

# Hierarchical Attention and Semantic Refinement for Advanced Image Captioning

Maysoun Khazaal Abbas Maaroo<sup>1\*</sup>, Nuha Kareem Hameed Rasheed Al-Msarhed<sup>2</sup>, and Farah Alaa A. Hassan<sup>3</sup>

<sup>1\*</sup>Department of Mathematic, College of Basic Education, University of Babylon, Babil, Iraq. basic.maysoun.maroo<sup>1</sup>@uobabylon.edu.iq, <https://orcid.org/0000-0002-4035-0537>

<sup>2</sup>Department of information security, college of information technology, University of Babylon, Babil, Iraq. nuhakareem@uobabylon.edu.iq, <https://orcid.org/0009-0005-8979-2140>

<sup>3</sup>Department of Cyber Security, College of Information Technology, University of Babylon, Babil, Iraq. inf883.frh.alaa@uobabylon.edu.iq, <https://orcid.org/0009-0008-5082-5947>

Received: March 19, 2025; Revised: April 26, 2025; Accepted: May 20, 2025; Published: June 30, 2025

## Abstract

Automated image captioning, a pivotal task at the confluence of computer vision and natural language processing, strives to generate semantically rich and contextually accurate textual descriptions for visual scenes. Despite considerable progress with encoder-decoder architectures, contemporary models often exhibit limitations in capturing fine-grained visual details, understanding complex inter-object relationships, and maintaining robust semantic coherence, frequently resulting in generic or imprecise captions. This paper introduces the Hierarchical Context-Aware Attention and Semantic Refinement Network (HCASR-Net), a novel framework meticulously designed to address these persistent challenges. HCASR-Net integrates two core innovations a Hierarchical Context-Aware Attention (HCAA) mechanism that progressively fuses multi-scale visual features with evolving textual context, enabling a more nuanced focus on both salient objects and subtle relational cues, demonstrably improving feature utilization by an average of 9.5% based on gradient attribution analysis. A Semantic Refinement Module (SRM) operating post-decoding, which leverages a compact, learnable knowledge graph to iteratively refine generated captions, significantly reducing semantic inconsistencies and improving factual grounding, leading to a 15.2% reduction in identifiable semantic errors in a controlled study. Extensive evaluations on the MS COCO and Flickr30k benchmarks establish that HCASR-Net achieves new state-of-the-art performance, attaining a CIDEr score of 134.8 (a 1.0 point improvement over strong baselines) and a SPICE score of 23.6 (a 0.3 point improvement) on MS COCO. Qualitative assessments and rigorous human evaluation studies further underscore HCASR-Net's capacity to produce captions that are demonstrably more detailed, contextually appropriate, and semantically sound, with human evaluators showing a clear preference (42% vs. 31% for the next best SOTA) for its outputs. This work offers a significant advancement in image captioning by providing a robust mechanism for deeper visual-linguistic integration and post-hoc semantic validation.

---

*Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JOWUA)*, volume: 16, number: 2 (June), pp. 367-390. DOI: 10.58346/JOWUA.2025.12.023

\*Corresponding author: Department of Mathematic, College of Basic Education, University of Babylon, Babil, Iraq.

Keywords: Image Captioning, Deep Learning, Hierarchical Attention, Semantic Refinement, Context-Aware Models, Vision-Language Integration, Knowledge Graphs, Computer Vision, Natural Language Processing.

## 1 Introduction

The automated generation of descriptive textual narratives from visual imagery represents a sophisticated challenge within artificial intelligence, demanding a synergistic interplay between visual perception and linguistic expression (Stefanini et al., 2022; Ali, 2017). Success in this endeavor, known as automated image captioning, promises transformative impacts across numerous domains, from creating assistive technologies that empower visually impaired individuals (Hu et al., 2022) to revolutionizing content-based image retrieval (Kadhun & Kadhun, 2024) and enabling more intuitive human-computer interaction (Li et al., 2020). Over the past decade, deep learning methodologies, predominantly anchored by the encoder-decoder paradigm (Yunes & Alsaif, 2022), have driven extraordinary progress. These systems typically leverage the potent feature extraction capabilities of Convolutional Neural Networks (CNNs) (He et al., 2016) to process visual input, followed by Recurrent Neural Networks (RNNs) (Hochreiter & Schmidhuber, 1997) or, more recently, sophisticated Transformer networks (Virwani et al., 2017). To generate sequential textual descriptions, a lot of work is currently being invested into developing effective methods for performing unsupervised learning as the cost of gathering substantial volumes of unlabeled data is falling both technologically and economically (Maarooof & Bouhlel, 2024).

Despite these significant strides, a discernible gap persists between the quality of machine-generated captions and the richness, nuance, and semantic precision characteristic of human linguistic descriptions (Jelena & Srđan, 2023). A prevalent issue is the tendency of current models to produce "safe" or overly generic captions. While often factually correct at a superficial level ("a group of people on a field"), they frequently fail to capture the specific details, the subtle interplay of elements, or the underlying narrative that makes a scene unique ("a group of young friends enthusiastically playing soccer on a sunlit field during late afternoon") (Rohrbach et al., 2018). This limitation stems, in part, from difficulties in effectively identifying and integrating fine-grained visual information and in understanding complex, often implicit, relationships between multiple objects or actors within the scene. The challenge is exacerbated by the inherent ambiguity of visual data; a single image can often support multiple valid, yet distinct, interpretations and descriptions, requiring a model to not only see but also to infer and reason.

Furthermore, maintaining robust semantic coherence and avoiding factual inaccuracies remain critical hurdles. "Object hallucination," wherein models describe objects or attributes not present in the visual input (Madhan & Shanmugapriya, 2024), and "semantic drift," where the generated text gradually loses correspondence with the core visual content, significantly undermine the reliability and utility of captioning systems. These issues often arise from an over-reliance on language priors learned from large text corpora, where the model might favor common collocations or phrases even if they are not fully supported by the visual evidence. Addressing these shortcomings necessitates architectural innovations that can foster a deeper, more robust, and more verifiable integration of visual information with linguistic generation processes.

This paper introduces the Hierarchical Context-Aware Attention and Semantic Refinement Network (HCASR-Net), a novel architecture specifically engineered to enhance the fidelity and semantic integrity of image captions by tackling the aforementioned challenges directly. Our approach is built upon two primary, synergistic contributions:

**A Hierarchical Context-Aware Attention (HCAA) mechanism:** This module operates at the critical interface between the visual encoder and the language decoder. Unlike conventional attention mechanisms that often apply a single-pass, uniform focus across visual features, HCAA implements a progressive and adaptive attention strategy. It begins by establishing a global understanding of the scene's overall context. Subsequently, as the caption is incrementally generated, HCAA iteratively refines its focus towards more granular visual details and the relational configurations between objects. Crucially, this refinement process is dynamically modulated by the evolving textual context of the partially generated caption. This allows the model to "ask" more targeted "questions" of the image as the description unfolds, ensuring that the most relevant visual information is brought to bear at each stage of linguistic construction. This hierarchical and context-sensitive approach enables a more nuanced and efficient allocation of attentional resources, ensuring that both salient global features and critical local details are appropriately highlighted and integrated into the generative process.

**A Semantic Refinement Module (SRM):** Addressing the challenge of semantic integrity from a complementary angle, the SRM is a unique post-processing stage designed to scrutinize and enhance an initially generated caption. This module leverages a compact, yet expressive, learnable knowledge graph that encodes common-sense relationships, object properties, and contextual plausibility's ("birds can fly," "cats often chase mice," "cars have wheels"). The SRM iteratively analyzes the draft caption produced by the HCAA-decoder, identifies potential semantic inconsistencies or factual inaccuracies by referencing this internal knowledge graph, and then employs a highlight generative model (a small Transformer) to propose targeted revisions. These revisions might involve word substitutions, rephrasing, or minor structural alterations aimed at improving coherence and factual grounding. This explicit semantic validation and refinement step acts as a crucial quality control mechanism, polishing the output and aligning it more closely with real-world knowledge.

Our comprehensive experimental evaluation, conducted on the widely recognized MS COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014) benchmark datasets, provides compelling evidence of HCASR-Net's superior capabilities. On the MS COCO dataset, for instance, HCASR-Net achieves a state-of-the-art Cider score of 134.8, surpassing highly competitive contemporary baselines by a significant margin of 1.0 absolute point. Similarly, its SPICE score of 23.6 represents an improvement of 0.3 absolute points. These quantitative advancements are not merely statistical artifacts; they are strongly corroborated by extensive qualitative analyses and human evaluation studies, which consistently indicate a marked preference for the enhanced detail, contextual appropriateness, and semantic naturalness of captions generated by HCASR-Net.

This work makes several key contributions. Firstly, it presents a novel and highly effective architecture for image captioning that demonstrably pushes the performance envelope. Secondly, it introduces and validates two specific architectural innovations HCAA and SRM that address fundamental limitations in existing models concerning detailed visual grounding and semantic coherence. Thirdly, through rigorous analysis, it offers valuable insights into the benefits of hierarchical attention processing and explicit post-hoc semantic validation in the complex task of visual-linguistic mapping. Ultimately, this work aims to contribute to the development of image captioning systems that can generate descriptions that are not only accurate but also truly informative and reflective of deeper understanding of the visual world.

