Technical Presentation: BLIP-2, Flamingo, MedCLIP, and LLaVA – Multimodal Learning Models

[1]Denis Samatov
[1]Department of Mathematics and Mathematical Physics,
Tomsk Polytechnic University, Russia
denissamatov470@gmail.com

April 20, 2025

Abstract

Multimodal learning, which jointly processes visual and linguistic data, has achieved remarkable progress with the emergence of the BLIP-2, Flamingo, MedCLIP, and LLaVA models. This presentation examines these four cutting-edge models, each addressing unique challenges in the field. BLIP-2 introduces an efficient bootstrapping strategy for vision-language pre-training, significantly reducing computational overhead. Flamingo implements few-shot in-context learning for multimodal tasks, enabling rapid adaptation with minimal data. MedCLIP tackles the shortage of paired data in the medical domain by leveraging innovative data scaling and semantic alignment techniques. LLaVA extends multimodal functionality to instruction following, creating an interactive AI assistant for image-based dialogues. By exploring their architectures, training approaches, and notable innovations, this work underscores the versatility and potential of multimodal learning to drive transformative AI applications across diverse domains.

1 Introduction

Multimodal learning refers to AI models that jointly learn from multiple data modalities, such as images and text. This is crucial for tasks like image captioning, visual question answering (VQA), and dialogue about images, where understanding both visual and linguistic information is required. In recent years, large vision-language models have achieved remarkable performance by combining advances in computer vision and natural language processing. However, training such models end-to-end is often resource-intensive, and new approaches are needed to make multimodal learning more efficient, adaptable, and specialized.

Key Recent Advances:

- **BLIP-2 (2023):** Introduced a strategy to **bootstrap** vision-language pre-training using *frozen* image encoders and language models, bridged by a lightweight *Querying Transformer (Q-Former)*. This drastically reduces training cost while achieving state-of-the-art performance with far fewer trainable parameters [1]. (For example, BLIP-2 outperforms DeepMind's 80B Flamingo model on VQAv2 with 54× fewer learnable parameters.)
- **Flamingo (2022):** Developed by DeepMind as a **few-shot** multimodal learner that combines a frozen vision encoder and a frozen language model via **Perceiver Resampler** and gated cross-attention layers. Flamingo was the first to demonstrate broad *in-context learning* ability in vision-language tasks, setting new state-of-the-art results on many benchmarks with minimal task-specific data [4].
- **MedCLIP (2022):** Pioneered a domain-specific multimodal model for medicine, addressing the unique challenges of medical imaging and text. It uses a **semantic alignment loss** (based on medical knowledge) and a novel data scaling approach with unpaired images and texts to overcome data scarcity and false negatives [3]. MedCLIP surpassed prior methods (e.g. GLoRIA, ConVIRT) in accuracy while using an order of magnitude less paired data.
- **LLaVA (2023):** Proposed as a "**Large Language and Vision Assistant**," LLaVA integrates a vision encoder with a language model and is **instruction-tuned** on GPT-4 generated image-question-answer data [2]. This approach endows the model with impressive multimodal conversational abilities, achieving near GPT-4 level performance on certain benchmarks (e.g. 92.5% accuracy on ScienceQA when paired with GPT-4 reasoning) [2]. LLaVA demonstrates the power of synthetic multimodal training data for zero-shot image understanding.

These four models highlight how researchers are pushing the boundaries of multimodal learning: improving efficiency (BLIP-2), enabling few-shot learning (Flamingo), specializing to high-stakes domains (MedCLIP), and aligning with instruction-following paradigms (LLaVA). Next, we discuss the specific problems each model addresses and the methods they employ.

