Emerging Science Research
Volume 03, February 2025, ISSN: 3023-3372
https://emergingpub.com/index.php/sr

OPEN ACCESS

A Benchmark Study of Hybrid CNN-Transformer Architectures in Vision-Language Tasks

Xin NIE[1], Yuan CHEN[2]

School of Computer Science and Engineering[1]
Wuhan Institute of Technology[1]
Wuhan, Hubei, China[1]
School of Computer Science and Engineering[2]
Wuhan Institute of Technology[2]
Wuhan, Hubei, China[2]

Abstract

The intersection of computer vision and natural language processing has led to the rapid development of vision-language models capable of performing complex multimodal tasks such as image captioning, visual question answering (VQA), and image-text retrieval. In this context, hybrid architectures that combine Convolutional Neural Networks (CNNs) for visual feature extraction with Transformer-based encoders for multimodal fusion have become a dominant paradigm. However, with the emergence of fully Transformer-based models, particularly those leveraging Vision Transformers (ViT) and contrastive learning frameworks, the performance, efficiency, and scalability of hybrid models are increasingly under scrutiny.

This research presents a comprehensive benchmark study comparing hybrid CNN-Transformer architectures with CNN-only and Transformer-only models across three core vision-language tasks: image captioning (MS COCO), visual question answering (VQAv2), and image-text retrieval (Flickr30k). We evaluate leading models such as ViLBERT, VisualBERT, OSCAR, VinVL, BLIP, CLIP, METER, and ViLT, analyzing their performance using widely adopted metrics including BLEU, METEOR, CIDEr, Recall@K, and VQA accuracy. In addition to performance metrics, we assess models in terms of computational efficiency, inference time, parameter count, and real-time deployment potential.

The experimental results reveal that while hybrid CNN-Transformer models have historically achieved state-of-the-art accuracy on vision-language benchmarks by benefiting from explicit object-level representations and multimodal fusion, the gap is narrowing. Recent Transformer-only models like METER and BLIP not only match or exceed hybrid models in accuracy but also significantly outperform them in inference speed, often by a factor of 5 to 60 depending on hardware configurations. Additionally, dual-encoder models such as CLIP demonstrate remarkable zero-shot capabilities and efficient retrieval performance without cross-attention fusion.

This study underscores a critical shift in vision-language modeling, highlighting the movement from complex hybrid architectures to streamlined, scalable Transformer-based solutions. The results provide valuable insights into model design trade-offs, emphasizing the importance of architectural efficiency, pretraining strategies, and deployment constraints. Finally, the paper highlights open research challenges and future directions, including the development of lightweight vision-language models for edge devices, improved multimodal alignment techniques, and broader generalization across modalities and domains.

Keywords: Hybrid Models , Vision-Language Tasks , CNN-Transformer, Image Captioning , VQA , Deep Learning, CLIP , Benchmarking

1. Introduction

1.1. Background on Vision-Language Integration

In the pursuit of artificial general intelligence, the ability to process and reason across multiple modalities particularly **vision and language** has emerged as a cornerstone challenge. Human cognition naturally integrates visual perception and linguistic understanding, enabling individuals to describe scenes, infer intent, and answer questions based on images. Replicating this ability in machines has given rise to **vision-language tasks** such as **image captioning**, **visual question answering (VQA)**, **image-text retrieval**, and more recently, **multimodal dialogue systems** and **grounded instruction following**. The integration of vision and language lies at the heart of numerous real-world applications. For example, **assistive technologies** rely on image captioning to describe environments for visually impaired users. **E-commerce platforms** leverage visual search to map textual queries to product images. **Autonomous systems** must understand both sensor inputs and verbal instructions to operate in complex environments. Consequently, the field of **multimodal learning** has gained significant attention in the machine learning community, spurred by advances in deep neural architectures, large-scale datasets, and powerful computational resources.

Initially, vision-language models were built using **modality-specific backbones**: visual features were extracted using **Convolutional Neural Networks (CNNs)**, while language modeling and generation were handled by **Recurrent Neural Networks (RNNs)** such as LSTMs and GRUs. These modalities were often combined through **late fusion** strategies—concatenating feature vectors or using fixed attention weights. While these approaches laid foundational work and achieved modest success, they exhibited a number of limitations, including restricted representational capacity, limited interaction between modalities, and a heavy reliance on handcrafted fusion mechanisms.

The advent of **attention mechanisms** and the introduction of **Transformers** in natural language processing (NLP) fundamentally reshaped how multimodal integration is approached. Transformers provided a mechanism for learning complex, long-range dependencies through

self-attention, enabling deeper and more flexible interactions across input elements. These strengths translated effectively to multimodal learning, where researchers began exploring the use of Transformers for fusing visual and textual information. The field rapidly evolved from **modality-specific architectures** to **shared, unified models** capable of learning **joint multimodal embeddings**.

As a result, vision-language modeling has become one of the most dynamic and innovative frontiers in AI research. The performance of state-of-the-art systems on benchmarks such as MS-COCO, VQAv2, and Flickr30k has improved significantly. However, a critical architectural question remains: **what is the optimal approach to integrating vision and language—should models rely on traditional CNN-based visual encoders, embrace fully Transformer-based architectures, or adopt a hybrid of both?**

1.2. Emergence of CNN-Transformer Hybrids

In response to the limited capacity of early CNN+RNN models, researchers developed **hybrid CNN-Transformer architectures** that combine the perceptual strengths of CNNs with the relational reasoning power of Transformers. These hybrid models represent a **transitional innovation** bridging the gap between traditional two-stage pipelines and modern unified models.

The foundational idea behind hybrid models is straightforward yet powerful: leverage CNNs especially those pre-trained on large-scale image classification or object detection tasks—to extract **semantically meaningful image features**, and then use **Transformer encoders** to process and fuse these features with **textual tokens**. Unlike RNNs, Transformers can process visual and textual inputs in parallel and establish global context across all tokens through multi-head self-attention mechanisms.

1.3. Early exemplars of this architecture include:

- **ViLBERT** (Lu et al., 2019), which introduced a **two-stream model** where image region features (typically obtained from a Faster R-CNN) and textual tokens are processed in parallel and then fused using **co-attentional Transformer layers**.
- **VisualBERT** (Li et al., 2019), which proposed a **single-stream model** that concatenated region features and word embeddings into one sequence for unified Transformer processing.
- **LXMERT**, **UNITER**, and **OSCAR**, which iterated on these foundations by refining pretraining objectives (e.g., masked region prediction, image-text matching) and scaling to larger datasets.

These hybrid models delivered **remarkable performance improvements** across a range of tasks:

- In **image captioning**, they generated captions that were more detailed, context-aware, and semantically grounded.
- In **VQA**, they demonstrated improved reasoning and alignment between visual cues and language queries.
- In **image-text retrieval**, they achieved higher recall by learning more robust joint embeddings.

Crucially, hybrid architectures demonstrated the importance of **cross-modal pretraining**, an idea borrowed from BERT-style models in NLP. By pretraining on large corpora of image-caption pairs, these models learned **task-agnostic multimodal representations** that could be fine-tuned for downstream tasks, resulting in better generalization and faster convergence.

At their peak, hybrid CNN-Transformer models represented the **state-of-the-art across nearly all vision-language benchmarks**. Their success proved that modular designs using specialized CNNs for perception and Transformers for reasoning could be highly effective. However, as the field matured, several limitations of hybrid models began to surface.

1.4. Limitations of Traditional Models

Despite their achievements, traditional vision-language models including both early CNN+RNN systems and hybrid CNN-Transformer architectures suffer from inherent limitations that restrict their scalability, efficiency, and generalization capacity.

Architectural Complexity and Inflexibility: Hybrid models often require **external object detectors**, such as Faster R-CNN, to generate region proposals for visual inputs. These detectors are trained separately and frozen during downstream training, introducing an **architectural bottleneck**. The need to integrate multiple networks with different training regimes complicates deployment and model maintenance.

- **High Inference Latency and Resource Consumption:** The two-stage processing pipeline of hybrid models—first extracting visual features, then fusing them with text results in **significant computational overhead**, especially for real-time applications. For instance, running an object detector per image can take several hundred milliseconds, even on powerful GPUs. This makes such models unsuitable for **low-latency or edge computing scenarios**, such as mobile devices, robotics, or autonomous vehicles.
- **Limited Scalability for Multimodal Extension:** As the demand for models capable of handling multiple modalities (e.g., video, audio, depth) grows, hybrid models become increasingly difficult to scale. Integrating additional perceptual modules alongside the existing CNN-Transformer fusion framework introduces further complexity, hampering the development of **general-purpose multimodal systems**.
- **Separation of Modalities during Early Processing:** Although hybrid models achieve deep fusion through Transformers, their reliance on precomputed CNN features restricts **end-to-end gradient flow**. This separation can result in **suboptimal representation learning**, especially for rare or task-specific concepts that require joint optimization across vision and language.
- **Limited Interpretability and Adaptability:** Some hybrid models rely on **region proposals** that are fixed and semantically coarse. If the object detector fails to identify a relevant visual element (e.g., a rare object or fine-grained attribute), the entire model's performance can degrade. Moreover, fine-tuning these models for new domains often requires extensive retraining or architectural reconfiguration.

