoning.",2020, [Online]. Available:10.48550/arXiv.2003.10065

[69] B. Manesha Samarasekara Vitharana Gamage and L. Chern Hong, "Improved RAMEN: Towards Domain Generalization for Visual Question Answering," 2021. [Online]. Available: https://visualqa.org/vqa

[70] M. G. Sethuraman, A. Payani, F. Fekri, and J. C. Kerce, "Visual Question Answering based on Formal Logic.", 2021, [Online]. Available: arXiv:2111.04785

[71] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. B. Tenenbaum, "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding," 2018, [Online]. Available: http://arxiv.org/abs/1810.02338

[72] T. Le, N. Tien Huy and N. Le Minh, "Integrating Transformer into Global and Residual Image Feature Extractor in Visual Question Answering for Blind People",pp.2164-2508, 2020, doi: 10.1109/KSE50997.2020.9287539

[73] Z. Kuang, C. Ma, J. Wang, Z. Zhao, and Y. Zhou, "Research on an End-to-End Assistive Model for the Visually Impaired Population," in *2024 5th International Conference on Computer Vision, Image and Deep Learning, CVIDL 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1018–1022. doi: 10.1109/CVIDL62147.2024.10603642.

[74] H. Pan, S. He, K. Zhang, B. Qu, C. Chen, and K. Shi, "AMAM: An Attention-based Multimodal Alignment Model for Medical Visual Question Answering," *Knowl Based Syst*, vol. 255, 2022, doi: 10.1016/j.knosys.2022.109763.

[75] T. Do, B. X. Nguyen, E. Tjiputra, M. Tran, Q. D. Tran, and A. Nguyen, "Multiple Meta-model Quantifying for Medical Visual Question Answering." [Online]. Available: https://github.com/aioz-ai/MICCAI21_MMQ.

[76] T. Van Sonsbeek, M. M. Derakhshani, I. Najdenkoska, C. G. M. Snoek, and M. Worring, "Open-Ended Medical Visual Question Answering Through Prefix Tuning of Language Models."

[77] Y. Bazi, M. M. Al Rahhal, L. Bashmal, and M. Zuair, "Vision–Language Model for Visual Question Answering in Medical Imagery," *Bioengineering*, vol. 10, no. 3, 2023, doi: 10.3390/bioengineering10030380.

[78] A. Verma, A. K. Yadav, M. Kumar, and D. Yadav, "Automatic image caption generation using deep learning," *Multimed Tools Appl*, vol. 83, no. 2, pp. 5309–5325, 2024, doi: 10.1007/s11042-023-15555-y.

[79] Gaurav and P. Mathur, "An Attention Mechanism and GRU Based Deep Learning Model for Automatic Image Captioning," *International Journal of Engineering Trends and Technology*, vol. 70, no. 3, pp. 302–309, 2022, doi: 10.14445/22315381/IJETT-V70I3P234.

[80] K. Ansari and P. Srivastava, "An efficient automated image caption generation by the encoder decoder model," *Multimed Tools Appl*, 2024, doi: 10.1007/s11042-024-18150-x.

[81] P. Singh, P. Gupta, and H. Jain, "A Comparative Study of Machine Learning Based Image Captioning Models," in *2022 6th International Conference on Trends in Electronics and Informatics, ICOEI 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1555–1560. doi: 10.1109/ICOEI53556.2022.9777153.

[82] J. Wang, W. Li, and S. Zeng, "Study on Image Caption Generation Model with Improved Attention Mechanism," in *Proceedings - 2022 International Conference on Intelligent Transportation, Big Data and Smart City, ICITBS 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 259–264. doi: 10.1109/ICITBS55627.2022.00063.

[83] S. Sheng and Z. Zhou, "Revolutionizing Image Captioning: Integrating Attention Mechanisms with Adaptive Fusion Gates", IAENG International Journal of Computer Science,vol 51,pp.212-221, 2024.

[84] X. Cao, Y. Zhao, and X. Li, "Optimizing image captioning algorithm to facilitate english writing," *Educ Inf Technol (Dordr)*, vol. 29, no. 1, pp. 1033–1055, 2024, doi: 10.1007/s10639-023-12310-6.