2 Problem Statement

Each model was designed to tackle specific challenges in vision-language learning, often limitations of prior approaches:

2.1 BLIP-2: Bridging Modalities with Limited Training Cost

- **High computational cost of VLP:** Before BLIP-2, most vision-language pre-training required end-to-end training of very large models on massive image-text datasets, which is *prohibitively expensive* in compute [1]. This made it hard for researchers with limited resources to train or fine-tune such models.
- **Underutilization of unimodal models:** Prior methods did not effectively leverage the rich knowledge in existing single-modal models (e.g. powerful vision encoders or language models). Training from scratch meant *relearning* vision and language representations that were already available in pretrained ViTs or LLMs [1].
- **Modality alignment gap:** Simply plugging a frozen visual encoder and a frozen language model together doesn't work out-of-the-box – their feature spaces are incompatible. Ensuring the image features align with what the language model expects is a core difficulty. Earlier attempts (e.g. Frozen, Flamingo) used an image-to-text generation loss to try to align modalities, but this proved *insufficient* to bridge the gap.

2.2 Flamingo: Few-Shot Learning in Vision-Language Models

- **Need for task-specific fine-tuning:** Previously, vision-language models typically required fine-tuning on each new task with substantial data. There was no architecture that could *learn new tasks from only a few examples* (in contrast to text-only GPT-style few-shot learning).
- **Handling interleaved image-text sequences:** Traditional models struggled to process inputs consisting of alternating images and text (e.g. a dialog about an image). Designing a model that could ingest a sequence of multiple images and text in a single context (for in-context learning) was unsolved.
- **Preserving pre-trained knowledge:** Combining a pretrained vision model with a pretrained language model is tricky – naïve approaches might degrade their knowledge. The challenge was to condition the language model on visual data *without* forgetting its language skills, and similarly incorporate images without re-training the vision encoder from scratch.

2.3 MedCLIP: Adapting to Medical Domain Constraints

- **Scarcity of paired data:** In the medical domain, large captioned image datasets are virtually nonexistent. Often images (like X-rays) have only short labels or no accompanying report. This leads to *orders of magnitude*

less data for vision-language training compared to web data [3]. For example, a dataset like CheXpert has 190K images but only binary pathology labels, not detailed captions.

- **False negatives in contrastive learning:** Standard contrastive VLP (like CLIP) assumes any non-matched image-text pair is negative. In medicine, two different patient's images and reports might describe the *same condition*, yet they are treated as negatives in training. This introduces harmful noise – the model learns to separate semantically identical concepts.

- **Fine-grained semantic nuance:** Medical images and text contain subtle distinctions (e.g. mild vs. moderate findings) and often multiple findings per image (*multi-morbidity*). Existing models weren't specialized to handle these nuances, leading to suboptimal representation of medical features. The challenge was to incorporate domain knowledge to align visual and textual semantics more intelligently than a generic method.

2.4 LLaVA: Vision-Language Models that Follow Instructions

- **Lack of multimodal instruction data:** Big language models became good at following instructions by training on human or AI-generated instruction-response pairs. However, vision-language models had mostly been trained on captions or VQA, which are *not* in an interactive instruction format. Datasets like COCO Captions describe images but do not ask or answer questions about them, leaving a gap for developing an AI assistant that can have a dialogue about an image.

- **Costly data annotation:** Creating high-quality, diverse vision-language instruction data (e.g. asking complex questions about images and providing answers) would normally require a *huge manual effort*, involving experts to annotate dialogues or QA pairs for many images. This was a bottleneck for instruction-tuned multimodal models.

- **Modality integration challenges:** Simply connecting a vision encoder to a language model does not guarantee it will follow instructions about images. Previous attempts (e.g. OpenFlamingo) showed that without careful tuning, the language model may ignore visual input or hallucinate. The key problems were how to inject visual information into the LLM effectively and how to train the combined model to reliably understand a user's question about an image.

