Name: Prayash Das
Project Advisor: Professor Hong Man

"I pledge to adhere to the Stevens Graduate Student Code of Academic Integrity, and I have discussed the report with my Project Advisor."

Name: Prayash Das
Signature: *Prayash Das*

ViTCap-R: Vision Transformer-Based Image Captioning With Dual-Encoder Retrieval and Patch Level Attention

Prayash Das
Dept. of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, USA
pdas4@stevens.edu

Professor Hong Man
Dept. of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, USA
hman@stevens.edu

***Abstract*—Image Captioning is a process to understand and get insight of visual scenes and generate meaningful and semantically aligned description of the image. Generating semantically grounded captions requires precise alignment between visual content and language, which remains a challenging problem. Existing approaches treat caption generation and cross-modal retrieval as separate tasks, thereby limiting the model's ability to learn robust visual-semantic associations. To address this, we propose ViTCap-R, a unified framework that integrates a Vision Transformer encoder, a patch-level attention decoder for interpretable captioning, and a contrastive dual-encoder for image-text retrieval. Evaluated on Flickr8k and MS COCO 2014 datasets, our model achieves a BLEU-1 score of 0.5124 and METEOR score of 0.3103 on Flickr8k, with retrieval Recall@10 reaching 1.02 for Image-to-Text and 0.80 for Text-to-Image Retrieval. Qualitative visualizations confirm the model's ability to localize and align language with image regions. These results demonstrate that ViTCap-R effectively unifies captioning and retrieval while improving interpretability.**

***Index Terms*—Image Captioning, Vision Transformers, Patch-Level Attention, Dual-Encoder Retrieval, Contrastive Learning, Multimodal Learning**

I. Introduction

Image captioning refers to the task of automatically generating semantically relevant and contextually accurate textual descriptions for visual scenes. It lies at the intersection of Computer Vision and Natural Language Processing, requiring a system to simultaneously understand visual semantics and articulate them in natural language. Typical works used Convolutional Neural Networks (CNN) for feature extraction and traditional Recurrent Neural Network (RNN) for caption generation. These frameworks have certain limitations such as understanding the global context and often struggling to capture long range dependencies leading to semantically misaligned captions generation.

Recent works have leveraged the use of Transformer architectures to tackle these problems. Transformer-based language models have shown promising results in generating coherent and semantically aligned captions. For example, the *ThaiTC* model showcased an end-to-end transformer-based captioning system that can outperform CNN-RNN baselines by capturing deeper semantic structure [1]. Other works have combined Vision Transformers (ViT) with GPT-style decoders, achieving strong performance on Flickr8k and Flickr30k using BLEU and METEOR metrics [2]. In a work by Biradar et al. [3], they have leveraged the use of hybrid models such as Xception+LSTM, achieving a BLEU-1 score of 0.755 on the Flickr8k dataset.

Alongside caption generation, contrastive learning has gained traction in image-text retrieval. The CLIP-inspired approach in *Image Caption Generation using Contrastive Language Image Pretraining* learns a joint image-text representation space without complex attention modules [4]. Similarly, Liu et al. [5] proposed a two-teacher one-student retrieval framework using knowledge distillation, yielding a 22% boost in mean average precision. However, these methods typically lack integration between retrieval and caption generation, and often do not offer interpretability features. Arman et al. [6] explored transformer-based architectures for medical image captioning by evaluating several multimodal configurations on the ROCO dataset, including ViT+BART, DEiT+MBART, and Swin Transformer+GPT-2. Among these, the ViT+BART setup demonstrated the most stable caption generation with minimal training loss. While their work shows the strength of multimodal transformers in specialized domains, it focuses solely on diagnostic reports and lacks interpretability and cross-modal retrieval mechanisms. Yang et al. [7] introduced BITA, a two-stage vision-language pretraining approach for remote sensing image captioning. Their method combines a frozen ViT-L/14 image encoder and an OPT-2.7B language model, bridged by a novel Interactive Fourier Transformer (IFT) that aligns visual and textual features in both the frequency and semantic domains. Evaluated on RSIC-specific datasets, the model outperformed existing methods in BLEU, METEOR, CIDEr, and SPICE. In another work, Huang et al. [8] proposed a global-local contrastive learning framework for video captioning. While it achieves strong performance on MSVD and MSR-VTT, its temporal modeling approach differs significantly from static image captioning, and it lacks patch-level interpretability and dual-encoder retrieval capabilities.

Eluri et al. [9] developed a hybrid model using Inception-ResNetV2 and Detection Transformers (DETR) for captioning occluded images. Although their system shows strong object-level precision, it focuses on detection metrics and overlooks retrieval and transformer-based decoding. A recent hybrid approach [10] combines CNNs with RNNs and attention for caption generation, evaluating variants including a ViT model and DenseNet201. On a Kaggle dataset, the DenseNet201-based model achieved the best results (BLEU: 0.668, ROUGE-L: 0.746). In another work by Li et al. [11], they introduced a ViT-based retrieval system for pedestrian images using dual-stage attention and external knowledge refinement. Despite strong retrieval metrics, its domain-specific design limits generalization and lacks a captioning component.

In this work, we present **ViTCap-R**, a unified image captioning and retrieval framework that integrates multiple advancements in transformer-based modeling and contrastive learning. The proposed system combines a Vision Transformer encoder for extracting spatially rich image representations with a patch-level attention mechanism in the LSTM decoder to generate interpretable and contextually grounded captions. In parallel, we design a dual-encoder contrastive retrieval module that projects both images and captions into a shared semantic space using cosine similarity. This retrieval module is enhanced through *hard negative mining*, allowing the model to learn finer distinctions between semantically similar pairs.

Our framework is initially prototyped on the Flickr8k dataset and subsequently to the MS COCO 2014 dataset. Caption generation on Flickr8k was performed using both greedy and beam search decoding, while beam search was exclusively used on MS COCO to explore multiple caption hypotheses. The dual encoder is further integrated into the captioning pipeline to re-rank beam candidates based on image-text alignment, improving final caption quality. For evaluation, we report Bilingual Evaluation Understudy (BLEU) and Metric for Evaluation of Translation with Explicit Ordering (METEOR) scores for captioning, Recall@K for retrieval, and provide qualitative insights via attention heatmaps, t-SNE visualizations, and embedding distance histograms.

