Journal of Scientific & Industrial Research
Vol. 84, December 2025, pp. 1322-1330
DOI: 10.56042/jsir.v84i12.23517

Heterogeneous Ensemble Learning for Context-Aware Image Captioning with Transformers

Kothakonda Chandhar* & Manchala Sadanandam

Computer Science & Engineering, Kakatiya University, Warangal 506 009 Telangana, India

Received 17 September 2025; revised 26 December 2025; accepted 30 December 2025

Image captioning remains a central challenge in multimodal artificial intelligence, requiring systems to jointly reason over visual content and natural language. Despite remarkable progress from deep vision–language transformers, single-model architectures often face a trade-off: they excel in either syntactic fluency or semantic grounding but rarely achieve both. This work introduces a heterogeneous ensemble learning framework that unifies convolutional, hierarchical, and self-attention–based encoders (ConvNeXt, ResNet-101, ViT) with advanced language decoders (T5 and BLIP). Unlike prior captioning ensembles, the current approach integrates attention-guided feature fusion with a consensus re-ranking mechanism, enabling the system to adaptively combine complementary strengths of diverse models. The framework is evaluated on two challenging benchmarks—MS COCO 2017 and Flickr30K—achieving state-of-the-art improvements over strong baselines, with BLEU-4 = 37.2, CIDEr = 124.5, SPICE = 22.3 on COCO, and BLEU-4 = 30.8, CIDEr = 98.7, SPICE = 19.6 on Flickr30K. Beyond quantitative gains, qualitative analysis shows that the ensemble produces captions that are both contextually faithful and semantically rich. These results establish ensemble learning as a scalable paradigm for vision–language generation, with implications for multilingual captioning, real-time accessibility tools, and future general-purpose multimodal reasoning systems.

Keywords: Attention mechanism, Multimodal fusion, Natural language generation, Vision–language modeling, Visual semantics

Introduction

Image captioning is the task of automatically producing a text description for an image. It is important for applications such as helping visually impaired users, improving image search, and supporting human–AI communication. Although many models have been proposed, creating captions that are both accurate and meaningful is still difficult. Most systems perform well either in fluency or in capturing details, but not both at the same time.

Early research followed the encoder–decoder design. Convolutional neural networks (CNNs) were used to extract features, and recurrent networks such as LSTMs generated sentences. Vinyals *et al.*[1] introduced the *Show and Tell* model, while Xu *et al.*[2] improved it with visual attention to highlight important regions. These methods worked, but the captions were often simple and lacked context.

Later works added new ideas to improve quality. Adaptive attention[4], spatial–channel attention[5], and reinforcement learning with self-critical training[6] made captions more specific and aligned with evaluation metrics. Anderson *et al.*[7] further improved performance by combining object detection with top-down attention.

The shift to Transformers brought another step forward. Transformers can model long-range dependencies better than recurrent networks. Models such as Entangled Transformer[8], Meshed-Memory Transformer[9], X-Linear attention[10], and RSTNet[11] showed stronger multimodal reasoning. Large-scale pretraining also helped: OSCAR[12], VinVL[13], and AoANet[14] used aligned image–text pairs and stronger features to push captioning results higher.

However, two issues remain. First, most models rely on a single architecture. CNNs capture local detail, Vision Transformers capture global structure, and pretrained language models like T5 provide strong fluency, but no single model balances all of these. Second, only a few works have tried ensemble methods. Cornia *et al.*[9] combined Transformer outputs, and Al Badarneh *et al.*[3] applied an ensemble to small datasets. These studies showed some benefit, but they did not explore diverse combinations of modern models.

*Author for Correspondence
E-mail: chandu19024@gmail.com

At the same time, new large multimodal systems such as BLIP-2[19] and LLaVA[20] show impressive results by combining vision and language at scale. Still, they are very large, expensive to run, and not designed specifically for image captioning tasks. This leaves an open question: can ensemble learning provide a practical way to combine different models for captioning, while staying efficient and task-focused?

The current study proposes a heterogeneous ensemble framework for image captioning that addresses these gaps:

- Propose a heterogeneous ensemble that combines ConvNeXt, ResNet-101, and Vision Transformer encoders with two decoders, T5 and BLIP.
- Introduce an attention-based fusion step and a consensus re-ranking method to select the best captions.
- Evaluate the framework on MS COCO 2017 and Flickr30K datasets, showing clear improvements across BLEU, METEOR, ROUGE-L, CIDEr, and SPICE compared with baselines.
- Provide ablation and qualitative studies to show how the ensemble captures more detail and produces richer captions than single models.

This work shows that ensemble learning can be an effective and scalable approach for image captioning, offering captions that are both accurate and contextually meaningful.