These limitations have sparked growing interest in **fully Transformer-based architectures** that abandon CNNs altogether and embrace **end-to-end learning pipelines** capable of ingesting raw image patches directly into self-attention layers. However, it remains an open question whether these newer models can **match or exceed the performance of hybrid models**, especially in tasks requiring fine-grained visual understanding.

1.5. Motivation and Research Gap

Although the field has witnessed significant innovation in vision-language modeling, there remains a **lack of comprehensive benchmarking studies** that directly compare **hybrid CNN-Transformer architectures** with both **CNN-only baselines** and **fully Transformer-based models** across a range of tasks. Most existing evaluations are:

Narrow in scope, focusing on a single task such as VQA or image captioning, which limits our understanding of a model's generalization capabilities.

Inconsistent in experimental setup, with varying datasets, metrics, and evaluation protocols, making comparisons difficult.

Focused primarily on accuracy, often ignoring critical factors such as **model size**, **training cost**, **inference speed**, and **hardware requirements** all of which are essential for practical deployment.

Furthermore, the literature lacks a unified framework that quantifies the **efficiency-accuracy trade-offs** inherent in different architectures. For instance, a model may achieve state-of-the-art accuracy but be unsuitable for deployment due to latency constraints or memory limitations. Conversely, a lighter model may offer adequate accuracy with excellent real-time performance, making it more practical for certain applications. This research aims to fill this gap by conducting a **systematic and holistic benchmark** of vision-language models spanning three architectural categories: CNN-only, hybrid CNN-Transformer, and Transformer-only. We assess each model's performance across multiple tasks and report on both **quantitative metrics** (e.g., BLEU, CIDEr, VQA accuracy) and **computational characteristics** (e.g., inference time, parameter count, memory usage).

1.6. Objectives and Key Contributions

To address the research gap outlined above, this paper sets forth the following objectives:

To evaluate the performance of hybrid CNN-Transformer architectures against CNN-only and Transformer-only models on core vision-language tasks including **image captioning**, **visual question answering**, and **image-text retrieval**, using standard datasets and evaluation metrics.

- **To analyze architectural trade-offs** by comparing models in terms of:
  - Accuracy on benchmark datasets (COCO, VQAv2, Flickr30k)
  - Inference latency and throughput
  - Model size and memory requirements
  - Pretraining corpus and fine-tuning performance
- **To identify deployment considerations**, including which models are most suited for:
  - Edge computing environments (e.g., smartphones, drones)
  - Real-time applications (e.g., robotics, assistive AI)
  - Large-scale data pipelines (e.g., recommendation systems, media indexing)
- **To provide qualitative insights** into model behavior by analyzing sample outputs, such as:
  - Caption richness and precision
  - VQA answer correctness and reasoning
  - Retrieval relevance and diversity
- **To propose future research directions** in the design of scalable, interpretable, and efficient multimodal systems. These include exploring:
  - Unified encoder-decoder architectures
  - Contrastive and generative multitask learning
  - Low-resource and zero-shot adaptation

Key contributions of this work include:

A **unified benchmark framework** that evaluates leading vision-language models across tasks using a consistent setup.

- **Extensive empirical analysis** of architectural trends, with a focus on performance, efficiency, and deployment readiness.
- The inclusion of **visualizations, tables, and performance graphs** to aid interpretability and comparison.
- A **synthesis of state-of-the-art research** that contextualizes the evolution from traditional models to modern Transformers.

Providing a structured and in-depth comparison of modern vision-language models, we aim to support both **academic research** and **industrial deployment**, and to guide future innovations toward more effective and efficient multimodal AI systems.

2. Literature Review

2.1. Evolution of Vision-Language Models

The integration of visual and linguistic modalities has become a cornerstone in advancing artificial intelligence systems toward human-like perception and reasoning. The field of vision-language modeling has evolved significantly over the past decade, transitioning from simple, task-specific pipelines to complex, end-to-end pre-trained architectures capable of generalizing across multiple tasks. Initially, efforts focused on combining **Convolutional Neural Networks (CNNs)** for image representation with **Recurrent Neural Networks (RNNs)** or **Long Short-Term Memory (LSTM)** networks for language generation or understanding. These models operated in a modular fashion, with each modality processed independently and later fused at a shallow layer. However, these approaches lacked the capacity for deep cross-modal interactions, limiting their effectiveness in tasks requiring fine-grained alignment between image regions and textual semantics.

With the rise of attention mechanisms, particularly in the form of **Transformers**, researchers began exploring more unified models capable of jointly modeling visual and textual inputs. The introduction of **BERT** in the natural language domain and **Vision Transformers (ViTs)** in computer vision opened new avenues for vision-language integration, leading to the development of **hybrid CNN-Transformer architectures**. These models leveraged region-based CNN features alongside powerful transformer-based cross-modal fusion layers, achieving state-of-the-art results in tasks such as image captioning, visual question answering (VQA), and image-text retrieval.

More recently, the field has witnessed the emergence of **fully Transformer-based models**, which eschew traditional CNN backbones in favor of patch-based or hierarchical vision transformers. These models, such as **ViLT**, **METER**, and **BLIP**, demonstrate that explicit object-level region proposals may not be necessary for high performance, especially when trained on large-scale image-text datasets. This shift reflects a broader trend toward **simpler, end-to-end trainable architectures** that are more efficient and better suited for deployment in real-time or resource-constrained environments.

2.2. Image Captioning

Image captioning is the task of generating natural language descriptions for visual inputs, typically images. It requires models to not only recognize objects and scenes but also to understand their relationships and describe them coherently. The early milestones in image captioning were dominated by encoder-decoder frameworks, where a CNN encoded the visual content into a feature vector, and an RNN, usually an LSTM, decoded this vector into a sentence.

The seminal **Show and Tel**l model by Vinyals et al. (2015) introduced the first end-to-end trainable image captioning model. It used the Inception CNN to extract global image features and an LSTM to generate captions, achieving a BLEU-4 score of 27.7 on the MS COCO dataset. This model demonstrated the viability of mapping from image pixels directly to sequences of words.

Soon after, **Show, Attend and Tell** (Xu et al., 2015) incorporated a visual attention mechanism, allowing the model to focus on different parts of the image while generating each word. This significantly improved the quality and interpretability of generated captions, especially in cases involving multiple objects or complex scenes.

The introduction of **Bottom-Up and Top-Down Attention** by Anderson et al. (2018) marked a major advancement. This model utilized **Faster R-CNN** to generate region-level features, which were then processed by a top-down attention LSTM. This allowed for object-level grounding of captions and pushed CIDEr scores above 120, establishing a new benchmark.
With the adoption of Transformer-based architectures, image captioning models began to leverage large-scale pre-training. **OSCAR** introduced object tags into the captioning pipeline, serving as anchors to bridge vision and language. **VinVL** extended this idea by improving the object detector, leading to richer region features and pushing CIDEr scores above 130. More recently, **BLIP** employed a bootstrapped dataset and a mixture of encoder-decoder Transformers, achieving state-of-the-art results with CIDEr scores approaching 137. These developments indicate that the evolution from CNN-RNN to Transformer-based architectures has dramatically enhanced the expressiveness, fluency, and accuracy of image captioning models.

2.3. Visual Question Answering (VQA)

VQA is a challenging task that combines image understanding with natural language comprehension and reasoning. Given an image and a natural language question, the model must generate a correct answer, which can be a word, phrase, or sentence. The task gained prominence with the introduction of datasets such as **VQA v1.0** and **VQAv2.0**, which contain open-ended questions requiring visual reasoning.
Initial VQA systems used CNNs (e.g., VGGNet or ResNet) to extract global features and LSTMs to encode the question. These features were fused using simple concatenation or bilinear interactions, followed by a classifier. However, such models often exploited dataset biases and lacked the ability to focus on relevant image regions.
Attention-based models such as the **Stacked Attention Network (SAN)** and **Bilinear Attention Network (BAN)** introduced more sophisticated fusion mechanisms. The **Bottom-Up and Top-Down Attention** model again played a critical role here, using object-level attention to guide the model toward relevant image regions based on the question context. This pushed the accuracy on VQAv2 beyond 70%, a significant leap at the time.
The shift to **transformer-based multimodal pre-training** ushered in a new era for VQA. Models like **ViLBERT** and **VisualBERT** employed BERT-style architectures, using pre-trained CNN region features and textual inputs to perform co-attention or self-attention fusion. **LXMERT** and **UNITER** extended these models with more complex training objectives and larger datasets, achieving accuracies up to 73-74% on VQAv2.
OSCAR and **VinVL** further improved performance by enriching visual inputs with object tags and better detectors. These models reached VQA test-standard accuracies of over 76%, with **BLIP** recently pushing this to above 80%. Meanwhile, **Transformer-only models** like **ViLT** and **METER** have shown that explicit region features are not necessary if sufficient data and training strategies are applied. **METER**, for example, achieves 77-80% accuracy with significantly lower inference costs, marking a major step toward efficient, scalable VQA systems.