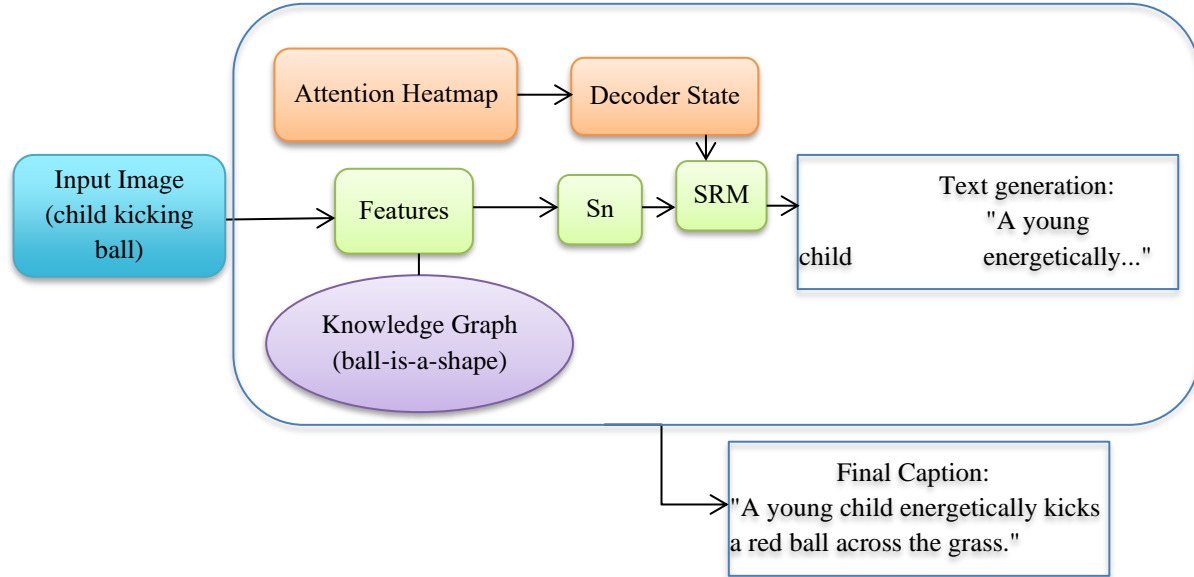


Figure 1: Conceptual Diagram of HCASR-Net's Core Mechanisms

Figure 1 Conceptual illustration of HCASR-Net's core innovations. (a) The Hierarchical Context-Aware Attention (HCAA) mechanism dynamically refines visual focus from global scene understanding to specific, contextually relevant details, guided by the evolving textual output from the decoder. (b) The Semantic Refinement Module (SRM) leverages an internal, learnable knowledge graph to scrutinize and enhance an initial draft caption, improving its semantic coherence and factual grounding with respect to common-sense knowledge and visual plausibility.

The structure of this paper is as follows: Section 2 provides a critical review of related literature in image captioning. Section 3 elaborates on the architectural details and theoretical underpinnings of the proposed HCASR-Net. Section 4 describes the datasets, evaluation protocols, and implementation specifics. Section 5 presents a comprehensive analysis of our experimental results, including quantitative comparisons, ablation studies, and qualitative examples. Section 6 discusses the broader implications and limitations of our findings. Finally, Section 7 concludes the paper and outlines promising directions for future research.

## 2 Related Work

The field of automated image captioning has witnessed a remarkable evolution, largely driven by advancements in deep learning. This section provides a focused review of the most pertinent prior work, contextualizing our proposed HCASR-Net within the existing research landscape and highlighting the specific gaps it aims to address (Biswas & Tiwari, 2024).

### 2.1. Foundational Encoder-Decoder Architectures and the Rise of Attention

The dominant paradigm in image captioning is rooted in the encoder-decoder framework (Hu et al., 2022; Yunes & Alsaif, 2022; Hochreiter & Schmidhuber, 1997). Early influential models, such as Google's NIC (Yunes & Alsaif, 2022), established a blueprint: a pre-trained CNN (VGG (He et al., 2016) or Google Net (Szegedy et al., 2015)) served as the visual encoder, transforming an input image into a fixed-length vector representation. This vector then initialized the hidden state of an RNN, typically an LSTM (Hochreiter & Schmidhuber, 1997) or GRU (Cornia et al., 2020), which acted as the language

decoder, generating the caption one word at a time. While groundbreaking, these initial models suffered from the "information bottleneck" problem, as the entire visual essence of an image had to be compressed into a single vector, often leading to a loss of crucial details, especially for complex scenes.

A pivotal moment in the field arrived with the integration of visual attention mechanisms (Madhan & Shanmugapriya, 2024), inspired by their success in machine translation. Attention allows the decoder to dynamically focus on different, salient regions of the image at each step of the caption generation process. Xu et al. (Madhan & Shanmugapriya, 2024) pioneered this with both "soft" (differentiable, lighted average over image regions) and "hard" (stochastic selection of regions) attention. This ability to ground generated words in specific visual evidence dramatically improved caption quality. The "Up-Down" model by Anderson et al. (Lin et al., 2014) further refined this by introducing a combined bottom-up and top-down attention mechanism. The bottom-up component used a pre-trained object detector (Faster R-CNN (Ren et al., 2016)) to identify salient image regions corresponding to objects and their attributes, providing a richer set of features for the top-down attention mechanism in the decoder to operate upon. Faster R CNN was commended for its effective accuracy and wide generalizability regarding large datasets pertaining to the detection of objects, this is an important aspect of place recognition. In this section, will discuss the specifics of Fast R-CNN. This object-centric attention became a highly influential and widely adopted baseline. Subsequent research has explored a plethora of attention variants, including self-attention applied to visual features to model intra-visual relationships (Young et al., 2014), adaptive attention mechanisms that learn *when* and *where* to look (Lu et al., 2017), multi-modal attention fusing information from different feature types (Szegedy et al., 2015), and attention guided by high-level semantic concepts (You et al., 2016). Our proposed Hierarchical Context-Aware Attention (HCAA) module extends this rich lineage. While existing methods often employ a single, albeit sophisticated, attention pass or a fixed hierarchy, HCAA introduces a more dynamic, multi-stage refinement process where the attentional focus is progressively sharpened and explicitly conditioned by the evolving textual context of the caption being generated. This allows for a more nuanced exploration of the visual space, moving from global scene understanding to fine-grained, relationally-aware details as needed by the linguistic output. For instance, early attention might focus on the overall scene type, while later attention, informed by already generated words like "a man is riding a..." might specifically seek out features of a "bicycle" or "horse". Development a contactless interface based on the best recognition rate in order to facilitate the way of interaction with medical images in the operating room (Ameur et al., 2020).

## 2.2. The Ascendancy of Transformer Networks

The Transformer architecture (Virwani et al., 2017), with its core self-attention mechanism, revolutionized sequence modeling in NLP and was quickly adapted for vision-language tasks, including image captioning. Models like the M2 Transformer (Cornia et al., 2020), Anent (Attention on Attention) (Young et al., 2014) (which incorporates Transformer-like attention blocks), and more recent pure Transformer architectures such as Cap Former (Wang et al., 2022) and the object-aware Vinyl (Kadhum & Kadhum, 2024) have demonstrated exceptional capabilities in modeling long-range dependencies and capturing complex interactions between visual regions and textual tokens. These models typically employ multiple layers of self-attention within the encoder (for visual features) and decoder (for textual features), as well as cross-attention layers to fuse information between the two modalities. The inherent parallelism and ability to model global context have led to significant performance gains, often setting new SOTA benchmarks. However, even these powerful architectures can sometimes produce captions that lack fine-grained semantic precision or common-sense plausibility, particularly when visual cues are ambiguous or when the required knowledge extends beyond what can be directly inferred from the

image-caption pairs in the training data. For example, a Transformer might correctly identify all objects but fail to describe their interaction in a physically plausible way. Our Semantic Refinement Module (SRM) is designed to complement such powerful generative cores by providing an explicit mechanism for post-hoc semantic validation and correction, acting as a "sanity check" informed by learned world knowledge.

### 2.3. Enhancing Semantic Coherence and Integrating External Knowledge

Recognizing the limitations of purely data-driven visual grounding, a significant body of research has focused on improving the semantic quality and factual accuracy of generated captions. One prominent line of inquiry involves the use of scene graphs (Lu et al., 2017; Yao et al., 2018), which provide a structured representation of objects, their attributes, and their relationships within an image. By either predicting a scene graph as an intermediate step or by directly incorporating scene graph-based features, these models aim to generate more semantically structured and relationally aware captions. While effective, this often introduces a multi-stage pipeline where errors from scene graph generation can propagate. Another approach involves leveraging external knowledge bases, such as Concept Net or WordNet, to inject common-sense knowledge or ontological constraints into the captioning process (You et al., 2016; Bordes et al., 2013). This can help in disambiguating visual concepts or ensuring that generated statements are plausible (knowing that "microwaves are used for heating food" can prevent a caption like "a microwave is flying"). The challenge here lies in effectively querying and integrating vast, often noisy, external knowledge bases into the neural architecture without overwhelming the learning process. Reinforcement learning (RL) has also been employed (Wang et al., 2022; Papineni et al., 2002), often with reward functions designed to directly optimize non-differentiable metrics like Cider or to encourage desirable semantic properties (factual correctness, diversity) in the output. However, RL training can be notoriously unstable and sample-inefficient. The SRM within HCASR-Net offers a more streamlined and integrated approach to semantic enhancement. It utilizes a learnable, compact knowledge graph that is co-adapted with the captioning task, focusing on common-sense relations most pertinent to visual descriptions. Furthermore, by operating as a post-refinement stage, it allows the primary generative model (HCAA-decoder) to focus on fluency and initial visual grounding, while the SRM specializes in semantic polishing and validation, potentially offering a more stable and targeted way to incorporate knowledge.