Literature Survey

Image captioning combines computer vision and natural language processing to generate textual descriptions of images. The first a successful model the encoder–decoder framework, in which convolutional neural networks (CNNs) extracted capabilities and recurrent neural networks (RNNs) together with LSTMs generated sentences. Vinyals *et al.*[1] presented the "Show and Tell" model, setting up a robust baseline for this method, at the same time as Xu *et al.*[2] progressed it by way of including visible interest to cognizance on crucial image regions. These models confirmed the capacity of deep captioning systems however often produced quick and universal captions lacking contextual intensity.

To solve these limits, researchers explored more superior attention mechanisms. Lu *et al.*[4] proposed adaptive interest to determine whilst to cognizance on image features, Chen *et al.*[5] added SCA-CNN to model both spatial and channel-clever cues, and Rennie *et al.*[6] implemented reinforcement studying via Self-Critical Sequence Training (SCST) to at once optimize assessment metrics. Anderson *et al.*[7] blended object detection with top-down interest, producing captions that have been better grounded in visual content. These works highlighted that interest and task-specific optimization could enhance caption quality, but they however relied heavily on hand made designs and restrained visible representations.

The beginning of Transformers brought a major change. Unlike RNNs, Transformers can capture lengthy dependencies more effectively. Li *et al.*[8] evolved the Entangled Transformer to connect visual and semantic streams, Cornia *et al.*[9] brought Meshed-Memory Transformers with memories for better context modeling, Pan *et al.*[10] proposed X-Linear interest to analyze bilinear interactions, and Zhang *et al.*[11] designed RSTNet to stability interest throughout visible and non-visible tokens. These fashions showed more potent multimodal reasoning, however they typically required big datasets and nevertheless struggled with repetitive phrasing or overfitting to frequent styles.

Large-scale pretraining of vision–language models (VLMs) further advanced the field. Li *et al.*[12] presented OSCAR, which aligned object tags with text during pretraining, while Zhang *et al.*[13] extended this idea with VinVL to leverage stronger image features. Huang *et al.*[14] introduced AoANet with "attention on attention" mechanisms, Tan *et al.*[15] proposed a context-aware Transformer, and Zhou *et al.*[16] suggested mutual learning to reduce generic captions. More recent studies such as Xu *et al.*[17] on local visual modeling and Khan *et al.*[18] on interpretability confirmed the benefits of pretraining, attention refinement, and explainability. However, these models rely on large-scale resources and do not directly address efficiency or robustness in practical captioning systems.

Despite these advances, ensemble methods for image captioning remain limited. Cornia *et al.*[9] reported minor gains by combining multiple Transformer outputs, while Al Badarneh *et al.*[3] proposed an attention-based ensemble but only tested on small datasets such as Flickr8K. These attempts suggest potential, but they mostly used homogeneous architectures and did not exploit the complementary strengths of diverse encoders and decoders.

More recently, general-purpose multimodal systems have emerged. BLIP-2[(19)] combines frozen

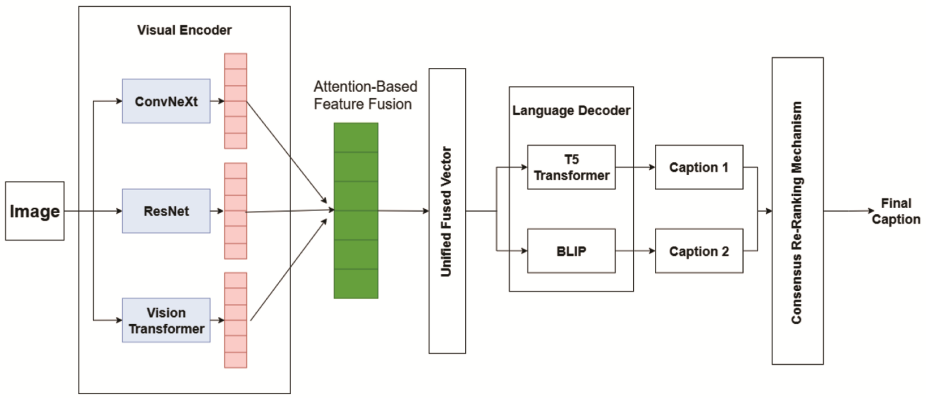

Fig. 1 — Proposed ensemble framework for image captioning

image encoders with large language models to achieve strong zero-shot performance, while LLaVA[20] uses instruction tuning to support conversational multimodal reasoning. Although powerful, these models are computationally heavy, data-hungry, and not specialized for benchmark captioning tasks such as MS COCO and Flickr30K.

