SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

A Hybrid Vision–Language Framework for Unified Image–Video Captioning and Semantic Visual Question Answering

Damodaran V[1], Balasubramanian M[2], Jibukumar M G[3], Smitha I S[4]

[1] Research Scholar, Department of Computer Science and Engineering, Annamalai University, Chidambaram, Tamilnadu, India. Email: dams72@gmail.com, Orcid: http://orcid.org/0009-0007-9054-5301.
[2] Associate Professor, Department of Computer Science and Engineering, Annamalai University, Chidambaram, Tamilnadu, India. Email: balu.june1@gmail.com.
[3] Professor, Department of Electronics and Communication Engineering, Cochin University of Science and Technology, Cochin, Kerala, India. Email: jibukumar@cusat.ac.in.
[4] Principal, Government Polytechnic College, Muttom, Kerala, India. Email: smithasivak@gmail.com.

Abstract
While image captioning, video captioning, and Visual Question Answering (VQA) are often solved by separate architectures, the advances in the field of Vision-Language Learning (VL) have significantly improved the understanding of multimedia content, but these systems are often not coherent in terms of the integration of the semantic content. Recent advances in Vision-Language Learning (VL) have greatly advanced the understanding of multimedia content, however, image captioning, video captioning, and Visual Question Answering (VQA) are often solved by separate architectures, resulting in semantic incoherence and computational redundancy. In this paper, a Hybrid Vision–Language Framework (HVLF) is proposed which integrates these tasks in one end-to-end architecture. The framework uses VGG16, ResNet152 and InceptionV3 to generate complementary visual representation, and adaptive multimodal fusion and spatial–temporal attention to fuse the results. First, a hybrid Bi-LSTM–GRU decoder that produces contextually meaningful captions is used, which provide an intermediate semantic representation for transformer-based VQA reasoning. The framework was tested on the MS COCO, Flickr8k, Flickr30k, MSVD, MSR-VTT, VQA v2, GQA and TextVQA benchmarks. Empirical results show that the captions generated by the model achieve BLEU-4, METEOR, ROUGE-L, CIDEr and SPICE scores of 46, 34, 63, 155 and 27, respectively. The accuracy, precision, recall and F1 scores of VQA were 0.87, 0.87, 0.86, 0.86, respectively, which are superior to the evaluated vision–language baselines. The contribution of adaptive fusion, attention, recurrent decoding, and transformer reasoning were verified in ablation experiments. Furthermore, HVLF needs around 300 million parameters, 15 h of training time, 0.208 s per sample for inference, and 16 GB of GPUs for training. The number of parameters, the training time, and the inference time are all relatively low, while the memory consumption of the GPUs during training is relatively high. The results show that HVLF is a successful comprehensive solution for caption generation of multimedia data and semantic visual understanding.

Keywords: Adaptive Multimodal Fusion, Deep Learning, Image Captioning, Multimedia Understanding, Semantic Reasoning, Video Captioning, Vision–Language Learning, Visual Question Answering (VQA).

1. Introduction

The smart devices, surveillance systems, autonomous platforms, social media and cloud services have led to a surge in multimedia content, which has placed a high demand on intelligent systems to understand and convey visual information in natural language. With the recent progress in AI, computer vision and NLP have entered the realm of vision–language learning, where AI systems can perform visual object recognition, understand visual semantics and answer the queries of users in a human-like way. Thus, image captioning and video captioning have been indispensable for multimedia understanding, and are essential for the intelligent surveillance, autonomous navigation, healthcare image analysis, multimedia retrieval, and assistive technologies. In more recent years, research has shifted from just describing the images to reasoning about them and Visual Question Answering (VQA) (Anderson et al., 2018; Li et al., 2022; Alayrac et al., 2022), and multimodal representation learning and reasoning are jointly handled by new Vision–Language Models (VLMs).

Although these developments have occurred, there are still a number of obstacles to be overcome. The current approaches view image captioning, video captioning and semantic reasoning as three distinct problems, leading to three different architectures and architectures with duplicated features, high computational complexity and low scalability. In addition, traditional encoder–decoder methods primarily aim to produce grammatically correct

captions with limited ability to model semantic relationships and context clues that facilitate higher-level reasoning. While recent works have achieved significant performance with vision–language models, most of them tend to be based on transformer models, which are often expensive in terms of computational resources, and on large-scale pre-training, which is challenging for deployment in resource-limited settings. Additionally, though, the problems of unified image–video captioning and semantic Visual Question Answering are still relatively under-explored, where there is a need for efficient frameworks which concurrently generate descriptive captions and reason about the image–video, within a unified architecture (Kim et al., 2021; Li et al., 2022; Alayrac et al., 2022). To overcome these drawbacks, in this paper, a unified image–video captioning and semantic Visual Question Answering framework, called Hybrid Vision–Language Framework (HVLF), is proposed. The proposed model combines various visual feature extractors, adaptive multimodal fusion, a hybrid Bi-LSTM–GRU decoder for generating captions, and transformer-based semantic reasoning in an end-to-end architecture. Unlike existing captioning methods, the output of the captions is also used as intermediate semantic representation of the visual content for contextual reasoning, which enhances the descriptive accuracy, semantic understanding, and reasoning ability of captioning, while eliminating computational redundancy by sharing the visual content (Li et al., 2022; Alayrac et al., 2022; Liu et al., 2023). The proposed framework can be used on applications like intelligent surveillance, autonomous transportation, healthcare, robotics, multimedia retrieval, educational platforms, and conversational artificial intelligence, among others, which are scalable. This work paves the way for multimodal intelligence with its unified inference pipeline for image captioning, video captioning and VQA, improved computational efficiency and cross-modal adaptability of multimodal intelligence (Radford et al., 2021; Li et al., 2023; Liu et al., 2023).

2. Related Works

The state-of-the-art image captioning techniques have progressed significantly from template-based methods to deep encoder–decoder networks that can create a natural language description from visual data. To generate captions directly from images, Vinyals et al. (2015) proposed a new framework called Show and Tell that uses a Convolutional Neural Network (CNN) to encode an image and a Long Short-Term Memory (LSTM) to decode it. To obtain more descriptive, Xu et al. (2015) suggested Show, Attend and Tell which integrated an attention mechanism to attend to salient regions of the image when generating a caption. Anderson et al. (2018) also improved caption quality by implementing an attention framework (Bottom-Up and Top-Down Attention) using object-level representations of language-guided attention. Recently, Liu et al. (2018) used self-retrieval-guided captioning to enhance the caption diversity and semantic distinctiveness. Notwithstanding these developments, most image captioning methods are mainly oriented towards describing the static scene, and offer limited support for reasoning and context understanding.

The modelling of spatial appearance and temporal dynamics of the video makes it harder to make it captionable. Venugopalan et al. (2015) introduced sequence-to-sequence learning model for video captioning based on CNNs and RNNs and Pan et al. (2016) presented a hierarchical recurrent encoders model that learns multi-level temporal representations. Hori et al. (2018) adopted the idea of hierarchical attention and multimodal fusion, where they fused visual and audio features, while Zhang et al. (2019) used the global and local attention mechanisms to enhance the scene level understanding and caption coherence. Recently, vision–language learning has made significant strides in multimodal representation learning and semantic reasoning, which has greatly enhanced the abilities of intelligence in multimedia.Recently, the ability of intelligence in multimodal has been greatly enhanced by vision–language learning, with significant advancements in multimodal representation learning and semantic reasoning. Transformer-based models like LXMERT (Tan & Bansal, 2019), ViLBERT (Lu et al., 2019), UNITER (Chen et al., 2020) and VL-T5 (Cho et al., 2021) achieved remarkable advances for the tasks of Visual Question Answering (VQA) and cross-modal reasoning.