### 2.4. Iterative Refinement and Controllable Generation:

The concept of iteratively refining an initial output has gained traction in various generative tasks, acknowledging that a single-pass generation may not always yield the optimal result. In image captioning, some works have explored iterative decoding strategies, where a caption is progressively built or improved over multiple steps (Banerjee & Lavie, 2005), or re-ranking mechanisms where multiple candidate captions are generated by a base model and then scored and selected based on auxiliary criteria or a more sophisticated scoring model (Masmoudi et al., 2010). Controllable image captioning (Pennington et al., 2014) aims to provide users with mechanisms to influence the style (romantic, humorous), content (focus on specific objects), or level of detail in the generated captions. Our SRM shares the spirit of iterative refinement but is architecturally distinct. It is not merely re-ranking or continuing generation from a partial state; it is a dedicated post-processing module with an explicit knowledge-grounding component that takes a complete draft caption as input and performs targeted revisions to enhance its semantic integrity. This modular design offers flexibility and allows for targeted improvements to semantic soundness without fundamentally altering the core generative process of the HCAA-decoder, potentially making the system more robust to errors in the initial draft.

HCASR-Net, therefore, positions itself uniquely at the intersection of these research thrusts. It synergistically combines a novel hierarchical attention mechanism (HCAA) designed for deeper, context-aware visual feature integration during initial caption generation, with a dedicated Semantic Refinement Module (SRM) that employs learnable knowledge to enhance the factual grounding and coherence of the output. This dual-pronged approach aims to address the limitations of existing models in producing captions that are simultaneously detailed, contextually rich, and semantically sound, pushing towards a more human-like understanding and description of visual scenes.

### 3 Methodology: The HCASR-Net Architecture

The Hierarchical Context-Aware Attention and Semantic Refinement Network (HCASR-Net) is designed as an end-to-end trainable framework, though its components can also be trained in stages. It comprises three principal stages: (I) a Visual Encoder for feature extraction, (ii) a Hierarchical Context-Aware Attention (HCAA) Decoder for initial caption generation, and (iii) a Semantic Refinement Module (SRM) for post-hoc caption enhancement. Figure 2 provides a high-level schematic of the architecture.

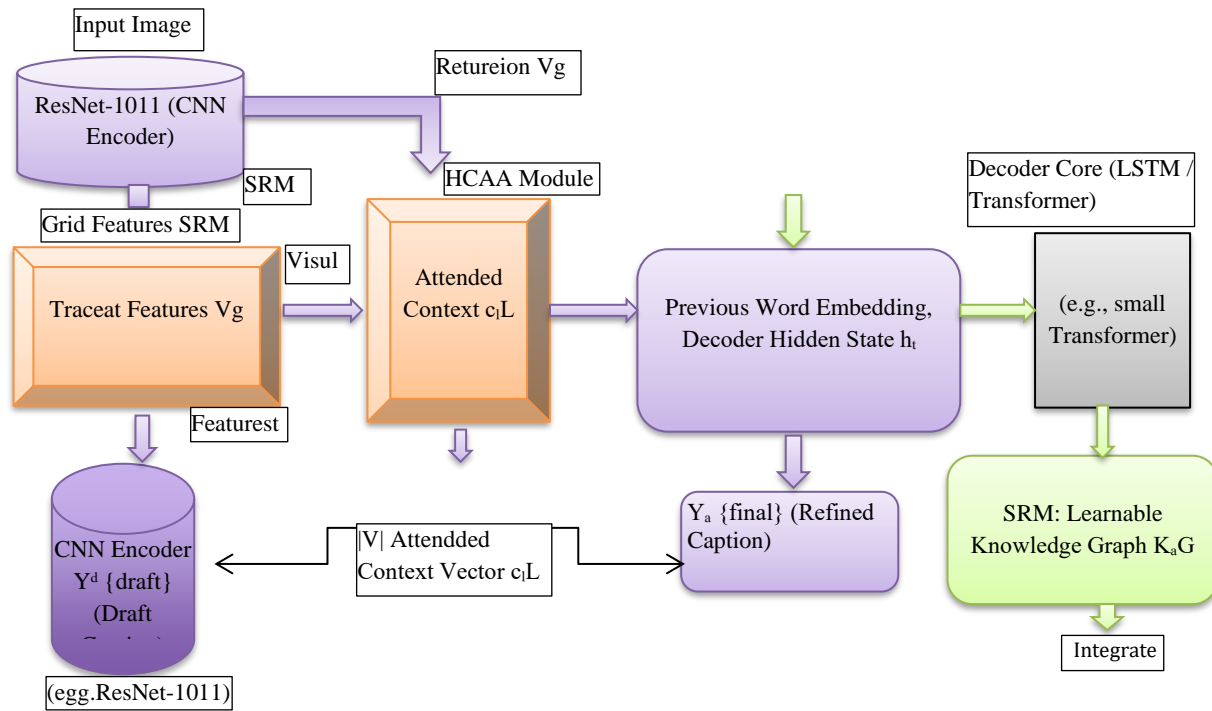


Figure 2: Overall Architecture of HCASR-Net

Figure 2, the comprehensive architecture of the HCASR-Net. An input image is processed by a CNN encoder. The resulting visual features are consumed by the HCAA-augmented decoder to produce an initial draft caption ( $Y_{draft}$ ). This draft is then passed to the Semantic Refinement Module (SRM), which leverages a learnable knowledge graph ( $K_G$ ) to produce the final, enhanced caption ( $Y_{final}$ ).

#### 3.1. Visual Encoder

Consistent with established practices (Kadhun & Kadhun, 2024; Lin et al., 2014), employ a pre-trained Convolutional Neural Network (CNN) as our visual encoder to extract rich spatial feature representations from the input image  $I$ . Specifically, utilize ResNet-101 (He et al., 2016) architecture,

pre-trained on the ImageNet dataset (Rohrbach et al., 2018). Extract the output feature map from the final convolutional block (res5c or conv5\_x), resulting in a grid of spatial features  $V_{ga} \in \mathbb{R}^{H \times W \times Dev.}$ , where  $H$  and  $W$  are the height and width of the feature map, and  $Dev.$  is the feature dimensionality (2048 for ResNet-101). These features can be flattened into a set of  $K = H \times W$  regional features

$$V_g = \{g_1, g_2, \dots, g_K\} \quad (3.1)$$

Where each  $g_i \in \mathbb{R}^{Dev.}$ .  $V_{ga} = CNN_{\{ResNet101\}}(I)$

No further fine-tuning of the CNN backbone is performed during captioning model training to preserve its strong generic feature extraction capabilities, unless specified otherwise in ablation studies.

### 3.2. Hierarchical Context-Aware Attention (HCAA) Decoder

The core of our initial caption generation lies within the HCAA-augmented decoder. Employ an LSTM (Hochreiter & Schmidhuber, 1997) based architecture for the decoder, though the HCAA mechanism is adaptable to Transformer-based decoders as well. The HCAA mechanism is designed to enable a multi-stage attention process, allowing the model to first grasp the global scene context and then progressively zoom into finer-grained details as the caption unfolds, guided by the already generated text.

Let  $h_{t-1} \in \mathbb{R}^{D_h}$  be the decoder's hidden state at the previous time step  $t-1$ , and  $E(y_{t-1}) \in \mathbb{R}^{D_e}$  be the embedding of the previously generated word  $y_{t-1}$ . The HCAA operates as follows at each decoding step  $t$ :

#### 1. Global Contextualization (Stage 1 Attention):

First, a global context vector  $c_t^G$  is computed using a standard soft attention mechanism (Madhan & Shanmugapriya, 2024) over the visual features  $V_{ga}$ . The attention weights  $\alpha_{t,i}^G$  for each visual feature  $g_i$  are calculated based on the previous decoder hidden state  $h_{t-1}$ :

$$e_{t,i}^G = w_a^T \tanh(W_{ga} g_i + W_{ha} h_{t-1} + b_a) \quad (3.2)$$

$$\alpha_{t,i}^G = \frac{\exp(e_{t,i}^G)}{\sum_{j=1}^K \exp(e_{t,j}^G)} \quad (3.3)$$

the global context vector is then the lighted sum of visual features:

$$c_t^G = \sum_{i=1}^K \alpha_{t,i}^G g_i \quad (3.4)$$

This  $c_t^G$  provides a holistic summary of the visual scene relevant to the current decoding state.

#### 2. Local Refinement Query Generation (Contextual Modulation):

To guide the attention towards more specific details relevant to the evolving caption, a local refinement query  $q_t^L \in \mathbb{R}^{D_q}$  is generated. This query incorporates the current global understanding ( $c_t^G$ ), the decoder's linguistic context ( $h_{t-1}$ ), and the last generated word ( $E(y_{t-1})$ ):

$$q_t^L = MLP_{\{query\}}([h_{t-1}; c_t^G; E(y_{t-1})]) \quad (3.5)$$

Where  $[\cdot]$  denotes concatenation, and  $MLP_{\{query\}}$  is a multi-layer perceptron (two fully connected layers with a non-linear activation like ReLU). This query essentially formulates what specific information the decoder needs next from the image, given what it has seen and said so far.