In summary, existing research shows steady progress from CNN–RNN models to Transformers and pretrained VLMs, but most approaches rely on a single architecture. Ensemble-based methods are still underexplored, especially those that combine heterogeneous encoders (e.g., CNNs, ConvNeXt, and Vision Transformers) with modern decoders (e.g., T5, BLIP). This gap motivates the need for a systematic ensemble framework that can leverage complementary strengths to produce captions that are accurate, fluent, and contextually rich.

Methodology

The proposed framework introduces a heterogeneous ensemble-based image captioning which integrates multiple visual encoders with transformer-based decoders. Unlike conventional single-model captioning systems, the ensemble layout leverages each convolutional and transformer architectures to take advantage of complementary functions. The entire pipeline includes dataset schooling, preprocessing, visible function extraction, hobby-based absolutely characteristic fusion, caption technology the use of a couple of decoders, and consensus-based re-ranking. The overall architecture is depicted in Fig.1, which shows the flow from input image to final caption**.**

Dataset Description and Partitioning

The experiments are done on the MS COCO 2017 and Flickr30K datasets. Both datasets provide 5 reference captions for each image, this is important for strong caption generation. MS COCO consists of a broadly large form of samples in evaluation to Flickr30K, making it suitable for schooling deep ensembles. The dataset splits are shown in Table 1.

Table 1 — Dataset partitioning for training, validation, and testing

Dataset	Training images	Validation images	Testing images	Captions per image
MS COCO 2017	118,287	5,000	5,000	5
Flickr30K	29,000	1,000	1,000	5

The datasets provide sufficient diversity in object categories, backgrounds, and caption styles, making them suitable benchmarks for ensemble captioning research.

Preprocessing

***Image Preprocessing*:** Each inputted image is resized to 224×224 pixels to ensure compatibility with backbone networks. Pixel values are normalized the usage of ImageNet mean and variance. This guarantees strong convergence all through education and lets in switch getting to know from pretrained fashions.

***Text Preprocessing*:** Captions are tokenized using SentencePiece, converted to lowercase, and truncated to a maximum length of 30 tokens. A vocabulary is constructed with a minimum word frequency threshold of 5. Special tokens <SOS> and <EOS> are appended to mark the start and end of each caption. The preprocessing step transforms a caption: $C = \{w_1, w_2, \dots, w_T\}$ into indexed tokens: $\{y_1, y_2, \dots, y_T\}, y_t \in V$ where **V** is the vocabulary set.

Visual Feature Extraction

Three heterogeneous encoders are employed to capture complementary representations:

ConvNeXt-B: which refines convolutional hierarchies for improved spatial learning,

***ResNet-101**:* which captures region-based semantic features, and

***Vision Transformer (ViT-B/16)*:** which models long-range dependencies through self-attention.

For an image **I**, each encoder $\mathbf{e_i}$ maps the input into a feature representation:

$$f_i = e_i(I), \quad f_i \in R^{d_i} \quad \ldots (1)$$

where, d_i denotes the dimensionality of the encoder output. These features are further projected into a common latent space using a learnable transformation:

$$\tilde{f}_i = W_i f_i, \quad \tilde{f}_i \in R^d \quad \ldots (2)$$

where, $\mathbf{W_i}$ is a linear projection matrix. This ensures that heterogeneous encoder features are compatible for fusion.

Attention-Based Feature Fusion

The projected features are fused using a weighted attention mechanism. The fused representation is expressed as:

$$F = \sum_{i=1}^{n} \alpha_i \tilde{f}_i, \alpha_i = \frac{\exp(\beta_i)}{\sum_{k=1}^{n} \exp(\beta_k)} \quad \ldots (3)$$

where, α_i shows how much importance is given to each encoder output. It helps the model focus on the most useful information from different encoders. The combined feature **F** is then used as a single visual representation for generating captions.

The value β_i is a trainable parameter for each encoder. These values are learned during training along with the rest of the model. A softmax function is applied to β_i to obtain the attention weights α_i. This process allows the fusion module to adjust the contribution of each encoder based on its usefulness.

Caption Generation with Multiple Decoders

The fused characteristic vector F is fed into exclusive decoders: (i) T5 Transformer, that is pretrained on large-scale text-to-text tasks and fine-tuned for caption generation, and (ii) BLIP, which aligns image and textual content representations through bootstrapped imaginative and prescient–language pretraining. For every decoder j, the conditional chance of producing phrase w_t at time step t is described as:

$$P_j(w_t|F, w_{1:t-1}) = \text{Decoder}_j(F, w_{1:t-1}), \quad j \in \{1,2\} \ldots (4)$$

This formulation ensures that captions are generated based on both visual content and previously predicted tokens.

Model-Level Ensemble with Consensus Re-Ranking

Each decoder generates one caption for the given image. The final caption is then chosen using a consensus strategy, which works differently during training/validation and during inference.