Main contributions of this work include:

- Development and integration of a unified framework combining captioning and retrieval using ViT, patch-level attention, and dual-encoder contrastive learning.
- Integration of hard negative mining into the dual encoder to improve retrieval accuracy on MS COCO dataset.
- Training and evaluation on both datasets with quantitative metrics as well as qualitative visualization for insights and interpretability.
- Introduction of a reranking mechanism using retrieval alignment scores for improved caption generation.

The remainder of the paper is organized as follows. Section II outlines the problem statement. Section III details the methodology, including overview of the dataset, dataset preprocessing, model architectures, and training strategy. Section IV presents evaluation results and visualizations. Section V provides the conclusion of the work, and the last section provides the references.

II. Problem Statement

The research problem addressed in this work is the automatic generation of semantically accurate and contextually grounded textual descriptions for visual scenes, while simultaneously enabling cross-modal retrieval between image and text modalities. Specifically, given an input image, the goal is to generate a natural language caption that accurately reflects its content, and to learn a shared semantic embedding space where both the image and its caption can be retrieved interchangeably with high precision.

Existing works often treat caption generation and image-text retrieval as individual problems. This hinders the model's potential to acquire robust visual-semantic associations. Currently, convolutional encoders struggle to utilize spatial encoders at the patch level completely while sequential decoders often lack detailed insights in their attention mechanism. Recent advances like attention mechanisms and transformer-based models offer improvements, yet most existing systems treat captioning and retrieval on an individual basis, leading to weak semantic alignment between the two modalities.

This paper addresses the unified problem of image-to-text captioning and image-text retrieval by designing a transformer-based framework that generates captions using patch-level attention and learns a contrastively aligned embedding space for retrieval. The proposed system takes as input a raw image and produces a meaningful caption sequence via a ViT-LSTM attention decoder, and on the other hand, a pair of image and text embeddings that are optimized to be semantically aligned using contrastive learning with hard negative mining.

The objective is to improve both caption generation quality and retrieval accuracy, validated through quantitative evaluation on benchmark datasets, using BLEU, METEOR, and Recall@K metrics.

III. Methodology

A. Design Rationale

The core objective of ViTCap-R framework is to have a unified integrated system capable of both generating semantically rich image captions as well as retrieving matched image-text pairs through shared embeddings. Unlike traditional captioning models that isolate caption generation from retrieval, ViTCap-R integrates both components into a cohesive system. The rationale behind this design lies in leveraging the global contextual understanding of Vision Transformers and enhancing interpretability via patch-level attention. To facilitate semantic alignment between modalities, we incorporate a dual-encoder retrieval module trained using contrastive learning with hard negative mining. This framework leverages ViTCap-R to address limitations in prior CNN-RNN pipelines, specifically in global feature understanding, alignment quality, and interpretability.

B. Dataset Overview

Our framework is initially trained and evaluated on the Flickr8k dataset. This dataset is a collection for sentence-based image description and search, consisting of 8,000 images that are paired with five different captions providing clear descriptions of the salient entities and events [12]. Next, we trained and evaluated our framework on the MS COCO 2014 dataset. It includes over 80,000 training images and 40,000 validation images, each associated with five descriptive captions [13]. The use of MS COCO introduces greater diversity and complexity in visual scenes and language constructs, enabling more rigorous performance evaluation.

C. Data Preprocessing

Captions: All captions are lowercased, stripped of punctuation and numerals, and tokenized using the Treebank tokenizer. Tokens with a frequency less than 10 are removed to reduce noise, resulting in a vocabulary of 1,848 unique words. Special tokens indicating the start and end sequence are added to the beginning and end of each caption. To reduce noise and improve model generalization, only tokens occurring more than ten times in the vocabulary, are kept in the vocabulary resulting in a final vocabulary size of 1848 unique tokens. Each word is mapped to a unique index and reverse-mapping is also implemented for sequence reconstruction. For initializing the embedding matrix of the decoder, we used 50 dimensional pre-trained GloVe embedding forming an embedding matrix of size (1848 x 50) for enhancing semantic representation.

Images: We employed the pre-trained Google Vision Transformer Base (ViT-B/16) model, pretrained on ImageNet-21k. Images are resized to 224×224 and divided into 16×16 patches, yielding 196 tokens per image. The output **[CLS]** token embedding, a 768-dimensional vector, is extracted as the global image representation. The patch tokens are retained for the attention mechanism.

D. Model Architecture

1) Baseline ViT + LSTM Decoder: The baseline architecture uses **Google's Vision Transformer Base Patch16 224 in21k Model** and a sequential language model to perform image captioning. The **[CLS]** token of the ViT model generates the visual features, which are then passed through a dropout layer and then a dense transformation is applied to the visual features to reduce the dimensionality to 256. At the same time, the input caption is processed using an embedding layer mapping words to 50-dimensional vectors. A dropout and a single-layer LSTM with 256 hidden units is incorporated into the processed input caption. The generated outputs of both image and text are merged and joined using an additive operation and the merged output are passed through an additional dense layer to generate a joint representation. A softmax layer then predicts the the next word in the caption sequences from the vocabulary.

2) Patch-Level Attention Decoder: To further enhance caption precision and get better insights, we extend our baseline model by incorporating a patch- level attention mechanism. The image input is represented by a sequence of 196 patch embeddings, each having 768 dimensions, from the pre-trained ViT model. These images are passed through a dropout layer and then a dense layer is incorporated to reduce the dimensionality to 256 per patch for managing dimensionality and encouraging efficient learning. Concurrently, the input caption is embedded and processed through a single-layer LSTM, where the LSTM's final hidden state summarizes the context. The hidden state is repeated across the number of patches concatenating it with the reduced image features to support dynamic focusing on image regions. The joint representation is passed through a two-layer feed- forward network to compute the raw attention scores. These raw attention scores are normalized further via a softmax activation function to produce the attention weights. The computed attention weights serve as a interpretable indicators of which image regions the model focuses on at each timestep. A context vector is computed as a weighted sum of image patches and combined with the LSTM output to generate the next word prediction. Introduction of the attention-augmented decoding extends the baseline model's capability by improving semantic alignment between the generated words and the visual content.