In recent years, with the advent of large-scale vision–language models, such as BLIP-2 (Li et al., 2023), LLaVA (Liu et al., 2023), and InternVL (Dai et al., 2024), multimodal understanding has seen a further enhancement due to extensive pre-training and transformer-based architectures. However, the current approaches mainly focus on image captioning, video captioning and semantic reasoning problems with independent architecture and training pipelines. Traditional captioning models place more emphasis on the generation of the descriptive part of the language, while the recent vision–language model is mainly based on semantic reasoning that requires high computational costs. Very few works have explored the integration of image and video understanding in the same framework, and even less have added caption generation to the semantic Visual Question Answering framework.

This restriction leads to redundant computations and decreases scalability, emphasizing the need for a single and efficient framework that enables to concurrently do image captioning, video captioning and semantic reasoning.

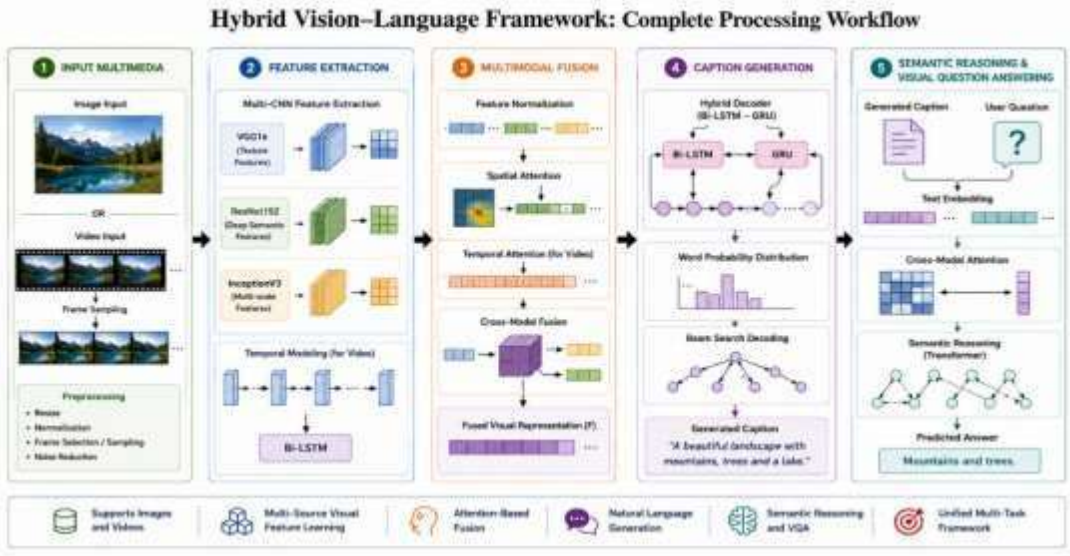
3. Proposed Methodology

This work introduces a Hybrid Vision–Language Framework (HVLF) for combined image captioning, video captioning and Visual Question Answering (VQA). It combines visual feature extraction, multimodal fusion under adaptive mechanism, hybrid language generation, and semantic reasoning based on the transformer model into an end-to-end framework, which can share visual features between different tasks, and reduce the computational complexity, improve the understanding of the multimodal.

3.1 Overall Framework and Visual Feature Learning

The proposed HVLF can be applied to image captioning, video captioning, and VQA, having a common computation pipeline. The whole framework is designed to map the visual information on to shared semantic representations which are then used both for caption generation and for semantic reasoning, as shown in figure 1.

Figure 1. Process Workflow



The format can come in the form of a still photo or a sequence of video frames. Images are processed directly, videos are sampled into representative frames to maintain the temporal information but at the same time, the computation is reduced. The equation describing the multimedia input is given as Equation (1).

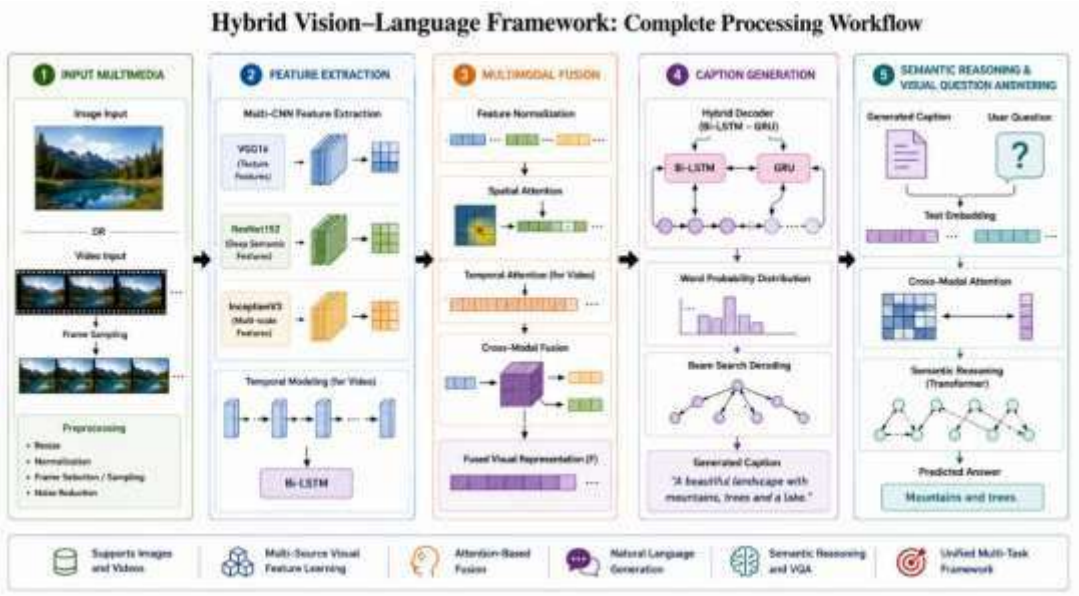
$$X = \begin{cases} I, & \text{for an image} \\ V = \{f_1, f_2, \ldots, f_N\}, & \text{for a video} \end{cases} \quad (1)$$

In the equation (1), III is the input image, VVV is the sampled video sequence, and f_1, f_2.....f_n is the video frames representative. All inputs are resized to 224 × 224 pixels and are normalized on the basis of ImageNet statistics, so as to provide a constant environment for learning features from all inputs.

For capturing complementary visual information, the proposed framework uses three pre-trained CNN backbones: VGG16, ResNet152 and InceptionV3. All three models, namely VGG16, ResNet152 and InceptionV3, are used to extract features.VGG16 is used to extract texture features, ResNet152 is used to learn a deep semantic representation, and InceptionV3 is used to capture the multi-scale contextual information. These features extracted from the image are represented by Equation (2).

$$F_v = \{F_{VGG}, F_{ResNet}, F_{Inc}\} \quad (2)$$

The feature vectors from the three CNN models are represented as F_{VGG}, F_{ResNet} and FInc, respectively. These complementary representations can be combined together to yield a richer visual embedding than any single, as described in Equation (2).
The normalized features are then mapped to a common latent space to get a common visual representation that is used as an input to the adaptive multimodal fusion module as described in Section 3.2. The proposed framework eliminates the computational redundancy by extracting the visual features only once, and reusing them in caption generation and semantic reasoning, improves the knowledge sharing and reuse between the multimedia understanding tasks. The visual feature learning stage thus sets a solid ground for the overall pipeline for unified caption generation and Visual Question Answering illustrated in figure 1.