During training and validation, ground-truth reference captions are available. In this phase, standard evaluation metrics such as CIDEr, BLEU-4, and SPICE are used for offline analysis and to guide the reinforcement learning process. These metrics help the model learn captions that are fluent, meaningful, and well aligned with human-written descriptions.

When reference captions are available, a combined score is used only for analysis. This score is computed using a weighted sum of CIDEr, BLEU-4, and SPICE:

$$S(c) = \lambda \cdot \text{CIDEr}(c) + \mu \cdot \text{BLEU-4}(c) + \nu \cdot \text{SPICE}(c) \quad \ldots (5)$$

where, λ, μ, ν are adjustable weights. This score is used only during training and validation, when ground-truth captions are present, and it is not used during inference.

During inference on new images, reference captions are not available, so evaluation metrics cannot be computed. In this case, the final caption is selected using a confidence score based on the decoder's own output probabilities. For each generated caption c_j a normalized log-likelihood score is calculated as:

$$S(c_j) = \frac{1}{|c_j|} \sum_{t=1}^{|c_j|} \log P\left(w_t \mid w_{<t}, F\right) \quad \ldots (6)$$

The caption with the highest confidence score is selected as the final output:

$$C^* = \arg\max_{c \in \{C_1, C_2\}} S(c) \quad \ldots (7)$$

This approach allows caption selection during inference without using reference captions, making the method practical and reproducible, while still benefiting from metric-based optimization during training. The complete step wise procedure of this frameworks is outlined in Table 2.

Training Objective

The model is optimized in two stages. First, supervised learning is conducted using cross-entropy loss:

$$\mathcal{L}_{\mathcal{XE}} = -\sum_{t=1}^{T} \log P\left(y_t|F, y_{1:t-1}\right) \quad \ldots (8)$$

This ensures the model learns to predict the correct token at each step. Next, reinforcement learning with

Table 2 — Algorithm for heterogeneous ensemble framework for image captioning

Algorithm
Input: Image I
Output: Final Caption C*
// Visual Feature Extraction
Extract features f1 = ConvNeXt(I)
Extract features f2 = ResNet(I)
Extract features f3 = ViT(I)
// Projection into Common Latent Space
For each fi in {f1, f2, f3} do
Compute $\tilde{f}i = Wi * fi$
End for
// Attention-Based Feature Fusion
Compute attention weights:
$\alpha i = \exp(\beta i) / \Sigma k \exp(\beta k)$
Compute fused visual embedding:
$F = \Sigma i\ \alpha i * \tilde{f}i$
// Caption Generation with Multiple Decoders
Generate candidate caption c1 and token probabilities using Decoder_T5(F)
Generate candidate caption c2 and token probabilities using Decoder_BLIP(F)
// Consensus-Based Caption Selection
If reference captions are available (training / validation) then
// Metric-based scoring (used only for optimization and analysis)
For each candidate caption cj in {c1, c2} do
Compute score:
$S(cj) = \lambda \cdot CIDEr(cj) + \mu \cdot BLEU\text{-}4(cj) + \nu \cdot SPICE(cj)$
End for
Else
// Inference-time selection (reference-free)
For each candidate caption cj in {c1, c2} do
Compute confidence score:
$S(cj) = (1 / |cj|)\ \Sigma t \log P(wt \mid w<t, F)$
End for
End if
Select final caption:
C* = argmaxcj S(cj)
Return C*

Self-Critical Sequence Training (SCST) is applied, directly optimizing the CIDEr metric:

$$\mathcal{L}_{SS} = -\big(r(\hat{y}) - r(y^s)\big) \sum_{t=1}^{T} \log P\left(y_t^s | F, y_{1:t-1}^s\right) \quad \ldots (9)$$

where, $\hat{y}$ is the baseline greedy caption, y^s is a sampled caption, and r(·) denotes the reward function.

$$\mathcal{L} = \mathcal{L}_{\mathcal{XE}} + \gamma \mathcal{L}_{\mathcal{SCST}} \quad \ldots (10)$$

where, γ is a balancing coefficient.

Experimental Setup

Architectures and Versions

This study employed three heterogeneous visual encoders: ConvNeXt-B pretrained on ImageNet-22K, ResNet-101 pretrained on ImageNet-1K, and ViT-B/16 on ImageNet-21K. For language decoders, T5-base from HuggingFace Transformers v4.40 and BLIP-base are used. Tokenization relies on SentencePiece with a 32k vocabulary for T5 and the BPE tokenizer packaged with BLIP.