3) Dual-Encoder Contrastive Retrieval: Inspired by the Karpathy-style image-text retrieval framework [14], the third architectural component of ViTCap-R is a contrastive dual-encoder designed to facilitate cross-model retrieval between image and text. This architecture involves the training of separate encoders for both image and text to project the inputs into a shared embedding space. The image encoder in our architecture consists of the baseline Vision Transformer model to extract rich global features from the input image. The **[CLS]** token output is projected to a 256-dimensional embedding space. To capture temporal dependencies in the caption sequence, the text encoder consists of an embedding layer followed by an LSTM network. The final hidden state of the LSTM network is projected into the same 256-dimensional embedding space. The embeddings are then normalized using L2 normalization to lie on a unit hemisphere, allowing effective similarity conversion. To minimize the distance between the matched image-caption pairs, the model has been trained using a symmetric contrastive InfoNCE loss across a mini-batch and to maximize the distance for mismatched image-caption pairs as well. The framework enhances alignment between text and image modalities and also allows effective semantic retrieval. It learns a shared embedding space making it effective to augment image understanding. To improve alignment, we employ hard negative mining by selecting the most semantically similar incorrect pairs within a batch. The dual encoder is also used at inference time to rerank beam search candidates based on their alignment score.

E. Training and Implementation Details

Training for Flickr8k was conducted locally on a standard workstation using Jupyter Notebook. This setup was used for prototyping and model validation during early experimentation phases. For MS COCO, we utilized the Jarvis High

Performance Computing (HPC) cluster at Stevens Institute of Technology, which provides large-scale GPU-enabled parallel computing infrastructure. Training jobs were submitted using SLURM with the configuration: `-G 1 -c32 -p gpu-l40s -N 1`, allocating one NVIDIA GPU and 32 CPU cores per job. All models were implemented in PyTorch within a dedicated virtual environment and trained using mixed-precision (fp16) to accelerate computation and reduce memory overhead. Checkpoints were saved based on the lowest average validation loss across epochs.

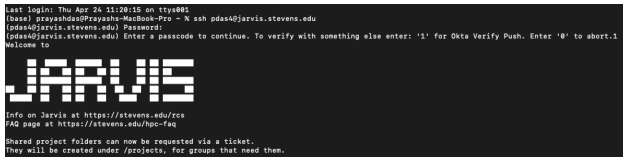

Fig. 1. SSH login to Jarvis HPC cluster at Stevens Institute of Technology.

Fig. 2. SLURM job submission on Jarvis cluster for ViTCap-R training.

Fig. 3. Contrastive dual-encoder training logs showing reduction in average loss over epochs.

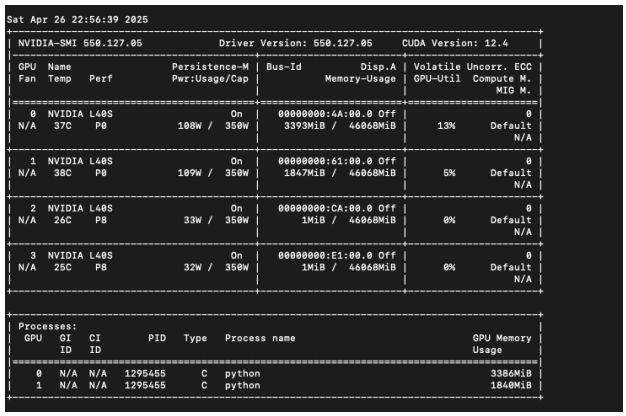

Fig. 4. GPU utilization and memory footprint during ViTCap-R training on NVIDIA L40S GPUs.

1) Training Configurations for Flickr8k: We implemented two core modules on the Flickr8k dataset: an image captioning model and a contrastive dual-encoder retrieval model. The captioning architecture combines a Vision Transformer (ViT) encoder with a patch-level attention decoder built on an LSTM. The retrieval model aligns image and text embeddings in a shared semantic space using contrastive loss.

1) Image Captioning Model: The image encoder uses the `ViT-Base Patch16 224` model, extracting both global and spatial patch embeddings. A dense layer projects the encoder output to 256 dimensions, which is merged with embedded caption tokens passed through a single-layer LSTM consisting of 256 hidden units. Additive attention is applied over 196 spatial patches. The decoder consists of two dense layers consisting of ReLU and Softmax for predicting the next word in the sequence. Word embeddings are initialized with 50-dimensional GloVe vectors , and special tokens are added to each caption. A dropout rate of 0.3 is applied.

The model was trained in two phases, one in 15 epochs and the other additional 10 epochs, gradually increasing the batch size from 3 to 6 due to GPU memory constraints. The loss function used is categorical cross-entropy, and optimization is done using Adam with a learning rate reduced to 0.0001 after initial tuning. The maximum caption length is set to 33 tokens, later extended to 35. The best model was saved for evaluation with loss of 1.4799.

2) Contrastive Dual-Encoder Retrieval Model: The dual-encoder architecture uses the ViT [CLS] token projected to 256 dimensions for image encoding, and a caption encoder comprising an embedding layer, LSTM layer consisting of 256 units, and a linear projection to match the image embedding size. Both the outputs are L2-normalized before similarity scoring.

This model is trained using symmetric InfoNCE contrastive loss. Optimization is done using Adam (learning rate = 10^{-4}) with a StepLR scheduler consisting a step size of 5 and a decay factor of 0.5. Training is performed for 25 epochs with a batch size of 64 to ensure a diverse set of negative samples. Captions are truncated to a maximum length of 33 tokens to ensure consistent input dimensions. The best model was saved with a loss of 3.2258.

2) Training Configurations for MS COCO: For large-scale evaluation and performance scaling, we trained our models on the MS COCO 2014 dataset. The captioning and retrieval modules were adapted to handle higher caption diversity and visual complexity while preserving interpretability and alignment objectives.

1) Image Captioning Model with Patch-Level Attention: We implemented the previous `PatchAttentionDecoder` architecture, using the `ViT-Base Patch16 224 (in21k)` model as the image encoder. The 768-dimensional patch embeddings were projected to 256 dimensions via a linear layer. Textual inputs were embedded with a trainable embedding layer with a dimension of 256 and passed into an LSTM decoder with 256 hidden units. The attention mechanism is additive, computing weights over the 196 patch features using a two-layer network (Linear $\rightarrow$ ReLU $\rightarrow$ Linear). The final output layer concatenates the attention context and LSTM output, then projects to the vocabulary space via a dense layer.