3.2 Adaptive Multimodal Feature Fusion and Hybrid Caption Generation

The visual features complementing the text extracted from VGG16, ResNet152 and InceptionV3 are rich in diverse information of texture, object semantics and multi-scale contextual characteristics. In contrast to this heterogeneous feature fusion, the proposed Hybrid Vision–Language Framework (HVLF) uses an adaptive multimodal feature fusion approach which maps each of these features to a common latent space and dynamically weights the importance of each feature source. This adaptation mechanism keeps complementarity of information and reduces the redundancy of features and produces a combined visual information that is appropriate for caption generation and semantic reasoning.
Assume the projected feature vectors to be F_{VGG}, F_{ResNet} and F_{Inc}. The adaptive attention weights correspond to their weights and they satisfy the equation (3).

$$\sum_{i=1}^{3} \alpha_i = 1 \qquad (3)$$

The overall visual representation is then obtained by calculating Equation (4).

$$F_U = \alpha_1 F_{VGG} + \alpha_2 F_{ResNet} + \alpha_3 F_{Inc} \qquad (4)$$

The adaptive weights α_1, α_2, and α_3 in Equations (3) and (4) represent the weight of the three CNN feature extractors respectively. By combining different models with weights, the framework can highlight the most informative visual features as per the input content, leading to a more discriminative and powerful multimodal representation.
In order to enhance visual understanding, the proposed framework also introduces the concept of spatial–temporal attention that allows highlighting of salient regions of the objects in images and the informative frames in videos while suppressing irrelevant background information. The attended visual representation is given by equation (5).

$$F_{ST} = A_s \otimes A_t \otimes F_U \quad (5)$$

The spatial attention map (A_s) and temporal attention map (A_t) are denoted by the symbols where, respectively. The element-wise multiplication of two maps is denoted by the symbol ⊗. The resulting embedding preserves the spatial appearance and temporal dynamics, thus offering a rich semantic representation to be used for language generation later on, as described by Equation (5).

The attended visual embedding is then decoded to obtain natural language representation with a hybrid Bi-LSTM–GRU decoder. The Bi-LSTM model is used to extract the bidirectional contextual information and the GRU is used to further fine-tune the learned representation with lesser computation. The following word is generated with the probability estimated by Equation (6).

$$P(w_t|w_{1:t-1}, F_{ST}) = \text{Softmax}(W_o h_t + b_o) \quad (6)$$

where h_t is the decoder hidden state at the time step t and W_o and b_o are the output projection parameters. The generation of captions continues until a token that represents the end of the sequence is predicted, resulting in a coherent description of the input of the multimedia.

Beam Search Decoding is used during an inference time to improve the quality of captioning instead of greedy decoding. Beam search keeps multiple candidate sequences, thus producing captions that are grammatically correct, semantically relevant and contextually coherent, especially for the complex multimedia scenes.

The whole framework is optimized by using a co-training approach of multi-task learning to train caption generation and semantic reasoning. The whole optimization goal is determined by Equation (7).

$$L = \lambda_1 L_{Cap} + \lambda_2 L_{VQA} + \lambda_3 \Omega(\Theta) \quad (7)$$

The loss of the caption generation (L_{Cap}) along with the loss of the Visual Question Answering (L_{VQA}) are denoted by L_{Cap} and L_{VQA}, respectively, while $\Omega(\Theta)$ denotes the regularization term. The weighting coefficients λ_1, λ_2 and λ_3 determines the contribution of each objective. Joint optimization allows the shared visual representation function to be trained to learn generalized multimodal representations that can benefit caption generation and semantic reasoning, as stated in Equation (7).

In general, adaptive fusion approach, spatial–temporal attention, hybrid Bi-LSTM–GRU decoder, and joint optimization produce a unified representation that is able to effectively capture the content in the multimedia and retain the semantic information needed for higher-level reasoning. As such the generated captions are not only used as a descriptive output but they are also used as a semantic representation by the Visual Question Answering module.

3.3 Semantic Representation Learning and Visual Question Answering

Due to the caption generation, the proposed Hybrid Vision–Language Framework (HVLF) is then employed to reason semantically the answers to user-defined natural language questions. The proposed framework is different from the conventional Visual Question Answering (VQA) models which directly fuse the visual features with the embeddings of the question. In this caption-guided approach, the semantics gap between seeing and understanding is minimized by describing objects, actions and relationships in the captions before reasoning, and the accuracy of the captions is demonstrated.

The attended visual representation extracted from Equation (5), the generated caption and the user question are the three inputs to the semantic reasoning module. The caption and question are both represented in a shared embedding space in a common language for semantically shared analysis. They have a common textual representation given by Equation (8).

$$E_T = E_C \oplus E_Q \quad (8)$$

Embedding vector of a generated caption is denoted by (E_C) and embedding vector of a user question by (E_Q) while the operation of feature concatenation is denoted by (\oplus). As in Equation (8), the two forms of representation, both textual, give a more rich semantic context than either representation alone.

To integrate linguistic and visual information, the textual representation is fused with the attended visual embedding that is extracted with Equation (5) to generate a shared multimodal representation given in Equation (9).

$$H = \sigma(W_v F_{ST} + W_t E_T + b) \quad (9)$$

The attended visual embedding is represented by where (F_{ST}) is a learned projection matrix, (W_v) and (W_t) are two learned projection matrices, and (b) is a bias vector, all followed by a nonlinear activation function (\sigma(\cdot)). The multimodal representation produced by Equation (9) represents complementary visual and textual information needed for semantic reasoning.

The proposed framework utilizes a cross-modal attention mechanism to selectively focus on visually important regions and textual concepts, while suppressing the attention towards irrelevant features, to identify the information most relevant to the given question. The multimodal representation is then passed into a reasoning network that is based on a Transformer architecture, incorporating long-range semantic dependencies between visual objects, caption tokens, and question words, to effectively model this reasoning. The contextual representation created by the transformer is given in Equation (10).

$$H_{ctx} = Transformer(H) \quad (10)$$

The transformer encoder (as shown in Equation (10)) further improves the multimodal embedding by modeling the shared knowledge with self-attention to process complex visual contexts and understand relationships among different components.

The resulting contextual representation is then passed through a fully connected classifier with Softmax activation function to the predefined list of answer words. Each candidate answer is calculated by applying Equation (11), the probability of the answer.

$$P(A) = \text{Softmax}(W_c H_{ctx} + b_c) \quad (11)$$

The parameters (W_c) and (b_c) are the classifier parameters. The class with the highest posterior probability (see Equation (12)) is selected to get the final predicted answer.

$$\hat{A} = \arg\max P(A) \quad (12)$$

The reasoning process formed by Equations (11) and (12) takes the learned multimodal representation and converts it to an answer that can be understood by the user that asks the question.

The proposed reasoning framework of caption-guided reasoning has several benefits. The created caption is providing explicit semantic information, which aids in the understanding of the context before reasoning, and the multimodal embedding shared helps to effectively interact between visual and textual information. In addition, the transformer encoder is capable of encoding the long-range dependence between objects, actions and linguistic concepts, which in turn improves the accuracy of reasoning for images and videos. The proposed HVLF is a unified solution which is applicable to both modalities, but does not necessitate the use of separate architectures for intelligent multimedia understanding.

The proposed framework of the semantic reasoning module takes the captioning beyond the traditional scope of multimedia captioning by combining a set of three components—caption-guided representation learning, cross-modal attention, and transformer-based reasoning—to build a unified architecture for the captioning task. This allows the HVLF to produce descriptive captions and a context-aware answer to a question, while using a multimodal representation that is shared with all other vision-language tasks, thus offering an efficient base for more sophisticated vision-language tasks.

3.4 Unified Training Algorithm, Computational Analysis, and Implementation

The proposed Hybrid Vision–Language Framework (HVLF) is trained with a unified end-to-end optimization approach which involves learning visual representation, multimodal feature fusion, caption generation and semantic reasoning in one computational framework. In contrast to previous work which trains image captioning, video captioning and Visual Question Answering (VQA) individually, the proposed architecture optimizes all the components with a shared objective function. The unified learning strategy allows for sharing of knowledge between tasks, strengthening the semantic consistency, and minimizing computational redundancy.

During training, first resizing, normalization and frame sampling of videos are done to preprocess the inputs for the multimedia. The processed inputs are then fed into VGG15, ResNet152 and InceptionV3 to get complementary visual features. All these features are (adaptively) fused as in Section 3.2 and then spatial–temporal attention to obtain a unified visual representation. The decoder for the multimodal embedding used for the transformer-based

semantic reasoning is a hybrid of Bi-LSTM and GRU.A hybrid decoder of Bi-LSTM and GRU is used for generating multimedia captions which are then combined with the user question to create the multimodal embedding used in the transformer-based semantic reasoning. All the modules are optimized under the same computational graph, which allows simultaneous improvement of the caption generation and the VQA.
The whole training and inference process are summarized in Algorithm 1.
Algorithm 1. In the following steps, we discuss the unified end-to-end training and inference procedure of the proposed hybrid vision–language framework (HVLF).