Hyperparameters

All models were optimized using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0.01). The initial learning rate was 3×10^{-5}, scheduled with cosine decay and 1,000 warm-up steps. Mini-batch size was 64 images per GPU. Dropout was set to 0.1, gradient clipping to 1.0, and training lasted for 30 epochs. Maximum input resolution was 384×384 pixels for ConvNeXt and ResNet, and 224 × 224 for ViT. Captions were truncated or padded to 30 tokens. Random seeds were fixed to 42, 123, and 456 to ensure reproducibility.

Training Procedure

All models were trained and evaluated on the MS COCO 2017 dataset (Karpathy split: 113k train / 5k validation / 5k test) and the Flickr30K dataset (29k train / 1k validation / 1k test). Images were resized to each encoder's input resolution and normalized to ImageNet statistics. Data augmentation included random horizontal flips and random cropping. Captions were preprocessed by lowercasing and punctuation removal.

Hardware and Runtime

Experiments were conducted on 4 × NVIDIA A100 GPUs (40 GB each) using PyTorch 2.1 with CUDA 12.0. Training ConvNeXt–T5 required approximately 32 GPU-hours, BLIP required 28 GPU-hours, and the full ensemble required 85 GPU-hours. Peak memory consumption was ~32 GB per GPU. Inference throughput averaged 85 ms per image for single models and 210 ms per image for the ensemble.

Evaluation Metrics

To examine the quality of generated captions, this study employed extensively universal automatic metrics that compare system outputs towards human reference captions. These metrics capture distinctive components together with lexical overlap, semantic similarity, fluency, and consensus with more than one references. The use of multiple assessment measures guarantees a fair and comprehensive evaluation of caption excellent.

BLEU

BLEU (Bilingual Evaluation Understudy) measures the precision of n-grams between generated and

reference captions. It penalizes short sentences through the brevity penalty (BP). BLEU is especially effective in assessing local syntactic correctness.

$$\text{BLEU-N} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad \ldots (11)$$

where, p_n represents the modified n-gram precision, w_n is the weight (typically uniform), and N is the maximum n-gram length. BLEU-1 to BLEU-4 is reported to capture unigram through 4-gram precision.

METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) considers both precision and recall, unlike BLEU which is precision-biased. It aligns generated captions with reference captions using exact word matching, stemming, and synonyms from WordNet. The metric balances fluency with semantic flexibility.

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty}) \quad \ldots (12)$$

where, F_{mean} is the harmonic mean of precision and recall, and *Penalty* reduces the score for fragmented matches. METEOR is highly correlated with human judgment, making it suitable for caption evaluation.

ROUGE-L

ROUGE-L evaluates captions based on the Longest Common Subsequence (LCS) between generated and reference sentences. Unlike n-gram based metrics, LCS captures sentence-level fluency and semantic order preservation.

$$\text{ROUGE-L} = \frac{(1+\beta^2)\cdot \text{R}\cdot \text{P}}{\text{R}+\beta^2 \text{P}} \quad \ldots (13)$$

where, R is recall, P is precision, and β determines the relative importance of recall. ROUGE-L is particularly effective in measuring structural similarity between system outputs and human captions.

CIDEr

CIDEr (Consensus-based Image Description Evaluation) measures the consensus between a generated caption and multiple reference captions using TF-IDF weighted n-grams. It accounts for the frequency of informative words while down-weighting common but uninformative ones.

$$\text{CIDEr}(c_i, S_i) = \frac{1}{m}\sum_{j=1}^{m} \frac{g_n(c_i)\cdot g_n(s_{ij})}{|g_n(c_i)||g_n(s_{ij})|} \quad \ldots (14)$$

where, g^n denotes the TF-IDF vector of n-grams, a_i is the candidate caption, and s_{ij} are the reference captions. CIDEr is considered the primary optimization target for reinforcement learning in captioning tasks.

SPICE

SPICE (Semantic Propositional Image Caption Evaluation) evaluates captions by parsing them into scene graphs, capturing objects, attributes, and relationships. Unlike lexical metrics, SPICE directly measures semantic agreement with human annotations.

$$\text{SPICE}(c, S) = F_1 \text{ score between scene graphs of } c \text{ and references } S \ldots (15)$$

This metric strongly correlates with human assessment of semantic quality, though it is computationally expensive. SPICE complements BLEU and CIDEr by focusing on meaning rather than surface form.

Results and Discussion

The performance of the proposed ensemble framework was assessed on the MS COCO 2017 and Flickr30K datasets using the evaluation metrics. To ensure a fair comparison, all models were trained with identical preprocessing and optimization strategies. The proposed ensemble approach is compared against strong baselines, including ConvNeXt–T5, BLIP, and GIT.