3 Methods

We now summarize the architecture and training methodology of each model, highlighting their notable design innovations:

3.1 BLIP-2: Bootstrapping with Frozen Encoders and Q-Former

BLIP-2 employs a **two-stage pre-training strategy** with almost all weights frozen [1]. First, a pre-trained vision encoder (e.g. a ViT visual backbone) and a pre-trained large language model (LLM such as OPT or FLAN-T5) are taken *as-is* (frozen). To connect them, BLIP-2 introduces a small learnable module called the **Querying Transformer (Q-Former)** [1]. The Q-Former is a lightweight transformer that uses a set of learnable query vectors to "query" the image encoder's output and produce a fixed-length embedding that the LLM can understand.

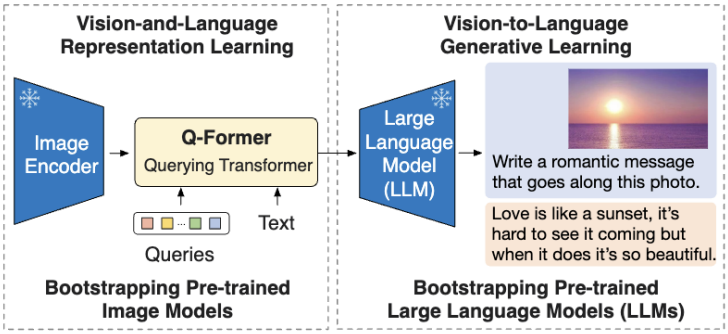

Figure 1: Overview of BLIP-2's framework. We pre-train a lightweight Querying Transformer following a two-stage strategy to bridge the modality gap. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen LLM, which enables zero-shot instructed image-to-text generation.

- **Two-stage training:** The Q-Former is trained in two phases. In Stage 1, it learns to align visual features with text by outputting embeddings that correspond to image captions (vision-language representation learning). The image encoder is frozen, but a dataset of image-text pairs is used to train Q-Former such that the queries extract features relevant to the accompanying text. In Stage 2, the Q-Former's output is fed into the frozen LLM, and the model is optimized on a generative task (predicting text) with the image in context [1]. This teaches Q-Former to produce visual tokens that the LLM can integrate for image-to-text generation.
- **Modality bridging:** By training this intermediate transformer, BLIP-2 effectively **bridges the modality gap** between vision and language. The frozen image encoder provides rich visual features, and the frozen LLM provides powerful language generation; Q-Former serves as the translator between the two. This setup avoids catastrophic forgetting in the LLM (since its weights stay fixed) and avoids full fine-tuning of the vision model, drastically cutting down learnable parameters.

3.2 Flamingo: Few-Shot Visual Language Model with Perceiver Resampler

Flamingo's architecture is designed to enable **few-shot learning** in a multimodal context [4]. It achieves this by integrating a frozen vision encoder and a frozen LLM, with a couple of trainable components connecting them:

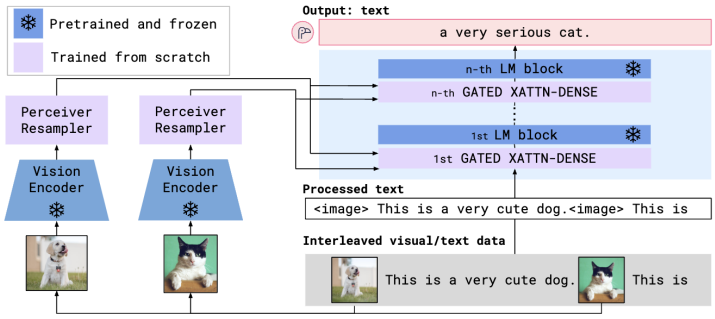

Figure 2: Flamingo architecture overview. Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

- **Frozen backbones:** Flamingo uses a pretrained **CLIP vision encoder** (e.g. a ViT) to handle images, and a large **language model** (70B Chinchilla in the full version) to handle text, both kept frozen during Flamingo training. This preserves the visual recognition abilities of CLIP and the linguistic knowledge of the LLM.