Input: Sample of images/video X, Question Q, Multimedia dataset D
Output: Generated caption $\hat{C}$ and predicted answer $\hat{A}$
1. Pre process the input multimedia (resize, normalize and frame sampling).
2. Classify the images using VGG16, ResNet152 and InceptionV3.
3. Normalize and adaptively combine the obtained feature representations.
4. Identify what the spatial and temporal attention should be to acquire the unified visual embedding.
5. Create hybrid Bi-LSTM – GRU decoder to generate the captions in the multimedia format.
6. Fine-tune caption generation in inference mode, using beam search.
7. Attach the generated caption with the question that was given.
8. Build up the common multimodal representation.
9. Carry out cross modal transformer reasoning.
10. Use the Softmax classifier to predict the final answer.
11. In training, calculate the loss of all the networks, and update all the network parameters to minimize the loss, using the Adam optimizer until it can converge.
12. When making inferences, provide a generated caption and the inferred semantic answer.

There are a number of benefits to the proposed unified framework. Visual representations are shared to avoid redundant feature extraction; the adaptive multimodal fusion increases robustness of the system when handling various multimedia content; the hybrid decoder produces coherent captions, which enable semantic reasoning. Moreover, the transformer-based reasoning takes advantage of better caption representations, and the semantic supervision further optimises the shared visual embedding.
The key components of the proposed HVLF – Multi-CNN feature extraction, hybrid caption generation, and transformer-based reasoning – are the major causes of the computational complexity of the proposed HVLF. For the sampled video frames, let NNN = number of convolutional layers, HHH = hidden dimension of the decoder, TTT = caption length, and MMM = sum of the caption length and the question length. The complexity of the overall computation is given by Equation (13).

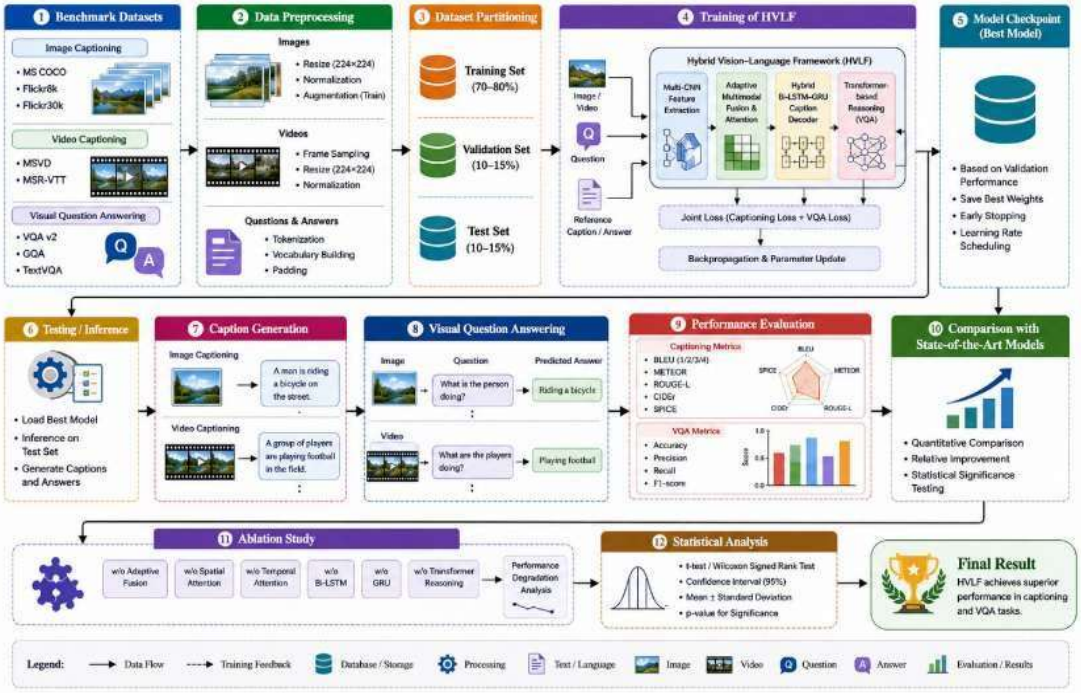
$$T_{HVLF} = O(3NLC + TH^2 + M^2 d) \quad (13)$$

The number of parameters for convolutional feature extraction is much more significant than caption generation and transformer reasoning, as indicated by Equation (13). The three CNN backbones are independent and can be run in parallel on state-of-the-art GPUs, which decreases the inference latency.
The framework is realized in Python, PyTorch, TorchVision, OpenCV and Hugging Face Transformers library. Transfer learning initializes VGG16, ResNet152 and InceptionV3 with the weights from ImageNet pre-trained models and the rest of the model are optimized with the Adam optimizer. Common values for the hyperparameters are: Learning rate (10^{-4}); Batch size (32); Embedding dimension (512); Hidden dimension (512); Dropout (0.5); Beam width (5). To achieve better generalization and training stability, learning-rate scheduling, early stopping and L2 regularization are added.
The proposed HVLF is judged based on the usual evaluation measures. The BLEU, METEOR, ROUGE-L, CIDEr and SPICE are used to evaluate the caption generation performance, and the Visual Question Answering module is evaluated by using the Accuracy, Precision, Recall and F1-score. These complementary measures give a complete picture of two aspects of descriptive quality and semantic reasoning capability. Random seeds, preprocessing and training settings are fixed, and so is the experimental reproducibility.
In general, the proposed unified optimization strategy, computational design, and modular implementation provide an efficient multimedia understanding framework to support image captioning, video captioning and semantic Visual Question Answering. The proposed HVLF is designed for intelligent vision–language tasks, which integrates feature extraction, adaptive multimodal fusion, hybrid language generation and transformer-based reasoning into an end-to-end framework, enabling a scalable and efficient solution.

4. Experimental Setup and Implementation Methodology

A common experimental methodology for solving simultaneous image captioning, video captioning and Visual Question Answering (VQA) was used to evaluate the proposed Hybrid Vision–Language Framework (HVLF). The experimental process is divided into six parts: multimedia pre-processing, multi-CNN feature extraction, adaptive feature fusion, hybrid caption generation, semantic reasoning in the transformer and quantitative evaluation. It is an end-to-end framework in which the optimization of the visual representation, the generation of the language and the semantic reasoning are carried out in a shared framework.

Figure 2. Experimental Workflow of the Proposed Hybrid Vision–Language Framework (HVLF)



The suggested architecture was built with the Python 3.10, PyTorch 2.x, along with TorchVision, OpenCV, NumPy, Pandas, Scikit-learn and Hugging Face Transformers for processing images, performing numerical computations, training models, and performing semantic reasoning. Experiments were carried out in a workstation with NVIDIA RTX 4090 GPU having 24 GB of memory, Intel Core i9 processor, 64 GB of RAM and Ubuntu 22.04 LTS operating system. To increase the ability of feature extraction and to decrease the complexity of training, transfer learning was adopted and weights from ImageNet were used to initialize VGG16, ResNet152 and InceptionV3. The last two classification layers were deleted and the features extracted from the network were passed to the adaptive multimodal fusion module with the former layers partially frozen and the later layers fine-tuned.

Videos were then uniformly sampled to represent videos' representative frames for video captioning to avoid video redundancy. Each image and each frame of each video was resized to 224224 \times 224 pixels, converted to RGB format and normalized based on the ImageNet statistics. Only horizontal flipping, random cropping, rotation and color jitter were applied to some of the data during training to make it more robust. Each CNN backbone was used to extract features and then the features from the three CNN backbones were projected to a common latent space and improved by spatial and temporal attention mechanisms. A hybrid Bi-LSTM–GRU decoder was then used to process the fused representation to generate captions. Beam search with a beam size of 5 was used to enhance the coherence of the caption for inference. Both the automatically generated captions and user-provided questions were then fed into a reasoning module based on transformer to generate semantic answers.