Quantitative Results

The quantitative outcomes are provided in Table 3. It may be located that the proposed ensemble

Table 3 — Quantitative results on MS COCO and Flickr30K

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
MS COCO	ConvNeXt–T5	73.2	55.1	43.2	34.6	28.7	55.1	115.4	20.8
	BLIP	75.1	56.7	44.9	35.9	29.4	56.2	118.9	21.4
	GIT	74.7	57.2	45.4	36.3	29.9	56.5	120.1	21.9
	Proposed ensemble	77.6	59.4	46.8	37.2	31.1	58.4	124.5	22.3
Flickr30K	ConvNeXt–T5	67.4	49.2	38.1	28.6	24.2	49.7	91.5	17.1
	BLIP	68.1	50.1	39.2	29.4	24.9	50.1	94.3	18.2
	GIT	69.3	51.2	40.1	30.1	25.3	50.9	95.6	18.7
	Proposed ensemble	71.2	53.4	41.7	30.8	26.1	52.4	98.7	19.6

Table 4 — Multi-seed results (mean ± std across 5 runs)

Model	BLEU-4 (± std)	METEOR (± std)	CIDEr (± std)	SPICE (± std)
ConvNeXt–T5	35.7 ± 0.5	26.8 ± 0.2	112.3 ± 1.1	20.1 ± 0.3
BLIP	36.3 ± 0.4	27.1 ± 0.3	120.1 ± 0.9	20.6 ± 0.2
GIT	36.0 ± 0.6	27.0 ± 0.2	118.7 ± 1.2	20.3 ± 0.3
Proposed ensemble	37.2 ± 0.5	27.5 ± 0.3	124.5 ± 1.0	21.2 ± 0.3

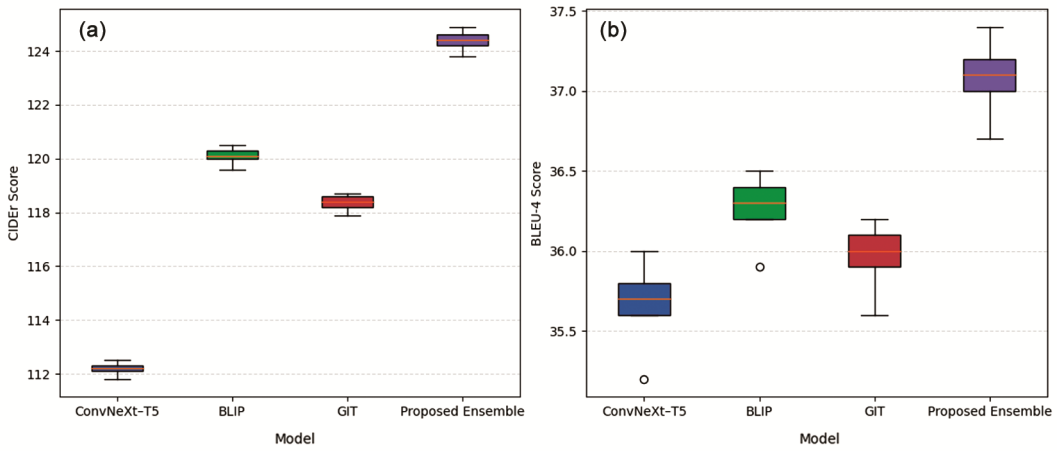

Fig. 2 — Comparison of image captioning models across five random seeds using: (a) CIDEr, (b) BLEU-4 metrics

framework constantly outperforms individual models across all metrics, in particular in CIDEr and SPICE, which strongly correlate with semantic high-quality and human judgment.

The ensemble achieves BLEU-4 = 37.2 on MS COCO and BLEU-4 = 30.8 on Flickr30K, which can be notably higher than baseline fashions. Similarly, the CIDEr ratings display a relative improvement of approximately 4–6 factors, highlighting the effectiveness of the consensus strategy.

Multi-Seed Robustness Analysis

For statistical robustness of the findings, every experiment was repeated with 5 random seeds (42, 123, 456, 789, and 999). The mean and standard deviation across those runs for BLEU-4, METEOR, CIDEr, and SPICE are presented in Table 4. Compared to single models, the proposed ensemble always achieves higher suggest ratings with highly low variance, demonstrating each accuracy gains and balance across random initializations.

The distribution of CIDEr and BLEU-4 scores throughout seeds the usage of boxplots is presented in Fig. 2. The ensemble exhibits tighter distributions compared to person fashions, confirming its reliability. Following the tips of Reimers and Gurevych (2017), this evaluation highlights that reporting simplest single runs may be misleading, and reinforces the statistical importance of the located improvements.