Captions are tokenized using the NLTK Treebank tokenizer, truncated at 35 tokens, and padded. The vocabulary is built using words appearing at least 10 times in a json file. The model was trained for 121 epochs using CrossEntropyLoss and was configured to ignore $<$**pad**$>$ tokens during opti-

mization. The Adam optimizer is used with a learning rate of 10^{-4} and a batch size of 64. Images are then resized to 224×224 and normalized to zero mean and unit variance. The model was trained on CUDA-enabled GPUs and saved to `models/patch_attention_best.pth` based on the lowest validation loss, which was 0.6885.

2) Dual-Encoder Contrastive Retrieval Model: Our baseline retrieval model used ViT image embeddings, projected to 256 dimensions and L2-normalized. The text encoder included a 256-dimensional embedding layer, followed by an LSTM and a linear projection to match the image feature space. Both embeddings were trained to align using the InfoNCE contrastive loss, computed as the cosine similarity between matched and mismatched pairs.

Training was performed over 120 epochs with a learning rate of 10^{-4} using Adam, and the best model with average contrastive loss of 0.1570 was saved to `models/dual_encoder_best.pth`. Vocabulary was loaded from `coco_vocab.pkl`, and preprocessing matched that of the captioning pipeline.

3) Dual-Encoder with Hard Negative Mining (HNM): To improve retrieval precision, we trained an enhanced dual-encoder variant with HNM. During training, for each image, the top two most similar captions based on cosine similarity, were identified, and the second-most similar caption was selected as the hard negative. This forced the model to learn finer-grained distinctions between true and near-matching samples, going beyond surface-level similarity, enhancing precision in retrieval where visually similar as well as semantically distinct pairs exist.

The architecture is aligned with the baseline dual encoder. HNM training was performed for 60 epochs. The model was trained using contrastive loss with a batch size of 32. The best model with average constrastive loss of 0.1457 was saved to `models/dual_encoder_hnm_best.pth`.

3) Training Objectives and Loss Functions: The ViTCap-R framework involves two core training objectives: caption generation using a patch-level attention decoder, and contrastive alignment for image-text retrieval. Each component is optimized with distinct loss functions suited to their learning tasks.

a) 1) Caption Generation Loss (Cross-Entropy):: The image captioning model is trained to maximize the likelihood of the target caption sequence $Y = \{y_1, y_2, \ldots, y_T\}$ given the image representation I. The objective minimizes the categorical cross-entropy loss over the predicted word probabilities:

$$\mathcal{L}_{\text{caption}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{1:t-1}, I) \tag{1}$$

Here, $P(y_t \mid y_{1:t-1}, I)$ is the probability of generating the next word y_t, conditioned on the image and the previous tokens. This loss is optimized using teacher forcing during training.

b) 2) Patch-Level Attention Mechanism:: The attention module computes alignment scores between the decoder's hidden state h_t and the encoder's patch features v_i extracted from the ViT backbone. We employ an additive attention formulation:

$$e_{t,i} = \mathbf{w}^\top \tanh(\mathbf{W}_v v_i + \mathbf{W}_h h_t + b) \tag{2}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{N} \exp(e_{t,j})} \tag{3}$$

$$\mathbf{c}_t = \sum_{i=1}^{N} \alpha_{t,i} v_i \tag{4}$$

Here, $N = 196$ corresponds to the number of spatial patches (14×14). The resulting context vector $\mathbf{c}_t$ is concatenated with the LSTM output to predict the next word token.

c) 3) Image-Text Similarity (Cosine Distance):: For retrieval, both image and text embeddings are L2-normalized and aligned in a shared 256-dimensional semantic space. The similarity between image embedding x and caption embedding y is defined via cosine similarity:

$$\text{sim}(x, y) = \frac{x \cdot y}{\|x\| \, \|y\|} \tag{5}$$

This metric is used for both loss computation and hard negative mining.

d) 4) Contrastive Retrieval Loss (InfoNCE):: The dual-encoder retrieval model is trained using symmetric InfoNCE loss. For each anchor such as image, the goal is to bring the matching positive (caption) closer than all negatives in the batch. The loss for an image–caption pair (i, j) is:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(i, j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(i, k)/\tau)} \tag{6}$$

where τ is the temperature hyperparameter . This formulation is applied in both directions (image-to-text and text-to-image) to optimize retrieval performance.

F. Summary

The ViTCap-R framework introduces several key innovations that enhance both image captioning and cross-modal retrieval by unifying them into a single pipeline, rather than treating them as separate tasks. It leverages Vision Transformers to generate spatially-aware image representations, which help in capturing a more comprehensive global context. Additionally, it incorporates patch-level attention within the caption decoder, enabling finer interpretability and stronger visual grounding. For retrieval, ViTCap-R employs a dual-encoder architecture alongside hard negative mining, which strengthens semantic alignment and overall retrieval accuracy. Furthermore, it integrates contrastive retrieval scores during the caption re-ranking process, fostering a tighter connection between the visual and language components.

IV. Results and Evaluation

Our proposed model has undergone a series of qualitative and quantitative evaluations using the Flickr8k and the MS COCO dataset.

A. Caption Generation Quality

Figures 5 and 6 display sample images along with their generated captions by the **baseline model using greedy decoding**. Captions were generated by a ViT+LSTM architecture trained on the Flickr8k dataset, where the [CLS] token from the ViT encoder was projected and concatenated with embedded caption sequences before being passed to the decoder. During inference, greedy decoding was applied by selecting the highest-probability word at each step. These qualitative examples highlight both the strengths and limitations of the baseline design: while the model can correctly identify prominent objects and actions in simple scenes, it often fails to capture deeper contextual or relational elements in more complex environments. These drawbacks motivates us the introduction of patch-level attention integrated with greedy and beam search decoding for improved visual grounding and better caption fluency.

Fig. 5. Greedy-decoded caption generated by the baseline ViT+LSTM model on a Flickr8k image. The model successfully identifies core objects but lacks contextual richness.

Fig. 6. Qualitative result from the baseline ViT+LSTM model using greedy decoding, illustrating its ability to recognize simple visual cues but limited sentence-level coherence.

Figures 7 and 8 display patch-level attention heatmaps overlaid on a sample image from the Flickr8k dataset, highlighting the regions attended to when generating the words *"little"* and *"arms"* using the **patch-level attention model with greedy decoding**. These heatmaps were obtained by capturing the attention weights from the decoder's alignment mechanism over the 196 patch embeddings extracted by the ViT encoder. The weights were then upsampled to the original image resolution and visualized as heatmaps. These visualizations demonstrate the model's ability to associate individual words with semantically relevant spatial regions, thereby improving both caption grounding and model interpretability.