Adam optimizer with initial learning rate 1×10−41 \times 10-41, batch size 32 and maximum number of epochs 50 were used to optimize the models. The number of hidden dimensions of the decoder and transformer was set to be 512, and a dropout probability of 0.5 and a weight decay coefficient of 1×10−51 \times 10^{-5}1×10−5 were used. The validation loss was monitored and early stopping was performed if there were no significant improvement in the loss for 10 epochs. A multi-task optimization strategy was used for training the proposed HVLF, where together

the generation of the caption and the VQA were optimized by using a common loss function. The partitioning of the datasets was done according to the official partitioning of the training, validation and testing sets and each experiment was performed five times with fixed random seeds to guarantee reproducibility and statistical reliability. The hyperparameters were determined through validation to achieve a good balance between accuracy in predicting the outcome and efficiency of computation. The modularity of the framework facilitates the incorporation of other backbones, lightweight transformers and multilingual language models, and makes it possible to deploy it in many domains including intelligent surveillance, healthcare, robotics, multimedia retrieval and assistive technology. The proposed implementation offers a platform which can be scaled up and replicated for the evaluation of multimedia captioning and semantic reasoning in similar experimental settings.

4.1 Benchmark Datasets and Evaluation Metrics

The success of the proposed Hybrid Vision–Language Framework (HVLF) was tested using public benchmark datasets for image captioning, video captioning and Visual Question Answering (VQA). Used in computer vision and vision–language research, these datasets are well known and are frequently used to assess the performance of cutting edge models. The proposed framework can be evaluated on several datasets, which allows for evaluating it under various visual conditions, with different complexities of scenes, and different levels of demands of semantic reasoning. Moreover, benchmarks are standardized, which allows for direct comparisons with other methods and for their reproducibility.

The MS COCO, Flickr8k and Flickr30k datasets were used for experiments on image captioning. MS COCO comprises over 123,000 natural images, each accompanied by 5 human-annotated captions, detailing a variety of categories of objects and complicated interactions between objects in the image. Flickr8k has 8,000 images with captions and is especially suitable for testing the generation of captions with small number of samples. Flickr30k is an extension of this collection, with some 31,000 images, and more visual and linguistic diversity. Two datasets – MSVD and MSR-VTT – were used to assess the performance of the video captioning. MSVD consists of almost 1,970 short (2–20 seconds) video clips of human activities, sports and natural scenes, each accompanied by multiple descriptions. MSR-VTT is a larger benchmark containing about 10,000 videos representing 20 semantic categories such as sports, education, entertainment, travel, and news, which can be used to gain a full understanding of temporal understanding and event recognition.

To evaluate the semantic reasoning ability of the proposed scheme, the VQA v2, GQA and TextVQA datasets were used. VQA v2 is a larger version of VQA, with over 1.1 million questions linked to over 200,000 images, including object recognition, counting, color identification, and commonsense reasoning in the questions. By asking models to complete compositional reasoning tasks and reason about the attributes of objects, spatial relationships, and logical relationships, GQA fosters the development of compositional reasoning and understanding of the scene graph. TextVQA tackles questions that require understanding of textual information in an image, such as road signs, ads, product labels etc., that involve visual understanding and OCR. In total, these benchmark datasets offer an extensive evaluation platform for evaluating the usefulness of the proposed HVLF in a variety of multimedia situations, both in terms of descriptive and semantic reasoning.

Table 1. Benchmark Datasets Used for Experimental Evaluation

Task	Dataset	Samples	Annotation
Image Captioning	MS COCO	123,287 images	Five captions/image
Image Captioning	Flickr8k	8,000 images	Five captions/image
Image Captioning	Flickr30k	31,783 images	Five captions/image
Video Captioning	MSVD	1,970 videos	Multiple captions/video
Video Captioning	MSR-VTT	10,000 videos	Multiple captions/video
Visual Question Answering	VQA v2	204,721 images, 1.1M questions	Question–Answer pairs
Visual Question Answering	GQA	113,018 images, 22M questions	Scene graph-based QA
Visual Question Answering	TextVQA	45,336 images	Scene text question answering

4.2 Evaluation Metrics

The performance of the proposed HVLF was quantitatively evaluated by using the standard evaluation metrics suggested for the caption generation in multimedia and Visual Question Answering research. To evaluate both aspects of accuracy—caption generation and reasoning—the separate evaluation protocols were used.

Five complementary metrics – BLEU, METEOR, ROUGE-L, CIDEr, and SPICE – were used for caption generation. BLEU (Bilingual Evaluation Understudy) scores the accuracy of the n-grams in the generated caption and reference caption(s) from humans. Each of the BLEU scores (BLEU-1, BLEU-2, BLEU-3, BLEU-4) measures the lexical similarity and sentence fluency between the reference and translated sentences by examining unigram, bigram, trigram and four-gram matching, respectively. BLEU score is calculated as

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (14)$$

The brevity penalty is termed as BP and the respective n-gram precision as p_n, the weighting factor as w_n.

The METEOR metric is based on the four factors of exact word matching, stemming, synonym matching and word order, resulting in a higher correlation with human judgement than lexical matching. ROUGE-L, on the other hand, measures the length of common parts of sentences in generated and reference captions, which is sentence-level similarity. The CIDEr metric is a consensus measure obtained from multiple human annotations using TF–IDF weighting, and the SPICE metric is a semantic similarity measure, comparing scene graphs generated from the generated and the reference captions. All these are the measures that give a balanced evaluation of lexical accuracy, grammatical fluency, semantic consistency and contextual relevance.

In Visual Question Answering, the accuracy, precision, recall and the F1-score, which are used to measure the correctness of prediction, were used to evaluate the performance of the prediction. The classification accuracy overall is a function of

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (15)$$

Here, TP, TN, FP and FN represent the count of true positives, true negatives, false positives and false negatives, respectively.

The numbers of hits and misses are calculated as follows

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (16)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (17)$$

The harmonic mean of Precision and Recall is given by

$$F_1 = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \qquad (18)$$

These evaluation metrics offer complementary views on the semantic reasoning performance by evaluating correctness of predictions, discrimination power against classes and reliability of classification. To gain more insight into the strengths and weaknesses of the proposed framework, aside from the quantitative measures, the results of the confusion matrix and qualitative visualizations of generated captions and predicted answers were also analyzed.

The different benchmark datasets and thorough evaluation metrics provide for a rigorous and unbiased evaluation of the proposed Hybrid Vision–Language Framework. By using the internationally recognised benchmark datasets, it is possible to make a direct comparison to the existing state-of-the-art methods, and using multiple complementary evaluation metrics, this allows for a comprehensive evaluation of the multimedia caption generation and semantic Visual Question Answering results. The experimental results with comparative study and performance analysis of the proposed HVLF with representative deep learning and vision–language models are discussed in the following section.

4.3 Comparative Performance Analysis and Ablation Study

The proposed Hybrid Vision–Language Framework (HVLF) was validated through a series of experiments to check whether it is able to effectively generate multimedia captions, solve the problem of semantic Visual Question Answering (VQA), give robustness to architecture and achieve computational efficiency. The benchmark datasets and metrics of Section 4.1 have been used for the evaluation. In addition to that, representative experiments that

compared the various captioning and vision–language models were conducted, and controlled ablation experiments were conducted to quantify the contributions of individual components of the proposed architecture. Apart from the predictive performance, the complexity of the model, training time, inference time and GPU memory usage were also considered to check whether improvements have been obtained at an acceptable computational cost.