- **Perceiver Resampler:** A novel module called the Perceiver Resampler takes the patch-wise image features from the vision encoder and **compresses them into a small fixed set of visual tokens** [4]. For example, regardless of an input having one or several images, the Resampler might output, say, 16 learned tokens that summarize the visual content. This learned compression enables the model to handle a variable number of images and varying image sizes, while interfacing with the language model in a consistent way [4]. The Perceiver Resampler is inspired by DeepMind's Perceiver IO architecture and is trained from scratch within Flamingo.

- **Gated cross-attention layers:** To feed visual information into the LLM, Flamingo inserts new **cross-attention layers** (with gating) at certain points in the language model's stack. These layers allow the LLM to attend to the visual tokens from the Resampler. In effect, as the language model generates text, it can query the image representations through these cross-attention blocks. The gating mechanism helps modulate how much influence the image has at each layer, ensuring that the LLM's original

language ability is not overwhelmed. Only these inserted layers (and the Resampler) are trained; the rest of the LLM remains untouched.

- **Training regimen:** Flamingo is trained on a massive set of **interleaved image-text sequences** – on the order of 1.8B examples drawn from web data (e.g. alt-text, captions, dialogues) according to DeepMind's paper. The model is given a few exemplars (image-question-answer triples, for instance) and then tested on a new query, all in a single sequence, to simulate few-shot learning. The objective is next-word prediction (as in language modeling), so Flamingo learns to utilize the provided examples to predict the correct output for the query. By training in this fashion, it learns to *learn within its context*, i.e. to perform few-shot reasoning.

3.3 MedCLIP: Medical Image-Text Contrastive Learning with Semantic Alignment

MedCLIP adapts the CLIP-style contrastive framework to the medical domain, introducing several important modifications to address the issues of data scarcity and semantic misalignment [3]:

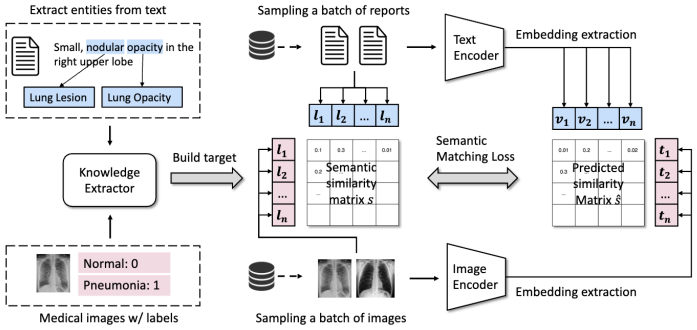

Figure 3: The workflow of MedCLIP. The knowledge extraction module extracts medical entities from raw medical reports. Then, a semantic similarity matrix is built by comparing medical entities (from text) and raw labels (from images), which enables pairing arbitrary two separately sampled images and texts. The extracted image and text embeddings are paired to match the semantic similarity matrix.

- **Architecture:** It consists of a **ResNet-50** CNN as the visual encoder and a **BioClinicalBERT** (a BERT model pre-trained on clinical text) as the text encoder. These encoders produce image and report embeddings, respectively. The model is trained to align embeddings of matching image-report pairs and separate those of non-matching pairs, similar to CLIP. MedCLIP doesn't use an extravagant architecture; instead, its novelty lies in *how it constructs training pairs and defines the loss function.*

- **Combinatorial data scaling:** To compensate for the lack of explicit image-report pairs, MedCLIP leverages metadata and labels to **generate additional training pairs**. Essentially, it *decouples images and texts* during training. For example, if an X-ray image is labeled with "pneumonia" and there is a report (from another patient) that discusses pneumonia, MedCLIP can treat them as a positive pair for training even though they didn't originally come together [3]. By doing this systematically using disease tags and medical ontology (UMLS concepts), the authors scaled the effective training set size **combinatorially** without new data collection. This approach provided far more image-text pairs (albeit noisy ones) to learn from, addressing the data shortage.