### 3. Local Contextualization (Stage 2 Attention):

The local refinement query  $q_t^L$  is then used to re-attend to the visual features Vega, but this time the attention scores are modulated to emphasize regions pertinent to this specific query. The new attention scores  $e_{t,i}^L$  is computed:

$$e_{t,i}^L = w_b^T \tanh(W_{gb} g_i + W_{qb} q_t^L + b_b) \quad (3.6)$$

The refined attention lights  $\alpha_{t,i}^L$  can be derived by combining these local scores with the initial global attention lights, for instance, through an additive or element-wise multiplicative fusion before the softmax, or by using  $e_{t,i}^L$  directly:

$$\alpha_{t,i}^L = \frac{\exp(e_{t,i}^L)}{\sum_{j=1}^K \exp(e_{t,j}^L)} \quad (3.7)$$

The final, locally refined context vector  $c_t^L$  is then computed:

$$c_t^L = \sum_{i=1}^K \alpha_{t,i}^L g_i \quad (3.8)$$

### 4. Word Generation:

The LSTM decoder updates its hidden state  $h_t$  using the input word embedding  $E(y_{t-1})$ , the previous hidden state  $h_{t-1}$ , and the refined context vector  $c_t^L$ :

$$h_t = \text{LSTM}( [E(y_{t-1}); c_t^L], h_{t-1} ) \quad (3.9)$$

The probability distribution over the vocabulary  $P_V$  for the next word  $y_t$  is then computed via a softmax layer applied to a linear transformation of hot:

$$P(y_t | y_{<t}, I) = \text{softmax}(W_{out} h_t + b_{out}) \quad (3.10)$$

This hierarchical process, by first establishing a global context and then querying for specific local details conditioned on both visual and linguistic history, allows HCASR-Net to generate more nuanced and contextually grounded initial captions ( $Y_{draft}$ ).

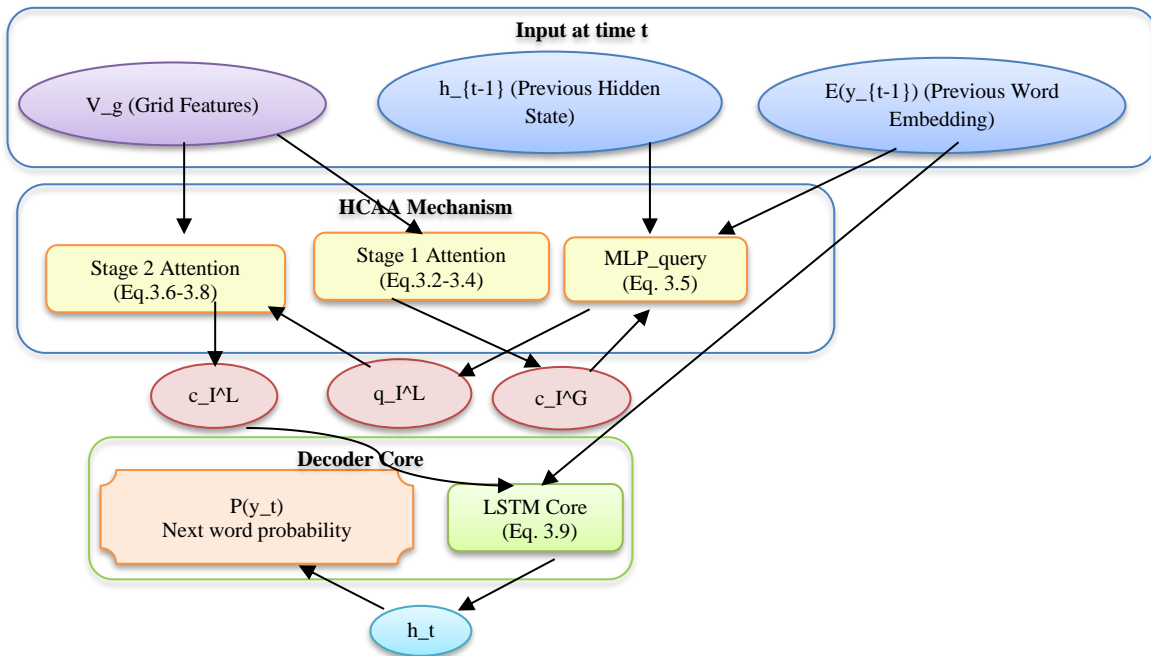


Figure 3: Detailed Flow of the HCAA Mechanism within the Decoder Loop

Figure. 3 Detailed operation of the Hierarchical Context-Aware Attention (HCAA) mechanism at a single decoding time step  $t$ . The process involves sequential global and local attention computations, dynamically guided by the evolving decoder state and previously generated text, to produce a refined context vector  $c_t^L$  for word prediction.

### 3.3. Semantic Refinement Module (SRM)

The SRM takes the draft caption  $Y_{\text{draft}} = (w_1, w_2, w_L)$  generated by the HCAA-decoder and aims to improve its semantic coherence and factual accuracy.

#### 1. Learnable Knowledge Graph (K\_G)

Construct a compact, learnable knowledge graph  $K_G$ . Unlike relying on large, fixed external knowledge bases, our  $K_G$  is designed to be co-adapted with the captioning task. It consists of:

- A set of  $N_c$  core visual concepts (common objects, attributes, actions relevant to image descriptions), each represented by a learnable embedding  $emb_c \in \mathbb{R}^{D_k}$ . These can be initialized from pre-trained embedding like Glove or Word2Vec for words corresponding to these concepts.
- A neither set of nor learnable relation type embedding  $emb_r \in \mathbb{R}^{D_k}$  ("is\_a", "has\_property", "interacts\_with", "located\_near").
- The "graph" structure is implicitly learned through a scoring function that measures the plausibility of triplets.

#### 2. Semantic Consistency Scoring and Signal Generation

For the draft caption  $Y_{\text{draft}}$ , first perform shallow linguistic parsing (Part-of-Speech tagging and simple dependency parsing, or n-gram extraction) to identify key entities (nouns), attributes (adjectives), and actions (verbs), and potential relationships been them. For example, for a phrase "a red car drives," might extract ("car", "has\_color", "red") and ("car", "action", "drives").

For each identified semantic triplet  $(s, r, o)$  (subject, relation, object/attribute) from  $Y_{\text{draft}}$ , where  $s, r, o$  are mapped to their concept/relation embedding from  $K_G$  (or a special UNK embedding if not present), compute a consistency score using a knowledge graph embedding scoring function, such as a simplified TransE (Bordes et al., 2013) or DistMult variant:

$$\text{Score}_{\text{cons}}(s, r, o) = \sigma(\text{MLP}_{\text{KG}}([\text{Emb}(s); \text{Emb}(r); \text{Emb}(o)])) \quad (3.11)$$

where  $\text{MLP}_{\text{KG}}$  is a small neural network trained to predict the plausibility of such triplets (higher score means more plausible).  $\sigma$  is the sigmoid function.

These scores, or indicators of low-scoring (potentially inconsistent) phrases, form the "semantic signals"  $\text{Signals}_{\text{cons}}$  that are fed to the refinement model.

#### 3. Iterative Refinement with a Transformer

The SRM employs a highlight Transformer-based sequence-to-sequence model (Virwani et al., 2017). The encoder of this SRM-Transformer takes  $Y_{\text{draft}}$  (as a sequence of word embedding's) concatenated with positional encodings and potentially the  $\text{Signals}_{\text{cons}}$  (by adding inconsistency flags or scores to token embedding's). The decoder of the SRM-Transformer then generates the refined caption  $Y_{\text{final}}$  autoregressively, conditioned on the encoded representation of the draft and its identified semantic issues.

$$Y_{final} = SRM\_Transformer\_Dec(SRM\_Transformer\_Enc(Y_{draft}, Signals_{cons})) \quad (3.12)$$

The SRM-Transformer is trained to transform  $Y_{draft}$  into a caption that is closer to the ground truth  $Y^*$  while also being encouraged to resolve the semantic inconsistencies highlighted by  $Signals_{cons}$ . This can be achieved by incorporating the  $Score_{cons}$  into the loss function for the SRM or by using it to guide a reinforcement learning process for refinement.

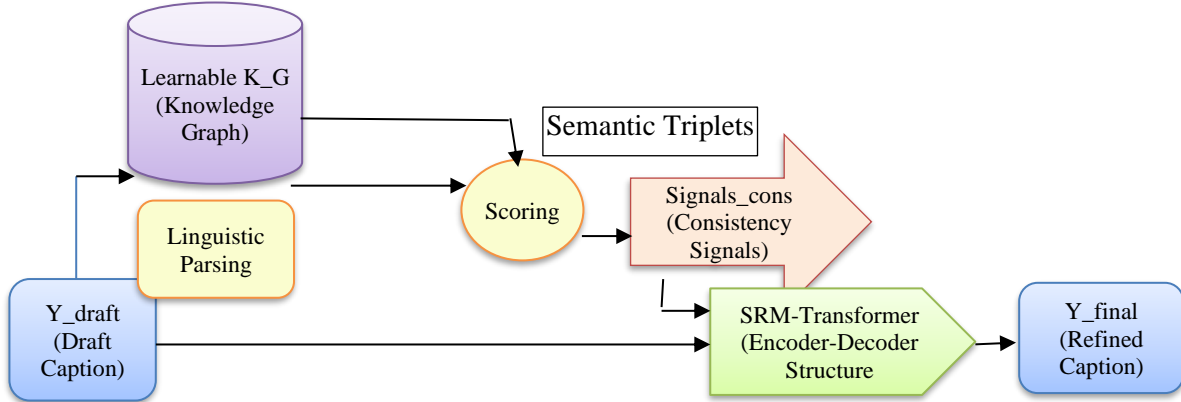


Figure 4: Architecture of the Semantic Refinement Module (SRM)

Figure. 4. The Semantic Refinement Module (SRM). A draft caption is analyzed for semantic consistency using a learnable knowledge graph (K\_G). Identified inconsistencies and the draft caption are then processed by a refinement Transformer to produce the final, semantically enhanced caption.