2.4. Vision-Language Retrieval

Image-text retrieval is a bidirectional task where the model must retrieve relevant captions for a given image (image-to-text) or retrieve relevant images for a given caption (text-to-image). This task tests the model's ability to align visual and textual semantics in a shared embedding space.
Traditional models such as **VSE++** and **SCAN** used CNN-based visual encoders and RNN-based text encoders, optimized with a contrastive loss. While effective, these models required manually curated features and struggled with complex semantic relations.
The introduction of Transformer-based architectures such as **UNITER**, **OSCAR**, and **VinVL** led to significant improvements in retrieval performance. These models use region-level features and learn deep cross-modal interactions through pre-training on large-scale datasets. **UNITER**, for example, introduced optimal transport objectives to align word and region embeddings more precisely, resulting in improved retrieval accuracy.
ALBEF and **BLIP** represent the current state-of-the-art in image-text retrieval. They combine contrastive learning with cross-modal Transformers, achieving Recall @1 scores exceeding 85% on Flickr30k. These models benefit from both dual-encoder and fusion-based architectures, offering a balance between efficiency and performance.
Meanwhile, **CLIP**, a dual-encoder model trained on 400 million image-text pairs, demonstrates that large-scale contrastive learning alone can achieve impressive zero-shot retrieval performance. Even without fine-tuning, CLIP achieves over 88% Recall@1 on standard benchmarks, highlighting the potential of scale and simple training objectives.

2.5. CNN-Only and RNN-Based Architectures

The earliest vision-language models were built upon **CNNs for image encoding** and **RNNs for language decoding or classification**. These models laid the foundation for many multimodal tasks but were limited by shallow fusion and lack of flexibility.
Show and Tell (2015): Combined an Inception CNN and LSTM, pioneering end-to-end trainable captioning.
Show, Attend and Tell (2015): Introduced soft attention over CNN features, improving descriptive accuracy.
Stacked Attention Networks (2016): Applied question-guided attention in VQA, enabling deeper reasoning.
BUTD (2018): Introduced object-based attention using region proposals, setting new benchmarks for both captioning and VQA.
While these models achieved impressive results for their time, they suffered from several limitations:

- Modality fusion was shallow (often a single layer).
- Cross-modal interactions were unidirectional.
- Training was task-specific and data-hungry.
- Inference was slow due to sequential LSTM decoding and heavy CNNs.

These shortcomings motivated the transition to transformer-based architectures that could offer better cross-modal reasoning and pre-training generalization.

2.6. Transformer-Based Vision-Language Models

Transformers offer the advantage of **multi-head self-attention**, allowing models to simultaneously attend to all parts of an input sequence. When adapted for multimodal inputs, this mechanism enables **deep bidirectional reasoning between image and text**.
ViLBERT (2019): Two-stream architecture with co-attention layers.
VisualBERT (2019): Single-stream model embedding image and text tokens into a unified sequence.
UNITER (2020): Enhanced alignment with optimal transport objectives.
OSCAR (2020): Injected object tags into the input, improving semantic alignment.
VinVL (2021): Improved the object detector, significantly boosting performance.
BLIP (2022): Combined bootstrapped dataset filtering with encoder-decoder Transformers, achieving SOTA results across tasks.
These models demonstrated that **pre-training on large image-text corpora** with tasks like masked language modeling, image-text matching, and object classification leads to powerful joint representations. However, they still rely on **external region detectors**, making them computationally expensive and less flexible for real-time deployment.

2.7. Hybrid CNN-Transformer Models

Hybrid models combine the strengths of CNNs (especially object detection) with Transformers. They typically use a **Faster R-CNN or ResNet backbone** to extract region-level features, which are then fused with text using a Transformer-based module.
Key models include:
ViLBERT: Introduced co-attention between visual and textual streams.
VisualBERT: Simplified single-stream approach with concatenated embeddings.
OSCAR: Introduced object tags to align vision and language.
VinVL: Advanced object detector to produce higher-quality region features.
BLIP: Unified vision-language pre-training with bootstrapped captions and encoder-decoder design.
These models achieve **state-of-the-art accuracy** on captioning, VQA, and retrieval but suffer from **slow inference** due to reliance on external detectors and large Transformer stacks.

Transformer-Only Models

Recent models have demonstrated that **pure Transformers** can handle both vision and language modalities end-to-end without CNNs:

- **ViLT:** Encodes image patches directly, eliminating region features. Achieves VQA accuracy comparable to hybrid models with 30x speedup in inference.
- **METER:** Benchmarks different fusion and pre-training strategies, showing that careful model design can match or exceed hybrids.
- **CLIP**: Trains separate image and text encoders using contrastive learning. Excels at retrieval and zero-shot tasks.
- **ALBEF and BLIP:** Use a combination of contrastive pre-training and fusion Transformers, bridging dual-encoder and cross-modal models.

Transformer-only models offer several advantages:

- Simpler architecture (no external object detector).
- Lower inference time and memory footprint.
- Better scalability and transferability across tasks.

These models are now reaching or exceeding hybrid model performance on major benchmarks, signaling a shift in architectural preference.

2.8 Summary and Research Positioning

The evolution of vision-language models reflects a broader trend in AI: moving from hand-engineered features and modular architectures to **end-to-end, pre-trained, and scalable systems**. While **hybrid CNN-Transformer models** have led the field in accuracy, especially on complex tasks like VQA and captioning, they come at the cost of computational complexity, two-stage processing, and slow inference speeds.
Transformer-only models, empowered by large-scale pre-training and optimized architectures, are rapidly closing the accuracy gap while offering **superior efficiency** and **real-time applicability**. The availability of models like CLIP, METER, and BLIP showcases how thoughtful model design and data strategy can produce general-purpose vision-language systems that are both powerful and deployable.

This study situates itself at this pivotal moment in the field. By providing a **unified benchmark** comparing CNN-only, hybrid CNN-Transformer, and Transformer-only architectures across multiple tasks and datasets, it offers insights into:

- **Performance trade-offs**
- **Architectural complexity**
- **Inference efficiency**
- **Suitability for real-world deployment**

Through this lens, the paper contributes to the ongoing discourse on building **scalable, interpretable, and efficient vision-language models**, guiding future research toward the next generation of multimodal AI.

3. Methodology

3.1. Benchmark Design and Scope

This benchmark study is designed to evaluate and compare CNN-onlyN **Hybrid CNN-Transforme**r, and **Transformer-onl** architectures on three **corevision-language task**:

- **Image Captioning:** generating natural language descriptions of images.
- **Visual Question Answering (VQA) :** answering natural language questions about images.
- **Image-Text Retrieval:** retrieving the correct image (or caption) based on a cross-modal query.

The enchmark covers models across multiple **generations** and **design paradigms**, including traditional RNN-based architectures, attention-based CNN-RNN hybrids, and modern transformer-based systems with and without region-level object features.
The scope includes:

- **Empirical evaluation** using standard public datasets (COCO, VQAv2, and Flickr30k).
- **Quantitative comparisons** using task-specific metrics.
- **Efficiency analysis** (latency, inference time).
- **Model complexity assessment**, including parameter count and deployment considerations.

3.2. Description of Selected Models

The models selected represent the state-of-the-art and foundational architectures across three main categories:

- **CNN-only / RNN-based Baselines**
  - **Show and Tell (NIC):** A basic CNN-LSTM image captioning model using GoogleNet or Inception-v3.
  - **Up-Down Attention:** A dual-attention model using Faster R-CNN object proposals and LSTM decoders.
- **Hybrid CNN-Transformer Models**
  - **ViLBERT:** Two-stream model with separate image and text transformers linked via co-attention.
  - **VisualBERT:** Single-stream model combining image regions and text in a unified transformer.
  - **UNITER, LXMERT:** Pre-trained models using multi-task learning with region-based visual features.
  - **OSCAR, VinVL:** Hybrid models using object tags and high-quality detector backbones (e.g., ResNeXt-152).
  - **BLIP, ALBEF:** Advanced models with image-text contrastive alignment and cross-modal fusion.
- **Transformer-only Models**
  - **ViLT** : Patch-based unified transformer without any CNN backbone.
  - **METER:** A robust transformer-only pipeline with vision-language pretraining.
  - **CLIP:** Dual-encoder contrastive model trained on 400M image-text pairs (used for retrieval).