- **Semantic alignment loss:** A standard InfoNCE loss would treat any unmatched image-report as negative. MedCLIP replaces this with a **semantic matching loss** that uses medical knowledge to weigh pairs appropriately [3]. In practice, the model avoids penalizing embeddings of an image and a report that share key medical findings. If two items are different patients but both mention "cardiomegaly" or "no abnormality," the loss function recognizes the shared semantics and does not push those embeddings apart as strongly (mitigating the false negative issue). This is implemented by incorporating similarity of *UMLS concept* annotations into the loss, so the model focuses on aligning disease-related content.

- **Efficient training and results:** With these improvements, MedCLIP achieved excellent performance on tasks like zero-shot medical image classification and retrieval. Notably, it reached higher accuracy than prior state-of-the-art methods (such as GLoRIA, which was another medical CLIP variant) while using only **20K training pairs vs. ∼200K** in those methods [3]. This 10× reduction in required data is extremely valuable in the medical context. MedCLIP's image embeddings also proved more clinically relevant, since the training better reflected medical semantics rather than just surface-level alignment. Overall, MedCLIP demonstrated how introducing domain knowledge and creative data augmentation can greatly enhance multimodal learning in specialized fields.

3.4 LLaVA: Large Language and Vision Assistant with Instruction Tuning

LLaVA extends the paradigm of instruction-tuned language models (like ChatGPT) to the **vision+language setting** [2]. Its method can be seen in two parts: (1) constructing a multimodal instruction dataset, and (2) building a model that learns from it by connecting a vision encoder to a language model.

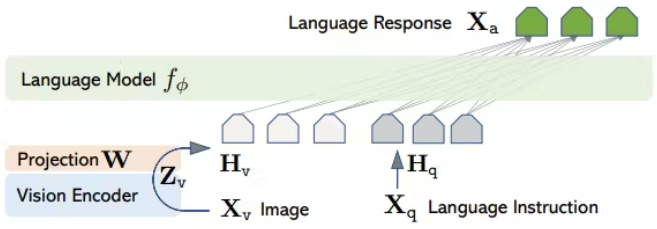

Figure 4: LLaVA network architecture.

- **GPT-4 generated training data:** A major contribution of LLaVA is showing that one can use a powerful LLM (GPT-4) to *generate training data* for multimodal tasks [2]. The authors took image-caption pairs from existing datasets (e.g. COCO) and prompted GPT-4 to act as a user and assistant, producing rich **dialogues and question-answer pairs** about the image. For instance, given an image and its caption, GPT-4 might generate a series of questions ("What is in the image? Why might the scene be happening?") and detailed answers. They curated a dataset of ∼158K such **instruction-following samples** covering various interaction types – conversations about the image, detailed descriptions, complex reasoning, etc. This became the training data for LLaVA's instruction tuning. The key here is that *no human annotation was needed*; GPT-4's outputs, while not perfect, were used as high-quality proxies to train the vision-language assistant.

- **Model architecture:** The LLaVA model connects a **ViT-L/14 image encoder (from CLIP)** to a **Vicuna-13B language model** (Vicuna is an open derivative of LLaMA). The image encoder is pre-trained (e.g. CLIP) and provides a visual feature vector for an input image. A small **projection layer** (a linear layer) then maps the image features into the embedding space of the language model. This projected image token is appended to the language model's input, effectively acting as a special token that informs the LLM about the image. The language model itself (Vicuna/LLaMA) is then fine-tuned (with gradients) on the multimodal instruction-response data. During training, it sees prompts like: " $<$ *Image* $><$ *User* $>$: [*question*]? $<$ *Assistant* $>$: " and learns to produce the answer, using the image token information when relevant. Because the image encoder is frozen or only lightly tuned, most of the learning happens in how the LLM incorporates visual context.