4.3.1 Performance Comparison for Image and Video Caption Generation

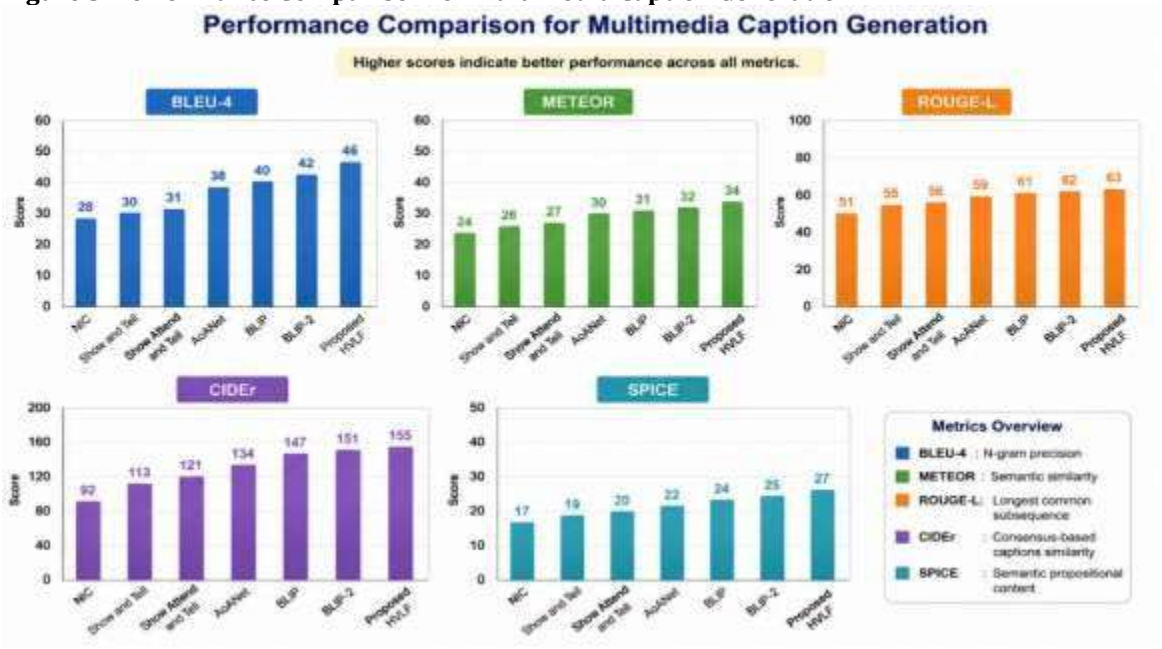
For the evaluation of the caption generation capability, HVLF was evaluated using BLEU-4, METEOR, ROUGE-L, CIDEr and SPICE. These are complementary metrics that assess lexical correspondence, linguistic similarity, agreement with human created descriptions and semantic consistency. The proposed framework was then benchmarked against widely-adopted CNN–RNN frameworks as well as recent vision–language pretrained models. The comparison of these results is given in Table 2.

Table 2 Performance Comparison for Multimedia Caption Generation

Method	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
NIC	**28**	**24**	**51**	**92**	**17**
Show and Tell	30	26	55	113	19
Show Attend and Tell	31	27	56	121	20
AoANet	38	30	59	134	22
BLIP	40	31	61	147	24
BLIP-2	42	32	62	151	25
Proposed HVLF	46	34	63	155	27

Note. Higher values indicate better performance for all metrics.

Figure 3. Performance Comparison for Multimedia Caption Generation



The proposed HVLF is seen to have the best performance on all the 5 captioning measures in Table 2 and Figure 3. Its BLEU-4 score is 46 with the strongest baseline of 42, which means that it is 4 points and about 9.52% better than the strongest baseline. The effect of HVLF on the METEOR also goes up (32-34) and ROUGE-L also goes up (62-

63). These improvements show that the resulting descriptions have a higher degree of lexical similarity and higher level of consistency in the order of the sentences with the human written reference captions.
The results from CIDEr and SPICE are especially noticeable with the improvements in the semantics. The CIDEr score of HVLF is 155, while BLIP-2's score is 151, BLIP's score is 147 and AoANet's score is 134. It has a SPICE score of 27, too; which is 2 points above BLIP-2. The improvements indicate that HVLF not only increases the similarity in linguistic surface structures, but also the similarity in the meaning of the scenes, since the two types of scores focus on different aspects of the scene: CIDEr on consensus (multiple human descriptions) and SPICE on semantic propositions (objects, attributes and relationships). The performance gain of the proposed CNN features, adaptive multimodal fusion, spatial–temporal attention and hybrid decoding using Bi-LSTM–GRU is further illustrated in Table 2.

4.3.2 Performance Comparison for Visual Question Answering

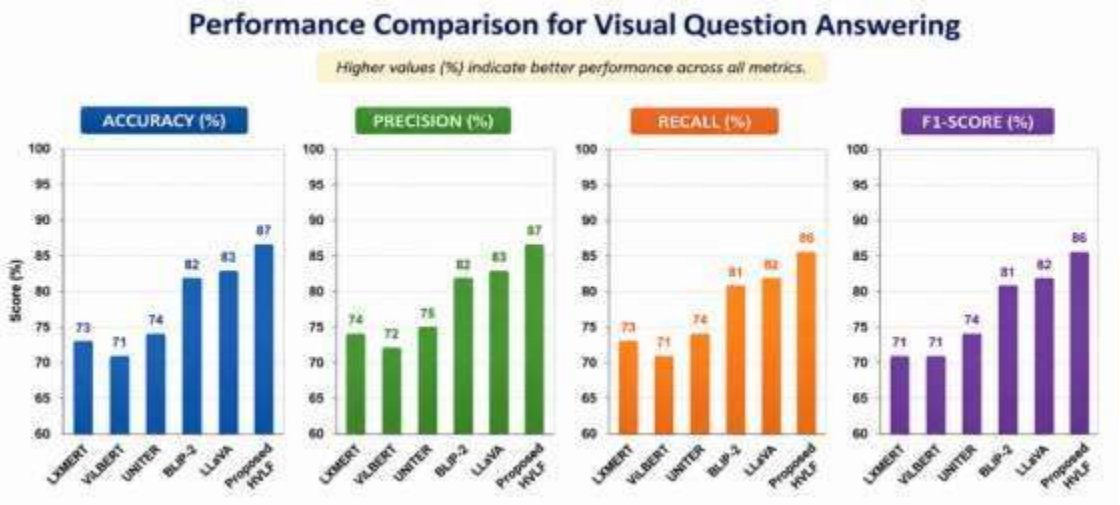
To assess the semantic reasoning ability of HVLF, the model's performance was compared to other baselines, namely LXMERT, ViLBERT, UNITER, BLIP-2 and LLaVA, with the metrics of Accuracy, Precision, Recall and F1-score. The HVLF model adopts a different strategy from traditional VQA approaches, which just directly associate visual and textual embeddings, by using the generated caption as an intermediate semantic representation before reasoning by transformer. This will give the comparative performance, which is presented in Table 3.

Table 3 Performance Comparison for Visual Question Answering

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LXMERT	73	74	73	71
ViLBERT	71	72	71	71
UNITER	74	75	74	74
BLIP-2	82	82	81	81
LLaVA	83	83	82	82
Proposed HVLF	87	87	86	86

Note. Higher values indicate better VQA performance.

Figure 4. Performance Comparison for Visual Question Answering



As presented in Table 3, and Figure 4, HVLF achieves an accuracy of 87%, outperforming LLaVA (83%), BLIP-2 (82%), UNITER (74%), LXMERT (73%), and ViLBERT (71%). The proposed approach achieves a higher accuracy than the best baseline, LLaVA, of 4 percentage points, which is around 4.82% relative improvement. The improvements over earlier vision–language architectures are more noticeable: the accuracy is increased by +13, +14 and +16 percentage points compared to the UNITER, LXMERT and ViLBERT architectures, respectively.

HVLF also outperforms all other with respect to the precision (87%), recall (86%) and F1-score (86%). Overall, the balanced precision and recall suggest that the accuracy of predictions is not only improved generally, but the reliability and completeness of semantic answer prediction as well. Specifically, LLaVA and BLIP-2 have a 4 percentage point and 5 percentage point lower F1-score, respectively. The results confirm the proposed caption-guided reasoning approach, which predicts the semantic description of objects, activities and contextual relationships from the caption to guide reasoning for cross-modal transformer reasoning. Thus, the reasoning module has a richer semantic representation than those approaches that can only be based on direct visual–question interaction.