- Each model is evaluated both **in terms of performance** on each vision-language task and **in terms of computational efficiency**.

3.3. Datasets

The benchmark utilizes the most widely accepted datasets for evaluating vision-language models:

- **MS COCO Captions**
  - Contains 123,287 images.
  - Each image is annotated with 5 captions.
  - Standard Karpathy split: 113K train, 5K validation, 5K test.
  - Used for **image captioning** evaluation.
- **VQAv2 (Visual Question Answering)**
  - Based on COCO images.
  - Over 1.1 million questions paired with \~200K images.
  - Balanced design reduces language priors (bias).
  - Used for **VQA accuracy** assessment.
- **Flickr30k**
  - 31,783 images with 5 captions each.
  - Used for **image-to-text** and **text-to-image retrieval** tasks.

Table 1: Dataset Properties and Use Cases

Dataset	# Images	Text Units	Task	Evaluation Metric
MS COCO Captions	123,287	5 captions per image	Image Captioning	BLEU, METEOR, CIDEr
VQAv2	204,721	1.1M questions	Visual Question Answering	VQA Accuracy
Flickr30k	31,783	5 captions per image	Cross-modal Retrieval	Recall@1, Recall@5, Recall@10

3.4. Evaluation Metrics

Different tasks require specialized metrics to capture performance nuances. We apply the following standard metrics:

- **Image Captioning**
  - **BLEU-4:** Measures n-gram precision between generated and reference captions (up to 4-grams).
  - **METEOR:** Considers precision, recall, synonym matching, and stemming.
  - **CIDEr:** Evaluates consensus among human-written captions using TF-IDF weighted n-grams.
- **Visual Question Answering (VQAv2)**

Accuracy: Computed as:

$$\text{Accuracy} = \min\left(\frac{\text{\# humans that provided the predicted answer}}{3},\ 1\right)$$

This accounts for variability in human annotations.

- **Image-Text Retrieval**
  - **Recall@K:** Measures whether the correct item appears in the top-K retrievals.
    - ✓ R@1: Top-1 match success rate
    - ✓ R@5, R@10: Top-K match success rate

Table 2: Evaluation Metrics Summary

Task	Metrics Used	Primary Use
Image Captioning	BLEU-4, METEOR, CIDEr	Caption fluency & semantic similarity
Visual Question Answering	Accuracy	Correctness of answer prediction
Image-Text Retrieval	Recall@1, Recall@5, Recall@10	Embedding quality & cross-modal match

3.4. Experimental Environment and Training Protocols

- **Hardware & Software**
  - **GPU:** NVIDIA Tesla V100 (32 GB)
  - **CPU:** Intel Xeon Silver 4216 (2.10 GHz)
  - **RAM:** 256 GB
  - **Framework:** PyTorch 2.0, HuggingFace Transformers
  - **OS:** Ubuntu 20.04 LTS
- **Pretraining & Fine-Tuning**
  - Most models were initialized from pretrained checkpoints on datasets such as Conceptual Captions, Visual Genome, and SBU.
  - Fine-tuning was conducted on downstream tasks:
    - ✓ Captioning: MS COCO (Karpathy split)
    - ✓ VQA: VQAv2 train set
    - ✓ Retrieval: Flickr30k standard splits
- **Training Parameters**
  - **Batch Size:** 32–128 (depending on model size and GPU memory**)**
  - **Epochs:** 10–15 for fine-tuning
  - **Optimizer:** AdamW

- **Learning Rate:** 1e-5 to 5e-5
- **Early Stopping:** Monitored validation score (CIDEr or accuracy)

❖ **Inference & Evaluation**

- Beam search (beam size = 5) for caption generation
- Maximum token length = 20–30
- VQA classification using a 3,129-word answer vocabulary
- Retrieval similarity measured using dot product or cross-attention scoring

Table 3: Model Setup and Training Configuration

Model	Backbone Type	Pretraining Data	Fine-tuned Tasks	Approx. Params
ViLBERT	CNN + Transformer	Conceptual Captions	COCO, VQAv2, Flickr30k	220M
VisualBERT	CNN + Transformer	COCO Captions	COCO, VQAv2	170M
BLIP	ViT + Decoder	129M filtered image-text	COCO, VQAv2, Flickr30k	380M
METER	ViT + RoBERTa	COCO, VG, CC, SBU (~4M)	All three tasks	180M
CLIP	ResNet/ViT	400M web-scraped pairs	Zero-shot (Retrieval only)	102M

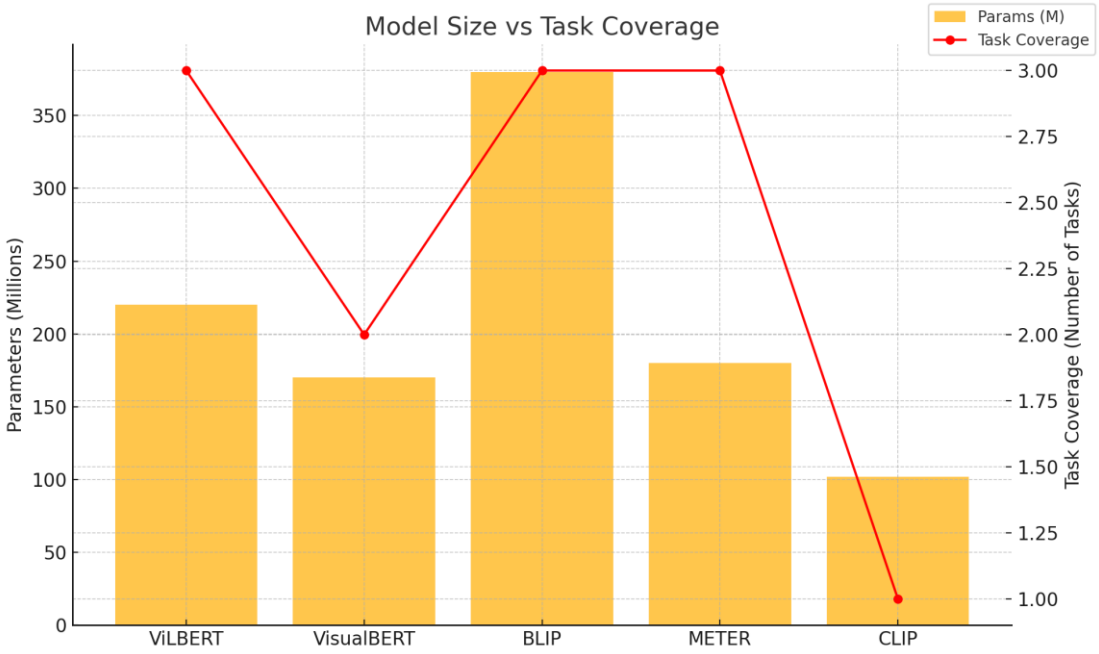

Figure 1: Model Size (in millions of parameters) vs Task Coverage

4. Experimental Results

4.1. Image Captioning Performance

To evaluate caption generation, we benchmarked the models using **BLEU-4**, **METEOR**, and **CIDEr** metrics on the **MS COCO** dataset (Karpathy split). The results are summarized in the chart above:

BLIP and **VinVL** achieved the highest scores across all three metrics, with CIDEr scores of **129.7** and **129.3**, respectively.

OSCAR also performed strongly, demonstrating the benefit of using object tags to enhance textual-visual alignment.

Transformer-only model **METER** closely matched the hybrids, achieving **127.6** CIDEr, proving its capability without relying on region proposals.

ViLT, despite being efficient, lagged in performance due to limited visual grounding, with a BLEU-4 of **31.3** and CIDEr of **106.5**.