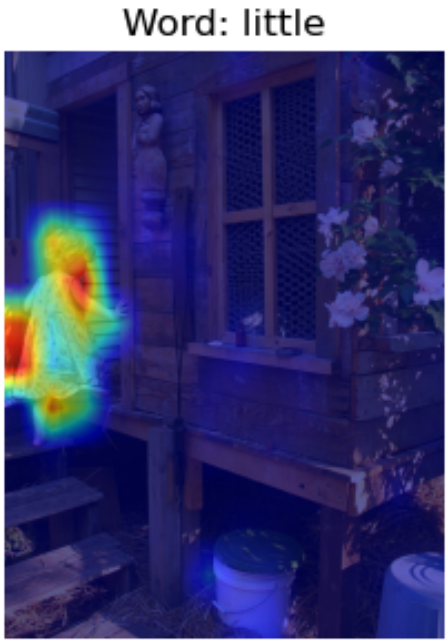

Fig. 7. Attention heatmap for the word *"little"* generated by the patch-level attention model using greedy decoding. The model focuses on the smaller human figure in the image, demonstrating word-level visual grounding.

Figures 9 and 10 illustrate patch-level attention heatmaps generated for the words *"people"* and *"street"* in a sample image from the Flickr8k dataset using the **patch-level attention model with beam search decoding**. To produce these heatmaps, we extracted the 196 image patch embeddings from the ViT encoder and computed attention weights at each decoding step for the target word using the decoder's attention mechanism. The attention weights were then upsampled and overlaid on the input image to visualize on the region of importance, the model focused on during generation of words. These visualizations demonstrate the model's ability to attend to semantically coherent regions—such as human clusters for *people* and ground-level spatial elements for *street*—while benefiting from the improved fluency provided by beam decoding.

1) Qualitative Results: Patch Attention with Beam Search (MS COCO): To evaluate the model's generalization beyond smaller datasets, we further present qualitative results on the MS COCO 2014 dataset using the patch-level attention model with beam search decoding. Figures 11 and 12 display generated captions for two representative samples from the `val2014` split. A beam size of 5 and a maximum decoding length of 35 tokens were used for generation.

These examples highlight the model's ability to anchor linguistic constructs in visually relevant patches across more

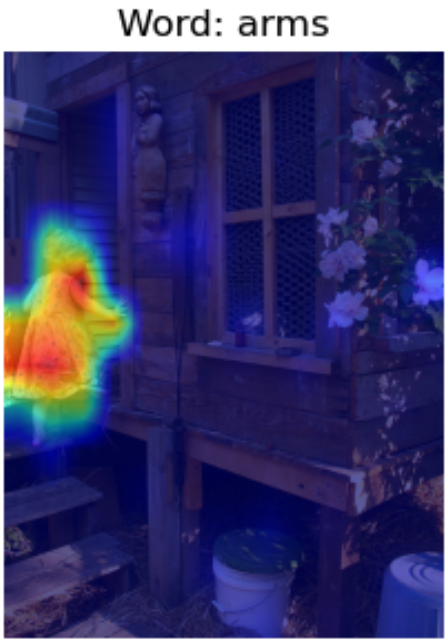

Fig. 8. Attention heatmap for the word *"arms"* generated by the patch-level attention model using greedy decoding. The highlighted patches align with the subject's limbs, demonstrating the model's spatial grounding for body-part nouns.

Fig. 9. Attention heatmap for the word *"people"* generated by the patch-level attention model using beam search decoding. The attention is concentrated on human clusters within the image, highlighting the model's ability to semantically ground plural subject references.

Fig. 10. Attention heatmap for the word *"street"* generated by the patch-level attention model using beam search decoding. The model attends to lower spatial regions corresponding to the road surface, demonstrating effective scene-level localization.

complex scenes. While captions reflect spatial fluency and concept grounding, some mild hallucinations such as object color may still emerge, pointing to the challenges of working with high-diversity datasets like MS COCO.

Fig. 11. Caption generated using patch-level attention and beam search decoding on an MS COCO validation image. The model accurately identifies the dog and its spatial position relative to the road, though the color is slightly hallucinated.

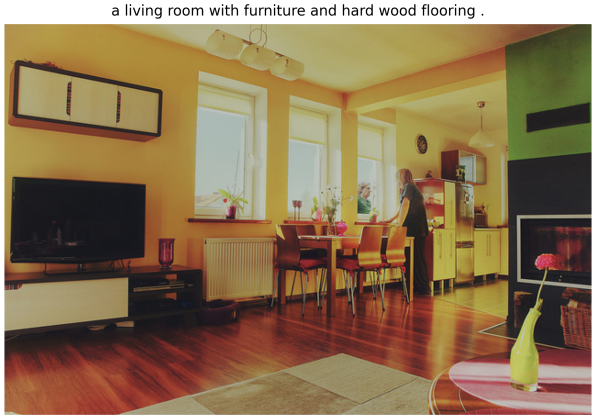

Fig. 12. Generated caption from the patch-level attention model with beam search, describing a well-lit indoor scene. The result demonstrates semantic fluency and strong spatial grounding.

2) Impact of Dual-Encoder Reranking on Caption Selection: To further enhance caption quality, we integrated a dual-encoder reranking module that reorders beam search outputs based on image–caption similarity in a learned joint embedding space. Figures 13 and 14 show reranked captions for the same MS COCO images as in the previous subsection.

In the first example (Figure 13), the model originally described the scene as *"a large brown dog laying on the side of the road"*. After reranking, the caption became *"a dog laying*

on the ground with a shoe on the sidewalk". This revised caption reflects a stronger alignment with typical urban settings seen during training, even introducing a hallucinated object (shoe) likely due to high cosine similarity with visually similar captions. The change from "road" to "sidewalk" improves context specificity, illustrating how the reranker emphasizes semantic plausibility over low-level fidelity.

In the second example (Figure 14), the original caption *"a living room with furniture and hard wood flooring"* is reranked to *"a spacious living room with furniture and windows"*. This version incorporates more visually grounded details such as windows and conveys a broader spatial layout, indicating that the dual encoder favors semantically richer and more descriptive sentences.