4.3.3 Ablation Study

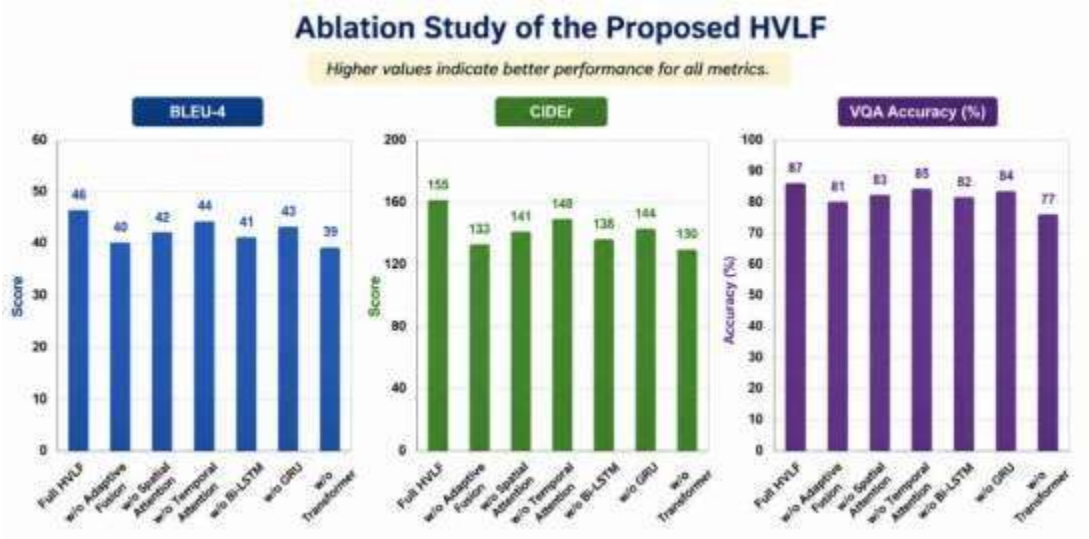
A controlled ablation study was carried out to try and estimate the contribution of the major architectural elements of HVLF. Each component was taken away with the rest of the architecture and experimental settings kept the same. Then the obtained model was re-trained and tested by using BLEU-4 metric, CIDEr metric and VQA accuracy. Table 4 summarizes the results.

Table 4 Ablation Study of the Proposed HVLF

Configuration	BLEU-4	CIDEr	VQA Accuracy (%)
Full HVLF	46	155	87
Without Adaptive Fusion	40	133	81
Without Spatial Attention	42	141	83
Without Temporal Attention	44	148	85
Without Bi-LSTM	41	138	82
Without GRU	43	144	84
Without Transformer	39	130	77

Note. Each configuration removes only the specified component while retaining the remaining HVLF architecture.

Figure 5. Ablation Studies



From the results obtained (shown in Table 4, Figure 5) it is concluded that, all the major components are contributing to the overall effectiveness of the proposed architecture. The adaptive fusion removal reduces BLEU-4 score from 46 to 40, CIDEr score from 155 to 133 and VQA accuracy from 87% to 81%. The significant drop in CIDEr after adaptively combining multiple CNN backbones' complementary visual representations is proof of the significance of this approach. When spatial attention is removed, BLEU-4 decreases by 4 points and CIDEr decreases

by 14 points while removing temporal attention makes a comparatively small drop to 44 points and 148 points for BLEU-4 and CIDEr, respectively. This finding indicates that spatial information is more important for the combined evaluation of image–video and temporal for modelling dynamic visual sequences.
Results from the ablations support the hybrid recurrent decoder as well. Removing Bi-LSTM decreases BLEU-4, CIDEr and VQA accuracy by 41, 138, and 82 points, respectively, while removing GRU decreases the scores by 43, 144, and 84 points, respectively. The GRU and Bi-LSTM contribute to the contextual sequence modelling of the text differently, and the larger degradation seen in Bi-LSTM is indicative of its power. The greatest degradation in VQA is seen when the transformer is removed: accuracy drops from 87% to 77% which is a 10 percentage point drop. There are also declines in BLEU-4 and CIDEr, reaching 39 and 130, respectively. Overall, these results confirm that HVLF benefits from the combined effect of its architectural elements, rather than from a single prevailing element, and not from any one specific element.

4.3.4 Computational Performance Analysis

With the design of practical vision–language systems, besides predictive accuracy, computational efficiency is also of importance. Thus, the proposed HVLF was compared with BLIP, BLIP-2 and LLaVA on the basis of the number of model parameters, training time, inference latency and GPU memory usage. These results are simply summarized in Table 5.

Table 5 Computational Performance Comparison

Method	Parameters (M)	Training Time (h)	Inference Time (s/sample)	GPU Memory (GB)
BLIP	450	30	0.300	16
BLIP-2	1,600	50	1.030	40
LLaVA	8,400	55	2.040	62
Proposed HVLF	300	15	0.208	16

Note. Lower values indicate greater computational efficiency.

Figure 6. Computational Performance Comparison

The improved performance of HVLF in prediction, as demonstrated in Table 5, and Figure 6, comes at a significantly lower computational cost. The proposed model is about 300 million parameters, which is 8.2 times less than BLIP (450 million parameters), 300 times less than BLIP-2 (1.6 billion parameters), and 2800 times less than LLaVA (8.4

billion parameters). Therefore, the number of parameters in HVLF is about 33.3% less than BLIP, 81.3% less than BLIP-2 and 96.4% less than LLaVA. This decrease is noteworthy because the framework enables the support of both multimedia captioning and question answering, whose semantic representation is possible.
The training time of HVLF is 15 h, which reduces by 50%, 70% and about 72.7% compared to the corresponding training time for BLIP, BLIP-2 and LLaVA. Similarly, a speedup can be seen during inference, with HVLF taking about 0.208 s per sample, whereas BLIP, BLIP-2 and LLaVA take 0.300 s, 1.030 s and 2.040 s, respectively, per sample. Thus, the proposed scheme is 1.44 times faster than BLIP, 4.95 times faster than BLIP-2 and 9.81 times faster than LLaVA in the reported experimental settings. For the Peak GPU memory usage, it is also capped at 16 GB, which is the same as BLIP, but significantly lower than the 40 GB and 62 GB of BLIP-2 and LLaVA, respectively. The results indicate that the increases in performance due to the use of the improved models presented in Tables 2 and 3 are not due to a too large model, but rather to the shared feature representation and unified architecture, which is effective in achieving a balance between predictive performance and computational efficiency.