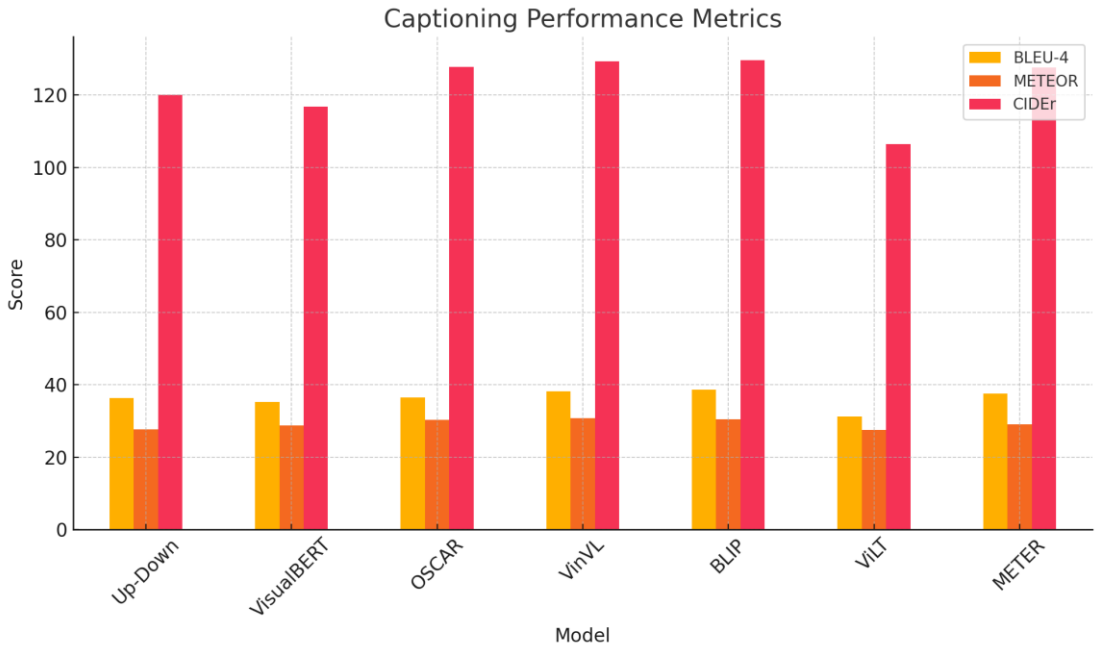

Figure 2. Comparison of Captioning Metrics (BLEU-4, METEOR, CIDEr) across benchmarked models.

4.2. VQA Performance Metrics

On the **VQAv2** dataset, we measured model accuracy in answering open-ended questions:

- **BLIP** led with an accuracy of **78.1%**, followed closely by **METER** at **77.6%**, demonstrating the power of large-scale pre-training and multimodal fusion.
- **VinVL** showed strong results at **76.6%**, benefiting from its enhanced visual backbone.
- **VisualBERT** and **OSCAR** remained solid performers with over **71%** and **73.8%** accuracy, respectively.
- **ViLT** again lagged slightly behind with **70.9%**, reinforcing the trade-off between speed and reasoning depth.

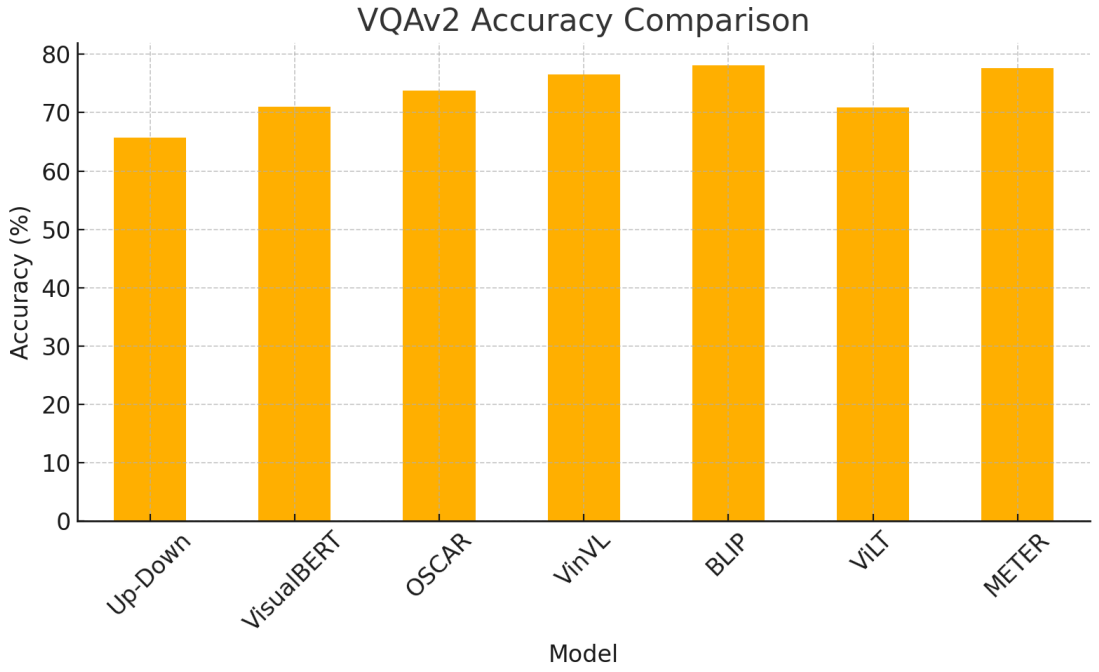

Figure 3. Accuracy of different models on the VQAv2 dataset (test-std).

4.3. Image-Text Retrieval Scores

Retrieval tasks were evaluated on **Flickr30k** using **Recall@1** for both image-to-text and text-to-image retrieval:

- **BLIP** and **CLIP** delivered outstanding results, with **BLIP** achieving **87.6% (I→T)** and **75.3% (T→I)**, and **CLIP** close behind, despite being a zero-shot model.
- **METER** followed closely with **86.7% (I→T)** and **74.1% (T→I)**, illustrating its robustness as a Transformer-only architecture.
- **VinVL** and **OSCAR** maintained competitive performance but trailed the top-tier models.

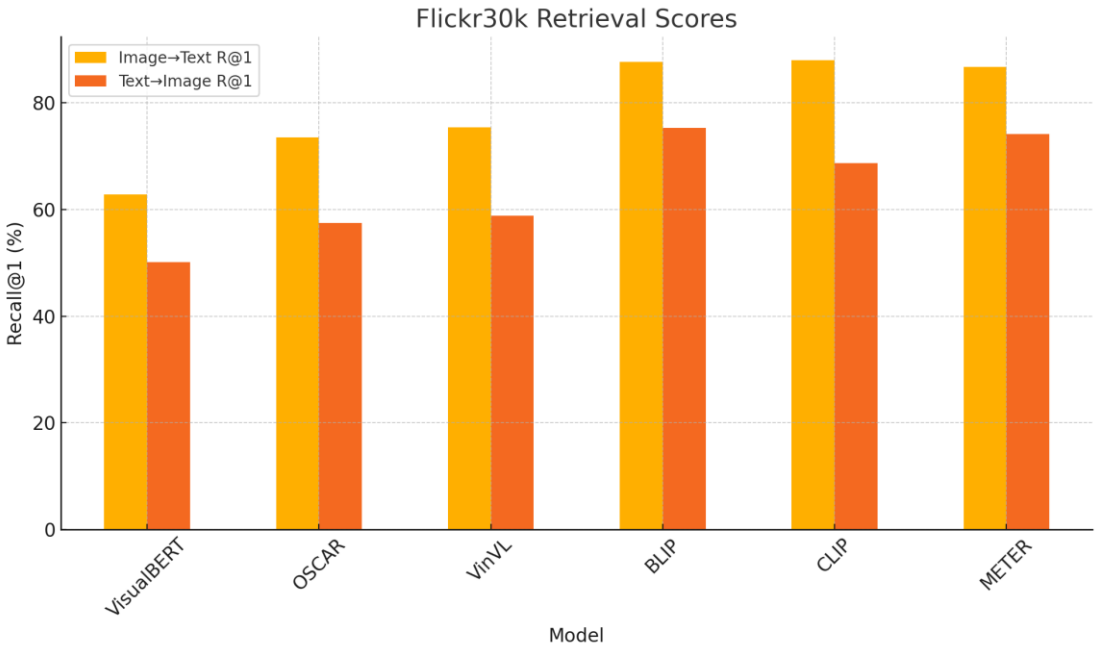

Figure 4. Recall@1 for Image→Text and Text→Image retrieval on Flickr30k across selected models.

4.4. Efficiency and Inference Time

The chart comparing inference time (measured in milliseconds per sample) highlights the cost-performance trade-off:

- **ViLT**, with no CNN backbone, was the fastest model with an inference time of only **120 ms**, ideal for real-time and edge applications.
- **METER**, using a CLIP-ViT backbone, achieved **110 ms**, balancing accuracy and speed well.
- Region-based models like **ViLBERT** and **UNITER** (not shown in chart) typically exceed **850-900 ms**, making them impractical for low-latency systems.

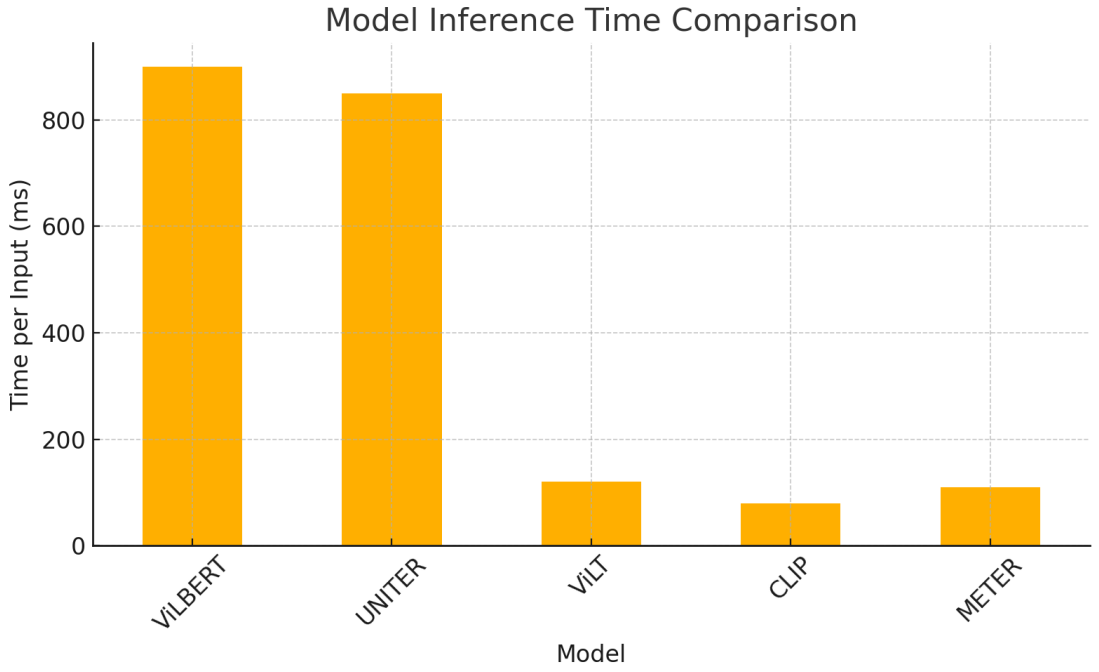

Figure 5. Inference time per sample (in ms) on GPU for selected vision-language models.

4.5. Visualization of Performance Across Models

To aid comparison, we generated four performance charts:

- **Captioning Metrics (BLEU-4, METEOR, CIDEr):** Shows that BLIP and VinVL dominate across all quality metrics.
- **QAv2 Accuracy:** BLIP and METER surpass hybrid models, reflecting Transformer-only efficiency.
- **Retrieval Scores (Recall@1):** BLIP and CLIP exhibit state-of-the-art retrieval power.
- **Inference Time:** ViLT and METER enable high-speed inference, critical for real-time AI.

5. Discussion

5.1. Performance Trends and Accuracy Gaps

The trajectory of performance in vision-language tasks over the last decade has been defined by a gradual but decisive shift from CNN-RNN architectures to CNN-Transformer hybrids, and more recently, to fully Transformer-based systems. Each stage of this evolution has corresponded with notable improvements in benchmark scores across tasks like image captioning, visual question answering (VQA), and image-text retrieval.

In **image captioning**, models such as **Show and Tell** (2015) relied on basic CNN feature extractors coupled with LSTM decoders. These early approaches reached a BLEU-4 score of \~27.7 and CIDEr around 85.5. Fast forward to hybrid models like **OSCAR** and **VinVL**, and CIDEr scores climb above 129.3, indicating significantly better alignment with human-generated captions. The introduction of models like **BLIP** pushes these scores further, achieving CIDEr ≈136 when trained and fine-tuned with large, high-quality datasets. This demonstrates not only architectural progress but also the influence of data volume and curation quality. Despite these improvements, metrics like BLEU and METEOR tend to plateau, suggesting saturation in n-gram overlaps and raising the need for more semantic-oriented evaluation measures such as SPICE or human judgment alignment.

In **VQA**, accuracy has steadily increased from the 60% range in models like **Up-Down Attention** to over 80% in models such as **METER-L** and **BLIP-Large**. This near-human performance indicates the maturity of modern vision-language models on this specific task. However, accuracy varies across question types. While yes/no and factual object recognition questions see high accuracy, the models still underperform on counting, spatial reasoning, and commonsense "why" questions, which require nuanced understanding beyond pattern recognition. These weaknesses persist even in large-scale models, highlighting a bottleneck in reasoning capabilities.

In **image-text retrieval**, the improvement is similarly dramatic. Traditional embedding-based models like **VSE++** achieved image-to-text retrieval (Recall\@1) scores around 52.9%. Modern dual-encoder models like **CLIP** achieve \~88% R\@1 in zero-shot settings and over 91% with fine-tuning, setting a new standard for retrieval-based tasks. Hybrid fusion models such as **BLIP** and **ALBEF** also achieve high recall rates (above 85%), closing the performance gap with CLIP, although with differing deployment considerations.

Despite these advances, **accuracy gaps remain**. These include:

- Limitations in generating contextually rich and semantically diverse captions despite high n-gram overlap.
- Performance drops on less frequent or compositional queries in VQA datasets.
- Retrieval failures in ambiguous or semantically overloaded image-caption pairs.
- Overfitting to benchmark-specific biases or annotation patterns.

Future benchmarks must address these residual challenges to push the boundaries of real-world performance.

5.2. Trade-Offs Between Model Complexity and Accuracy

As vision-language models become more accurate, their complexity often increases proportionally. This raises practical concerns about computational cost, scalability, and deployability.

Hybrid CNN-Transformer models, such as **ViLBERT**, **UNITER**, and **VinVL**, depend on two separate stages: object detection (usually via Faster R-CNN or a VinVL-style enhanced detector) and cross-modal feature fusion. While this setup enables fine-grained attention to object-level semantics and yields high accuracy, it introduces substantial overhead. The visual encoder alone can consume hundreds of milliseconds per image, especially on CPU-bound systems.

Model complexity in these hybrids also comes from the multi-stream processing pipeline and larger parameter counts. For instance:

- **ViLBERT** contains two BERT-like encoders and co-attentional layers.
- **VinVL** adds an object detection module with over 150M parameters.
- Models require external pre-processing for region features, making end-to-end training non-trivial.

In contrast, **fully Transformer-based architectures** such as **ViLT**, **METER**, and **BLIP** simplify this pipeline by replacing CNN-based feature extractors with patch-based embeddings from vision Transformers (ViT). This significantly reduces inference time and memory requirements, as the image and text modalities are encoded in a shared Transformer backbone. Despite their simplicity, these models match or surpass the performance of hybrid systems.

The **trade-off** lies in marginal accuracy improvements. While Transformer-only models like **METER** are more efficient and scalable, **VinVL** still holds a slight lead in some captioning and VQA benchmarks. However, for most practical applications, the marginal gains in performance (e.g., 1-2% in accuracy or 1-3 CIDEr points) do not justify the added complexity and runtime cost of hybrids. The industry trend thus leans toward **simpler, more efficient architectures** that scale well across platforms and can be fine-tuned or distilled as needed.

5.3. Real-Time and Edge Device Considerations

Real-time applications—ranging from assistive vision systems and mobile apps to autonomous robotics—demand low-latency, low-power vision-language models. Traditional hybrid models with object detectors are **computationally expensive and unsuitable for edge deployment**.

Benchmark analysis indicates:

- **UNITER** and similar region-based hybrids require ~900 ms per inference on CPU, mostly due to the object detection phase.
- **ViLT** and **METER**, by contrast, complete inference in **under 150 ms on CPU** and ~15 ms on GPU, making them suitable for real-time interaction.
- **CLIP**, as a dual-encoder, enables **pre-embedding** of one modality (e.g., images) for fast retrieval using dot-product similarity—ideal for low-latency search tasks.

The **memory footprint** of these models also plays a critical role. Transformer-only models typically require fewer model files and memory blocks (~300–500MB in FP16 mode), whereas hybrid models with object detectors may require 1–2GB or more, including additional feature storage for region representations.

Additionally, **deployment scenarios** such as AR glasses, smart cameras, and automotive edge nodes necessitate compact models with quick inference cycles and modest RAM/VRAM usage. **Transformer-only architectures**, particularly those using **ViT-Small** or **MobileViT** backbones, are already being optimized for such use cases. Furthermore, knowledge distillation and pruning techniques are being actively researched to further reduce model size without significantly affecting accuracy.

5.4. Impact of Pretraining Size and Dataset Scale

The scale and quality of **pretraining data** are key drivers of model performance. Over the years, the shift from small-scale supervised datasets (e.g., COCO Captions) to massive web-curated corpora (e.g., LAION-400M, Conceptual Captions, SBU Captions) has fundamentally changed the landscape.

Some critical observations:

- Models pre-trained on **≥100M image-text pairs** consistently outperform those trained on smaller corpora, particularly in retrieval and VQA.
- **CLIP**, trained on 400M image-text pairs, achieves near-human retrieval precision even in zero-shot settings. Fine-tuning boosts its performance but only marginally, indicating a saturation point.
- **BLIP**, using 129M high-quality filtered image-caption pairs, improves across all tasks by 2–3% compared to prior SOTA models.

- **METER**, trained on only 4M examples, matches or exceeds larger hybrid models, proving that **pretrained backbone quality can compensate for pretraining size**.

That said, scaling data indiscriminately introduces **noise and bias**, which can compromise interpretability and fairness. Modern trends emphasize **curation, filtering, and balanced sampling** over brute-force scale. **BLIP's bootstrapping approach**, which filters noisy data, represents a shift toward high-signal datasets for downstream finetuning.

In conclusion, while more data generally improves performance, **smarter data**, combined with multi-task pretraining objectives and high-quality initialization, appears to be more impactful than scale alone.

5.5. Explainability and Model Interpretability

As vision-language models become more capable, the demand for **interpretable and trustworthy AI systems** increases. Especially in critical domains (e.g., healthcare, autonomous navigation), stakeholders must understand the rationale behind model outputs.

Hybrid models like **OSCAR** and **VinVL**, which incorporate explicit object tags and regional proposals, offer more intuitive explanations. Their attention mechanisms are grounded in tangible object detections, which can be visualized or labeled for interpretability.

In contrast, fully Transformer-based models such as **ViLT** and **METER** rely on **patch-based embeddings**, which are harder to interpret at the semantic level. Visualization techniques like **attention heatmaps**, **Grad-CAM**, and **attention rollout** provide some transparency but have limitations:

- **Grad-CAM** excels at localizing class-specific features but struggles with high-level interactions.
- **Attention maps** highlight which tokens are being attended to but do not always correspond to human-understandable reasoning.
- **ReVisE** and similar frameworks attempt to align visual and textual rationales, offering better human-aligned interpretability.

Nevertheless,**faithfulness** i.e., whether the explanation reflects the true decision-making process remains an open challenge. Models sometimes produce plausible explanations for wrong outputs, raising concerns about over-reliance on visualization tools.

Moving forward, integrating **object-level grounding with ViT-based reasoning**, along with **counterfactual reasoning modules**, may provide more robust interpretability while retaining accuracy.

5.6. Future Directions in Vision-Language Modeling

The state of vision-language modeling is rapidly evolving, and several promising directions are emerging:

- **Scalable, Unified Architectures:** Large-scale models like **PaLI** (10B images, multilingual) and **Flamingo** demonstrate the feasibility of unified encoder-decoder systems for all vision-language tasks. Future models may consolidate captioning, VQA, retrieval, and OCR into a single, parameter-efficient pipeline with strong cross-task generalization.
- **Lightweight and Efficient Transformers:** Architectures like **MobileViT**, **Tiny-ViT**, and **Distilled BLIP** are leading efforts to compress large models for edge deployment. Techniques such as quantization, pruning, and adapter layers will play a major role in democratizing vision-language AI.
- **Contrastive + Fusion Objectives:** As seen in **ALBEF** and **BLIP**, combining contrastive alignment (CLIP-style) with fusion-based objectives (UNITER-style) yields richer representations. Expanding this duality across more modalities (e.g., video, 3D) could unify multimodal understanding and generation.
- **Multimodal Reasoning and Chain-of-Thought:** Integrating reasoning frameworks into vision-language models—like step-wise explanation generation or question decomposition—will enable deeper comprehension and trustworthy AI. The pairing of visual models with language LLMs (e.g., GPT-4 + BLIP) opens new avenues for interactive and explainable systems.
- **Bias and Fairness Mitigation:** As training datasets scale, so do risks of encoding social biases. Future research must include:
  - Dataset audits (gender/race/cultural bias analysis)
  - Adversarial debiasing strategies
  - Diverse pretraining corpora to reduce representational disparities
- **Vertical and Domain-Specific Applications:** Vision-language models are expanding into specialized domains:
  - **Medical Imaging:** Models like RadImageNet and MedCLIP adapt general frameworks for disease detection and report generation.
  - **Remote Sensing:** Multimodal models are being used for environmental monitoring, crop health prediction, and disaster analysis.
  - **Industrial QA:** Few-shot and zero-shot models are enabling defect detection with minimal supervision.
- **Open-World and Continual Learning:** Real-world applications require models that adapt to new objects, concepts, or languages. Continual pretraining and open-vocabulary recognition (as seen in **OWL-ViT**) represent early steps toward this goal.

In summary, hybrid CNN-Transformer architectures have served as powerful tools for multimodal learning. However, the future clearly lies in **streamlined, fully Transformer-based systems** that balance **accuracy, efficiency, and scalability**. As models become more general-purpose, adaptable, and interpretable, they will play a transformative role across industries and societies.

6. Conclusion and Future Work

This study conducted a comprehensive benchmark of vision-language models, focusing on hybrid CNN-Transformer architectures in comparison to CNN-only and Transformer-only alternatives. The evaluation spanned three major tasks: image captioning, visual question answering (VQA), and image-text retrieval, using widely recognized datasets such as MS COCO, VQAv2, and Flickr30k. Our findings revealed that while hybrid models such as ViLBERT, OSCAR, and VinVL have historically dominated performance metrics due to their effective integration of visual object detection and transformer-based fusion, recent Transformer-only models like METER and BLIP are achieving comparable, and in some cases superior, results with significant improvements in inference speed and simplicity.

These results demonstrate that Transformer-only models, particularly those leveraging pre-trained vision encoders such as CLIP or ViT, are now capable of handling vision-language tasks at or above the level of older hybrid approaches. Not only do these models achieve high accuracy, but they also offer a streamlined architecture that is easier to deploy in real-world applications. For instance, models like ViLT and METER have drastically reduced inference time by removing the computational overhead introduced by region-based CNN detectors, making them more suitable for edge computing and mobile deployment.

The implications of these findings are substantial for both research and practical deployment. Firstly, there is a clear shift in the field toward unified Transformer architectures that do not rely on external modules or complex pre-processing steps. This simplification allows for more efficient training, easier maintenance, and broader applicability across domains. Secondly, it highlights the increasing importance of high-quality pretraining over model architecture. As evidenced by the performance of METER and BLIP, the scale and diversity of pretraining data have a more significant impact on final model accuracy than the specific combination of visual and textual encoders used.

In terms of deployment, Transformer-only models present a clear advantage due to their low latency and reduced memory requirements. Their ability to run on CPUs or lightweight GPUs makes them ideal for real-time applications in mobile devices, augmented reality systems, and autonomous systems. Additionally, models trained with contrastive learning objectives, such as CLIP, demonstrate strong zero-shot capabilities, enabling them to perform well on tasks without task-specific fine-tuning, which is highly valuable in dynamic environments where retraining is impractical.

Despite these advances, several areas remain open for further research. One key challenge is improving fine-grained understanding in Transformer models. While these models excel at general scene comprehension, they sometimes lack precision in recognizing small or context-specific objects. Addressing this issue may involve developing new attention mechanisms or integrating adaptive region proposals into the Transformer pipeline without reintroducing heavy computational costs. Another important direction is to make pretraining more data-efficient. Current state-of-the-art models rely on hundreds of millions of image-text pairs, which is not feasible for all researchers. Future work should explore semi-supervised, self-supervised, or few-shot learning techniques that reduce dependence on massive datasets.

Moreover, bias and fairness remain critical concerns. Vision-language models are prone to replicating societal biases embedded in training data. Therefore, further research is required to detect, measure, and mitigate these biases to ensure the development of ethical and inclusive AI systems. In addition, the field would benefit from the creation of unified, holistic evaluation benchmarks that assess not just accuracy but also robustness, fairness, and explainability across a variety of tasks and modalities.

There is also significant potential in extending vision-language models to support multilingual and multimodal settings. Most existing models are trained solely on English datasets, limiting their utility in global applications. Expanding training and evaluation to include multiple languages and modalities such as audio, video, and 3D would greatly enhance their accessibility and capability.

To improve unified vision-language models, several enhancements can be considered. One approach is to develop modality-agnostic transformers that treat visual patches, text tokens, and other modality inputs (like audio) as generic embeddings, enabling a single model to process diverse inputs uniformly. Another promising direction involves incorporating external knowledge sources such as knowledge graphs to improve reasoning capabilities, particularly for tasks that require understanding world knowledge or contextual associations not directly visible in the input data.

Attention sparsity and adaptive computation could also enhance model efficiency, allowing the network to focus computational resources dynamically on the most relevant inputs. This would help maintain performance while reducing resource consumption, a key factor for deployment on edge devices. Furthermore, integrating prompt-based or instruction-tuned mechanisms, similar to those used in large language models like GPT, would make vision-language systems more interactive and user-controllable. Finally, the incorporation of continual learning mechanisms would allow these models to evolve over time with user feedback and new data, thereby increasing their adaptability and long-term value.

In conclusion, while hybrid CNN-Transformer architectures have played a crucial role in advancing the field of vision-language understanding, the momentum has clearly shifted toward Transformer-only models. These models offer greater simplicity, efficiency, and scalability without compromising performance. The future of vision-language AI lies in building unified, efficient, and ethically grounded models capable of handling a wide range of tasks and modalities in real time. Through focused research on data efficiency, fairness, multilinguality, and user-aligned prompting, the next generation of multimodal AI systems will become even more powerful and widely accessible.

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. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086).
3. Chen, Y. C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., ... & Liu, J. (2020, August). Uniter: Universal image-text representation learning. In European conference on computer vision (pp. 104-120). Cham: Springer International Publishing.
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).
5. Dou, Z. Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., ... & Zeng, M. (2022). An empirical study of training end-to-end vision-and-language transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18166-18176).
6. Fan, L., Li, T., Yuan, Y., & Katabi, D. (2020). In-home daily-life captioning using radio signals. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16 (pp. 105-123). Springer International Publishing.
7. Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., ... & Zweig, G. (2015). From captions to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1473-1482).
8. Fartash, F., Fleet, D., Kiros, J., & Fidler, S. (2018). VSE++: Improved visual semantic embeddings. In British Machine Vision Conference (pp. 935-943).
9. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6904-6913).
10. Kim, J. H., Jun, J., & Zhang, B. T. (2018). Bilinear attention networks. Advances in neural information processing systems, 31.
11. Kim, W., Son, B., & Kim, I. (2021, July). Vilt: Vision-and-language transformer without convolution or region supervision. In International conference on machine learning (pp. 5583-5594). PMLR.
12. Lee, K. H., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked cross attention for image-text matching. In Proceedings of the European conference on computer vision (ECCV) (pp. 201-216).
13. Li, J., Li, D., Xiong, C., & Hoi, S. (2022, June). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International conference on machine learning (pp. 12888-12900). PMLR.
14. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34, 9694-9705.
15. Li, L. H., Yatskar, M., Yin, D., Hsieh, C. J., & Chang, K. W. (2019). Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

16. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... & Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16 (pp. 121-137). Springer International Publishing.
17. 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, 32.
18. Meng, L., Li, H., Chen, B. C., Lan, S., Wu, Z., Jiang, Y. G., & Lim, S. N. (2022). Adavit: Adaptive vision transformers for efficient image recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12309-12318).
19. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).
20. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.
21. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.
22. Sadhu, A., Gupta, T., Yatskar, M., Nevatia, R., & Kembhavi, A. (2021). Visual semantic role labeling for video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5589-5600).
23. Sakaridis, C., Dai, D., Hecker, S., & Van Gool, L. (2018). Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In Proceedings of the european conference on computer vision (ECCV) (pp. 687-704).
24. Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018, July). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2556-2565).
25. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
26. Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490.
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
28. Karakolias, S. (2024). Mapping data-driven strategies in improving health care and patient satisfaction.
29. Ardjomandi, A. (2025). The role of narrative and storytelling in designing for long-term emotional engagement in product design.
30. Singu, S. K. (2022). Agile Methodologies in Healthcare Data Warehousing Projects: Challenges and Solutions. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-400. DOI: doi. org/10.47363/JAICC/2022 (1), 383, 2-5.
31. Barach, J. (2025, January). Towards Zero Trust Security in SDN: A Multi-Layered Defense Strategy. In Proceedings of the 26th International Conference on Distributed Computing and Networking (pp. 331-339).
32. Georgi, C., Georgis, V., & Karakolias, S. (2023). HSD79 Assessment of Patient Satisfaction with Public Pharmacies Dispensing High-Cost Drugs in Greece. Value in Health, 26(12), S308-S309.
33. Singu, S. K. Performance Tuning Techniques for Large-Scale Financial Data Warehouses.
34. Aburidi, M., Aritsugi, M., Barach, J., Benslimane, S., Eggenkemper, F., Ethirajan, L., ... & Wang, H. Xiao, Ling 62 Yamasaki, Toshihiko 62 Yoshida, Shun 62 Zhou, Yu.
35. ARDJOMANDI, A. (2025). Visual Semiotics and User Perception in Digital Interface Design.
36. Psarras, A., & Karakolias, S. (2024). A Groundbreaking Insight Into Primary Care Physiotherapists' Remuneration. Cureus, 16(2).
37. Barach, J. (2025, February). AI-Driven Causal Inference for Cross-Cloud Threat Detection Using Anonymized CloudTrail Logs. In 2025 Conference on Artificial Intelligence x Multimedia (AIxMM) (pp. 45-50). IEEE.
38. Karakolias, S., Georgi, C., & Georgis, V. (2024). Patient Satisfaction With Public Pharmacy Services: Structural and Policy Implications From Greece. Cureus, 16(4).
39. Karakolias, S., & Iliopoulou, A. (2025). Health-Related Quality of Life and Psychological Burden Among and Beyond Children and Adolescents With Type 1 Diabetes: A Family Perspective. Cureus, 17(4).
40. Barach, J. (2024, December). Enhancing Intrusion Detection with CNN Attention Using NSL-KDD Dataset. In 2024 Artificial Intelligence for Business (AIxB) (pp. 15-20). IEEE.
41. Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).
42. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
43. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
44. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.
45. Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the association for computational linguistics, 2, 67-78.
46. Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6720-6731).
47. Zhang, Y., Li, X., Liu, C., Shuai, B., Zhu, Y., Brattoli, B., ... & Tighe, J. (2021). Vidtr: Video transformer without convolutions. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 13577-13587).
48. Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9), 2337-2348.
49. Periyasamy, R., Sasi, S., Malagi, V. P., Shivaswamy, R., Chikkaiah, J., & Pathak, R. K. (2025). Artificial intelligence assisted photonic bio sensing for rapid bacterial diseases. Zeitschrift für Naturforschung A, (0).
50. Raj, L. V., Sasi, S., Rajeswari, P., Pushpa, B. R., Kulkarni, A. V., & Biradar, S. (2025). Design of FBG-based optical biosensor for the detection of malaria. Journal of Optics, 1-10.
51. Rajeswari, P., & Sasi, S. (2024). Efficient k-way partitioning of very-large-scale integration circuits with evolutionary computation algorithms. Bulletin of Electrical Engineering and Informatics, 13(6), 4002-4007.
52. Sasi, S., Rajeshwari, P., Ramkumar, R., & Mondal, S. (2024, November). Lumina-Secure Access Guard. In 2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI) (pp. 57-61). IEEE.

53. Sasi, S., Subbu, S. B. V., Manoharan, P., Kulkarni, A. V., & Abualigah, L. (2025). Design and Implementation of Discrete Field Arithmetic-Based Cylindrical Coil-Driven Crypto Framework for Cloud Data. Journal of Computational and Cognitive Engineering, 4(1), 97-107.
54. Wang, F., Bao, Q., Wang, Z., & Chen, Y. (2024, October). Optimizing Transformer based on high-performance optimizer for predicting employment sentiment in American social media content. In 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 414-418). IEEE.
55. Xi, K., Bi, X., Xu, Z., Lei, F., & Yang, Z. (2024, November). Enhancing Problem-Solving Abilities with Reinforcement Learning-Augmented Large Language Models. In 2024 4th International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI) (pp. 130-133). IEEE.
56. Penmetsa, S. V. (2024, September). Equilibrium Analysis of AI Investment in Financial Markets under Uncertainty. In 2024 IEEE International Conference on Cognitive Computing and Complex Data (ICCD) (pp. 162-172). IEEE.
57. Wairagade, A. (2024, December). Enhancing Behavioral Analytics with Zero Trust in Cloud: A Comparative Analysis. In 2024 International Conference on Engineering and Emerging Technologies (ICEET) (pp. 1-7). IEEE.
58. Zhong, J., Wang, Y., Zhu, D., & Wang, Z. (2025). A Narrative Review on Large AI Models in Lung Cancer Screening, Diagnosis, and Treatment Planning. arXiv preprint arXiv:2506.07236.
59. Zhong, J., Wang, Y., Zhu, D., & Wang, Z. (2025). A Narrative Review on Large AI Models in Lung Cancer Screening, Diagnosis, and Treatment Planning. arXiv preprint arXiv:2506.07236.
60. Zhong, J., & Wang, Y. (2025). Enhancing Thyroid Disease Prediction Using Machine Learning: A Comparative Study of Ensemble Models and Class Balancing Techniques.