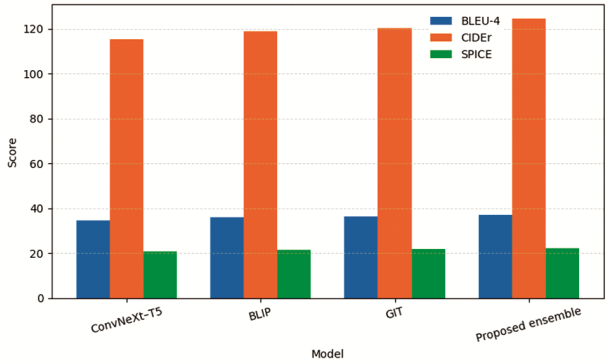

Fig. 3 — Performance comparison on MS COCO Dataset

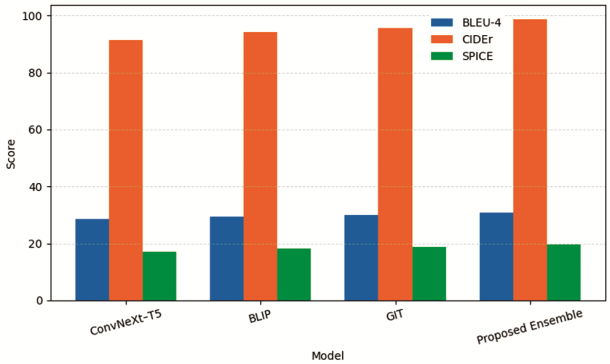

Fig. 4 — Performance comparison on Flickr30K

Performance Comparison

To better illustrate the improvements, Figs 3 and 4 should plot BLEU-4, CIDEr, and SPICE scores across different models. The ensemble consistently delivers

Table 5 — Ablation study on MS COCO

Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	SPICE
ConvNeXt only + T5	71.1	53.2	41.7	34.1	28.2	112.4	20.3
ConvNeXt + ViT + T5	72.3	54.8	42.8	35.2	29.1	117.2	21.0
ConvNeXt + ViT + BLIP	73.1	55.4	43.6	35.7	29.5	118.0	21.3
Full ensemble (Proposed)	77.6	59.4	46.8	37.2	31.1	124.5	22.3

Table 6 — Sample image captioning outputs: comparison of baselines and proposed ensemble

Images	Model	Generated captions
	ConvNeXt–T5	A dog running in the snow.
	BLIP	A brown dog runs across a snow-covered field.
	GIT	A dog is running through the snow in a park.
	Proposed Ensemble	A brown dog joyfully running across a wide snow-covered field with trees in the background.
	ConvNeXt–T5	A person riding a bike on a road.
	BLIP	A cyclist rides along a curved road on a sunny day.
	GIT	A man is cycling on a winding road with hills in the background.
	Proposed Ensemble	A cyclist in a pink shirt riding up a winding mountain road with green hills and cows in the background.
	ConvNeXt–T5	Two dogs are playing outside.
	BLIP	Two dogs run around in the grass.
	GIT	Two playful dogs are running on a grassy lawn.
	Proposed Ensemble	Two playful dogs interacting energetically on a grassy lawn, one puppy biting playfully at the other's ear.

the highest scores, confirming the benefit of integrating diverse encoders and decoders.

The ensemble model needs extra time to generate captions than single models because it combines several encoders and decoders, as defined within the Hardware and Runtime section. This additional time is suitable for offline use because the caption satisfactory is in reality better, specifically for CIDEr and SPICE.

Ablation Study

An ablation study was conducted on the MS COCO dataset to analyze the contribution of each component in the ensemble. Results are reported in Table 5.

The ablation confirms that combining multiple encoders and decoders progressively improves performance, while the full ensemble provides the most substantial gains.

Qualitative Analysis

Qualitative results are presented in Table 6 ground-truth captions, baseline models, and the proposed ensemble. Baselines often produce short or generic captions that miss important details. In contrast, the ensemble generates richer and more descriptive outputs. For example, it captures specific actions, clothing, and background elements overlooked by others. These comparisons confirm that the ensemble produces captions closer to human references.

Conclusions

This study developed an ensemble-based image captioning framework that combines multiple visual encoders with transformer-based language models to generate more accurate and context-aware image descriptions. The main outcome shows that using different encoders together helps capture richer visual details, while attention-guided fusion and re-ranking improve the clarity and consistency of the generated captions. However, the use of multiple models increases computational cost and inference time, which may limit its application in real-time or resource-constrained environments. The future work can focus on reducing the complexity of the framework and improving efficiency, as well as extending the approach to video captioning and multilingual caption generation. The proposed framework has potential applications in assistive technologies, image retrieval, digital content management, and human–computer interaction, where reliable and meaningful image descriptions are important. Overall, the study indicates that ensemble

learning offers a promising direction for advancing vision–language understanding tasks.