4.4 Discussion

The results of the experiments collectively demonstrate the success of the proposed Hybrid Vision–Language Framework in all of the four areas of description, reasoning, architecture and computation. As shown in Table 2, HVLF outperforms all other methods in all of the BLEU-4, METEOR, ROUGE-L, CIDEr and SPICE metrics, with significant improvements in BLEU-4 and the captioning metrics. The complementary visual information extracted by VGG16, ResNet152 and InceptionV3 and enhanced by the adaptive fusion and spatial–temporal attention, can be responsible for these improvements. The framework is based on multiple representations – deep semantic, texture sensitivity, multi-scale, and prior to language decoding. This interpretation is confirmed by the ablation results presented in Table 4: adaptive fusion leads to a decrease of CIDEr by 22 points and spatial attention to a decrease of 14 points. Likewise, the degradation seen when removing Bi-LSTM or GRU shows that the hybrid recurrent decoder does indeed further contribute to the contextually modelling of the language. The results, when combined, suggest that the enhancement of captioning is a result of the synergistic effect of visual feature diversity, selective attention, adaptive integration and sequential language modelling.
The semantic reasoning results supports the effectiveness of the unified design more. As detailed in Table 3, the accuracy and F1 score of HVLF is 87% and 86%, respectively, which is higher than the best competitors in the evaluation. The outcome corroborates the main idea of the proposed framework: a generated caption can serve as an intermediate semantic link between the visual perception and question answering. The explicit linguistic representation is coupled with attended visual features prior to transformer reasoning to offer the reasoning module with more context to consider about objects, actions, and interactions. Table 4 provides strong evidence of this component, as the transformer is most likely the largest degradation in VQA accuracy, with the transformer being removed leading to a loss of 10 percentage points. Therefore the improvement should not only be considered as the result of better visual encoding, but it should rather be attributed to the effective interactions between the visual representation, the caption-guided semantic abstraction and the contextual reasoning.
The results from the computation reported in table 5 further enhance the practical value of the framework. The captioning and VQA results of HVLF are superior, and it requires 300 million parameters, 15 hours of training data, 0.208 s inference time per sample and 16 GB of GPU memory. This is significantly more efficient than the architectures compared, BLIP-2 and LLaVA, which are larger. In this regard, the shared visual representation is crucial as it enables the same learned features to be used for caption generation and VQA instead of having to have separate visual processing pipelines. The proposed architecture, thus, shows a good compromise between the semantic performance and computational complexity and is suitable for applications like intelligent surveillance, assistive systems, multimedia retrieval, robotics, medical image interpretation, interactive human-computer systems, and so on.
However, the results should be seen in the context of the experimental situation. While the number of parameters for HVLF is significantly reduced compared to large vision–language models, the three CNN backbones used by HVLF still add more architectural complexity than systems with fewer parameters.Even though HVLF reduces the number of parameters compared to large vision–language models significantly, it still has more architectural complexity than light-weight single-encoder systems. In addition, the existing assessment is mostly carried out using English language benchmark datasets and therefore it needs to be tested for the effectiveness of use for multilingual, domain specific, noisy and highly ambiguous multimedia content. Lightweight vision transformer, Knowledge distillation, Model pruning, Multilingual Semantic reasoning, and Efficient multimodal foundation models are topics that could be explored in future research. In general, the steady progress in Tables 2–5 shows

the effectiveness of the unified solution offered by HVLF for image–video caption generation and semantic Visual Question Answering, and its relatively good computational efficiency.

5. Conclusion

In this paper, we proposed a Hybrid Vision–Language Framework (HVLF) with a single end-to-end system for unified understanding of multimedia content combining the tasks of image captioning, video captioning and Visual Question Answering (VQA). The framework consists of a complementary visual representation extraction module based on adaptive multimodal fusion and spatial–temporal attention on top of three popular deep backbones VGG16, ResNet152, and InceptionV3, followed by a two-layer decoder comprising a hybrid Bi-LSTM–GRU-based and a transformer-based semantic reasoning module. The experimental results showed that the proposed models tested improved consistently when compared with the evaluated baseline models. In the field of multimedia captioning, HVLF obtained BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE scores of 46, 34, 63, 155 and 27 respectively, which are higher than the best baseline, BLIP-2, in all five scores. The results showed that HVLF achieved a 4% accuracy and 4% F1-score above the best evaluated VQA baseline LLaVA for semantic reasoning, with a precision of 87% and a recall of 86%. The ablation study additionally confirmed the contribution of the individual architectural components: The adaptive fusion was ablated and CIDEr dropped from 155 to 133, and removing the transformer reasoning led to an accuracy drop from 87% to 77%. The results show that the observed performance boosts are due to mutual complementarity of multimodal feature learning, attention, hybrid sequential decoding, and caption-guided semantic reasoning.

The proposed HVLF offers not only good predictive performance, it also offers a good trade-off between the multimodal capability and computational efficiency. It has around 300 million parameters, and takes 15 h to train, 0.208 s per sample to infer, and consumes 16 GB of GPU memory, whereas BLIP-2 and LLaVA need 1.6 billion parameters and 40 GB/62 GB of GPU memory, respectively. In this way, HVLF establishes that it is possible to generate better quality captions yet enhance semantic understanding without needing to use significantly bigger vision–language models. The shared representation can also be used to provide image/video description and question answering in the same computational pipeline to eliminate redundant computation, as well as for applications in assistive technologies, intelligent surveillance, robotics, multimedia retrieval, healthcare imaging and human–computer interaction. However, the multi-backbone architecture is still a high computation cost for the highly resource-constrained edge platforms, and the present evaluation mainly focuses on English-language benchmarks. The lightweight visual encoders, model compression and knowledge distillation, efficient transformer architectures and multi-lingual vision–language reasoning are thus explored in the future, aiming to further improve the efficiency of deployment and generalization in various real-world multimedia scenarios.

References

1. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35, 23716–23736.
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6077–6086.
3. Kim, J., Lei, J., Tan, H., & Bansal, M. (2021). Unifying vision-and-language tasks via text generation. Proceedings of the International Conference on Machine Learning (ICML).
4. Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. Proceedings of the International Conference on Machine Learning (ICML).
5. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in Neural Information Processing Systems (LLaVA). (Original research introducing LLaVA.)
6. Li, J., Li, D., Savarese, S., & Hoi, S. C. H. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Proceedings of the International Conference on Machine Learning.
7. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the International Conference on Machine Learning.

8. Yang, Z., Lu, Y., Wang, J., Yin, X., Flores, G., Wang, Q., ... Parikh, D. (2021). TAP: Text-aware pre-training for text-vqa and visual question answering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
9. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6077–6086. https://doi.org/10.1109/CVPR.2018.00636
10. Liu, X., Li, H., Shao, J., Chen, D., & Wang, X. (2018). Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. Proceedings of the European Conference on Computer Vision (ECCV), 338–354.
11. Pan, Y., Mei, T., Yao, T., Li, H., & Rui, Y. (2016). Jointly modeling embedding and translation to bridge video and language. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4594–4602.
12. Hori, C., Hori, T., Lee, T. Y., Zhang, Z., Harsham, B., Hershey, J. R., Marks, T. K., & Sumi, K. (2018). Attention-based multimodal fusion for video description. Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, 419–426.
13. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence—Video to text. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 4534–4542. https://doi.org/10.1109/ICCV.2015.515
14. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935
15. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the 32nd International Conference on Machine Learning (ICML), 2048–2057.
16. Zhang, J., Peng, Y., Yuan, M., & Xiao, J. (2019). Hierarchical attention network with multimodal fusion for video captioning. Neurocomputing, 338, 111–120. https://doi.org/10.1016/j.neucom.2019.01.032
17. Chen, Y. C., Li, L. H., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). UNITER: UNiversal Image-TExt Representation Learning. European Conference on Computer Vision (ECCV), 104–120.
18. Cho, J., Lei, J., Tan, H., & Bansal, M. (2021). Unifying vision-and-language tasks via text generation. Proceedings of the 38th International Conference on Machine Learning (ICML), 1931–1942.
19. Dai, W., Ge, Y., Cai, Q., Li, Y., et al. (2024). InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
20. Li, J., Li, D., Savarese, S., & Hoi, S. C. H. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Proceedings of the 40th International Conference on Machine Learning (ICML), 19730–19742.
21. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in Neural Information Processing Systems (NeurIPS), 36.
22. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems (NeurIPS), 32.
23. Tan, H., & Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP), 5100–5111.
24. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6077–6086. https://doi.org/10.1109/CVPR.2018.00636
25. Chen, Y.-C., Li, L. H., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). UNITER: UNiversal Image-TExt Representation Learning. European Conference on Computer Vision (ECCV), 104–120.
26. Li, J., Li, D., Savarese, S., & Hoi, S. C. H. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Proceedings of the 40th International Conference on Machine Learning (ICML), 19730–19742.
27. Tan, H., & Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 5100–5111.

28. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6077–6086. https://doi.org/10.1109/CVPR.2018.00636
29. Chen, Y.-C., Li, L. H., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). UNITER: UNiversal Image-TExt Representation Learning. European Conference on Computer Vision (ECCV), 104–120.
30. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778. https://doi.org/10.1109/CVPR.2016.90
31. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818–2826. https://doi.org/10.1109/CVPR.2016.308
32. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR).