# Toward Human-Level Understanding: A Systematic Review of Vision-Language Models for Image Captioning

Ansar Hani
Department of Computer Engineering and Mathematics, University of Sfax, Tunisia
ansar.hani@stud.enis.tn

Najiba Tagougui
Department of Computer Science Engineering, University of Sfax, Tunisia
najiba.tagougui@isims.usf.tn

Monji Kherallah
Department of Physics University of Sfax, Tunisia
monji.kherallah@fss.usf.tn

**Abstract:** *Large Language Models (LLMs), particularly multimodal LLMs, have significantly enhanced image captioning in recent years, producing output that is more descriptive, detailed, and context-aware. However, differences in architecture and training data lead to captions that vary in length, style, and level of detail, offering flexibility for diverse applications. In this survey, we provide a comprehensive overview and comparative analysis of prominent Vision-Language Models (VLMs) for image captioning, with a focus on their performance in zero-shot settings on the Microsoft Common Objects in Context (MS-COCO) dataset. We evaluate these models using both human assessments (fluency, groundedness, relevance) and automatic metrics Contrastive Language–Image Pretraining Score (CLIPScore). Our findings reveal trade-offs between efficiency and performance, linking architectural decisions to issues such as hallucinations and caption grounding. Beyond benchmarking, we propose a human evaluation to capture nuances like fluency, factual grounding, and stylistic preferences, leading to recommendations for selecting VLMs based on different use cases.*

**Keywords:** *Image captioning, vision-language models, multimodal learning, large language models.*

*Received May 10, 2025; accepted September 1, 2025*
*https://doi.org/10.34028/iajit/23/1/8*

## 1. Introduction

The image captioning task involves generating accurate, relevant, and human-like sentences by blending Computer Vision (CV) and Natural Language Processing (NLP). This fundamental visual understanding task has attracted a lot of interest due to its substantial implications in real-world applications like data labeling, accessibility aids, and content production. The emergence of deep learning, in particular Convolutional Neural Networks (CNNs) [26] for image encoding and Recurrent Neural Networks (RNNs) [4] such as Long Short-Term Memory Networks (LSTMs) for sequence generation, revolutionized captioning task under flexible and context-aware encoder-decoder framework.

The introduction of attention mechanisms [53] improved both relevance and coherence through dynamic area focusing while generating captions. The transformer architecture [44] replaced recurrence with self-attention to achieve better scalability and parallelization capabilities. Then, Vision Transformers (ViTs) [19] revolutionized the field by using image patches as sequences to merge visual and textual data processing. Despite these advances, progress in image captioning has faced persistent challenges. Many models produce overly generic captions, lacking specificity and informativeness. This stems from limitations in training data, model architecture, and decoding strategies, which often prioritize syntactic fluency over factual grounding and semantic precision.

Recent developments in Large Language Models (LLMs), when integrated with vision encoders, have revitalized the field. These multimodal architectures, combining rich linguistic priors with visual understanding, have shown potential for generating captions that are more abstract, context-aware, and diverse. However, the performance of these models still varies considerably depending on architecture, training objectives, and modality integration strategies.In this survey, we first present a systematic comparison of state-of-the-art Vision-Language Models (VLMs) on the Microsoft Common Objects in Context (MS-COCO) dataset [32] under zero-shot settings. Second, we propose a multi-perspective human evaluation approach through diverse lenses, including stylistic adaptability (descriptive vs. concise), syntactic structure (grammaticality and fluency), and use-case suitability (accessibility, creative generation, and technical applications). This approach allows us to uncover nuanced performance trade-offs, such as the tension between creativity and correctness, that are often missed by standard benchmarks.

## 2. Literature Review

### 2.1. Standard Image Captioning

From early approaches that relied on retrieval-based and template-based methods to advanced deep learning paradigms, image captioning has undergone meaningful change. While retrieval-based models [28] searched for similar images in a database and reused the original captions, they don't adapt well to new examples and don't provide much creativity. Template-based methods [27], on the other hand, employ fixed sentence structures with object labels to generate captions that are inflexible and unnatural. These methods were straightforward; however, they were limited in their flexibility and generalization capabilities. Numerous techniques have been proposed in the era of deep learning. Subsequent research led to the exploitation of encoder-decoder architectures, where a CNN is used to encode the visual input, and an RNN is used to condition the generation process. RNNs are often used with Long Short-Term Memory (LSTM) networks for decoding and generating the caption from the visual features [18, 25, 47]. Then LSTMs were replaced by Gated Recurrent Units (GRUs) [22] and provide performant results also.

Image captioning methods have been further improved by introducing attention mechanisms that allow the model to focus on important parts of the image when generating each word. The Show, Attend and Tell model [53] added a soft visual attention mechanism that improved the quality of captioning and its alignment with human descriptions. Follow-up work like the bottom-up and top-down attention model [2] extended this idea through object-level attention using parts of region proposal networks, such as Faster R-CNN, to provide richer and fine-grained features. Despite their effectiveness, these models frequently struggle to incorporate broader domain knowledge and fail to adapt to diverse contexts, which limits their ability to manage complex visual scenes and capture long-range dependencies within captions.

Driven by the success of transformers in NLP, recent image captioning research leverages transformers to model intra-modal interactions for automatic caption generation [15, 16, 23]. The initial adoption of transformers replaced RNNs in the decoder, capitalizing on parallel training capabilities. Recent work has also explored transformer-based approaches for image captioning, demonstrating the effectiveness of multi-encoder architectures in improving semantic coherence and contextual alignment [39]. Visual representations are typically derived using either a pre-trained object detector or a vision transformer, which can be applied directly to image patches, reducing or eliminating the reliance on convolutional operations. ViTs are becoming more popular, thus they generate contextually rich, coherent captions that better capture nuanced scenes, even with complex or lengthy descriptions. This enables them to adapt better to different contexts and tasks, and to produce more accurate and richer captions than early methods. Despite being relatively effective, encoder-decoder architectures have issues regarding reasoning, situatedness, and the level of semantics, especially in complex scenes. This raised the possibility of developing vision-language pretraining methods [12, 48, 54].

### 2.2. Multimodal Large Language Models (MLLM) for Image Captioning

Recent advancements in image captioning have demonstrated how LLMs are able to assist in understanding the visual signal. Therefore, image captioning combines LLMs and vision encoders to produce informative and accurate image descriptions. These models consider both visual and textual sources, facilitating an understanding of complex features to develop a full interpretation of the content. Bidirectional Encoder Representations from Transformers (BERT) [17] and Generative Pre-trained Transformer (GPT) [56] demonstrated the initial potential of LLMs, achieving significant advancements in few-shot and zero-shot learning and inspiring scaling efforts that yielded models like T5 [43], GPT-3 [8], Flan-T5 [13], and PaLM [14]. In the past year, large-scale Multimodal Large Language Models (MLLMs) have exhibited remarkable performance across a wide range of downstream tasks like visual dialogue, image captioning, and visual question answering [9].

Building on this progress, these MLLMs typically bridge visual and language modalities by connecting a pre-trained LLM with a large-scale visual encoder, such as Contrastive Language-Image Pretraining (CLIP) [42] or its variants. These models interpret both text and images, providing them with background knowledge to generate high-quality captions that refer to the objects and scenes depicted in the image while embedding contextual information and conveying a deeper understanding of the visual content. These models have been shown to have superior performance on several image captioning benchmarks and are also capable of changing the landscape of computer vision.

MLLMs are often categorized by their multimodal connection type, with many, like the Large Language-and-Vision Assistant (LLaVA) series [35, 36, 37], using an MLP [7, 52] or linear layer [11, 33] to establish multimodal connections. Several variations have been introduced, such as LLaMA-Adapter [21] that uses a zero-gating attention mechanism, while Cha *et al*. [10] replace linear layers with convolutions. Q-Former-based models [30] represent another major category.

Consequently, mPLUG-Owl [55] streamlines Q-Former with a visual abstractor, condensing visual information into trainable tokens. Qwen-VL [5] similarly uses a single-layer cross-attention module with learnable queries to compress visual features. Alternatively, some methods integrate dense cross-attention blocks within pre-trained LLM layers [1, 3], often employing a Perceiver model [24] to reduce visual tokens before integration. While MLLMs are undergoing rapid changes, they have not yet been explored in image captioning. There are a few MLLMs that have been specifically trained and evaluated using standard benchmarks, and most of the work has treated image captioning as an intrinsic capability. Recent work [34, 49, 57] has begun to measure hallucination of MLLMs, an important consideration for the detailed usage of MLLMs to create image captions. This paper assesses the performance of standard MLLMs in creating image captions and a number of fine-tuning methods to assist in adapting to this task, which includes a clear differentiation from existing literature. Standard image captioning approaches developed foundation models by learning to map visual input to textual output using explicit alignment mechanisms. However, these models need task-specific training, and they cannot generalize in a zero-shot setting. In contrast, VLMs can generalize well, learn about semantic grounding, and afford flexibility of tasks because of extensive pre-training on web-scale data. A significant shift occurred by moving from RNNs to transformers, from supervised training to contrastive and generative pre-trained training, and from isolated image encoders to unified multimodal architectures. As researchers continue to advance the future of VLMs, we expect the integration of richer lexical knowledge, enhanced grounding, improved reasoning capabilities, and support for multilingual and multimodal inputs, enabling deeper alignment between visual and linguistic understanding.

# 3. Methodology

## 3.1. Models' Selection

Our selection consists of eight VLMs developed from 2022 to 2024, representing both the chronological and the conceptual evolution in image captioning. These models have different strengths in linguistic fluency, visual grounding, task generalization, and computational efficiency; they also reflect different architectures from early encoder-decoder baselines to MLLMs, as depicted in Figure 1, including contrastive pre-training, instruction tuning, modular LLM integration, grounded generation, and efficient decoding. Models like ViT-GPT2 [38] are adopted on minimalist architectures, thus they provide greater accessibility and simplicity. This model consists of a ViT connected to a GPT-2 decoder, with a linear projection layer acting as a bridge in order to create a simple-to-train, good-performing baseline for image captioning. Although simple and effective, its generality and lack of spatial awareness revealed the need for more complex architectures that better combined visual and textual data. OFA [51] was selected as a foundational model for its pioneering unification of vision-language tasks, which emphasized capability over computational efficiency. GIT [50] achieved architecture simplification and top performance even on images with a large amount of text, though it is less intuitive for some applications. BLIP-2 [30] was favoured as it has the best zero-shot potential, allowing for deployment without large language model fine-tuning considerations. LLaVA [37] was Selected due to its prompt-based system of allowing users flexibility, though it raises some uncertainties in output rigour.

Kosmos-2 [41] provided spatial grounding for localising objects in 3D space, crucial for scene understanding, but increased complexity. Fuyu-8B [6] is notable because it has efficient performance in processing high-resolution data valuable for applications like digital agents; however, it lacks any aspects of dynamic representation. Moodream-2 [46] was tentatively included as an exploratory emerging model, representing a forward-looking perspective, though its speculative nature, and is thus included for exploratory comparison.

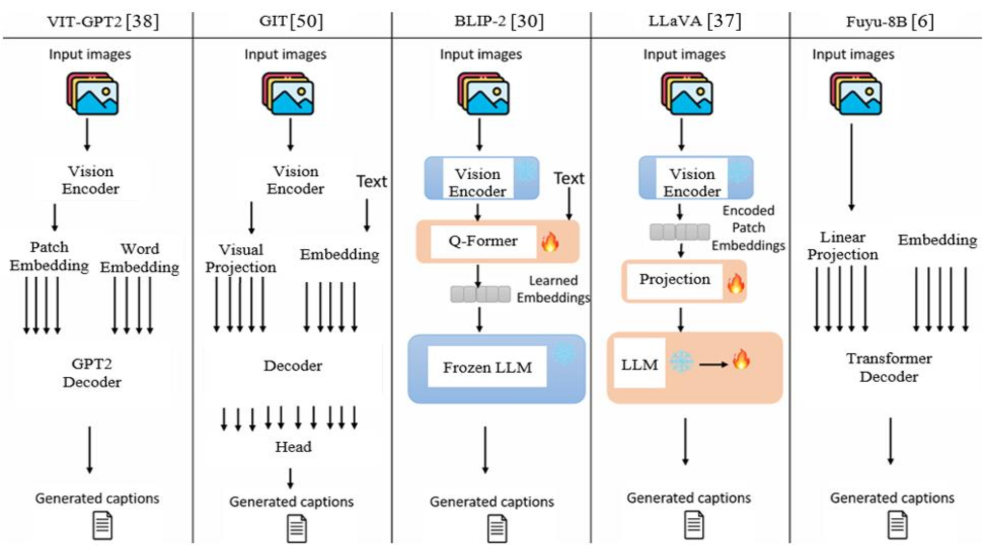

Figure 1. Image captioning pipelines in vision-language models: a structural overview.

Overall, the development from early models that rely largely on alignment to modular, grounded, and hybrid generative systems demonstrates a consistent effort to balance fluency, grounding, task generalization, and computational efficiency. Each model implements its own responses to their predecessors' limitations but also illustrates the increasing scope and ambition of vision-language integration for the task of image captioning.

A brief comparison of the selected models is presented in Table 1. This table analyses their architecture, performance, and contributions to vision-language understanding, particularly in image captioning.

Table 1. Comprehensive comparison of vision-language models for image captioning.

| Model | Year | Architecture | Training data | Parameters | Vision encoder/tokenizer | Pre-trained backbone model |
|---|---|---|---|---|---|---|
| **ViT-GPT2 [38]** | 2021 | Encoder-Decoder (ViT + GPT-2) | MS COCO, Flickr30k | ~124M (ViT + GPT-2) | ViT (Vision Transformer) + GPT-2 Tokenizer | ViT (ImageNet-21k), GPT-2 |
| **OFA [51]** | 2022 | Unified Transformer | Multi-task (COCO, VQA, NLVR, etc.) | ~930M | ResNet-101 + Transformer | BART, ResNet |
| **GIT [50]** | 2022 | Encoder-Decoder (ViT + Transformer Decoder) | 800M image-text pairs (filtered) | GIT-Base: 345M | ViT (Huge) + BERT Tokenizer | ViT-Huge (CLIP pre-training) |
| **BLIP-2 [30]** | 2023 | Two-Stage (Image Encoder → Q-Former → LLM) | 129M image-text pairs + synthetic data | BLIP-2 OPT2.7B / FLAN-T5 XXL | ViT-G / Q-Former | ViT-G, OPT/FLAN-T5 |
| **LLaVA [37]** | 2023 | Vision Encoder + connector + LLM (Vicuna) | COCO, Visual Genome, synthetic instruction tuning | ~13B (with Vicuna) | CLIP ViT-L/14 | CLIP, Vicuna |
| **Kosmos-2 [41]** | 2023 | Multimodal LLM with visual grounding | Web-scale multimodal data | ~1B–1.6B | Patch embedding → Linear projection | BERT-like encoder |
| **Fuyu-8B [6]** | 2023 | Decoder-Only Transformer (GPT-style) | Public image-text datasets + Optical Character Recognition (OCR) documents | 8B | Vision tokenizer into sequences | GPT-style pre-trained transformer |
| **Moondream2 [46]** | 2024 | two major components: SigLIP, Phi-1.5 | LLaVa training dataset | 2B | SigLIP as the vision encoder and Phi-1.5 as the text encoder | SigLIP, Phi-1.5 (LLM) |

## 3.2. Experimental Process

To evaluate the performance of recent VLM on image captioning, we use various pretrained models that have been officially released. This included ViT-GPT2 [38], OFA [51], GIT [50], BLIP-2 [30], LLaVA [37], Kosmos-2 [41], Fuyu-8B [6], and Moondream2 [46]. All models were evaluated in a zero-shot setting, and all models were officially released and used inference pipelines. Code implementation is available on my GitHub repository https://github.com/ansar2019/image-captioning.

To systematically evaluate the generalization performance of VLMs, we carefully created a comprehensive evaluation set of 1,000 images sampled frome the MS COCO 2014 test set, as shown in Figure 2. This subset was built around a category-aware sampling process that increases both the semantic span and diversity while controlling the variability. This was achieved by balancing representations from seven semantic groups:

1. People, including portraits and social gatherings.
2. Animals, including wild-life and pets.
3. Scenes, both indoor and outdoor.
4. Food and meal contexts.
5. Places, including natural and built landmarks.
6. Types of vehicles, including cars, planes, and boats.
7. Sport and activity scenarios. By bringing this level of semantic coverage.

Figure 2. Sample images from the MS COCO test dataset.

We anticipated that the evaluation set would allow us to include the richness of real-world photo content that captioning models might encounter. The image selection also sought to maximize object category variation while retaining the true distributions of scenes, to facilitate representativity and challenge. That ultimately provides a more robust evaluation of models across varied and realistic contexts.

Five expert annotators independently evaluated VLM-generated captions for 1,000 MS-COCO images, manually assessing fluency, grounding, richness, relevance, and error types using standardized rubrics.

Following individual scoring, trainers participated in structured discussion sessions to resolve discrepancies, focusing on three key criteria:

1. Syntactic validity (grammar and coherence).
2. Semantic alignment (object/action fidelity to the image).
3. Descriptive utility (detail appropriateness for target applications).

Disputed captions (18.3% of cases, primarily in crowded scenes) underwent iterative review until consensus was reached, with deliberation notes cataloging recurring failure modes like spatial relation errors (“man left of tree” vs. “man beside tree”). This consensus-driven approach yielded refined evaluation guidelines that informed our proposed VLM output structure taxonomy, categorizing errors into hallucination subtypes (attribute, object, or relation) and omission tiers (primary object vs. contextual detail). The results of this work will be presented in detail in the next section in a structured qualitative analysis of the VLM-generated caption judgment.

## 3.3. Zero-Shot Inference Implementation Details

To guarantee fairness and reproducibility, all models were assessed under zero-shot settings, indicating that no fine-tuning, supplementary supervision, or domain-specific adaptation was utilized. We employ only the officially released pretrained checkpoints and public inference APIs or repositories made available by the original authors or developers.

We followed the recommended inference pipeline for each model, which is available on open-source platforms like GitHub, HuggingFace, and model-specific demo APIs. This included using tokenizers, vision encoders, and decoding strategies. We used prompt templates where necessary, as explained in the model documentation. This was especially true for instruction-tuned or conversational models like LLaVA, Kosmos-2, Fuyu-8 Band Moondream2, which use task-specific prompt formatting to guide the generation. Table 2 shows a summary of the evaluation platform, programming libraries, and model-specific dependencies.

Table 2. Summary of VLMs, inference tools, and prompt usage.

| Model | Platform/Repository | Inference API/Library | Prompt used |
|---|---|---|---|
| **ViT-GPT2 [38]** | HuggingFace (nlpconnect/vit-gpt2-image-captioning) | transformers pipeline (image-captioning) | No explicit prompt (internal defaults) |
| **OFA [51]** | OFA-Sys GitHub (OFA, OFA-Large) | Official PyTorch/fairseq-based framework | Prompt “What does the image describe?” |
| **GIT[50]** | Hugging Face (microsoft/git-base) | transformers image-captioning pipeline | No prompt needed |
| **BLIP-2 [30]** | HuggingFace (ethzanalytics/BLIP-2-flan-t5-xl-sharded) | transformers: BlipProcessor+BLIP-2ForConditionalGeneration | Prompt=“Describe this image.” |
| **LLaVA [37]** | LLaVA GitHub (llava-v1.5) | Uses LLaVA GitHub with quantized model weights (llava-13b-4bit) | Prompt=(“Describe this image.” or “What’s in the image?”) |
| **Kosmos-2 [41]** | Hugging Face/Microsoft/Transformers docs | ONNX runtime or PyTorch | Prompt=“Describe the image: [IMAGE]” |
| **Fuyu-8B [6]** | HuggingFace (adept/fuyu-8b) | Transformers using decoder-only architecture | Prompt=“Describe the image in detail.” |
| **Moondream2 [46]** | HuggingFace (vikhyatk/moondream2) | Transformers (AutoModelForCausalLM+tokenizer) | Prompt=“What’s in this image?” |

Figure 3 illustrates the captions generated by the selected VLMs for a sample image taken from the training dataset, comparing their outputs to the corresponding ground-truth caption and highlighting differences in descriptive richness and semantic alignment across models. In contrast, Figure 4 presents the captions produced under zero-shot settings for multiple images from the test dataset, providing a broader view of each model’s generalization capability when applied to unseen visual samples.

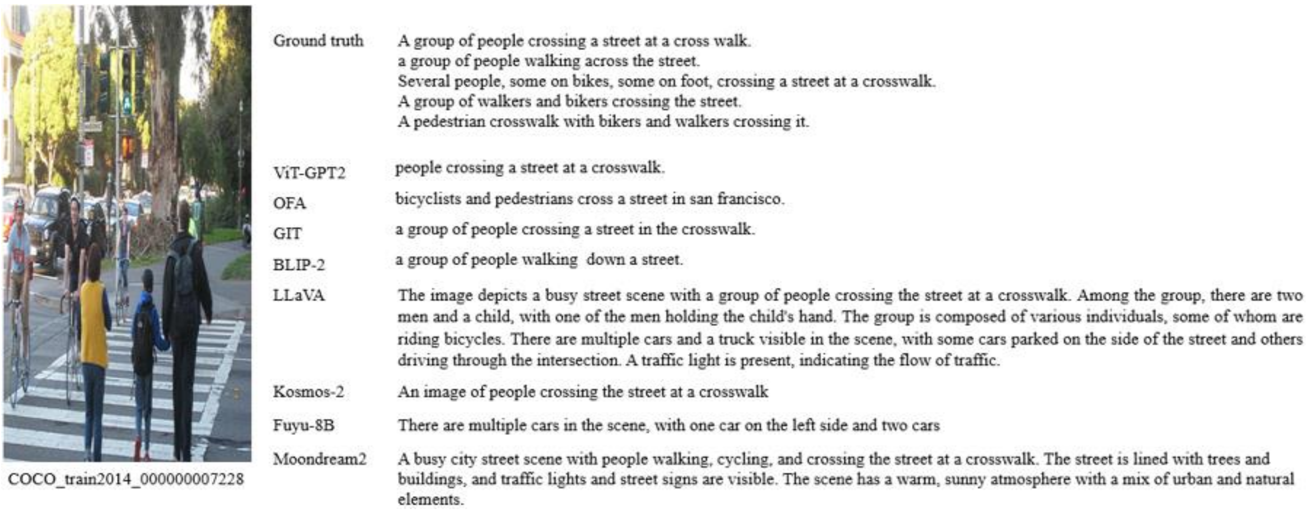


Figure 3. Model-specific caption outputs compared to ground truth descriptions.

COCO_test2014_000000104076

ViT-GPT2: people walking down a street with umbrellas.

OFA : a group of people walking down a street with umbrellas.

GIT: people walking on a street with umbrellas.

BLIP-2 : group of people walking down a street with umbrellas.

LLaVA : The image depicts a group of people walking down a street in the rain. They are using umbrellas to protect themselves from the rain, with a total of five umbrellas visible in the scene. The people are carrying various items, such as handbags and a backpack. There are at least nine people in the scene, with some walking closer to the camera and others further away. The street is lined with stores, and the overall atmosphere suggests a typical rainy day in a city.

Kosmos-2: An image of people crossing the street in the rain.

Fuyu-8B :The image features a group of people walking down a street while holding umbrellas to protect themselves from the rain.

Moondream2: A busy street scene in Paris, France, with people walking on the sidewalk and crossing the street. Many of them are holding umbrellas, likely to protect themselves from the rain. The street is lined with shops, including a H&M store. The buildings have large windows and awnings, typical of Parisian architecture. The street is wet, reflecting the city lights. The image has a slightly overexposed or washed-out look, giving it a vintage or nostalgic feel. The colors are mostly muted, with some bright umbrellas standing out. The Overall atmosphere is bustling and lively.

COCO_test2014_000000090294

ViT-GPT2: a plate of food with a salad and a fork.

OFA : a plate of food with broccoli and vegetables and a bowl of fruit.

GIT : a plate of food with a side of vegetables and a fork.

BLIP-2: a plate of food with a lemon on it.

LLaVA: The image features a dining table with a white plate filled with a variety of food items. The plate contains a mix of vegetables, including broccoli and peas, as well as a lemon wedge. There is also a piece of meat on the plate. The table setting includes a fork, a knife, and a spoon placed around the plate. In addition to the main plate, there are two bowls on the table, one on the left side and the other on the right side. The left bowl contains strawberries, while the right bowl contains a different type of food. The table setting appears to be a complete and balanced meal, with a focus on healthy and nutritious food choices.

Kosmos-2 : An image of a plate of food and a cup of coffee.

Fuyu-8B : The image features a dining table with a white plate of food, silverware, and a bowl of fruit. On the table, there is a bowl of strawberries, a fork, a knife, and a spoon. The fork and knife are positioned on the left side of the plate, while the spoon is on the right.

Moondream2: A plate of healthy food on a light blue countertop. The plate contains a variety of colorful vegetables, including broccoli, peas, and roasted squash. There are also lemon wedges and sliced chicken on the plate. To the left of the plate is a bowl of fresh strawberries. On the right side of the plate is a bowl of orange soup. The table is set with silverware, including a fork, knife, and spoon. The overall scene has a warm, inviting atmosphere.

COCO_test2014_000000098429

ViT-GPT2: a large building with a clock tower.

OFA: a view of big ben and the big ben clock tower.

GIT: a large clock tower in the city.

BLIP-2 : a city with a clock tower and a river.

LLaVA : The image is a black and white photo of a city skyline featuring a large clock tower, which is likely the Big Ben in London. The clock tower is prominently visible in the center of the scene, towering over the city. In addition to the clock tower, there are several other clocks scattered throughout the cityscape. The skyline is filled with various buildings, some of which have clocks on their exteriors. The image also captures a busy street with numerous cars and buses, including a red double-decker bus. The cars are parked or driving along the street, while the buses are likely providing public transportation for the city's residents. The overall atmosphere of the photo is bustling and dynamic, showcasing the city's lively urban environment.

Kosmos-2: An image of the famous Big Ben and other London landmarks from a high vantage point.

Fuyu-8B : The image features the Big Ben clock tower in London, England.

Moondream2: A panoramic view of the London skyline, with the iconic Big Ben clock tower in the center, surrounded by various historic buildings and modern structures. The sky is overcast, creating a soft, diffused light over the scene. The image is taken from a high vantage point, providing a sweeping view of the cityscape. The colors are muted, with a warm, golden hue from the Big Ben and the stone facades of the buildings. The overall atmosphere is atmospheric and atmospheric, with a sense of depth and scale.


COCO_test2014_000000508138

ViT-GPT2 : a sign that says "no swimming" on a beach".

OFA: a sign on a river with a person on a surfboard.

GIT: a sign warning people to surf on the beach.

BLIP-2 : a sign that says surfers and beware of the water.

LLaVA : The image features a beach scene with a group of people enjoying various water activities. There are at least nine people in the scene, with some of them surfing on waves and others standing or walking nearby. The surfers are using surfboards, with at least two surfboards visible in the scene. A sign is also present in the image, which is written in German. The sign is located near the water, possibly providing information or warnings for beachgoers. The overall atmosphere of the scene is lively and fun, with people engaging in water sports and enjoying the beach environment.

Kosmos-2: An image of a sign warning of the dangers of surfing in the ocean

Fuyu-8B : The image features a sign on a rocky beach that warns people about surfing and wading in shallow water. There are also two surfboards visible in the scene, with one being closer to the sign and another slightly further away.

Moondream2: A sign in German warning about the dangers of water sports, including surfing and bodysurfing, with a warning about the risk of drowning and the importance of wearing life jackets. The sign is located near a river with people surfing on the waves, and there are warning symbols for a skull and a red circle with a line through it. The background shows a river with a waterfall and trees, creating a natural setting. The overall color palette is muted, with the sign in white and red standing out against the natural background.

Figure 4. Comparative captions from vision-language models for diverse visual inputs.

## 4. Experiments and Discussion

The evaluation of generated captions is a challenging task, as it requires assessing both semantic accuracy) and linguistic quality. This analysis explores both quantitative and qualitative evaluations for selected image captioning models.

Previous studies on image captioning have adapted numerous types of evaluation metrics, from traditional, reference-based metrics like BLEU [40], METEOR [29], ROUGE [31], and CIDEr [45], to more recently developed, reference-free metrics such as Contrastive Language–Image Pre-training Score (CLIPScore) based on vision-language alignments. While these automated metrics provide quantitative insights, they are not always effective at capturing variation in the quality of generated descriptions.

We incorporate a detailed human evaluation protocol focusing on syntactic complexity, grammatical correctness, and context awareness to provide a deeper understanding of model performance that cannot always be obtained or measured automatically.

### 4.1. Quantitative Analysis

As part of our evaluation of VLMs for image captioning, we performed a quantitative analysis using four complementary metrics: CLIPScore, Perplexity, Lexical Diversity, and Caption Length. These measures offer a multi-faceted approach for assessing each model's performance in terms of semantic alignment, linguistic fluency, textual diversity, and verbosity.

Our experiments on the MS COCO test set demonstrate substantial variation in performance across architectures. As summarized in Table 3, the eight evaluated models exhibit distinct trade-offs across these dimensions, reflecting the impact of their underlying design choices on caption quality.

Table 3. Evaluation metrics for vision-language models on image captioning.

| Model | CLIPScore | Perplexity | Diversity (4-grams) | Caption length stats (Min/Max words) |
|---|---|---|---|---|
| **ViT-GPT2 [38]** | 0.7061 | 178.82 | 0.57 | 6/16 words |
| **OFA [51]** | 0.4702 | 128.99 | 0.66 | 6/16 |
| **GIT [50]** | 0.7153 | 51.86 | 0.68 | 4/17 |
| **BLIP-2 [30]** | 0.7183 | 177.09 | 0.70 | 3/15 |
| **LLaVA [37]** | 0.7568 | 15.86 | 0.63 | 37/153 |
| **Kosmos-2 [41]** | 0.7446 | 61.44 | 0.63 | 5/28 |
| **Fuyu-8B [6]** | 0.7070 | 26.00 | 0.70 | 9/84 |
| **Moondream2 [46]** | 0.4558 | 39.34 | 0.79 | 16/43 |

Based on CLIPScore, a metric that assesses semantic alignment between the generated caption and the visual content, LLaVA (0.757), Kosmos-2 (0.745), and BLIP-2 (0.718) demonstrated the strongest performance, exceeding the predefined robust performance threshold of 0.70. In contrast, Moondream2 (0.456) and OFA (0.470) scored considerably lower, indicating suboptimal visual-textual alignment.

When evaluating a language model with perplexity, which measures how confidently a language model can predict the next tokens, LLaVA again came out at the top (15.86), followed by Fuyu-8B (26.00), Moondream2 (39.34), and GIT (51.86). ViT-GPT2 (178.82) and BLIP-2 (177.09) had high perplexity, indicating low fluency or less confident word predictions during the caption generation process. This contrast places some weight on the observation that high CLIPScore does not equal fluent language generation, as seen with BLIP-2.

Lexical diversity, operationalized through 4-gram diversity, captures the model's ability to avoid textual repetitiveness. Moondream2 demonstrates the strongest performance (0.79), despite its relatively low CLIPScore. Other models with decent levels of diversity were Fuyu-8B and BLIP-2 (0.70 both), and ViT-GPT2 had a low level of diversity (0.57), meaning there is some repetition or templating in its outputs, which is not evident in the comparison against diversity. Descriptions that are detailed but do not burden their readers.

The evaluation of recent VLMs for image captioning, presented in Figure 5, emerges as a nuanced spectrum of performance, reflecting trade-offs between fluency, semantic alignment, and lexical diversity. LLaVA produces semantically aligned, fluent, and descriptively rich captions, but its tendency to overgenerate makes it less suitable for constrained caption use cases. Fuyu-8B exhibited a typical performance for VLMs referenced here, delivering high fluency and descriptiveness, yet produced captions that were more compact and perhaps more adaptable to constrained caption tasks in lieu of much of the expressiveness from LLaVA. BLIP-2 presents a middle path, balancing conciseness and diversity of captions. However, its high perplexity indicates it can be linguistically unpredictable, suggesting that its outputs were less polished or coherent with syntactical multi-variant linguistics.

In contrast, Moondream2 had a clear advantage in generating lexically varied captions with modest fluency loss, but its lower CLIPScore illustrates challenges it faces in terms of text-to-visual content alignment (probably due to limited exposure to or framework use of ground vision+language training data).

The GIT model has a reliable profile, performing below-average across all dimensions. The equality in performance makes it most adopted for use cases where no single captioning quality is deemed of greatest priority. In stark contrast, ViT-GPT2 underperforms in both fluency and lexical richness, but significant lexical alignment is not accounted for in the contribution to overall captioning potential.

Finally, OFA appears under-optimized and trailing across key metrics, including alignment, fluency, and diversity. Unless the model is assessed with the potential of massive retraining or architectural changes, it has limited chances of applicability as it currently stands.

LLaVA and Fuyu-8B were able to produce excellent fluency and grounding in a semantic sense, as well as very well-formed systemic structure, where verbosity is acceptable or even desirable. Occasionally informative but necessarily concise outputs are more applicable to GIT and BLIP-2. For diversity and lexical creativity, either as advertisements or storytelling, Moondream2 is quite strong, but it must be emphasized that careful management or additional fine-tuning will have to account for a noticeable decline in semantic accuracy in exchange for lexical diversity.

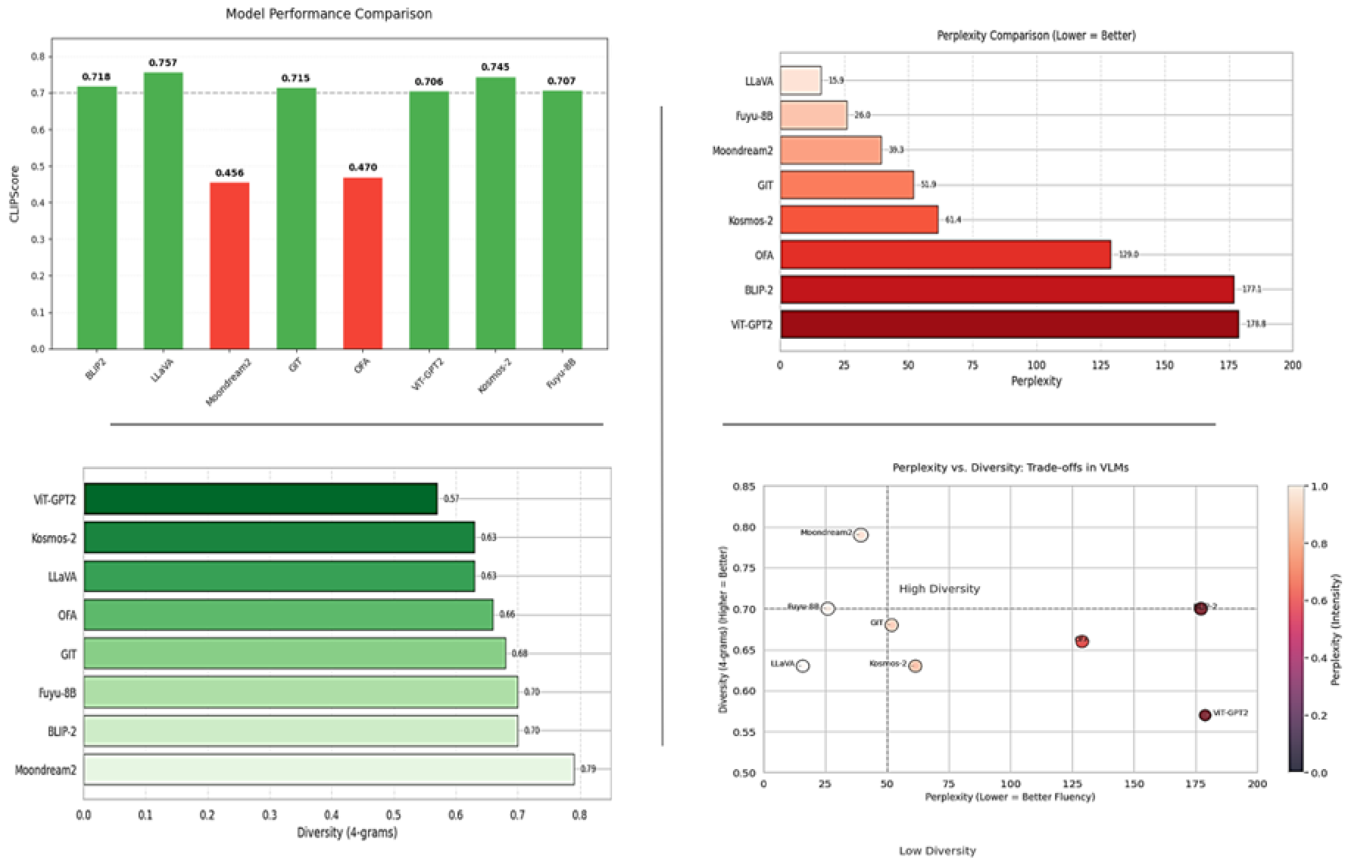


Figure 5. Comparative evaluation of vision-language models across semantic alignment, fluency, and lexical diversity dimensions.

## 4.2. Qualitative Analysis

### 4.2.1. Structured Analysis of Captioning Models Via WH-Components

To analyze the effectiveness and accuracy of each tested model, we perform a qualitative study of the generated captions. We adopted spaCy [20], which is a powerful, open-source NLP library used to analyze text structure, extract linguistic features, and derive insights from unstructured text. In the context of caption analysis, spaCy helps break down sentences into their grammatical components to answer WH questions (who, what, where, how, why) and quantify structural patterns. The analysis of generated caption's structure is presented in Table 4.

The evaluation was conducted on eight of the most advanced VLMs on six key dimensions of WH-question assessment: Subject Detail, Action Detail, Location Detail, Time Detail, Manner Detail, and Purpose/Reason. These factors reflect both important types of semantic detail and contextual understanding for producing captions, which then allow for a more thorough comparative understanding of the strengths and weaknesses of each model.

In regard to subject identification ("Who/What"), all models perform adequately at a basic level, while newer models are increasingly better at describing the characteristics of entities and connecting down to contextual recognition for multiple entities. For action recognition ("Doing What"), there is a meaningful change in previous basic verb use to recent models being able to indicate interaction and intent as these models can derive the purpose for actions. Improvement in spatial understanding ("Where") has occurred, moving from environmental location to spatial structures and useable space with the meaning associated with that location within scenes. Temporal understanding ("When") continues to be the most limited dimension, although there is some growth regarding previous a lack of recognition to an early development of visual time; recent models can now include visuals that indicate time. When describing manner ("How"), earlier models used next to no adverbs, while most recent models give a usable, integrated way to describe which conveys sometimes emotional tones. The most evident advancement is seen in the dimension of purpose or reasoning ("Why") whereby earlier models completely excluded causal understanding while the latest models are able to infer motivations and goals across visual events. Collectively, these advancements illustrate an increasing depth and coherency of model responses to WH-questions similar to the general improvement of semantic and contextual reasoning in vision-language models. The structural composition of image captions generated by VLM reveals a clear evolutionary trend in linguistic sophistication and contextual richness across six WH question dimensions: subject detail, action precision, location specificity, temporal awareness, manner description, and purpose/reasoning. All model tiers, the inclusion of core components such as Subject and Action is nearly universal. According to Figure 6, models like Fuyu-8B, LLaVA, Kosmos-2, and Moondream2, demonstrate a perfect or almost perfect detection rate in the location category and suggest that existing architectures are highly tuned for spatial scene recognition, possibly due to the fact that the visual datasets that are available for training these types of models tend to heavily focus annotation on object localization. However, a stark contrast emerges when assessing the manner, reason, and time dimensions. For instance, while fuyu-8B achieves an unusual score in both location and manner, it lacks temporal capabilities, reflecting a possible design bias towards descriptive features without deeper contextual modeling. Of all the models, LLaVA and Moondream2 represent the two most balanced semantic profiles, with relatively high location scores and moderate performance in time. These two models are unique in their ability to take on complex multi-modal inference problems, integrated descriptive and causal and temporal aspects simultaneously. LLaVA also demonstrates superior abilities overall-although it lags slightly from other models in Purpose/Reasoning category. On the other hand, ViT-GPT2 as an earlier generation model represents the weakest overall in many categories. ViT-GPT2, and GIT show serious limitations in reasoning, while ViT-GPT2, and OFA more serious limitations in reasoning, reflecting that they smaller semantic bandwidth. Models like BLIP-2 and OFA are more mixed-achieving decent Subject Detail scores, but demonstrating deficient performance in reasoning-based categories, consequently. Thus, while the models demonstrate similarities in high spatial awareness, few extend this capability to encompass richer, understanding. multi-dimensional semantic Model like LLaVA is currently best suited for tasks requiring diverse semantic interpretations, whereas others remain confined to more surface-level scene understanding. This analysis underscores the need for more holistic training approaches and benchmark datasets that go beyond object detection to include causal and temporal reasoning.

Table 4. Model-specific structural analysis.

| **Models** | **Caption structure tendencies** | | | | | |
|---|---|---|---|---|---|---|
| | **Subject identification (Who/What)** | **Action recognition (Doing what)** | **Location description (Where)** | **Temporal awareness (When)** | **Manner description (How)** | **Purpose/Reasoning (Why)** |
| **ViT-GPT2 [38]** | Basic subject identification with simple attributes | Limited verb vocabulary, mainly present progressive | Generic locations with minimal context | Almost no temporal indicators | Minimal, typically omits how actions are performed | Almost entirely absent, rarely speculates on intentions |
| **OFA [51]** | Multiple subject recognition with improved attributes | More diverse verbs with object interactions | More detailed than earlier models, contextualizes subjects | Basic temporal context recognition | Improved, sometimes includes adverbial descriptions. | Basic purpose of common activities |
| **GIT [50]** | Good attribute recognition with contextual relevance | Good verb variety with subject-object interactions | Contextual and often integrated with subjects | Limited, typically implied rather than stated | Moderate inclusion of descriptive elements | Limited purpose recognition |
| **BLIP-2 [30]** | Excellent, with detailed attribute recognition | Precise actions with contextual appropriateness | Well-integrated spatial awareness | Moderate temporal context recognition | Good inclusion of descriptive adverbs | Improved function and purpose recognition |
| **LLaVA [37]** | Contextually rich subject identification | Nuanced actions with contextual interpretation | Rich environmental context with function | Improved explicit/implicit time awareness | Rich manner with emotional understanding | Notable improvement in reasoning about intent and causation |
| **Kosmos-2 [41]** | Detailed identification with visual grounding | Precise verbs with spatial understanding | Excellent spatial relationships between objects | Good recognition of visual time cues | Strong spatial-manner integration | Improved reasoning about function and purpose |
| **Fuyu-8B [6]** | Efficient but precise subject identification | Context-appropriate action description | Effective spatial awareness and scene composition | Selective inclusion of temporal information | Contextual manner descriptions when relevant | Balanced purpose recognition |
| **Moondream2 [46]** | Focused subject identification with key attributes | Contextually appropriate actions and states | Concise but effective spatial descriptions | Efficient inclusion of key time indicators | Selective inclusion based on relevance | Efficient inclusion of key purposes |

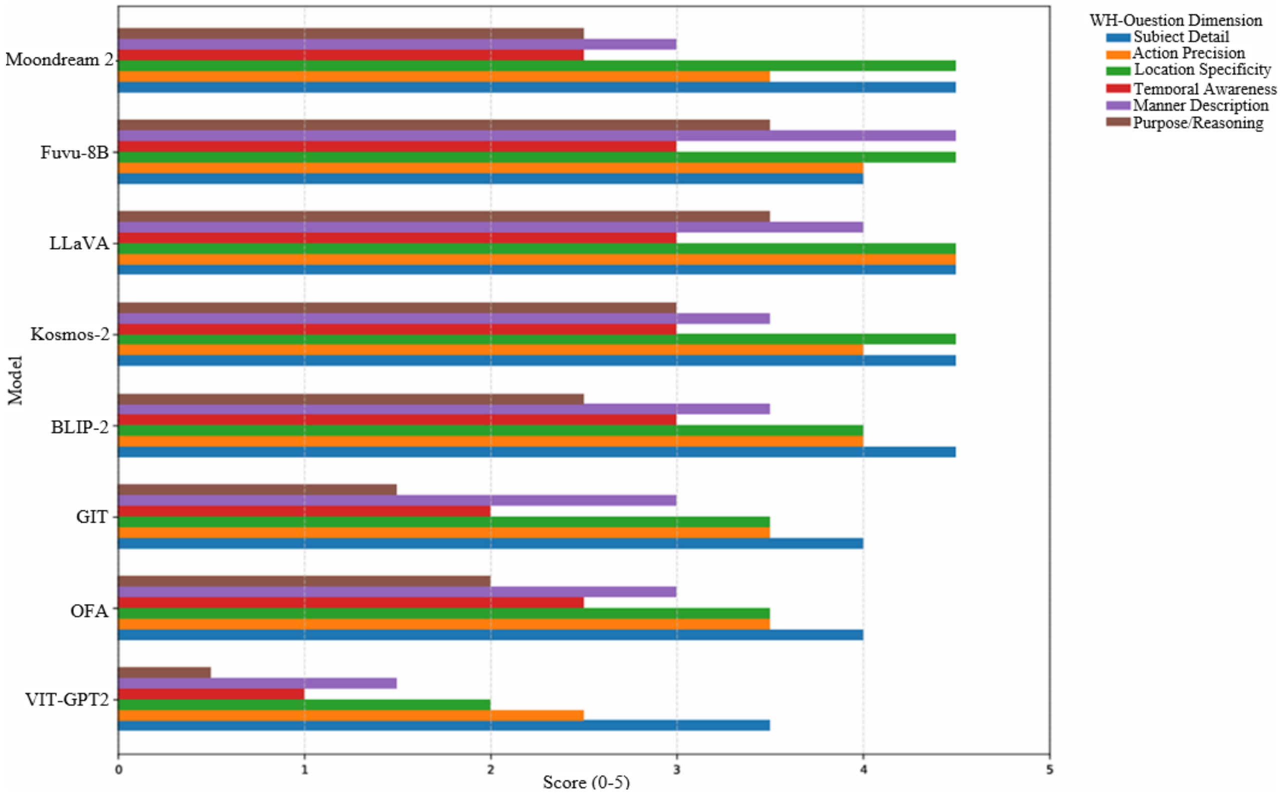


Figure 6. Multimodal model comparison: WH-Question capabilities (1,000 captions per model).

#### 4.2.2. Architectural Influence on WH-Question Coverage

The ability of VLM to adress WH-questions (who, what, where, when, why, and how) is bound to architectural developments and training paradigms. Comparative analysis indicates that newer models like LLaVA and Kosmos-2 demonstrate illustrates how newer models provide better and a more equal coverage of WH questions at a more abstract level, such as Purpose and Reasoning. This can be explained again with the size of the model and quality of the underlying language model. Larger, and more capable language backbones allow for the generation of more complex and contextually nuanced responses related to WH-questions. However, the key relationship affecting performance is the depth of the visual-linguistic interface: models like BLIP-2 and LLaVA use state-of-the-art cross-modal fusion so are able to merge spatial, contextual, and referential information better than other models built on basis of knowledge in literature. In addition, pre-trainings on multiple and semantical rich image-text datasets can help a model comprehend temporally and causally whilst architectures that employ explicit forms of visual grounding, as in Kosmos-2, enable models to accurately interpret scene-based dependence and juxtaposition of spatial locations. In terms of linguistic expressiveness, models are additionally impacted by the ability of the language decoder: the greater the language modules, usually, the larger and more varying and fluent the sentences produced by the models. Nearing captioning organization, in a lot of instances, models are fairly predictable in that they typically follow a cognitive hierarchy where they provide what the VLM perceives as the important notion of Subject→Action→Object→Context, with some more advanced models exhibiting variations in this pattern and beyond based on image salience and therefore semantics. In general, improvements in architecture and training enable a model to answer. Table 5 shows explicit patterns of increasing language and context sophistication closely correlated with model architecture and methods of training. The earliest-generation models such as VIT-GPT2, OFA, and GIT are motivated by a basic subject-action location framework and produce captions that are similar to bare factual statements with little contextual detail/attachment.

As models advanced to the intermediate tier, including BLIP-2 and LLaVA, their outputs began to incorporate a modest increase in descriptive richness, although they still adhered to a relatively formulaic syntactic structure. In contrast, the most recent and advanced models-Kosmos-2, Fuyu-8B, and Moondream2-demonstrate a substantial leap in caption complexity. These models generate multi-component narratives that not only contain elements of manner, degree, and purpose, but also generate descriptions that are similar to human interpretations in terms of elaborative scene descriptions. This evolution not only highlights a transition from basic object identification to more nuanced scene understanding but also underscores a growing capacity for contextual reasoning and semantic coherence.

Table 5. Characteristic caption patterns of vision-language models.

| Models | Characteristic caption pattern |
|---|---|
| **ViT-GPT2 [38]** | Formulaic structure: "[Subject] [simple verb] [object/location]", shorter captions with straightforward constructions<br>Limited handling of complex scenes with multiple subjects or actions |
| **OFA [51]** | More varied sentence structures than earlier models, can manage compound subjects and multiple actions<br>Pattern: "[Detailed subject] [action verb] [object] [prepositional phrase for location/manner]"<br>Better at capturing interactions between multiple entities |
| **GIT [50]** | More naturalistic language than earlier models, often begins with subject-focused descriptions before actions<br>Typical structure integrates location with subject or action<br>Example: "A [detailed subject description] [verb-ing] [object] in [detailed location]" |
| **BLIP-2 [30]** | Flexible structures with improved contextual awareness, rich descriptions with better relationships<br>"A [detailed subject with multiple attributes] is [specific action verb-ing] [object] in a [detailed environment] with [specific features]." |
| **LLaVA [37]** | More conversational and natural language, oomplex sentences with causal or temporal relationships<br>Better at abstract concepts and implied information, often includes evaluative or interpretive elements beyond description<br>"A [specific] [subject] that appears to be [contextual description] is [nuanced action] [object] [manner] [apparent purpose] in what appears to be a [specific environment type] with [contextual details]." |
| **Kosmos-2 [41]** | Strong spatial relationships and positioning, often includes relative positioning of elements<br>More sophisticated object attribute descriptions<br>"A [specific type] of [object] with [distinctive features] [precise action] [object] [precise spatial relation] to [another object] in a [specific environment]." |
| **Fuyu-8B [6]** | Efficient but informative descriptions, efficient balance between detail and conciseness<br>Less template-like, more adaptable sentence structures<br>Example structure: "The image shows a [subject with key attributes] [specific verb] [object] [essential qualifier] in a [relevant environment descriptor] [key spatial relationship]." |
| **Moondream2 [46]** | Concise, information-dense descriptions, more straightforward structures optimized for efficiency, prioritizes key elements over exhaustive description, often follows template: "A [distinctive attribute] [subject] [position/state] [focused action verb] [qualifier when relevant] in/on [concise location description]." |

Overall, these findings show that advancements in model design have increased the capability of vision language systems to produce captions that demonstrate a more comprehensive understanding of visual elements. ViT-GPT2 is frequently used to produce brief and direct descriptions as single sentences. These descriptions are framed in terms of observable entities without inference or context. The structure is also fairly standard and follows a "subject-action-location" format. In general, the descriptions show accuracy but often miss out on more subtle details and relationships among the visual elements. GIT increases the breadth of the description by incorporating a more attributes and relationships in it. The captions provided by GIT often begin with the main subject and then build out describing anything else in proximity. Overall, it is a more descriptive description than the ViT-GPT2 overall, although the overall description retains a mechanical structure, where the description first deals with the primary object in view and then the secondary objects. OFA represents a step forward in natural language generation. Its captions typically open with a scene overview before diving into specifics. The model creates more cohesive narratives by linking observations with transitional phrases, though it can sometimes be overly verbose in its attempt to be comprehensive. BLIP-2 provides impressive caption generation with coherence between sentence structure. It also has a narrative arc within descriptions, beginning with primary elements within scenes to contextual information. The strength of the model lies in its description of actions and relationships between elements in the scene. LLaVA considers captioning as a more conversational approach to description. The descriptions tend to capture direct observations and inferred context in both observation and narrative. LLaVA creates more of a narrative crossing stage directions and narrative point of view. When generating captions, LLaVA connect visual elements to implied purpose or context, which can sometimes lead to over-interpretation. Kosmos-2 considers heading spatial relations when structuring captions. Most descriptions begin with a description of the scene, then detail the spatial arrangements and relations between the elements. The model excels at creating a coherent mental image by paying attention to relative positioning. The Fuyu-8B model generates very structured captions, balancing directness with readability. Fuyu-8B systematically describes primary, secondary, and contextual elements, while keeping the natural flow of language. The model performs well in organizing multiple observations into coherent narratives. Despite its lightweight design, Moondream2 is able to produce effective and focused captions. Its descriptions prioritize key elements while maintaining coherence. Moondream2 tends towards being concise, but complete, in its captions, but as shown in an example above. It might sacrifice some nuance as a result of being more informationally efficient. To sum-up, this comparative analysis of image captioning models outlines some important trends in the trajectory that caption generation has evolved into. First, there is a developing sophistication from simple subject verb-object to richer scene semantics that use complicated, multi-layered, and layered categorization. More recent models clearly incorporate context clues and spatial relationship semantics, allowing them to produce captions that resemble a more proper human description. Additionally, some advanced architecture capabilities produce multi-sentence outputs and effectively describe primary elements and secondary details. There is a considerable range in how the models treat attributes and modifiers - either with succinct descriptions or entirely detailed with elaborate captions. It seems as though this trend relates to a generalized increase in line length and depth of information with complexity in a model. Finally, instruction-tuned are more flexible and presented in different ways in their output, finding that fine-tuning with some language guidance enables both lexicon flexibility and offers a broad range of expression shown in the generated output.

Table 6. Comparative analysis of vision-language models: performance across dynamic and behavioural understanding, object-level perception, and specialized recognition.

| Evaluation dimensions | | ViT-GPT2[38] | OFA [51] | GIT [50] | BLIP-2[30] | LLaVA [37] | Kosmos-2[41] | Fuyu-8B [6] | Monndream2[46] |
|---|---|---|---|---|---|---|---|---|---|
| **Scene and semantic understanding** | **Scene understanding** | Limited by small-scale pre-training but performs adequately on simple image-caption tasks. | Good at scene understanding due to large-scale pre-training, | moderate, but less tested on complex relational scenes | achieving robust performance on scene comprehension | indicating excellent scene comprehension with holistic contextual | Robust performance with detailed scene description | Strong in real-world scenarios, suggesting good scene understanding, but less tested on complex scenes. | Strong scene understanding with effective detail capture |
| | **Visual reasoning** | Limited reasoning capabilities beyond simple descriptions | Good reasoning from unified task training, as evidenced by zero-shot performance. | Moderate reasoning with occasional inconsistencies | Indicating strong reasoning. | Exceptional reasoning about visual relationships and implications | Advanced reasoning via grounding and causal language modeling | Solid reasoning capabilities across various complexity levels | Strong reasoning for a lightweight model |
| | **Spatial relation** | Limited spatial reasoning capabilities | Good spatial reasoning due to unified architecture | moderate; not spatial-focused | Strong spatial reasoning with good directional awareness | Superior spatial relationship understanding | Excellent spatial reasoning and relationship description | Good spatial relationship description capabilities | Good spatial relationship understanding |
| **Dynamic and behavioral understanding** | **Action recognition** | Limited action recognition capabilities | Adequate action recognition with some limitations | Basic understanding of actions portrayed in images. | Good perception of activities and actions in images | Strong action recognition and contextual interpretation | Advanced action recognition with contextual understanding | Decent action recognition capabilities | Good action recognition capabilities in common scenarios |
| | **Object interaction** | Limited interaction recognition capabilities | Good performance in identifying object relationships | Moderate; captioning focus limits interaction. | Strong interaction understanding with contextual interpretation | Excellent recognition of complex object interactions | Superior interaction analysis with detailed descriptions | Good interaction recognition capabilities | Reasonable interaction understanding. |
| **Object-level perception** | **Object identity** | Basic object identification with occasional errors | Good performance in identifying object relationships | Adequate interaction description with some limitations | Strong object recognition across diverse categories | Excellent object identification with fine-grained distinctions | Superior object identification with contextual understanding | Strong object identification capabilities | Strong object identification capabilities across diverse categories |
| | **Object attribute** | Basic attribute recognition for prominent features | Good attribute identification performance | Adequate attribute description with moderate detail | Strong attribute identification across object types | Excellent attribute recognition and description | Superior attribute detection with detailed descriptions | Strong attribute recognition capabilities | Strong attribute recognition for common object properties |
| | **Object counting** | Limited counting abilities, especially in complex scenes | Basic counting capabilities with some inconsistencies | Moderate; captioning may include counts but not explicit. | Decent counting but struggles with crowded scenes | Good counting accuracy across various scenarios | Strong counting abilities with spatial awareness | Reliable object counting capabilities | Reasonable counting accuracy for common scenarios |
| | **Object localization** | Limited Localization capabilities. | Good localization abilities due to unified architecture | Basic localization with occasional imprecision. | Good object localization with contextual understanding | Strong localization capabilities with good spatial language | Excellent spatial localization with precise descriptions | Capable object localization across image regions | Adequate object localization in standard compositions |
| **Specialized recognition** | **Text recognition** | Limited text recognition abilities | Good text recognition with contextual integration | Basic text recognition capabilities. | Adequate text recognition but struggles with complex layouts | Good text recognition with context integration | Superior text recognition and integration into understanding | Strong text recognition capabilities across different formats | Adequate text recognition for standard text formats |
| | **Landmark recognition** | Limited landmark recognition capabilities | Adequate landmark recognition for common locations. | Basic landmark recognition for well-known sites | Good Landmark identification capabilities | Good landmark recognition with contextual knowledge | Advanced landmark identification abilities | Decent landmark recognition with some limitations | Reasonable landmark recognition for common locations |
| | **Food recognition** | Basic food recognition for common items | Good food recognition across diverse cuisines | Adequate food identification with basic descriptions | Good food recognition with contextual understanding | Strong food identification with detailed descriptions | Advanced food identification capabilities | Strong food recognition capabilities | Strong food recognition with the ability to identify various dishes |

### 4.2.3. Structured Analysis of Captioning Models via Human Evaluation

Image captioning models are evaluated across three key aspects: dynamic and behavioral understanding, object-level perception, and specialized recognition. These aspects are assessed through twelve dimensions, as outlined in the table, to comprehensively gauge the models' visual understanding capabilities, particularly in terms of scene and semantic comprehension.

Overall performance of recent VLMs, as shown in Table 6, we get a clear stratification that reflects quite different strengths and design trade-offs at all model tiers. Best performing methods (i.e., LLaVA, Kosmos-2, BLIP-2) show enormous potential in multimodal understanding due to their sophisticated architectures and training techniques. LLaVA outperforms by a significant margin in visual and interaction reasoning, object identity, and interaction understanding, which is mainly due to its instruction-tuning method and strong language model backbone. Kosmos-2 is closely followed by Kosmos, which performs well in text recognition and spatial localization by considering its text-grounding framework and large-scale multimodal training that scopes a server room to integrate textual with visual semantics.

BLIP-2 shares the same consistency between evaluation metrics, and the consistency between scene understanding and landmark recognition, which benefits from the two-stage pipeline and the Q-Former module that well connects the vision encoder and the language model.

We classify a group of models, like Fuyu-8B, OFA, Moondream2, and GIT, to the middle tier as the models do not have ultra-specialized but relatively competent ability. By exploiting its efficient architectural design and its end-to-end pipeline training strategy, Fuyu-8B achieves remarkable performance and tends to show fine promise in text recognition as well as object counting.

OFA takes advantage of its end-to-end sequence-to-sequence framework and achieves superior performance in food recognition and attribute analysis. An apparent art-humanist implementation, Moondream2, it orally competes with general methods in tasks of image captioning and scene understanding at a certain scale of I model sizes. Although coarser, GIT retains rudimentary object competences and seems to be an earlier design.

On the other hand, lower-tier ViT-GPT2 models reveal the shortcomings of the earlier VLM designs. Although it does have some basic object detection capabilities, it struggles with more advanced reasoning, understanding, detecting spatial relationships, and recognizing text. This fine-grained performance summary not only describes how the architecture evolves and how the training varies among VLMs but also demonstrates the increasing significance of customizing pre-training and integration methods to yield comprehensive visual-linguistic representation.

The analysis indicates a wide disparity in performance across VLMs, with newer architectures typically outpacing older architectures. Among the VLMs, LLaVA, Kosmos-2, and Fuyu-8B consistently outperform the other models across several characteristics, as they represent advancements in architectural and training methodologies, catalysing the linking of vision and language modalities.

Performance gaps are most pronounced in complex reasoning tasks (e.g., visual reasoning, spatial relations, object interactions) and in granularity of understanding (e.g., text recognition, object counting, locating objects). Each of these aspects likely relies on complex interactions between visual perception and language understanding, illustrating the opportunities offered by thoughtful architectural development and an expansive training corpus.

Fuyu-8B's strong performance despite its simpler architecture suggests promising directions for architectural streamlining without significant capability loss. Similarly, Moondream2 represents an impressive achievement in efficient model design, demonstrating that lighter models can achieve surprisingly robust performance in image understanding tasks, particularly in image captioning and object identification.

Moondream2's effectiveness challenges assumptions about size-performance trade-offs in VLMs, suggesting that careful dataset curation and training optimization can produce highly efficient models with capabilities approaching those of much larger counterparts in many practical scenarios.

Future VLM development would benefit from exploring Moondream2's efficient design principles while continuing to improve fine-grained spatial understanding, complex visual reasoning, and better integration of text recognition with general visual perception capabilities across all model scales.

The comparison shows how each model has unique approaches to structuring and prioritizing information in captions. However, several challenges remain unresolved, particularly the issue of generic bias, which limits the informativeness of captions. Existing models often generate overly broad descriptions, failing to capture the unique aspects of images. This issue is exacerbated by the inherent nature of training data and generation mechanisms.

Our findings reveal a persistent challenge in vision-language models: the tendency to produce generic captions that lack domain-specific nuance. Motivated by this limitation, we are working on a centered context-aware regularization, a novel training paradigm that dynamically penalizes overly generic outputs while preserving fluency. This approach integrates two key innovations:

1. Domain-adaptive priors that steer descriptions toward application-critical details.
2. Contrastive concept learning to sharpen distinctions between visually similar but semantically distinct elements.

## 5. Conclusions

This study makes three key contributions to advancing image captioning evaluation. First, we establish a comprehensive human assessment framework that evaluates VLMs across multiple critical dimensions, including fluency, specificity, grounding accuracy, and stylistic appropriateness, revealing important limitations that standard automated metrics fail to capture. Our large-scale comparative analysis of state-of-the-art architectures uncovers fundamental trade-offs, such as the tension between descriptive richness and factual reliability, while identifying persistent challenges in spatial reasoning and domain adaptation. Most significantly, we develop and validate a standardized evaluation protocol featuring a detailed error taxonomy, domain-specific criteria, and robust quality control mechanisms. These contributions collectively demonstrate that while modern models excel at conventional benchmarks, they still struggle with contextual specificity and often produce generic descriptions. Our findings not only provide practical guidance for model selection in real-world applications but also establish essential groundwork for future research directions, including multilingual evaluation, bias mitigation, and the development of hybrid human-AI assessment systems. By integrating rigorous human evaluation with technical analysis, this work moves the field toward more dependable, adaptable, and contextually aware caption generation.

## References


[1] Alayrac J., Donahue J., Luc P., Miech A., and et al., "Flamingo: A Visual Language Model for Few-Shot Learning," *in Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 23716-23736, New Orleans, 2022. https://dl.acm.org/doi/10.5555/3600270.3601993

[2] Anderson P., He X., Buehler C., Teney D., and et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077-6086, 2018. https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00636

[3] Awadalla A., Gao I., Gardner J., Hessel J., and et al., "OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models," *arXiv Preprint*, pp. 1-20, 2023. https://ui.adsabs.harvard.edu/link_gateway/2023arXiv230801390A/doi:10.48550/arXiv.2308.01390

[4] Bahdanau D., Cho K., and Bengio Y., "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv Preprint*, pp. 1-15, 2014. https://arxiv.org/abs/1409.0473v7

[5] Bai J., Bai S., Yang S., Wang S., and et al., "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond," *arXiv Preprint*, vol. arXiv:2308.12966v3 pp. 1-24, 2023. https://arxiv.org/abs/2308.12966v3

[6] Bavishi R., Elsen E., Hawthorne C., Nye M., and et al., Fuyu-8B: A Multimodal Architecture for AI Agents, www.adept.ai/blog/fuyu-8b, Last Visited, 2025.

[7] Bo Z., Boya W., Muyang H., and Tiejun H., "SVIT: Scaling up Visual Instruction Tuning," *arXiv Preprint*, vol. arXiv:2307.04087v3, pp. 1-18, 2023. https://arxiv.org/abs/2307.04087v3

[8] Brown T., Mann B., Ryder N., Subbiah M., and et al., "Language Models are Few-Shot Learners," *in Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, pp. 1877-1901, 2020. https://dl.acm.org/doi/abs/10.5555/3495724.3495883

[9] Caffagni D., Cocchi F., Barsellotti L., Moratelli N., and et al., "The Revolution of Multimodal Large Language Models: A Survey," *in Proceedings of the Association for Computational Linguistics*, Bangkok, pp. 13590-13618, 2024. https://doi.org/10.18653/v1/2024.findings-acl.807

[10] Cha J., Kang W., Mun J., and Roh B., "Honeybee: Locality-Enhanced Projector for Multimodal LLM," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, pp. 13817-13827, 2024. https://openaccess.thecvf.com/content/CVPR2024/papers/Cha_Honeybee_Locality-enhanced_Projector_for_Multimodal_LLM_CVPR_2024_paper.pdf

[11] Chen J., Zhu D., Shen X., Li X., and et al., "MiniGPT-v2: Large Language Model as a Unified Interface for Vision-Language Multi-Task Learning," *arXiv Preprint*, vol. 2310.09478v3, pp. 1-20, 2023. https://arxiv.org/abs/2310.09478v3

[12] Cheng K., Song W., Ma Z., Zhu W., and et al., "Beyond Generic: Enhancing Image Captioning with Real-World Knowledge Using Vision-Language Pre-Training Model," *in Proceedings of the 31st ACM International Conference on Multimedia*, Ottawa, pp. 5038-5047, 2023. https://doi.org/10.1145/3581783.3611987

[13] Chowdhery A., Narang S., Devlin J., Bosma M., and et al., "PaLM: Scaling Language Modeling with Pathways," *Journal Machine Learning Research*, pp. 1-87, 2022. https://arxiv.org/abs/2204.02311v5

[14] Chung H., Hou L., Longpre S., Zoph B., Tay Y., and et al., "Scaling Instruction-Finetuned Language Models," *Journal Machine Learning Research*, vol. 25, no. 70, pp. 1-53, 2024. https://arxiv.org/abs/2210.11416v5

[15] Cornia M., Baraldi L., and Cucchiara R., "Explaining Transformer-Based Image Captioning Models: An Empirical Analysis," *AI Communications*, vol. 35, no. 2, pp. 111-129, 2022. https://doi.org/10.3233/AIC-210172

[16] Cornia M., Baraldi L., and Cucchiara R., "SMArT: Training Shallow Memory-Aware Transformers

for Robotic Explainability," *in Proceedings of* the *IEEE International Conference on Robotics and Automation*, Paris, pp. 1-25, 2019. https://doi.org/10.1109/ICRA40945.2020.9196653

[17] Devlin J., Chang M., Lee K., and Toutanova K., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *arXiv Preprint*, 2018. https://arxiv.org/abs/1810.04805v2

[18] Donahue J., Hendricks L., Rohrbach M., Venugopalan S., and et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 677-691, 2017. https://doi.org/10.1109/TPAMI.2016.2599174

[19] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., and et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *in Proceedings of* the *International Conference on Learning Representations*, Vinna, pp. 1-22, 2020. https://arxiv.org/abs/2010.11929v2

[20] Fantechi A., Gnesi S., Livi S., and Semini L., "A spaCy-Based Tool for Extracting Variability from NL Requirements," *in Proceedings of the 25th ACM International Systems and Software Product Line Conference*, Leicester, pp. 32-35, 2021. https://doi.org/10.1145/3461002.3473074

[21] Gao P., Han J., Zhang R., Lin Z., and et al., "LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model," *arXiv Preprint*, vol. 2304.15010v1, pp. 1-15, 2023. https://arxiv.org/abs/2304.15010v1

[22] Hani A., Tagougui N., and Kherallah M., "Image Caption Generation Using a Deep Architecture," *in Proceedings of the International Arab Conference on Information Technology*, Al Ain, pp. 246-251, 2019. https://doi.org/10.1109/ACIT47987.2019.8990998

[23] Huang L., Wang W., Chen J., and Wei X., "Attention on Attention for Image Captioning," *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, pp. 4633-4642, 2019. https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00473

[24] Jaegle A., Gimeno F., Brock A., Zisserman A, and et al., "Perceiver: General Perception with Iterative Attention," *arXiv Preprint*, vol. arXiv:2103.03206v2, pp. 1-43, 2021. https://arxiv.org/abs/2103.03206v2

[25] Karpathy A. and Fei-Fei L., "Deep Visual-Semantic Alignments for Generating Image Descriptions," *IEEE Conference on Computer Vision and Pattern Recognition,* vol. 39, pp. 3128-3137, 2015. https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298932

[26] Krizhevsky A., Sutskever I., and Hinton G., "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, pp. 84-90, 2012. https://doi.org/10.1145/3065386

[27] Kulkarni G., Premraj V., Dhar S., Li S., and et al., "Baby Talk: Understanding and Generating Simple Image Descriptions," *in Proceedings of the Computer Vision and Pattern Recognition*, Colorado Springs, pp. 1601-1608, 2011. https://doi.org/10.1109/CVPR.2011.5995466

[28] Kuznetsova P., Ordonez V., Berg T., and Choi Y., "TreeTalk: Composition and Compression of Trees for Image Descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 351-362, 2014. https://doi.org/10.1162/tacl_a_00188

[29] Lavie A. and Agarwal A., "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments," *in Proceedings of the 2nd Workshop on Statistical Machine Translation*, Prague, pp. 228-231, 2007. https://dl.acm.org/doi/proceedings/10.5555/1626355

[30] Li J., Li D., Savarese S., and Hoi S., "BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models," *International Conference on Machine Learning*, Honolulu, pp. 19730-19742, 2023. https://dl.acm.org/doi/10.5555/3618408.3619222

[31] Lin C., "ROUGE: A Package for Automatic Evaluation of Summaries," *Association for Computational Linguistics*, Barcelona, pp. 74-81, 2004. https://aclanthology.org/W04-1013/

[32] Lin T., Maire M., Belongie S., Bourdev L., and et al., "Microsoft COCO: Common Objects in Context," *in Proceedings of the Computer Vision-ECCV 13th European Conference*, Zurich, pp. 740-755 2014. https://doi.org/10.1007/978-3-319-10602-1_48

[33] Lin Z., Liu C., Zhang R., Gao P., and et al., "SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-Modal Large Language Models," *arXiv Preprint*, vol. 2311.07575v1, pp. 1-24, 2023. https://arxiv.org/abs/2311.07575v1

[34] Liu F., Lin K., Li L., Wang J., and et al, "Mitigating Hallucination in Large Multi-Modal Models Via Robust Instruction Tuning," *arXiv Preprint*, pp. 1-45, 2023. https://arxiv.org/abs/2306.14565v4

[35] Liu H, Li C., Li Y., Li B., and et al., LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge, https://llava-vl.github.io/blog/2024-01-30-llava-next/, Last Visited, 2025.

[36] Liu H., Li C., Li Y., and Lee Y., "Improved Baselines with Visual Instruction Tuning," *in*

*Proceedings of the EEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, pp. 26286-26296, 2023. https://doi.org/10.1109/CVPR52733.2024.02484

[37] Liu H., Li C., Wu Q., and Lee Y., "Visual Instruction Tuning," *in Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, pp. 34892-34916, 2023. https://dl.acm.org/doi/abs/10.5555/3666122.3667638

[38] Nlpconnect/it-gpt2-image-captioning, https://huggingface.co/nlpconnect/vit-gpt2-image-captioning, Last Visited, 2025.

[39] Osman A., Shalaby M., Soliman M., and Elsayed K., "Ar-CM-ViMETA: Arabic Image Captioning Based on Concept Model and Vision-based Multi-Encoder Transformer Architecture," *The International Arab Journal of Information Technology*, vol. 21, no. 3, pp. 458-465, 2024. DOI: 10.34028/iajit/21/3/9

[40] Papineni K., Roukos S., Ward T., and Zhu W., "BLEU: A Method for Automatic Evaluation of Machine Translation," *in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, pp. 311-318, 2001. https://doi.org/10.3115/1073083.1073135

[41] Peng Z., Wang W., Dong L., Hao Y., and et al., "Kosmos-2: Grounding multimodal Large Language Models to the World," *arXiv Preprint*, vol. arXiv:2306.14824v3, pp. 1-20, 2023. https://arxiv.org/abs/2306.14824v3

[42] Radford A., Kim J., Hallacy C., Ramesh A., and et al., "Learning Transferable Visual Models from Natural Language Supervision," *arXiv Preprint*, vol. arXiv:2103.00020v1, pp. 1-48, 2021. https://arxiv.org/abs/2103.00020v1

[43] Raffel C., Shazeer N., Roberts A., Lee K., and et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal Machine Learning Research*, vol. 21, pp. 1-67, 2019. https://arxiv.org/abs/1910.10683v4

[44] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., and et al., "Attention is all you Need," *arXiv Preprint*, vol. arXiv:1706.03762v7, pp. 1-15, 2017. https://arxiv.org/abs/1706.03762v7

[45] Vedantam R., Zitnick L., and Parikh D., "CIDEr: Consensus-Based Image Description Evaluation," *in Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, pp. 4566-4575, 2014. https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7299087

[46] Vikhyatk/Moondream2, https://huggingface.co/vikhyatk/moondream2, Last Visited, 2025.

[47] Vinyals O., Toshev A., Bengio S., and Erhan D., "Show and Tell: A Neural Image Caption Generator," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, pp. 3156-3164, 2015. https://doi.org/10.1109/CVPR.2015.7298935

[48] Wang A., Zhou P., Shou M., and Yan S., "Position-Guided Text Prompt for Vision-Language Pre-Training," *arXiv Preprint*, vol. arXiv:2212.09737v2, pp. 23242-23251, 2022. https://arxiv.org/abs/2212.09737v2

[49] Wang B., Wu F., Han X., Peng J., Zhong H., and et al., "VIGC: Visual Instruction Generation and Correction," *AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, pp. 5309-5317, 2024. https://doi.org/10.1609/aaai.v38i6.28338

[50] Wang J., Yang Z., Hu X., Li L., and et al., "GIT: A Generative Image-to-Text Transformer for Vision and Language," *arXiv Preprint*, vol. arXiv:2205.14100v5, pp. 1-49, 2022. https://arxiv.org/abs/2205.14100v5

[51] Wang P., Yang A., Men R., Lin J., and et al., "OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework," *arXiv Preprint*, vol. arXiv:2202.03052v2162, pp. 1-49, 2022. https://arxiv.org/abs/2202.03052v2

[52] Wang W., Lv O., Yu W., Hong W., and et al., "CogVLM: Visual Expert for Pre-Trained Language Models," *arXiv Preprint*, vol. arXiv:2311.03079v2 pp. 1-17, 2023. https://arxiv.org/abs/2311.03079v2

[53] Xu K., Ba J., Kiros R., Cho K., and et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *in Proceedings of the 32nd International Conference on Machine Learning*, Lille, pp. 2048-2057, 2015. https://dl.acm.org/doi/10.5555/3045118.3045336

[54] Yao L., Chen W., and Jin Q., "CapEnrich: Enriching Caption Semantics for Web Images Via Cross-Modal Pre-Trained Knowledge," *in Proceedings of the ACM Web Conference*, Austin, pp. 2392-2401, 2022. https://doi.org/10.1145/3543507.3583232

[55] Ye Q., Xu H., Ye J., Yan M., and et al., "mPLUG-OwI2: Revolutionizing Multi-Modal Large Language Model with Modality Collaboration," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, pp. 13040-13051, 2023. https://doi.org/10.1109/CVPR52733.2024.01239

[56] Yenduri G., Ramalingam M., Selvi G., Supriya Y., and et al., "Generative Pre-Trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions," *IEEE Access*, vol. 12, pp. 54608-54649 2023. https://doi.org/10.1109/ACCESS.2024.3389497

[57] Yin S., Fu C., Zhao S., Xu T., and et al., "Woodpecker: Hallucination Correction for

Multimodal Large Language Models," *Science China Information Sciences*, vol. 67, no. 12, 2024. https://doi.org/10.1007/s11432-024-4251-x

**Ansar Hani** received her Master's degree in Information Systems and New Technologies in 2016 from the Faculty of Economics and Management of Sfax, Tunisia. She is currently a Ph.D. candidate in Computer Engineering Systems at the National Engineering School of Sfax. Her research interests include Computer Vision, Multimodal Deep Learning, and Context-Aware Image Captioning. She focuses her research on developing models that bridge visual understanding and natural language for accurate and context-aware image descriptions.

**Najiba Tagougui** is actually Assistant Professor at the Higher Institute of Computer Sciences and Multimedia of Sfax, Sfax University Tunisia from where she graduated in Computer Sciences in 2005. She obtained a Master's degree in News Technologies of Dedicated Computer Systems in 2007 and a Ph.D. in Computer Systems Engineering in 2014 at the National Engineering School of Sfax. Her research interest includes Applications of Intelligent Methods to Pattern Recognition. She focuses her research on intelligent pattern recognition, especially online handwriting recognition and image caption generation.

**Monji Kherallah** received the Ing. Diploma degree, the Ph.D. and HU in Electrical Engineering, respectively in 1989, 2008 and 2012, from University of Sfax (ENIS). For fourteen years ago, he was an engineer in Biotechnologie Center of Sfax. Now he is an associate professor in Faculty of Science of Sfax and member in Research Group of Intelligent Machines: REGIM-Lab. His research interest includes the Handwritten Documents Analysis and Recognition. The techniques used are based on intelligent methods, such as neural network, logic fuzzy, and genetic algorithm. He is one of the developers of the ADAB-Database (used by more than 50 research groups from more than 10 countries). He is co-organizer of the online Arabic handwriting recognition competitions at ICDAR 2009 and ICDAR 2011. He has more than 70 papers, including journal papers and book chapters. He is a member of IEEE and IEEE AESS Tunisia Chapter Chair, 2010 and 2011. He is reviewer of several international journals.