- **Instruction tuning and capabilities:** By training on GPT-4's Q&A, LLaVA learns to follow a wide range of **natural language instructions about images**. After tuning, one can ask LLaVA things like "Describe this image in detail" or "What might happen next in the scene?" and it will generate a coherent answer referencing the image content, much like one would expect from ChatGPT with vision. It effectively **mimics the**

behavior of multimodal GPT-4 on many tasks [2]. The results were impressive: without any human-labeled data, LLaVA achieved strong zero-shot performance on benchmarks. When further fine-tuned on a specific evaluation (e.g. ScienceQA, a challenging multimodal science question dataset), LLaVA (plus GPT-4's reasoning) reached *92.5% accuracy*, a new state-of-the-art [2]. This shows the synergy of combining a vision model with an instruction-tuned LLM.

4 Model Comparison Table

To provide a clear side-by-side view of the key characteristics of the discussed models, here is a comparative table. This helps highlight the main differences in architecture, training approaches, and strengths of each model.

Model	Architecture	Training Data	Parameters	Key Innovation	Main Strength
Flamingo	Frozen Vision Encoder (CLIP ViT) + **Perceiver Resampler** (trainable) + Frozen LLM (Chinchilla 70B) with inserted **Gated Cross-Attention** layers (trainable)	Massive sets of interleaved image-text sequences (web data, ∼1.8B examples for in-context training objective)	∼80B (mostly frozen)	Perceiver Resampler for visual token compression; Gated Cross-Attention for LLM integration	**Few-shot in-context learning** across a wide range of tasks
BLIP-2	Frozen Vision Encoder (ViT) + **Q-Former** (trainable, lightweight) + Frozen LLM (OPT/FLAN-T5)	Standard image-text pairs (e.g., COCO, Visual Genome, CC); uses **bootstrapping** strategy	Significantly fewer *trainable* params (Q-Former, ∼188M–360M)	**Q-Former** as a lightweight bridge between frozen encoders; **Two-stage** pre-training strategy	High **efficiency** in training and resource usage; **Modularity**
MedCLIP	Vision Encoder (ResNet-50) + Text Encoder (**BioClinicalBERT**) + Contrastive learning	Medical images and texts; uses **unpaired** data and **combinatorial scaling** via metadata/labels (UMLS)	∼400M (estimated)	**Semantic Alignment Loss** using medical knowledge (UMLS); Overcoming false negative pairs	**Adaptation to medical domain specificity**; Accuracy with limited paired data
LLaVA	Vision Encoder (CLIP ViT-L/14, frozen) + **Linear Projection Layer** (trainable) + LLM (**Vicuna-13B**, fine-tuned)	**Instructions generated by GPT-4** based on image-caption pairs (∼158k instructions); Multimodal dialogues & Q&A	∼13B (primarily the Vicuna LLM)	**Visual Instruction Tuning** using synthetic data from GPT-4	Following **natural language instructions** in a multimodal context; Conversational abilities

Table 1: Comparison of Multimodal Learning Models

Table Summary:

- **Flamingo** stands out for its few-shot learning capability, enabled by the Perceiver Resampler and its massive scale.
- **BLIP-2** demonstrates how to achieve high performance with minimal training cost by effectively leveraging frozen components and the Q-Former.
- **MedCLIP** shows the importance of domain adaptation and clever data utilization (including unpaired data) for specialized tasks like medicine.

- **LLaVA** paves the way for multimodal assistants capable of following instructions, using data generated by other powerful models (GPT-4) for training.

5 Repository Links

For those interested in exploring or reproducing these models, the public codebases are:

- **BLIP-2:** Salesforce's *LAVIS* implementation[1].
- **Flamingo:** LAION's *OpenFlamingo* re-implementation[2].
- **MedCLIP:** Official PyTorch code[3].
- **LLaVA:** Complete codebase[4] accompanying the paper [2].

Each repository contains further documentation, instructions for training or using the models, and in some cases pretrained weights or demo applications. These resources enable researchers and developers to experiment with the models and build upon them in future work.

6 Research Gap Analysis and Recommendations

Building on the comparative study of BLIP-2, Flamingo, MedCLIP and LLaVA, this chapter distils the most pressing open problems and presents concrete research directions that I recommend pursuing over the next 18 months.

6.1 Key Gaps Still to Be Addressed

1. **Spatial Grounding Deficit**
   Current models rely on global or heavily – compressed visual tokens. None natively output region masks or bounding boxes, limiting clinical explainability and safety-critical deployment.

2. **Long-Horizon Multimodality**
   Real-world scenarios demand fusion of video, audio and longitudinal text records. Architectures remain image-centric; temporal reasoning is largely absent.

3. **Lack of Calibrated Uncertainty**
   Free-form text answers offer no confidence intervals. Medical decision support requires formally calibrated risk estimates.

[1] `https://github.com/salesforce/LAVIS`
[2] `https://github.com/mlfoundations/open_flamingo`
[3] `https://github.com/RyanWangZf/MedCLIP`
[4] `https://github.com/haotian-liu/LLaVA`

4. **Privacy and Governance**
   Large-scale scraping (Flamingo) and synthetic dialogues (LLaVA) risk memorising personal data. Differentially private multimodal learning is still an open field.

5. **Compute Inequality**
   Even "efficient" BLIP-2 variants assume multi-GPU clusters. Edge devices in low-resource clinics cannot host such models.

6.2 Recommended Research Directions

1. **Region-Aware Tokenisation**
   Embed DETR-style object queries into Q-Former and Perceiver Resampler modules. Train with grounding supervision so that each visual token maps to a specific salient region.

2. **Hierarchical Memory for Long Context**
   Combine a retrieval-augmented vector store with LLaVA to persist multi-visit patient histories and hour-long video-chat transcripts without GPU-heavy attention over the full sequence.

3. **Probabilistic Decoding**
   Integrate conformal prediction or Monte-Carlo dropout into the language head; output calibrated confidence scores alongside narrative answers.

4. **Federated Multimodal Training**
   Prototype a privacy-preserving BLIP-2 variant that learns from on-device images and texts via secure aggregation, eliminating raw data transfer.

5. **Mixture-of-Experts Distillation**
   Distil Flamingo/BLIP-2 into sparsely-activated experts (<1 B active parameters) for Jetson-class hardware, reducing inference latency by $2\times$.

7 Conclusion

This presentation has examined four pioneering multimodal models—**BLIP-2**, **Flamingo**, **MedCLIP**, and **LLaVA**—each representing important advancements in integrating vision and language:

- **BLIP-2** drastically cuts compute by pairing *frozen* image encoders with large language models via a lightweight *Querying Transformer (Q-Former)*, outperforming larger architectures with far fewer trainable parameters.

- **Flamingo** introduces *few-shot in-context learning*, efficiently adapting to new tasks with minimal data by using a *Perceiver Resampler* and *gated cross-attention layers*.

- **MedCLIP** addresses medical data scarcity using a *semantic alignment loss* and *combinatorial data scaling*, achieving robust performance with far fewer paired samples than prior methods.
- **LLaVA** merges *instruction-following* with multimodal inputs, training on synthetic GPT-4 data to create a conversational AI assistant that excels in interactive, image-based queries.

Collectively, these models reduce training costs, support few-shot learning, handle domain-specific constraints, and facilitate interactive dialogues. They spotlight the future of multimodal AI: more accessible, specialized, and user-centric systems applicable to fields like healthcare, education, robotics, and beyond.

References

[1] Li, J., Li, D., Savarese, S., Hoi, S. BLIP-2: Bootstrapping Language–Image Pre-training with Frozen Image Encoders and Large Language Models. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023).

[2] Liu, H., Mao, J., Bai, Y., et al. Visual Instruction Tuning. In Advances in Neural Information Processing Systems (NeurIPS 2023).

[3] Wang, Z., Wang, H., Wang, L., Yang, L. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In Empirical Methods in Natural Language Processing (EMNLP 2022).

[4] Alayrac, J.-B., Donahue, J., Simonyan, K., et al. Flamingo: A Visual Language Model for Few-Shot Learning. In Advances in Neural Information Processing Systems (NeurIPS 2022).