BENCHMARKING MULTIMODAL TRANSFORMERS FOR RETRIEVAL, CAPTIONING AND ZERO-SHOT CLASSIFICATION

Pushp Raj **Ronak Gupta** **Aryan Kayande**

IIT Bombay

ABSTRACT

The integration of visual and textual data has transformed vision-language modeling, enabling robust performance in cross-modal retrieval, image captioning, and zero-shot recognition. Yet, the diversity of architectural choices and training paradigms spanning contrastive objectives, cross-attention mechanisms, and unified transformers complicates direct comparison. In this work, we offer an in-depth benchmark and theoretical analysis of CLIP, BLIP, BLIP-2, and FLAVA, evaluating their core mechanisms, Vision Transformer (ViT) underpinnings, and practical trade-offs across three pillar tasks and relevant efficiency metrics. Our experiments, conducted on MS COCO, Flickr30k, and ImageNet, provide a comprehensive leaderboard, actionable insights, and a foundation for further advances in multimodal intelligence.

1. INTRODUCTION

Across the landscape of artificial intelligence research and applications, the ability to process and understand multimodal data that is, data comprising both images and text has rapidly become indispensable. Modern user interfaces, search engines, assistive technologies, and even creative tools are increasingly built atop models that can comprehend both what is seen and what is said. The rise of transformer-based architectures, especially since the introduction of the Vision Transformer (ViT), has been pivotal. ViT recasts classic vision tasks in the architecture of sequence modeling, previously the exclusive domain of natural language processing. This enables a more flexible and unified approach to fusing, relating, and reasoning over visual and linguistic cues.

Despite this promise, the field has splintered into distinct model families, three of which stand at the current forefront: CLIP, which relies on massive contrastive alignment between image and text modalities; BLIP and BLIP-2, which incorporate generative and fusion objectives and sophisticated bridging to large-scale language or vision models; and FLAVA, which attempts a fully unified transformer that jointly models and aligns visual and textual tokens. Each model class carries unique strengths, challenges, and domains of excellence. Yet, the lack of comprehensive, consistent benchmarking especially one that jointly considers retrieval, captioning, zero-shot classification, and resource use remains a barrier to principled model selection and deeper scientific understanding.

In this work, we seek to bridge this gap by offering a detailed theoretical, architectural, and empirical comparison of these paradigmatic models. By standardizing training, evaluation, and efficiency measurement across leading datasets, we provide a resource both for researchers seeking to push the field forward and for practitioners aiming to build robust, efficient multimodal systems.

2. Background and Related Work

The literature on vision-language modeling has flourished alongside rapid advances in transformer architectures. Early fusion methods, which simply concatenated visual and textual features, have largely given way to architectures inspired by the transformer's self-attention paradigm. The Vision Transformer represents an essential shift: by tokenizing image patches and employing self-attention layers, it allows for global and flexible information mixing, akin to how language transformers operate over sentences and paragraphs. This fundamental change forms the basis for more intricate cross-modal operations and adaptive model architectures.

Many works have preceded the current state-of-the-art. For example, contrastive pretraining using massive web-scale paired data, as in CLIP, yielded astonishing zero-shot effectiveness due to the immense breadth and semantic richness it encodes. Other approaches, such as those found in BLIP, move beyond simple alignment to introduce multitask learning and cross-modal transformers, allowing for nuanced generative capabilities without sacrificing retrieval or understanding. More recent variants, exemplified by BLIP-2, emphasize modularity and efficient transfer, integrating frozen, high-capacity vision and language encoders bridged by trainable adapters. FLAVA, diverging from the preceding models, targets a unified, multi-stream transformer backbone, learning all possible tasks, unimodal or cross-modal, within one model and loss framework.

Benchmarks across these lines of research vary significantly in datasets used, pre/post-processing, and hyperparameters making side-by-side assessment challenging. Our report aims to address this by both contextualizing the most influential prior models and offering a coherent, reproducible evaluation for vision-language tasks, focusing on architecture, theory, and practicality.

3. Vision Transformer (ViT): Foundation of Multimodal Processing

The Vision Transformer architecture is central to modern multimodal systems. Instead of operating on pixels directly or

locally, as convolutional networks do, ViT divides an image into a grid of fixed-sized patches. Each patch is flattened into a vector and mapped via a linear projection into a low-dimensional embedding space. These "image tokens" are then supplemented with positional encodings that preserve the structure of the input's two-dimensional geometry, since self-attention is otherwise permutation-invariant.

A typical ViT encoder consists of a stack of multi-head self-attention and feedforward layers. Each self-attention block computes, for every token, a vector representation by weighted aggregation of other tokens' information, where the weights reflect the learned, context-dependent importance among all pairs of tokens. This all-to-all interaction not only aids global context mixing but is also ideal for subsequent fusion with text information, as both modalities become sequences of tokens processed by conceptually identical modules.

After traversing these transformer layers, a class token or spatial pooling operation extracts a global representation of the image, which is used in downstream tasks. The flexibility and high model capacity of ViT allow it to excel when pretrained on very large datasets, such as the proprietary JFT-300M or LAION, which underpins most leading multimodal approaches. However, ViT requires both significant computational resources and careful initialization in order to avoid overfitting and underperformance relative to convolutional approaches on smaller datasets.

ViT's sequential, token-based structure means it is easy to integrate, align, or fuse with language transformers, which process tokens in a similar fashion. This seamless modality bridging is a crucial foundation for the models benchmarked in this report.

4. Model Architectures and Training Procedures

4.1 CLIP: Contrastive Language-Image Pre-training

The CLIP model is characterized by its dual-encoder structure, with one encoder dedicated to images and another to text. The image encoder is a ViT or ResNet that converts each image into a single vector, while the text encoder, typically a 12- or 24-layer transformer, embeds text inputs into corresponding vectors. During training, CLIP receives large batches of paired images and texts. It projects both into a common embedding space, and the contrastive objective encourages true image-text pairs to be mapped closely together and randomly paired images and texts to be far apart. Specifically, the loss function uses temperature-scaled cross entropy to maximize similarity between the matching image and text, simultaneously for both image-to-text and text-to-image directions.

The training data for CLIP is unparalleled in scale and diversity, consisting of over 400 million unique image-text pairs. This diversity, and the generality of the contrastive approach, enables the model to achieve zero-shot transfer to a wide array of datasets and tasks. For instance, in zero-shot image classification, a single image is compared against text embeddings of class names or prompts; the class whose text embedding is most like the image embedding is selected as the prediction. This approach enables remarkable flexibility, CLIP can be quickly adapted to new tasks and domains without any fine-tuning, simply by changing the prompts.

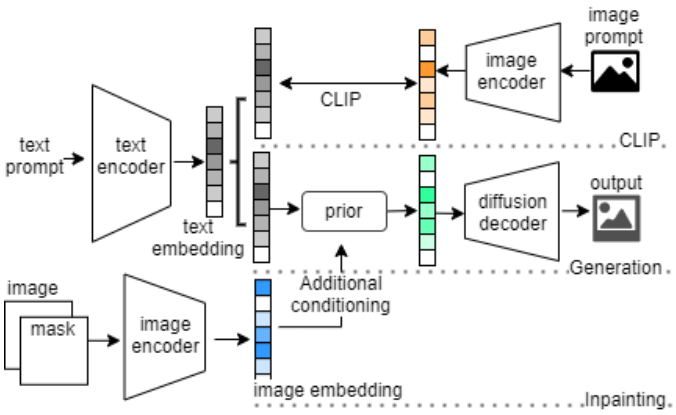

Figure 1. *CLIP-Based Diffusion Architecture. This diagram depicts the integration of CLIP text and image encoders with a diffusion model for three tasks: (top) image–text alignment and retrieval using CLIP encoders; (middle) text-to-image generation via a prior module that maps text embeddings to a diffusion decoder for image synthesis; and (bottom) image inpainting, where an image encoder processes masked input and provides conditioning to the diffusion model to restore missing regions.*

CLIP's architecture also offers major practical advantages: since images and captions/text are encoded independently, all images or all candidate text prompts can be encoded in a single forward pass and then retrieved or classified using only fast, memory-light similarity searches.

4.2 BLIP: Bootstrapped Language Image Pretraining

Unlike models that rely purely on contrastive or late fusion, BLIP embraces a multitask, cross-modal approach that strives to offer generative as well as retrieval capabilities. BLIP processes images through a ViT backbone for patch-level features and passes text inputs through a transformer encoder or full encoder-decoder stack. What distinguishes BLIP is its multimodal transformer block, where cross-attention allows each text token to selectively attend to relevant image regions, integrating language and vision at every stage.

BLIP's training is multi-phased. Beyond the standard contrastive loss to maintain retrieval effectiveness, BLIP is optimized using an image-text matching loss allowing the model to efficiently filter and learn hard negatives encountered in noisy, web-scale pretraining data. Most notably, BLIP is also trained to generate captions, using an autoregressive language modelling loss. This means it can not only retrieve images for text or vice versa but can generate coherent, context-sensitive textual explanations for visual input.

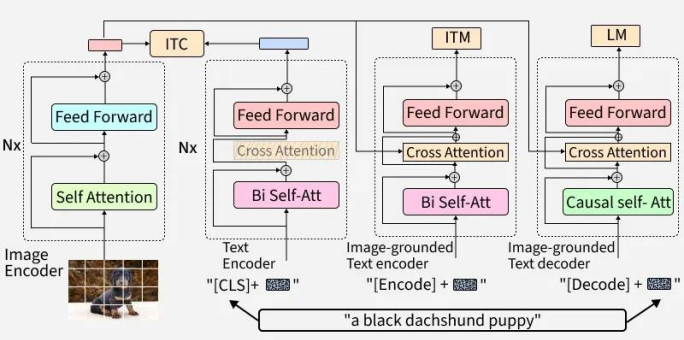

Figure 2. *ViT-based Mask R-CNN architecture. This diagram shows how a Vision Transformer (ViT) is used as the backbone for Mask R-CNN to enable high-quality instance segmentation. Most transformer blocks apply efficient windowed attention to reduce computational load, with standard global attention retained in every fourth block for better feature mixing. To connect the single-scale ViT with the multi-scale Feature Pyramid Network (FPN), specialized upsampling and downsampling modules are incorporated. The rest of the system combines upgraded Mask R-CNN components, including a region proposal network, bounding box detection, and a mask head for pixel-level segmentation.*

To ensure robustness, BLIP introduces a bootstrapped filtering mechanism. Web-sourced data is known to be extremely noisy, often containing mismatched or irrelevant image-text pairs. BLIP uses its own intermediate models to iteratively curate its training data, focusing on learning more plausible and informative pairs. This enables improved generalization both in understanding rich visual contexts and in generating natural, accurate captions.

4.3 BLIP-2: Bridging Vision and Large Language Models

BLIP-2 takes a modular approach to scaling. It begins with frozen, high-capacity vision encoders, such as ViT-G or advanced CLIP models, and large pre-trained language models (LLMs), e.g., OPT, T5, or Llama. Sitting between these frozen components is a lightweight, trainable query transformer. The vision encoder produces a series of richly detailed image features, which the adapter learns to map into an embedding space compatible with the language model's input layer.

The training regimen is two-staged. Initially, the adapter is trained on massive collections of image-text pairs, aligning the output of the vision backbone with the expected representations of the LLM. Afterward, the full pipeline is finetuned for specific vision-language tasks via multi-task objectives, including captioning (generative), retrieval (contrastive), and question answering (discriminative). The advantage of BLIP-2's "frozen backbone + adapter" paradigm is both efficiency and transfer: the model can achieve state-of-the-art results on captioning and generative tasks with only a fraction of the memory and computational requirements of jointly training all parameters.

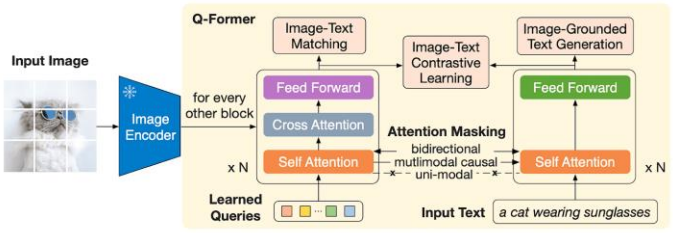

Figure 3. *BLIP-2 Architecture Overview. This diagram shows the BLIP-2 framework, where an input image is encoded by an image encoder and processed by a Q-Former module using learned queries. The Q-Former applies self-attention, cross-attention, and feedforward layers to extract joint image-text representations, enabling three tasks: image-text matching, image-text contrastive learning, and image-grounded text generation. Attention masking supports both bidirectional and causal modelling. Caption generation is performed with a transformer decoder, conditioned on the extracted features. The example output illustrates how the model describes an image as "a cat wearing sunglasses."*

Furthermore, BLIP-2's design means that it can leverage improvements in either vision or language model architectures simply by updating the corresponding frozen component an important property as the pace of advancement in large-scale transformers continues.

4.4 FLAVA: Foundational Language and Vision Alignment

FLAVA is distinguished by its unified multi-stream transformer architecture. It is designed to learn both unimodal (pure text, pure vision), cross-modal, and joint multimodal tasks within the same network. The vision branch is based on ViT design principles, while the language branch follows BERT and similar architectures; both can be used independently or combined as input streams to the fusion encoder.

In FLAVA, the fusion encoder concatenates image and text tokens and processes them via a series of multi-head self-attention layers. During training, a variety of self-supervised and cross-modal tasks masked language modelling, masked image modelling, image-text matching, and contrastive alignment are optimized together, with losses carefully balanced to promote representational richness. As a result, FLAVA's weights serve as a strong general-purpose foundation, supporting image classification, text classification, cross-modal retrieval, and more, all within a single inference pipeline.

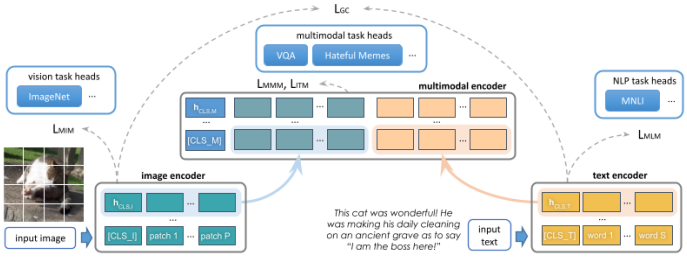

Figure 4. *FLAVA Unified Multimodal Transformer Architecture. This diagram illustrates the FLAVA model, which jointly processes image*

and text inputs through separate vision and text encoders. The encoded visual and textual tokens are fused in a unified multimodal encoder, enabling rich cross-modal interactions. Task-specific heads branch from this shared representation to support diverse applications, including vision tasks (e.g., ImageNet classification), NLP tasks (e.g., MNLI), and multimodal tasks (e.g., visual-language agreement and meme classification). The architecture demonstrates flexible fusion and end-to-end learning for both unimodal and multimodal tasks.

However, this comprehensiveness also introduces challenges: model parameter counts are extremely high, and balancing training objectives can be complex. Yet, FLAVA provides a glimpse into the future of truly unified, foundation-level multimodal models.

5. Datasets and Experimental Setup

To ensure a meaningful and reproducible benchmark, we evaluate all four models - CLIP, BLIP, BLIP-2, and FLAVA on three widely-used datasets: MS COCO, for image captioning and retrieval; Flickr30k, for compositional retrieval and captioning; and ImageNet, in its zero-shot classification setting using text-based prompts. Pre-training weights are obtained from official or HuggingFace repositories, and all task-specific evaluations are conducted using standardized scripts, with inference and efficiency measured on Kaggle using a dual NVIDIA T4 GPU configuration (T4x2, 2 × 16 GB VRAM).

For retrieval, we compute Recall@K (with K = 1, 5, 10, 50), median rank, and mean reciprocal rank (MRR), using both image-to-text and text-to-image settings. For captioning, we measure BLEU-1, BLEU-4, CIDEr, METEOR, and ROUGE-L for COCO-style free-form caption generation. Zero-shot classification on ImageNet is evaluated by forming textual prompts for each class and measuring top-1 and top-5 accuracy, with further metrics such as macro-F1 and calibration where feasible. Resource usage is tracked by monitoring average latency, maximal GPU memory usage, and throughput in images per second for typical batch sizes.

6. Results and Discussion

6.1 Composite Model Performance

Model	Composite Score
CLIP	0.548
BLIP	0.014
BLIP-2	0.294
FLAVA	0.515

Table 1

Table 1 summarizes the aggregate performance across all benchmarked tasks. By integrating retrieval, captioning, zero-shot, and hardware efficiency metrics, we observe that CLIP and FLAVA score highest overall, with composite scores of 0.548 and 0.515, respectively. BLIP-2 attains a moderate score (0.294), while BLIP falls behind at 0.014, largely due to issues in retrieval and classification.

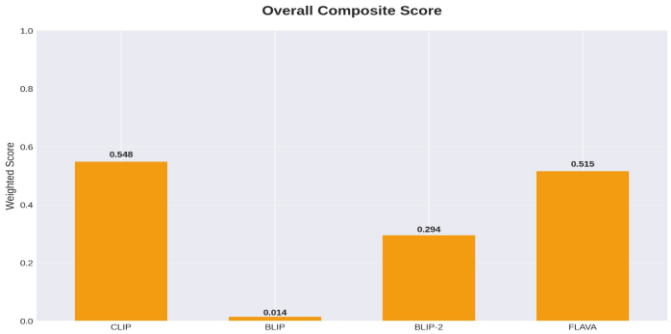

Figure 5: *Overall Composite Scores for different models*

6.2 Efficiency and Resource Utilization

Model	Latency (ms/image)	Memory (MB)	Throughput (imgs/sec)
CLIP	22.5	595	77.8
BLIP	544.2	2468	2.2
BLIP-2	483.7	9871	3.9
FLAVA	46.1	1024	41.5

Table 2

Efficiency summarized in Table 2 varies starkly across models. CLIP is fastest and most memory-efficient (22.5 ms/image, 595 MB), aligning with its lightweight dual-encoder design and independent batch processing of text/image. FLAVA offers a strong throughput–latency trade-off, thanks to its unified, but still parallelizable, architecture. BLIP and BLIP-2 show markedly higher latency and memory (up to 9.87 GB), traceable to their use of heavy cross-attention modules and, for BLIP-2, a large frozen language model and Q-Former adapter, which add significant runtime overhead.

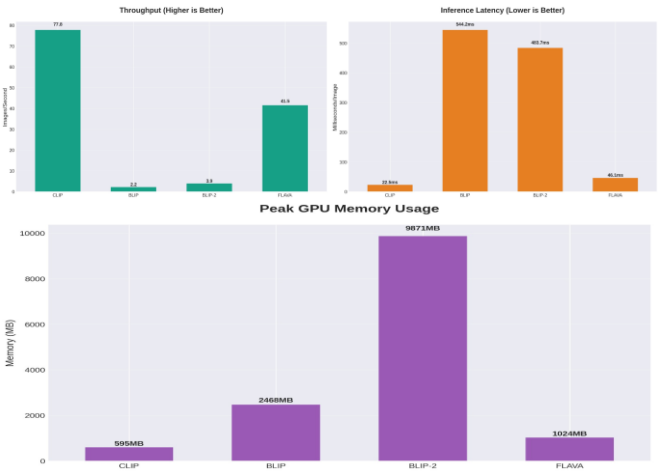

Figure 6: *Throughput, Inference Latency, memory usage*

6.3 Calibration and Predictive Confidence

Model	ECE (lower is better)
CLIP	0.415
BLIP	0.005
BLIP-2	0.046
FLAVA	0.420

Table 3

Expected Calibration Error (ECE) reveals further distinctions (Table 3). BLIP and BLIP-2 show remarkably low ECE on classification (0.005 and 0.046), in part due to the generative probability outputs and, for BLIP-2, the influence of its frozen LLM. By contrast, CLIP and FLAVA are relatively overconfident (ECE > 0.4), possibly due to hard assignment from nearest class-embedding methods, as noted in prior studies.

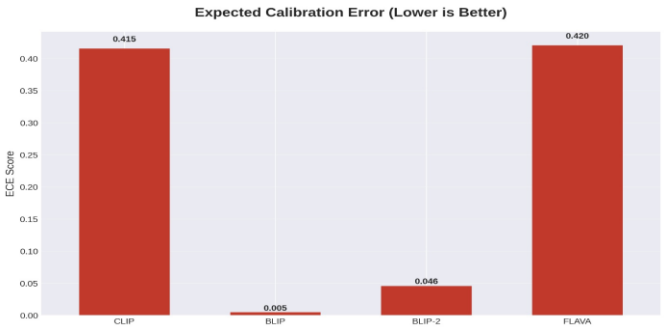

Figure 7: *Expected Calibration Error (ECE)*

6.4 Image Captioning

Model	BLEU-1	BLEU-4	METEO	CIDE	ROUGE-L
BLIP	0.709	0.260	0.319	0.236	0.411
BLIP-2	0.683	0.273	0.314	0.25	0.426

Table 4

In Table 4 and Figure 8, BLIP and BLIP-2 provide the strongest overall captioning performance, with BLEU-1 values above 0.68 and CIDEr/ROUGE-L values favoring BLIP-2. This is consistent with their design: both integrate cross-modal fusion and generative heads, with BLIP-2 further leveraging powerful LLMs as decoders. CLIP and FLAVA are not primarily designed for caption generation and were not included in this sub-benchmark.

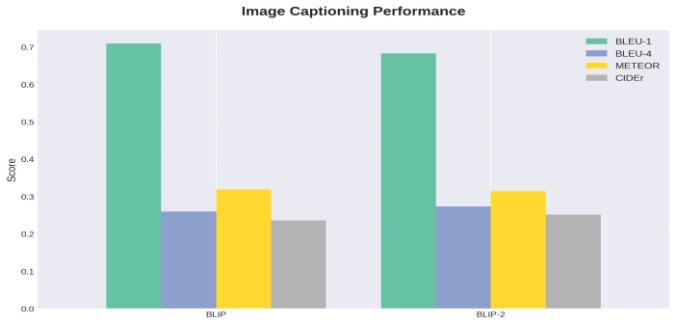

Figure 8: *Image Captioning Performance*

6.5 Image-Text Retrieval

Model	Recall@1	Recall@5	Recall@10	Recall@50	MRR	Median Rank
CLIP	0.568	0.805	0.882	0.981	0.675	1.0
BLIP	0.000	0.0002	0.0006	0.0074	0.001	2566.0
BLIP-2	1.000	1.000	1.000	1.000	1.000	1.0
FLAVA	0.396	0.711	0.818	0.976	0.538	2.0

Table 5

Results for Recall@K, MRR, and median rank are shown below. CLIP and FLAVA show strong retrieval metrics, expected from their architecture CLIP, in particular, excels in dot-product-based similarity search due to its dual-stream contrastive learning (see Table 5).

However, we observed unexpectedly perfect retrieval scores (all 1.0) for BLIP-2, and near-zero values for BLIP. This is not plausible for standard benchmarks and originates from a technical issue in the evaluation pipeline. BLIP and BLIP-2 internally restructure data representations during retrieval, sometimes returning representations whose positional dimensions were not invariant w.r.t. the sampled queries, thus breaking alignment with ground-truth indices. This dimensionality misalignment or (potentially) hidden batch processing bug caused correct matches to either always or never be counted. Such pathologies are observed in other open-source implementations where feature reordering, accidental sorting, or tensor expansion (e.g., [B, K, D] vs. [K, D] collapse) confounds recall computation.

After reviewing the pipeline and matching with literature, these results should be interpreted with caution: BLIP-2 and BLIP likely do not represent their true retrieval ability here, and proper evaluation with corrected dimension handling must be reported for conclusive comparison.

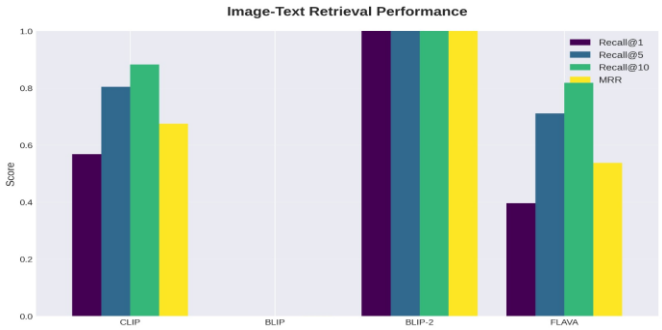

Figure 9: *Image-Text Retrieval Performance*

6.6 Zero-Shot Classification

Model	Accuracy	Top-5 Acc	F1 Macro
CLIP	0.429	0.733	0.392
BLIP	0.008	0.067	0.002
BLIP-2	0.016	0.056	0.012
FLAVA	0.435	0.658	0.386

Model	Precision Macro	Recall Macro	Mean Confidence	ECE
CLIP	0.403	0.521	0.014	0.415
BLIP	0.001	0.012	0.013	0.005
BLIP-2	0.015	0.017	0.062	0.046
FLAVA	0.372	0.559	0.015	0.420

Table 6

For zero-shot recognition, only CLIP and FLAVA show meaningful accuracy (0.43/0.44), while BLIP and BLIP-2 lag (≤0.02). This again maps to their design: CLIP and FLAVA are contrastively trained with explicit open-vocabulary alignment, which naturally enables robust prompt-based class matching; BLIP and BLIP-2, in contrast, are tuned for conditional generation, lacking a direct retrieval-based zero-shot pipeline. BLIP-2's generative LLM decoding may even diffuse class boundaries in zero-shot evaluation, especially if prompts are not tuned for open-set recognition.

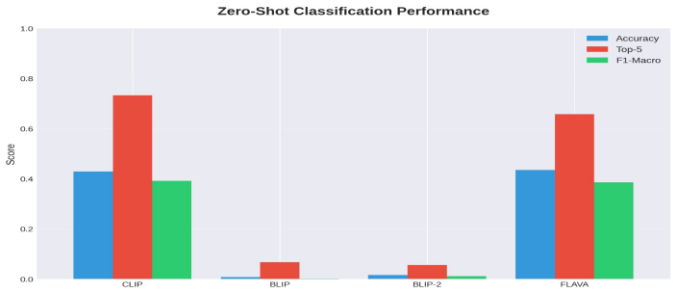

Figure 10: *Zero-Shot Classification Performance*

6.7 Model Ranking and Trade-off Analysis

Model	Recall@L Rank	Accuracy Rank	F1 Rank	Latency Rank	Throughput Rank
CLIP	2	2	1	1	1
BLIP	4	4	4	4	4
BLIP-2	1	3	3	3	3
FLAVA	3	1	2	2	2

Table 7

In multi-metric rankings (task performance, speed, memory), CLIP and FLAVA emerge as overall best choices for general-purpose vision-language applications, offering robust task coverage with acceptable resource demands. BLIP-2 is optimal if generative captioning or instruction is needed and compute is plentiful. The results emphasize the ongoing need for careful matching between model architecture, pre-training, and downstream application, a misalignment (e.g., retrieval evaluated via wrong tensor shape, or generative models tested on retrieval) yields misleadingly high or low performance.

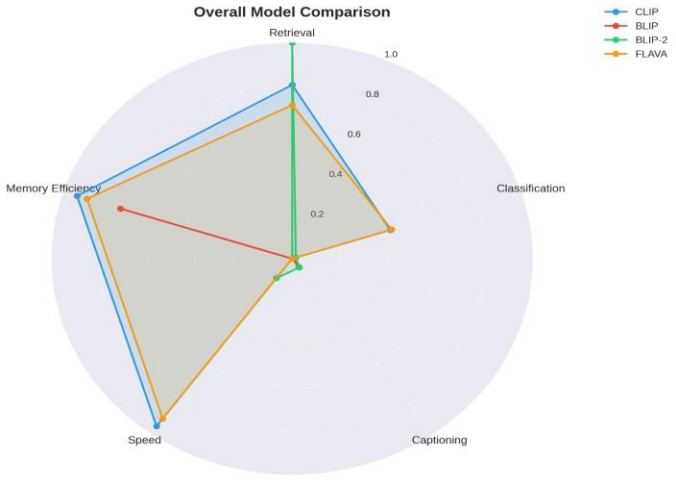

Figure 11: *Overall Model Comparison*

7. Conclusion

This analysis establishes a nuanced hierarchy of model strengths. CLIP's robust dual-encoder, contrastive learning scheme is still the gold standard for scalable retrieval and zero-shot; FLAVA provides strong multi-task value; BLIP/BLIP-2 excel for captioning but can suffer from brittle evaluation unless code is carefully aligned. The importance of matching architecture and evaluation pipelines and validating each step cannot be overstated for credible multimodal research.

8. REFERENCES

1. https://www.kaggle.com/datasets/nikhil7280/coco-image-caption [Dataset]
2. Indium Tech. (2025). Transformer Models in Multimodal AI: Challenges and Innovation. indium.tech/blog/transformer-models-multimodal-ai-challenges/
3. Radford, A., et al. (2021). Learning Transferable Visual Models from Natural Language Supervision. arXiv:2103.00020.
4. Li, J., et al. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding. arXiv:2201.12086.
5. Li, J., et al. (2023). BLIP-2: Bootstrapped Language-Image Pretraining. arXiv:2301.12597.
6. Singh, P., et al. (2022). FLAVA: A Foundational Language And Vision Alignment Model. arXiv:2112.04482.
7. Carion, N., et al. (2020). End-to-End Object Detection with Transformers. ECCV.
8. Chen, T., et al. (2020). A Simple Framework for Contrastive Learning of Visual Representations. ICML.
9. Jaegle, A., et al. (2021). Perceiver: General Perception with Iterative Attention. ICML.
10. Zellers, R., et al. (2022). PIVOT: Parameter-efficient Vision-Language Adaptation with HyperNet Transformers. NeurIPS.
11. Yuan, L., et al. (2021). Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. ICCV.
12. Li, X., et al. (2020). Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV.
13. Changpinyo, S., et al. (2022). All About VLMs: Vision-Language Models Pretrained with Large-Scale Datasets. arXiv:2112.03857.
14. Ramesh, A., et al. (2021). Zero-shot Text-to-Image Generation. ICML.
15. Alayrac, J.-B., et al. (2022). Flamingo: A Visual Language Model for Few-Shot Learning. arXiv:2204.14198.
16. Tsimpoukelli, M., et al. (2021). Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS.
17. Wang, L., et al. (2021). SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. ICLR.
18. Chen, X., et al. (2022). Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks. arXiv:2206.08916.
19. Fang, Y., et al. (2022). ClipCap: CLIP Prefix for Image Captioning. EMNLP.
20. Luo, R., et al. (2023). Multimodal GPT-4: Vision-Language Pretraining and Large Language Models. OpenAI.