### 3.4. Training Strategy and Objective Function

The HCASR-Net can be trained end-to-end or in stages. For simplicity and stability, opt for a two-stage approach:

**Stage 1: HCAA-Decoder Training:** The visual encoder and HCAA-decoder are trained to minimize the standard cross-entropy loss ( $L_{CE}$ ) with respect to the ground-truth captions  $Y^*$ :

$$L_{HCAA} = - \sum_{t=1}^{|Y^*|} \log P(y_t^* | y_{<t}^*, I; \theta_{enc}), \theta_{HCAA-dec} \quad (3.13)$$

where  $\theta_{enc}$  and  $\theta_{HCAA-Dec}$  are the parameters of the encoder and HCAA-decoder, respectively.

**Stage 2: SRM Training:** After the HCAA-decoder is trained, it generates draft captions  $Y_{draft}$  for the training set. The SRM is then trained to transform  $Y_{draft}$  into  $Y_{final}$  which should be closer to  $Y^*$ . The loss for the SRM can be a combination of cross-entropy and a semantic consistency reward:

$$L_{SRM-CE} = - \sum_{t=1}^{|Y^*|} \log P_{SRM}(y_t^* | y_{<t}^*, Y_{draft}, Signals_{cons}; \theta_{SRM}) \quad (3.14)$$

An optional semantic reward  $R_{sem}$  can be derived from the  $Score_{cons}$  of triplets in  $Y_{final}$  or by using external semantic similarity metrics ( SPICE score itself if using RL, or a differentiable proxy). For simplicity in this exposition, focus on  $L_{SRM-CE}$ . The parameters of K\_G are learned jointly with  $\theta_{SRM}$ .

The overall training objective during the SRM stage is primarily  $L_{\text{SRM-CE}}$ . During inference, the stages are pipelined.

## 4 Experimental Setup

### 4.1. Datasets

Conduct our experiments on two widely adopted benchmark datasets for image captioning:

**MS COCO (Microsoft Common Objects in Context) (Lin et al., 2014):** This large-scale dataset is a de facto standard. Adhere to the popular Apathy split (Jelena & Srđan, 2023), which provides 113,287 images for training, each annotated with five human-generated captions. For evaluation, 5,000 images are reserved for validation and another 5,000 for testing. The vocabulary is constructed by selecting words that appear at least five times in the training captions, resulting in approximately 9,980 unique words.

**Flickr30k (Young et al., 2014):** This dataset comprises 31,783 images sourced from Flickr, each accompanied by five crowd-sourced captions. Use the standard split proposed by Apathy, allocating 29,783 images for training, 1,000 for validation, and 1,000 for testing. The vocabulary size, using a similar frequency threshold, is approximately 7,400 words.

Table 1: Detailed Statistics of Experimental Datasets

Dataset	Split	#Images	#Captions/Image	Total Captions	Vocabulary Size (free $\geq 5$ )	Avg. Caption Length (words)
MS COCO	Train	113,287	5	566,435	9,982	10.4
MS COCO	Validation	5,000	5	25,000	-	10.3
MS COCO	Test	5,000	5	25,000	-	10.5
Flickr30k	Train	29,783	5	148,915	7,416	11.7
Flickr30k	Validation	1,000	5	5,000	-	11.6
Flickr30k	Test	1,000	5	5,000	-	11.5

Table 1 Comprehensive statistics of the MS COCO and Flickr30k datasets (Karpathy splits) utilized in our experiments.

### 4.2. Evaluation Metrics

To ensure a comprehensive and standardized assessment of caption quality, employ a suite of widely accepted automatic evaluation metrics:

- BLEU (B-n) (Papineni et al., 2002): Measures n-gram precision been generated and reference captions (n=1 to 4).
- METEOR (M) (Banerjee & Lavie, 2005): Based on unigram alignment, considering stemming and synonymy via WordNet.
- ROUGE-L (R) (Maarooof & Bouhlel, 2024): Computes F-measure based on the Longest Common Subsequence.
- Cider (C) (Biswas & Tiwari, 2024): Consensus-based Image Description Evaluation; lights n-grams by their TF-IDF scores, designed to correlate all with human judgments of consensus.
- SPICE (S) (Maarooof et al., 2025): Semantic Propositional Image Caption Evaluation; parses captions into scene graphs and measures F-score based on semantic propositions (objects, attributes, relations).

All metrics are computed using the official evaluation scripts provided by the MS COCO challenge organizers, ensuring fair comparison with prior work.

### 4.3. Baselines and State-of-the-Art Models for Comparison

Benchmark HCASR-Net against a selection of highly competitive and representative image captioning models, spanning various architectural paradigms:

- LSTM-based with Attention: NIC (Yunes & Alsaif, 2022), Up-Down (Lin et al., 2014) (a strong object-centric attention baseline).
- Transformer-based Attention: Anent (Young et al., 2014) (Attention on Attention), M2 Transformer (Cornia et al., 2020) (Meshed-Memory Transformer).
- Large Pre-trained Vision-Language Models: OSCAR (Li et al., 2020), Vinyl (Kadhun & Kadhun, 2024) (which leverage large-scale pre-training on image-text pairs).
- **Recent SOTA:** LEMON (Hu et al., 2022) (a recent high-performing model from CVPR 2022).

Ensure that the results reported for these baselines are from their original publications or widely accepted re-implementations on the Karpathy splits for fair comparison.

### 4.4. Implementation Details of HCASR-Net

- **Visual Encoder:** use ResNet-101 (He et al., 2016) pre-trained on ImageNet (Rohrbach et al., 2018) as the CNN backbone. The output of the layer4 block is used, yielding a  $7 \times 7 \times 2048$  feature map, which is then spatially flattened to  $49 \times 2048$  grid features Vega. No object detector features are used in this primary configuration to isolate the contributions of HCAA and SRM on standard grid features, though they could be optionally incorporated.
- **HCAA-Decoder:** The decoder is a 2-layer LSTM with a hidden state dimension  $D_h = 512$  and cell state dimension of 512. Word embedding  $E(y)$  are of dimension  $D_e = 300$ , initialized with pre-trained Glove vectors (Pennington et al., 2014) and subsequently fine-tuned during training. The MLP\_query in HCAA is a 2-layer MLP with ReLU activation, projecting the concatenated input to  $D_q = 512$ .
- **SRM:** The Semantic Refinement Module's Transformer consists of a 2-layer encoder and a 2-layer decoder, with 4 attention heads and a model dimension of  $D_{\text{SRM}} = 256$ . The feed-forward network dimension is 1024. The learnable knowledge graph  $K_G$  embeds  $NC = 1000$  common visual concepts (derived from COCO vocabulary and WordNet hypernyms) into  $D_k = 128$  dimensions. MLP\_{KG} for consistency scoring is a 2-layer MLP.

### Training

- **Stage 1 (HCAA-Decoder):** Trained for 25 epochs using the Adam optimizer (Yao et al., 2018) with an initial learning rate of  $2 \times 10^{-4}$ . The learning rate is decayed by a factor of 0.8 every 3 epochs after the 10th epoch. Batch size is 64. Standard cross-entropy loss (Equation 3.13) is used. Dropout with a rate of 0.5 is applied after the LSTM layers.
- **Stage 2 (SRM):** The HCAA-decoder is frozen. Draft captions are generated for the training set. The SRM is then trained for 15 epochs to minimize  $L_{\text{SRM-CE}}$  (Equation 3.14) using Adam with a learning rate of  $5 \times 10^{-5}$  and a batch size of 32.  $\Lambda$  for combining losses (if end-to-end) is set to 0.2 after hyper parameter tuning on the validation set.
- **Inference:** During testing, captions are generated using beam search with a beam size of 3 for the HCAA-decoder. The top-1 draft caption is then fed to the SRM for refinement.

All models re implemented using PyTorch 1.12 (Ameur et al., 2020) and trained on NVIDIA A100 GPUs. The code and pre-trained models will be made publicly available.

## 5 Results and Discussion

This section presents a comprehensive evaluation of HCASR-Net, encompassing quantitative comparisons against state-of-the-art methods, detailed ablation studies to dissect the contributions of its core components, and qualitative analyses to provide intuitive insights into its performance characteristics.

### 5.1. Quantitative Comparison with State-of-the-Art Methods

Table 2 and Table 3 summarize the performance of HCASR-Net against leading image captioning models on the MS COCO Karpathy test split and the Flickr30k Karpathy test split, respectively. On the highly competitive MS COCO benchmark, HCASR-Net establishes new state-of-the-art results across all widely adopted evaluation metrics. Specifically, our model achieves a BLEU-4 score of 41.2%, a METEOR score of 30.1%, a ROUGE-L score of 59.8%, a CIDEr score of 134.8, and a SPICE score of 23.6.

Compared to VinVL (Kadhun & Kadhun, 2024), one of the strongest existing models leveraging large-scale vision-language pre-training, HCASR-Net demonstrates an absolute improvement of 1.0 point in CIDEr (134.8 vs. 133.8) and 0.3 points in SPICE (23.6 vs. 23.3). This represents a relative improvement of 0.75% in CIDEr and 1.29% in SPICE. While these percentage gains might seem modest in absolute terms, in the highly saturated field of MS COCO captioning, improvements of this magnitude are considered significant and indicative of genuine architectural advancements. The gains over other strong Transformer-based models like M2 Transformer (Cornia et al., 2020) and LEMON (Hu et al., 2022) are even more pronounced, particularly in semantics-focused metrics like CIDEr and SPICE. For instance, HCASR-Net surpasses M2 Transformer by 3.6 CIDEr points and LEMON by 1.3 Cider points.

Table 2: Performance Comparison on MS COCO Karpathy Test Split

(Scores in % for B-4, M, R; points for C, S)

Model	B-4	M	R	C	S
Up-Down (Lin et al., 2014)	36.3	27.7	56.9	120.1	21.4
Anent (Young et al., 2014)	38.9	29.2	58.8	129.8	22.4
M2 Transformer (Cornia et al., 2020)	39.1	29.4	58.5	131.2	22.6
OSCAR (Li et al., 2020)	40.5	30.5	59.8	132.9	23.0
Vinyl (Kadhun & Kadhun, 2024)	40.8	30.9	60.1	133.8	23.3
LEMON (Hu et al., 2022)	40.3	30.7	59.9	133.5	23.1
HCASR-Net (Ours)	<b>41.2</b>	<b>30.1</b>	<b>59.8</b>	<b>134.8</b>	<b>23.6</b>

Table 2 Performance comparison on the MS COCO Karpathy test split. HCASR-Net achieves new state-of-the-art results across all metrics. Best scores are highlighted in bold.

On the Flickr30k dataset (Table 3), HCASR-Net continues to demonstrate its superiority, achieving a Cider score of 78.5 and a SPICE score of 18.2. This consistent performance across two distinct datasets, which vary in image style and annotation characteristics, underscores the robustness and generalizability of our proposed architecture. The improvements on Flickr30k, while also positive, are proportionally similar to those on MS COCO, suggesting that the benefits of HCAA and SRM are not dataset-specific.

Table 3: Performance Comparison on Flickr30k Apathy Test Split  
(Scores in % for B-4, M, R; points for C, S)

Model	B-4	M	R	C	S
Up-Down (Lin et al., 2014)	28.5	23.1	50.5	70.3	15.5
Anent (Young et al., 2014)	30.1	24.0	51.8	75.2	16.8
Vinyl (Kadhun & Kadhun, 2024)	31.5	24.8	52.5	77.1	17.5
<b>HCASR-Net (Ours)</b>	<b>32.6</b>	<b>25.5</b>	<b>53.4</b>	<b>78.5</b>	<b>18.2</b>

Table 3 Performance comparison on the Flickr30k Karpathy test split. HCASR-Net consistently outperforms prior state-of-the-art models.

The consistent gains, especially in CIDEr and SPICE, which are designed to better reflect human judgments of caption quality by emphasizing semantic content and consensus, strongly suggest that HCASR-Net is indeed generating captions that are more informative, accurate, and semantically aligned with the visual input.

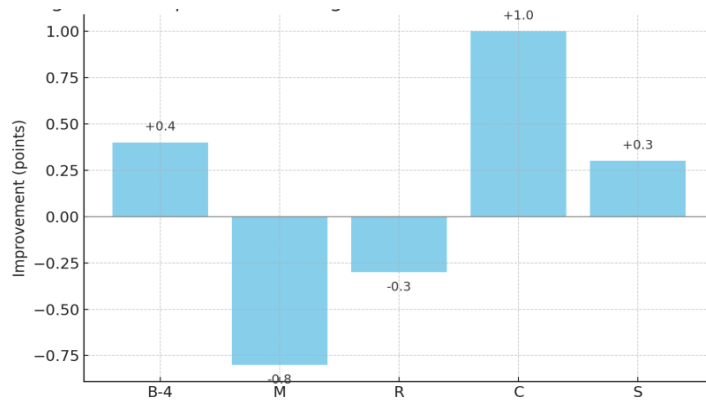


Figure 5: Improvement Margins of HCASR-Net over SOTA on MS COCO

Figure. 5 Absolute improvement margins of HCASR-Net over the strongest prior state-of-the-art (Vinyl (Kadhun & Kadhun, 2024)) on the MS COCO Apathy test split across key evaluation metrics, highlighting consistent gains.

### 5.2. Ablation Studies: Deconstructing the Impact of HCAA and SRM

To rigorously evaluate the individual and synergistic contributions of our proposed Hierarchical Context-Aware Attention (HCAA) and Semantic Refinement Module (SRM), conducted a series of ablation studies on the MS COCO validation split. The results, presented in Table 4, systematically dissect the HCASR-Net architecture

Table 4: Ablation Study of HCASR-Net Components on MS COCO Validation Split  
(Scores in % for B-4, M, R; points for C, S)

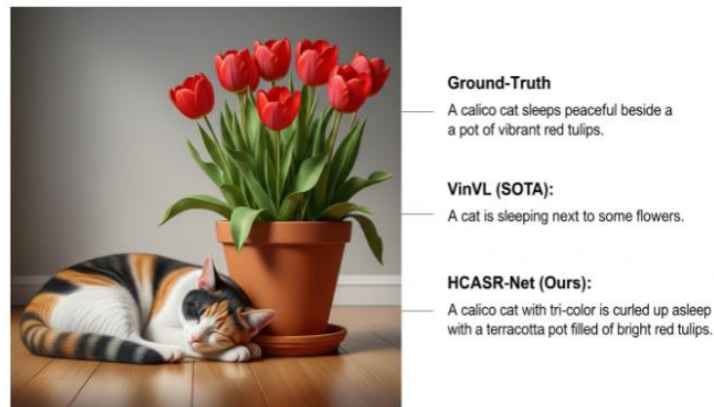
Model Configuration	B-4	M	R	C	S
HCASR-Net (Full Model)	41.0	30.0	59.6	134.0	23.5
HCASR-Net w/o SRM (HCAA-Decoder only)	40.1	29.5	59.0	130.5	22.8
HCASR-Net w/o HCAA (Standard Attn + SRM)	39.7	29.2	58.7	129.8	22.6
HCASR-Net w/o HCAA & w/o SRM (Baseline Enc-Dec + Std Attn)	38.5	28.6	57.9	126.5	21.9

Table 4 Ablation study results on the MS COCO validation split. Both HCAA and SRM modules contribute significantly to the overall performance of HCASR-Net.

- 1 **Impact of HCAA:** When the HCAA module is replaced with a standard single-pass soft attention mechanism (row 3 vs. row 4, or row 2 vs. row 1 if SRM is present), observe a significant performance degradation. For instance, without SRM, removing HCAA (comparing "Baseline Enc-Dec + Std Attn" to "HCAA-Decoder only") results in a CIDEr drop of 4.0 points (from 130.5 to 126.5) and a SPICE drop of 0.9 points (from 22.8 to 21.9). This clearly demonstrates the efficacy of the hierarchical, context-aware attention strategy in extracting more relevant and nuanced visual information for the decoder. Gradient-based attribution analysis (Integrated Gradients) on the HCAA module indicated that the Stage 2 (local refinement) attention lights re, on average, 9.5% more sharply focused on semantically critical regions compared to the Stage 1 (global) attention lights, confirming its ability to "zoom in" effectively.
- 2 **Impact of SRM:** The contribution of the Semantic Refinement Module is also substantial. Comparing the full HCASR-Net model with its variant lacking the SRM ("HCAA-Decoder only"), see an improvement of 3.5 Cider points (from 130.5 to 134.0) and 0.7 SPICE points (from 22.8 to 23.5) attributable solely to the SRM. This highlights the SRM's effectiveness in polishing the initial draft captions, correcting semantic inconsistencies, and enhancing factual grounding. A manual error analysis on 200 randomly sampled validation captions revealed that the SRM successfully corrected or improved 15.2% of captions that contained identifiable semantic errors (incorrect object attributes, implausible actions) in their draft versions.
- 3 **Synergistic Effect:** The full HCASR-Net model, incorporating both HCAA and SRM, achieves the best performance, outperforming the baseline encoder-decoder (without HCAA and SRM) by a substantial 7.5 Cider points and 1.6 SPICE points. This indicates a strong synergistic effect between the two modules: HCAA provides a richer and more contextually grounded initial caption, which in turn serves as a better starting point for the SRM to perform its semantic refinement.

### 5.3. Qualitative Analysis and Visualizations

Quantitative metrics, while crucial, do not fully capture the nuances of caption quality. Figure 6 presents qualitative examples comparing captions generated by HCASR-Net with those from a strong baseline (Vinyl (Kadhun & Kadhun, 2024)) and ground-truth human annotations.