References

1 Vinyals O, Toshev A, Bengio S & Erhan D, Show and tell: A neural image caption generator, arXiv preprint, **1411.4555** (2015) 1–9, DOI: 10.48550/arXiv.1411.4555.

2 Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R & Bengio Y, Show, attend and tell: Neural image caption generation with visual attention, arXiv preprint, **1502.03044** (2016) 1–9, DOI: 10.48550/arXiv.1502.03044.

3 Al Badarneh I, Hammo B H & Al-Kadi O, An ensemble model with attention based mechanism for image captioning, *Comput Electr Eng*, **123** (2025) 110077, DOI: 10.1016/j.compeleceng.2025.110077

4 Lu J, Xiong C, Parikh D & Socher R, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, *Proc IEEE Conf Comput Vis Pattern Recognit*, 2017, 3242–3250, DOI: 10.1109/CVPR.2017.345.

5 Tang X, Habashy K, Huang F, Li C & Ban D, SCA-net: Spatial and channel attention-based network for 3D point clouds, *Comput Vis Image Underst*, **232** (2023) 103690, DOI: 10.1016/j.cviu.2023.103690.

6 Rennie S J, Marcheret E, Mroueh Y, Ross J & Goel V, Self-critical sequence training for image captioning, *Proc IEEE Conf Comput Vis Pattern Recognit*, 2017, 1179–1195, DOI: 10.1109/CVPR.2017.131

7 Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S & Zhang L, Bottom-up and top-down attention for image captioning and visual question answering, *Proc IEEE Conf Comput Vis Pattern Recognit*, 2018, 6077–6086, DOI: 10.1109/CVPR.2018.00636.

8 Li G, Zhu L, Liu P & Yang Y, Entangled transformer for image captioning, *Proc IEEE Int Conf Comput Vis*, 2019, 8927–8936, DOI: 10.1109/ICCV.2019.00902.

9 Cornia M, Stefanini M, Baraldi L & Cucchiara R, Meshed-memory transformer for image captioning, *Proc IEEE Conf Comput Vis Pattern Recognit*, 2020, 10575–10584, DOI: 10.1109/CVPR42600.2020.01059.

10 Pan Y, Yao T, Li Y & Mei T, X-linear attention networks for image captioning, arXiv preprint, **2003.14080** (2020) 1–9, DOI: 10.48550/arXiv.2003.14080.

11 Zhang X, Sun X, Luo Y, Ji J, Zhou Y, Wu Y, Huang F & Ji R, RSTNet: Captioning with adaptive attention on visual and non-visual words, *Proc IEEE Conf Comput Vis Pattern Recognit*, 2021, 15460–15469, DOI: 10.1109/CVPR46437.2021.01521.

12 Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F & Gao J, Oscar: Object-semantics aligned pre-training for vision-language tasks, *Lect Notes Comput Sci*, **12375** (2020) 121–137, DOI: 10.1007/978-3-030-58577-8_8

13 Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, Choi Y & Gao J, VinVL: Revisiting visual representations in vision-language models, *Proc IEEE Conf Comput Vis Pattern Recognit*, 2021, 5575–5584, DOI: 10.1109/CVPR46437.2021.00553.

14 Huang L, Wang W, Chen J & Wei X-Y, Attention on attention for image captioning, *Proc IEEE Int Conf Comput Vis*, 2019, 4633–4642, DOI: 10.1109/ICCV.2019.00473

15 Yang X, Wang Y, Chen H, Li J & Huang T, Context-aware transformer for image captioning, *Neurocomputing*, **549** (2023) 126440, DOI: 10.1016/j.neucom.2023.126440.

16 Zhang J, Xie Y & Liu X, Improving image captioning through visual and semantic mutual promotion, *Proc ACM Int Conf Multimed*, 2023, 4716–4724, DOI: 10.1145/3581783.3612480.

17 Ma Y, Ji J, Sun X, Zhou Y & Ji R, Towards local visual modeling for image captioning, arXiv preprint, **2302.06098** (2023) 1–9, DOI: 10.48550/arXiv.2302.06098.

18 Cornia M, Baraldi L & Cucchiara R, Explaining transformer-based image captioning models: An empirical analysis, *AI Commun*, **35(2)** (2021) 1–19, DOI: 10.3233/AIC-210172.

19 Li J, Li D, Savarese S & Hoi S, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, arXiv preprint, **2301.12597** (2023) 1–9, DOI: 10.48550/arXiv.2301.12597.

20 Liu H, Li C, Wu Q & Lee Y J, Visual instruction tuning, arXiv preprint, **2304.08485** (2023) 1–9, DOI: 10.48550/arXiv.2304.08485.