Overall, dual-encoder reranking helps refine caption outputs by promoting candidates that are globally consistent with image features in the learned semantic space. While this often improves fluency and relevance, it may occasionally introduce hallucinated details, highlighting a trade-off between semantic alignment and factual precision.

Fig. 13. Caption after reranking: *"a dog laying on the ground with a shoe on the sidewalk."* Compared to the original, this output is semantically aligned with urban context but introduces a hallucinated object.

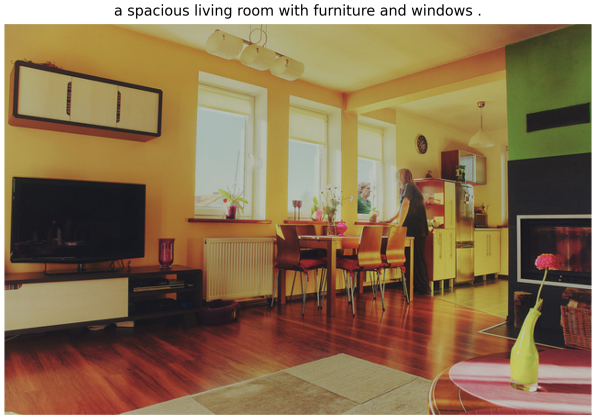

Fig. 14. Caption after reranking: *"a spacious living room with furniture and windows."* This version improves spatial fluency and includes grounded visual details.

B. Quantitative Evaluation

We evaluated the ViTCap-R framework across two primary tasks: image caption generation and cross-modal image-text retrieval. Experiments were conducted on both the Flickr8k and MS COCO 2014 datasets using established metrics including BLEU, METEOR, and Recall@K.

1) Caption Generation: Flickr8k Results: **Baseline Model (ViT + LSTM, Greedy Search):** Using greedy decoding over a subset of 100 validation images, the baseline model achieved BLEU-1 of 0.5124, BLEU-2 of 0.2901, BLEU-3 of 0.1180, and BLEU-4 of 0.0376. On a larger evaluation set of 1,000 images, the METEOR score reached 0.1568. These results reflect the expected behavior of a baseline model using greedy decoding. The high BLEU-1 indicates the model's ability to detect salient objects and keywords, but the substantial drop in BLEU-4 suggests a lack of coherence in multi-word phrase generation. Greedy decoding contributes to this decline by committing early to high-probability tokens without exploring alternatives, leading to repetitive or syntactically incomplete outputs. The moderate METEOR score supports this observation, showing partial semantic overlap but limited fluency and contextual richness. These outcomes highlight the limitations of the baseline architecture and decoding strategy in capturing complex linguistic structures.

Patch-Level Attention Model: Flickr8k Results The addition of patch-level attention improved semantic focus and interpretability. With greedy search decoding, the model achieved a BLEU average of 0.0200 and METEOR of 0.1568. Beam search decoding significantly enhanced fluency, yielding a BLEU average of 0.1039 and a METEOR score of 0.3103. While the introduction of patch-level attention enhances spatial interpretability, its effectiveness under greedy decoding remains limited. The low BLEU average of 0.0200 suggests that despite improved focus on visual regions, greedy decoding fails to construct coherent or diverse output sequences. However, when combined with beam search, the model significantly improves both syntactic fluency and semantic richness. The BLEU average rises to 0.1039 and the METEOR score nearly doubles to 0.3103, indicating that the model not only selects more relevant tokens but also forms more natural language constructions. These results demonstrate that patch-level attention provides the necessary visual grounding, but decoding strategy plays a critical role in unlocking its full potential for caption generation.

2) Cross-Modal Retrieval: Flickr8k Results: **Contrastive Dual-Encoder Retrieval:** The dual-encoder model was evaluated using Recall@K and median rank metrics.

- **Image-to-Text Retrieval:** Recall@1 = 0.07, Recall@5 = 0.60, Recall@10 = 1.02; Median Rank = 702

- **Text-to-Image Retrieval:** Recall@1 = 0.05, Recall@5 = 0.37, Recall@10 = 0.80; Median Rank = 775

The dual-encoder model achieves moderate performance on the Flickr8k dataset, with Recall@10 values of 1.02 for Image-to-Text and 0.80 for Text-to-Image retrieval. These scores indicate that the model is generally capable of semantically aligning image and caption embeddings within a top-10 retrieval window. However, the lower Recall@1 (0.07 for Image-to-Text and 0.05 for Text-to-Image) and high median ranks suggest limited precision at the top of the retrieval list. This can be attributed to the relatively small and semantically overlapping nature of the Flickr8k dataset, as well as the model's reliance on local batch-level contrastive signals without broader hard negative exposure. Despite this, the dual-encoder successfully demonstrates coarse semantic alignment, validating its role as a retrieval backbone within the ViTCap-R framework.

3) Caption Generation with Patch-Level Attention Model: MS COCO Results: The MS COCO evaluation used 100 randomly sampled images from the validation set. Beam search decoding was used with the patch-level attention model. With this setup, the model achieved a BLEU-1 score of 0.0037, BLEU-2 score of 0.0010, BLEU-3 of 0.0007, and a BLEU-4 score of 0.0006. A METEOR score of 0.0156 was also achieved with this framework. The BLEU and METEOR scores on the MS COCO dataset reflect the increased difficulty and variability present in this benchmark compared to Flickr8k. Despite using patch-level attention and beam search decoding, the model achieves a BLEU-1 score of 0.0037 and a METEOR score of 0.0156. This can be attributed to the greater diversity and complexity of scenes in MS COCO, which demands more advanced semantic reasoning and syntactic control than the current architecture provides. Additionally, the evaluation was conducted on a limited sample of 100 validation images, which may not sufficiently capture the model's generalization capability. These results highlight the challenges of scaling caption generation models from constrained datasets to broader, real-world settings without additional optimization or model capacity.

4) Retrieval Performance with and without Hard Negative Mining (MS COCO): We evaluated the contrastive dual-encoder retrieval model on the MS COCO `val2014` set, both before and after introducing Hard Negative Mining (HNM). The results are reported in terms of Recall@K for Image-to-Text (I2T) and Text-to-Image (T2I) retrieval.