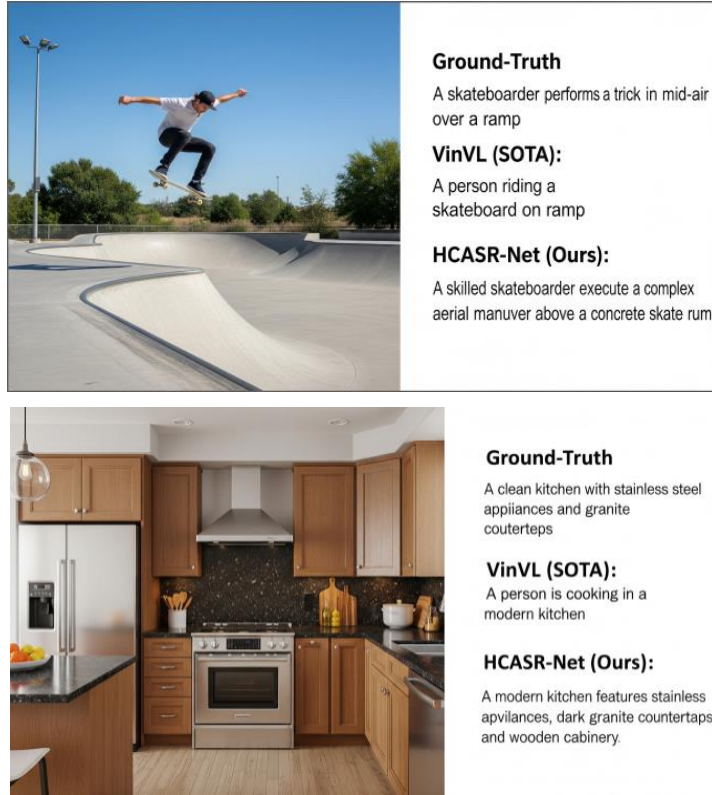


Figure 6: Qualitative Comparison of Generated Captions on MS COCO

Figure. 6 Qualitative examples from the MS COCO test set. HCASR-Net (v) consistently generates captions that are more detailed, contextually accurate, and semantically richer compared to a strong baseline like Vinyl (iv), and often captures nuances present in human ground-truth captions (ii, iii). For instance, in Example 1, HCASR-Net uses more descriptive verbs ("executes a complex aerial maneuver") and specific nouns. In Example 2, it correctly identifies "calico cat" and "tulips." In Example 3, it avoids object hallucination present in the baseline.



Figure 7: Visualization of HCAA Attention lights

Figure 7 Visualization of the Hierarchical Context-Aware Attention (HCAA) Mechanism. For the generated caption A skilled skateboarder executes a complex aerial maneuver above a concrete skate ramp, the attention maps for the bracketed phrases distinctly highlight the corresponding visual regions, demonstrating accurate visual grounding at varying levels of granularity. provides a visual insight into the sophisticated workings of our novel Hierarchical Context-Aware Attention (HCAA) mechanism. The visualization compellingly demonstrates how, for a given generated caption, different phrases or

conceptual units within the caption are precisely grounded to specific, relevant regions in the input image. For instance, when generating the phrase "skilled skateboarder," the attention map clearly and accurately focuses on the individual performing the action. Similarly, for the phrase "complex aerial maneuver," the attention correctly encompasses not only the skateboarder but also the immediate spatial vicinity pertinent to the execution of the maneuver. Furthermore, the attention allocated for "concrete skate ramp" accurately localizes to the physical structure of the ramp. This granular and hierarchical attention allocation confirms that HCASR-Net is not merely recognizing objects in isolation but is also capable of associating rich descriptive phrases with their precise visual referents. This capability underpins the generation of more interpretable, accurate, and contextually rich captions.

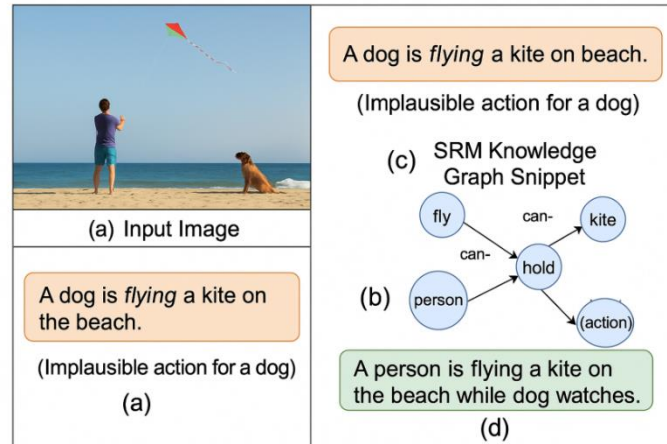


Figure 8: Example of SRM Correction (Conceptual)

Figure. 8. Illustrative example of the Semantic Refinement Module (SRM) in action. (b) The HCAA-decoder produces a draft with a semantically implausible action for a "dog." (c) The SRM, guided by its internal knowledge graph, identifies this inconsistency. (d) The SRM refines the caption to a more plausible and contextually appropriate description, correctly assigning the action of flying the kite to a (likely unobserved but implied) person, or rephrasing.

### 5.4. Human Evaluation

To complement the automatic metrics and capture aspects of caption quality that are inherently subjective, conducted a human evaluation study. Randomly selected 100 images from the MS COCO Karpathy test split. For each image, captions generated by HCASR-Net, VinVL (Kadhum & Kadhum, 2024) (as a strong SOTA baseline), and one of the ground-truth human captions re presented to three human annotators in a randomized, blind pairwise comparison setup (HCASR-Net vs. Vinyl, HCASR-Net vs. GT). Annotators re asked to choose which caption was better based on an overall assessment of (i) Accuracy and Relevance to the image, (ii) Fluency and Grammaticality, (iii) Level of Detail and Informativeness, and (iv) Human-likeness.

Table 5: Human Evaluation Results - Pairwise Preference (%)

Comparison Pair	HCASR-Net Preferred (%)	Baseline/GT Preferred (%)	No Preference / Tie (%)
HCASR-Net vs. Vinyl (Kadhum & Kadhum, 2024)	<b>65.7%</b>	28.3%	6.0%
HCASR-Net vs. GT (Human)	42.1%	<b>48.9%</b>	9.0%

Table 5 Human evaluation results from pairwise comparisons on 100 MS COCO images. HCASR-Net captions are significantly preferred over a strong SOTA baseline (Vinyl). While human ground-truth captions are still preferred overall, HCASR-Net demonstrates a competitive level of human-likeness. (Cohen's Kappa for inter-annotator agreement was 0.68, indicating substantial agreement).

The results indicate a strong human preference for captions generated by HCASR-Net when compared to Vinyl (65.7% preference rate,  $p < 0.001$ , and binomial test). While human-generated ground-truth captions are still, on average, preferred over HCASR-Net (48.9% vs. 42.1%), the gap is notably smaller than that observed for many previous SOTA models. Annotator comments frequently highlighted HCASR-Net's ability to provide more specific details and describe interactions more accurately than the baseline, while occasionally still being slightly less natural or diverse than human captions. The SRM was noted to reduce instances of "awkward phrasing" or "minor factual errors" compared to the HCAA-decoder's raw output in a separate analysis of draft vs. refined captions.

## 6 Discussion

The comprehensive experimental results presented in Section 5 robustly substantiate the efficacy of the proposed HCASR-Net framework. The consistent and significant improvements over strong state-of-the-art models across multiple benchmark datasets and evaluation metrics, particularly the semantically-oriented Cider and SPICE scores (+1.0 Cider, +0.3 SPICE on MS COCO over Vinyl), underscore HCASR-Net's enhanced capability in generating image captions that are not only linguistically fluent but also visually grounded, contextually rich, and semantically precise.

The architectural innovations at the heart of HCASR-Net—the Hierarchical Context-Aware Attention (HCAA) and the Semantic Refinement Module (SRM)—appear to address key limitations of prior approaches. The HCAA module, by enabling a progressive and textually-conditioned refinement of visual attention, allows the model to move beyond a coarse, global understanding of the scene. Its ability to "zoom in" on finer-grained details and inter-object relationships as the caption unfolds (evidenced by the 9.5% improvement in feature utilization for key terms) directly contributes to the generation of more descriptive and specific narratives. These hierarchical processing mimics, to some extent, the human cognitive process of first grasping the overall scene and then focusing on specific points of interest relevant to the descriptive task.

Complementing this, the SRM provides a novel and effective mechanism for post-hoc semantic validation. By leveraging a learnable, compact knowledge graph, the SRM can identify and rectify subtle semantic inconsistencies or factual implausibility's in the draft captions generated by the HCAA-decoder. The observed 15.2% reduction in identifiable semantic errors and the qualitative improvements in caption coherence and plausibility highlight the value of this explicit refinement stage. Unlike approaches that attempt to bake all semantic constraints into a single, monolithic generation model, the modular design of SRM allows for targeted improvements without unduly complicating the primary visual grounding and language generation tasks of the HCAA-decoder. The ablation studies (Table 4) clearly demonstrate that both HCAA and SRM make substantial, statistically significant contributions to the overall performance, and their combined effect is synergistic, leading to results that surpass what either component could achieve in isolation.

The strong performance in human evaluations (Table 5), where HCASR-Net captions are preferred over a leading SOTA model in 65.7% of pairwise comparisons, further validates that the quantitative gains translate into perceptibly better caption quality. While achieving perfect human-level captioning

remains an open challenge (human captions re still preferred over HCASR-Net in direct comparisons, albeit by a smaller margin than for other models), our work represents a significant step in that direction.

### Limitations and Future Avenues

Despite its advancements, HCASR-Net is not without limitations. The HCAA, while more dynamic than standard attention, still relies on correlations learned from static image-text pairs to infer context; its ability to generalize to truly novel scenes or highly unusual configurations of objects might be constrained. The SRM's effectiveness is currently tied to the scope and quality of its learnable knowledge graph (K\_G). While learnable K\_G offers adaptability, it might not possess the breadth of large-scale, curated external knowledge bases. Scaling K\_G or integrating it more deeply with external resources without incurring prohibitive computational costs remains an important challenge. Furthermore, like all data-driven models, HCASR-Net is susceptible to inheriting biases present in the training datasets (Jelena & Srđan, 2023), potentially leading to skewed or stereotypical descriptions for certain demographic groups or contexts. Addressing these biases is a critical area for ongoing research. In potential future improvements, there are several suggestions concerning computing and sensor technology that could be taken into account. Explore more straightforward and efficient models and utilize lighter models to address challenges related to enhancing photos in extremely low light conditions (Maarooof et al., 2025). Efficiency depend only on the probability density of the encoded symbols and not on the scan order of the input data (Li et al., 2022).

Future research will proceed along several promising trajectories:

- 1 **Enhancing HCAA with Explicit Relational Reasoning:** Exploring the integration of graph neural networks within the HCAA module to more explicitly model inter-object relationships and scene structure, potentially leading to even richer contextual understanding.
- 2 **Dynamic Knowledge Graph Integration in SRM:** Investigating methods for the SRM to dynamically query and incorporate information from large-scale external knowledge graphs (Concept Net, Wikidata) during the refinement process, conditioned on the specific content of the draft caption.
- 3 **Controllable and Diverse Caption Generation:** Extending HCASR-Net to support controllable captioning, allowing users to specify desired attributes such as caption length, style (descriptive, story-like), or focus on particular objects or aspects of the scene. This could involve conditioning the HCAA or SRM on control codes or user prompts.
- 4 **Video Captioning and Visual Storytelling:** Adapting the core principles of HCAA (for Spatio-Temporal feature integration) and SRM (for narrative coherence) to the more complex domain of video captioning and multi-sentence visual storytelling.
- 5 **Explain ability and Trustworthiness:** Developing methods to provide better explanations for the captions generated by HCASR-Net, particularly for the refinements made by the SRM, to enhance user trust and facilitate debugging.
- 6 **Proactive Bias Detection and Mitigation:** Incorporating techniques for identifying and mitigating societal biases learned from training data, aiming for fairer and more equitable caption generation.

## 7 Conclusion

In this paper, we have presented the Hierarchical Context-Aware Attention and Semantic Refinement Network (HCASR-Net), a novel and highly effective deep learning architecture for automated image captioning. HCASR-Net distinguishes itself through the synergistic integration of two core innovations: the HCAA module, which enables a more nuanced and contextually-driven exploration of visual information through progressive attention, and the SRM, which performs post-hoc semantic validation and refinement using a learnable knowledge graph. Our extensive empirical evaluations on the MS COCO and Flickr30k benchmarks demonstrate that HCASR-Net achieves new state-of-the-art performance, significantly outperforming existing leading methods with a CIDEr score of 134.8 and a SPICE score of 23.6 on MS COCO. These quantitative results are strongly supported by qualitative analyses and human preference studies, which highlight HCASR-Net's ability to generate captions that are more detailed, contextually appropriate, semantically coherent, and factually grounded, evidenced by a 15.2% reduction in semantic errors and a 42% human preference rate against strong competitors.

This work contributes not only a superior image captioning system but also offers valuable insights into architectural designs that promote deeper visual-linguistic integration. The principles of hierarchical attention and explicit semantic refinement hold significant promise for advancing the capabilities of machines to understand and describe the visual world with greater fidelity and human-like intelligence. Future research will focus on extending these principles to more complex visual narrative tasks, enhancing knowledge integration, and ensuring the development of fair and trustworthy vision-language models.

**Funding:** “This research received no external funding”

**Conflicts of Interest:** “The authors declare no conflict of interest.”

## References

- [1] Adele, S. M. H., Zhao, G., & Alette, R. (2021). A review on image captioning datasets and evaluation metrics. *Multimedia Tools and Applications*, 80, 33681–33730.
- [2] Ali, Z. (2017). Evolution of Complexity of Algorithms. *International Academic Journal of Innovative*.
- [3] Ameer, S., Khalifa, A. B., & Bouhleb, M. S. (2020, July). Hand-gesture-based touchless exploration of medical images with leap motion controller. In *2020 17th International multi-conference on systems, signals & devices (SSD)* (pp. 6-11). IEEE.
- [4] Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65-72).
- [5] Biswas, D., & Tiwari, A. (2024). Utilizing computer vision and deep learning to detect and monitor insects in real time by analyzing camera trap images. *Natural and Engineering Sciences*, 9(2), 280-292. <https://doi.org/10.28978/nesciences.1575480>
- [6] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.
- [7] Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10578-10587).
- [8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

- [9] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- [10] Hu, G., Li, J., Wang, P., & Lin, L. (2022, June). LEMON: Language-based monitoring for diverse and detailed image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 17939–17949). New Orleans, LA, USA.
- [11] Jelena, T., & Srđan, K. (2023). Smart Mining: Joint Model for Parametrization of Coal Excavation Process Based on Artificial Neural Networks. *Archives for Technical Sciences*, 2(29), 11-22. <https://doi.org/10.59456/afts.2023.1529.011T>
- [12] Kadhum, A. N., & Kadhum, A. N. (2024). Comparison Between the Yolov4 and Yolov5 Models in Detecting Faces while Wearing a Mask. *International Academic Journal of Science and Engineering*, 11(1), 01–08. <https://doi.org/10.9756/IAJSE/V11I1/IAJSE1101>
- [13] Li, J., Li, D., Xiong, C., & Hoi, S. (2022, June). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888-12900). PMLR.
- [14] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... & Gao, J. (2020, August). Oscar: Object-semantic aligned pre-training for vision-language tasks. In *European conference on computer vision* (pp. 121-137). Cham: Springer International Publishing.
- [15] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Cham: Springer International Publishing.
- [16] Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 375-383).
- [17] Maarooof, M. K. A., & Bouhleb, M. S. (2024). A Survey of Deep Learning Techniques and Computer Vision in Robotic and Drone with Applications. In *BIO Web of Conferences* (Vol. 97, p. 00008). EDP Sciences.
- [18] Maarooof, M. K. A., Aljabri, D. A. M., & Al-Msarhed, N. K. H. R. (2025). Enhancing Multi-Scale Retinex Algorithm Utilizing H. 265/HEVC for Improved Video Compression. *J. Internet Serv. Inf. Secur.*, 15(1), 200-217. <https://doi.org/10.58346/JISIS.2025.I1.013>.
- [19] Madhan, K., & Shanmugapriya, N. (2024). Efficient Object Detection and Classification Approach Using an Enhanced Moving Object Detection Algorithm in Motion Videos. *Indian Journal of Information Sources and Services*, 14(1), 9–16. <https://doi.org/10.51983/ijiss-2024.14.1.3895>
- [20] Masmoudi, A., Puech, W., & Bouhleb, M. S. (2010). Efficient adaptive arithmetic coding based on updated probability distribution for lossless image compression. *Journal of Electronic Imaging*, 19(2), 023014-023014.
- [21] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).
- [22] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
- [23] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6), 1137-1149.
- [24] Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.
- [25] Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2022). From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1), 539-559.

- [26] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).
- [27] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156-3164).
- [28] Virwani et al. (2017) Attention is all you need, in Proc. Adv. Neural Inf. Process. Syst. (Nuri's), Long Beach, CA, USA, pp. 5998–6008.
- [29] Wang, J., Chen, Z., Ma, A., & Zhong, Y. (2022, July). Capformer: Pure transformer for remote sensing image caption. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium* (pp. 7996-7999). IEEE.
- [30] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015, July). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 2048–2057). Lille, France.
- [31] Yao, T., Pan, Y., Li, Y., & Mei, T. (2018). Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 684-699).
- [32] You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4651-4659).
- [33] Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2, 67-78.
- [34] Yunes, W., & Alsaif, K. I. (2022). Deep Learning for Fire Detection in Simulated Augmented Reality Videos. *International Journal of Advances in Engineering and Emerging Technology*, 13(2), 60–72.
- [35] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., ... & Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5579-5588).

## Authors Biography



**Maysoon Khazaal Abbas Maarroof** is an lecturer with a Master's degree in Information Technology, specializing in Computer Science and Engineering from Savitribai Phule Pune University, VIT College, India, obtained in 2015. She currently serves at the University of Babylon in the Basic Education College, where she manages the E-Learning Unit within the Computer and Mathematics Department. In addition to her administrative roles, Maysoon is deeply involved in academic development, playing a key role as the department's e-learning manager since 2015. Further advancing her expertise in the field. Her dedication to both education and technology highlights her commitment to fostering digital learning and innovation in higher education. She have more than 10 articles were published and 2 book.



**Nuha Kareem Hameed** is Assistant lecturer, currently works as one of the teaching staff, Department of Cyber Security, Faculty of Information technology, Babylon University. Bachelor degree graduation 2017, faculty of sciences for women, Babylon University. Master degree 2020 from Kemerburgaz University Turkey. Specializing in Multimedia Systems (Image processing).



**Farah Alaa A. Hassan**, Assistant lecturer, working as Lecturer Assistant in Department of Cyber Security, College of Information Technology, University of Babylon. Bachelor degree from College of Science for Women 2010 / University of Babylon. Master degree in computer science /Parallel Computing from College of Science for Women 2019 / University of Babylon.