Before Hard Negative Mining:

- **Image-to-Text (I2T):** Recall@1 = 0.0044, Recall@5 = 0.0198, Recall@10 = 0.0357
- **Text-to-Image (T2I):** Recall@1 = 0.0044, Recall@5 = 0.0182, Recall@10 = 0.0347

After Hard Negative Mining:

- **Image-to-Text (I2T):** Recall@1 = 0.0031, Recall@5 = 0.0195, Recall@10 = 0.0385
- **Text-to-Image (T2I):** Recall@1 = 0.0035, Recall@5 = 0.0146, Recall@10 = 0.0274

The retrieval performance on the MS COCO validation set reveals key insights before and after the introduction of Hard Negative Mining (HNM). Prior to HNM, the dual-encoder model achieves modest retrieval scores, with Recall@10 reaching 0.0357 for Image-to-Text and 0.0347 Text-to-Image retrieval. These results reflect the model's ability to broadly align matching image-caption pairs, but its discriminative power remains limited due to the lack of challenging negative examples during training. Without hard negatives, the model primarily learns to separate easy mismatches, which leads to weaker fine-grained distinction between closely related but incorrect captions or images—especially in a high-overlap dataset like MS COCO.

After applying HNM, Recall@10 improves slightly for Image-to-Text retrieval (from 0.0357 to 0.0385), suggesting better separation at larger retrieval depths. However, Recall@1 and Recall@5 decline slightly, indicating that the model becomes stricter in its margin enforcement—occasionally penalizing semantically similar near-matches. This trade-off illustrates the delicate balance between enhancing negative discrimination and maintaining tolerance for contextually close positives. In highly diverse datasets like MS COCO, where multiple captions or images may plausibly describe the same content, overly rigid margins can degrade top-rank precision. These effects are further analyzed using t-SNE and cosine similarity visualizations to interpret how HNM alters the embedding space.

C. Failure Cases and Qualitative Limitations

Despite overall retrieval alignment, certain failure cases reveal limitations in semantic discrimination. Figures 15 and 16 show examples where the dual-encoder model retrieves contextually mismatched captions, often due to weak global grounding and visually misleading distractors.

D. Embedding Space Visualizations and Retrieval Diagnostics

To assess the quality of semantic alignment learned by the contrastive dual-encoder, we employed multiple visualization techniques across both datasets. These include Principal Component Analysis (PCA) for Flickr8k and t-distributed Stochastic Neighbor Embedding (t-SNE) projection along with cosine similarity histograms for MS COCO after Hard Negative Mining.

The **image-text embedding space** is visualized through Figure 17, learned by the dual-encoder model using Principal Component Analysis (PCA). To generate this plot, we extracted 256-dimensional embeddings from the ViT-based image encoder and the LSTM-based text encoder after contrastive training on the Flickr8k dataset. Both image and caption embeddings were L2-normalized prior to dimensionality reduction. PCA was then applied to reduce the high-dimensional representations to two principal components. The resulting 2D embeddings were visualized with image embeddings shown in blue and caption embeddings in green. This figure highlights the semantic alignment achieved between visual and textual modalities, as similar structures appear in the shared space

Fig. 15. Failure case from MS COCO where the model retrieved an incorrect caption for a street sign image. The true caption (green) accurately describes the image, while the retrieved hard negative (red) is semantically unrelated. This highlights a failure in semantic discrimination despite visual cues.

Fig. 16. Failure case showing a kite-flying scene where the model retrieved a mismatched kitchen-related caption. Such errors suggest difficulty in capturing global scene context in complex outdoor environments.

despite the distinct modalities. Such alignment is critical for effective cross-modal retrieval, and this projection qualitatively supports the success of our contrastive training objective.

Figure 18 presents a t-SNE projection of the dual-encoder image-text embedding space after training with Hard Negative Mining. To generate this visualization, we first extracted 256-dimensional L2-normalized embeddings from the ViT-based

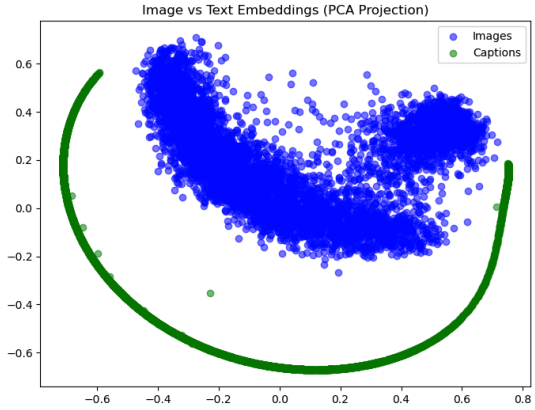

Fig. 17. PCA projection of the image-text embedding space learned by the dual-encoder model on the Flickr8k validation set. Image embeddings (blue) and caption embeddings (green) form semantically coherent clusters in the reduced 2D space, demonstrating the effectiveness of contrastive training in aligning multimodal representations for retrieval tasks.

image encoder and LSTM-based text encoder on the first 300 image-caption pairs from the MS COCO validation set. These embeddings were then concatenated and projected into a two-dimensional space using t-SNE with a perplexity of 30 and fixed random seed. Each blue point represents an image embedding, and each red point denotes a caption embedding. Gray lines connect true image-caption pairs, highlighting the alignment learned by the model. The dense web of connections reveals both aligned pairs as well as semantically challenging cases.

Notably, the figure shows relatively tight clusters of captions and images forming distinct modality regions, yet with consistent cross-modal pairings appearing close in the projected space. This distribution reflects the dual-encoder's ability to learn a shared semantic space where aligned image-caption pairs are close, while hard negatives (e.g., visually similar but semantically distinct samples) are pushed apart. However, overlapping clusters and nearby false matches also underscore the complexity of MS COCO and the delicate trade-offs introduced by HNM. The visual spread affirms that while HNM sharpens inter-modal discrimination, it requires careful calibration to prevent over-separation of semantically adjacent concepts.

To further quantify semantic separation, Figure 19 presents the cosine similarity distributions for positive and hard negative pairs.

V. Conclusion

This paper introduced ViTCap-R, a unified and extensible framework for image captioning and cross-modal retrieval that integrates Vision Transformers, patch-level attention, and dual-encoder contrastive learning. Building upon a ViT+LSTM baseline, we enhanced semantic grounding and interpretability by incorporating spatial attention over visual patches and improved language fluency through beam search decoding. A dual-encoder retrieval module, trained with symmetric InfoNCE loss and hard negative mining, aligned image and

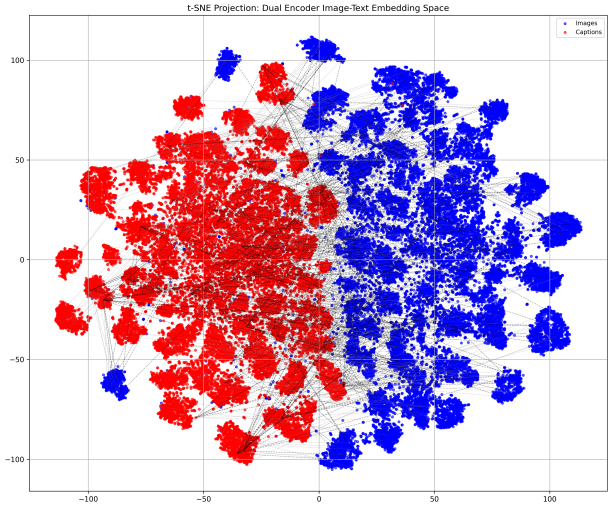

Fig. 18. t-SNE projection of the image-text embedding space learned by the dual-encoder model after training with Hard Negative Mining, computed on the first 300 image-caption pairs from the MS COCO dataset. Each blue point denotes an image embedding, and each red point represents a caption embedding. Dotted gray lines connect true image-caption pairs. This visualization highlights the model's ability to align semantically related modalities while maintaining structural separability, a key property for effective cross-modal retrieval and contrastive learning.

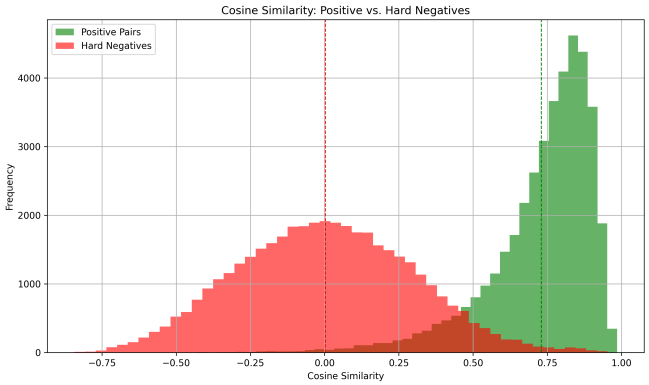

Fig. 19. Histogram of cosine similarities between image-text pairs from the MS COCO dataset, comparing true positive pairs (green) with hard negative samples (red). Hard negatives were mined based on high similarity yet incorrect alignment. Despite overlap, the model maintains clear separation, with positive pairs peaking near 0.8 and hard negatives clustering around zero. This distribution supports the model's ability to effectively distinguish semantically aligned pairs, even in the presence of challenging negatives.

text embeddings within a shared semantic space, enabling reranking of generated captions and supporting retrieval tasks. Experimental results on Flickr8k and MS COCO validate the effectiveness of our approach across BLEU, METEOR, and Recall@K metrics, while qualitative visualizations—such as attention heatmaps, PCA projections, and failure cases—offer interpretability and diagnostic insights. Overall, ViTCap-R demonstrates a scalable path toward explainable multimodal systems and provides a solid foundation for future research in unified vision-language understanding.

REFERENCES

[1] T. Jaknamon and S. Marukatat, "ThaiTC: Thai Transformer-based Image Captioning," in *Proc. 17th Int. Joint Symp. Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2022, pp. 1–4.

[2] S. Mishra, S. Seth, S. Jain, V. Pant, J. Parikh, R. Jain, and S. M. N. Islam, "Image Caption Generation using Vision Transformer and GPT Architecture," in *Proc. 2nd Int. Conf. Advancement in Computation & Computer Technologies (InCACCT)*, 2024, pp. 1–6.

[3] V. G. Biradar, M. G, S. Agarwal, S. K. Singh, and R. U. Bharadwaj, "Leveraging Deep Learning Model for Image Caption Generation for Scenes Description," in *Proc. Int. Conf. Evolutionary Algorithms and Soft Computing Techniques (EASCT)*, 2023, pp. 1–5.

[4] G. Bharathi Mohan, R. Harigaran, P. Sri Varshan, R. Srimani, R. Prasanna Kumar, and R. Elakkiya, "Image Caption Generation using Contrastive Language Image Pretraining," in *Proc. 15th Int. Conf. Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–5.

[5] J. Liu, M. Yang, C. Li, and R. Xu, "Improving Cross-Modal Image-Text Retrieval With Teacher-Student Learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3242–3253, 2021.

[6] M. Arman, M. K. Jahan, A. F. H. Dhrubo, M. M. Rhaman, S. B. Z. Choya, D. M. Dohan, M. A. U. Islam Sajid, and M. G. R. Alam, "Optimizing Multimodal Transformers for Medical Image Captioning: Enhancing Automated Descriptions via AI Systems," in *Proc. 6th IEEE Int. Conf. Image Process., Appl. and Syst. (IPAS)*, 2025, vol. CFP2540Z-ART, pp. 1–5.

[7] C. Yang, Z. Li, and L. Zhang, "Bootstrapping Interactive Image–Text Alignment for Remote Sensing Image Captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–12, 2024.

[8] Q. Huang, B. Fang, and X. Ai, "A Global-Local Contrastive Learning Framework for Video Captioning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2023, pp. 2410–2414.

[9] Y. Eluri, V. N, J. M, S. B. S. N, and G. S. Abhiram, "Image Captioning using Visual Attention and Detection Transformer Model," in *Proc. IEEE Int. Conf. Electronics, Computing and Communication Technologies (CONECCT)*, 2024, pp. 1–4.

[10] A. Raphael, A. S, E. Anitha, R. S, and M. Venugopalan, "Attention Based CNN-RNN Hybrid Model for Image Captioning," in *Proc. 5th IEEE Global Conf. for Advancement in Technology (GCAT)*, 2024, pp. 1–5.

[11] H. Li, S. Yang, Y. Zhang, D. Tao, and Z. Yu, "Progressive Feature Mining and External Knowledge-Assisted Text-Pedestrian Image Retrieval," *IEEE Trans. Multimedia*, vol. 27, pp. 1973–1987, 2025.

[12] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, May 2013.

[13] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," arXiv:1405.0312, 2015. [Online].

[14] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, Apr. 2017.