Visual Foundation Models

Introduction

- Visual foundation models bridge vision and language tasks.
- CLIP and Visual Question Answering (VQA) are state-of-the-art models enabling:
  - Image-text retrieval.
  - VQA.
  - Image captioning.

CLIP: Contrastive Language-Image Pre-training

- Developed by OpenAI for learning visual concepts from text.
- Trained on 400M image-text pairs from the web.
- Contrastive learning aligns image and text embeddings.

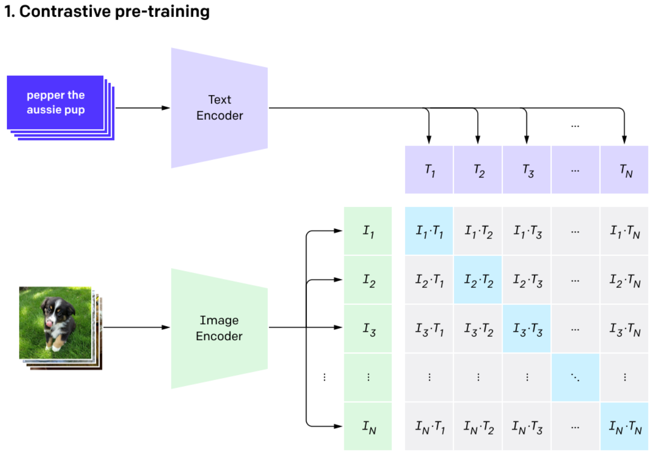


Figure: CLIP Architecture: Dual encoders for image and text.

Why CLIP Works

- Leverages diverse, large-scale datasets.
- Embedding alignment generalizes across tasks.
- No task-specific fine-tuning required.

Key Idea

Maximizes similarity for matching image-text pairs while minimizing it for non-matching pairs.

Applications of CLIP

- Zero-shot classification (e.g., "Is this a cat?").
- Text-to-image retrieval ("Find an image of a dog.").
- Visual concept understanding.

BLIP: Bootstrapping Language-Image Pre-training

- Unified vision-language model for understanding and generation..
- Trained on a mix of 14M curated and bootstrapped image-text pairs.

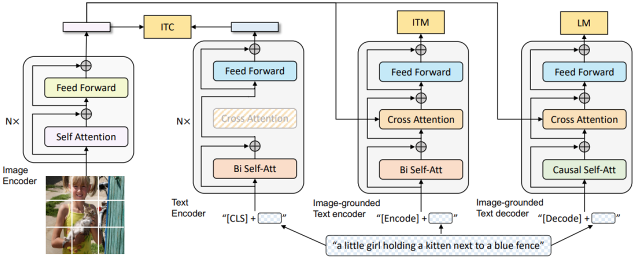


Figure: BLIP Architecture

BLIP Architecture: Multimodal Mixture of Encoder-Decoder (MED)

- Functions as:
  - **Unimodal encoder**: which separately encodes image and text.
  - **Image-grounded text encoder:** which injects visual information by inserting one additional cross-attention (CA) layer between the self-attention (SA) layer and the feed forward network (FFN) for each transformer block of the text encoder
  - **Image-grounded text decoder:** which replaces the bidirectional self-attention layers in the image-grounded text encoder with causal self-attention layers.

BLIP Architecture: Multimodal Mixture of Encoder-Decoder (MED)

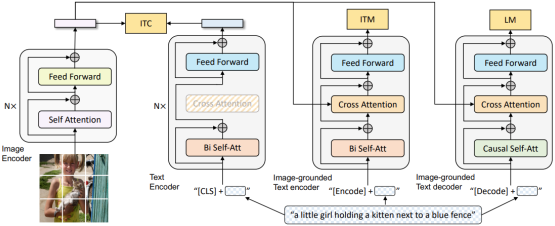


Figure: BLIP Architecture

BLIP is Pre-trained by :

- **Image-Text Contrastive (ITC) loss:** It aims to align the feature space of the visual transformer and the text transformer by encouraging positive image-text pairs to have similar representations in contrast to the negative pairs
- **Image-Text Matching (ITM) loss.**
- **Language Modeling (LM) loss.**

BLIP Architecture: Multimodal Mixture of Encoder-Decoder (MED)

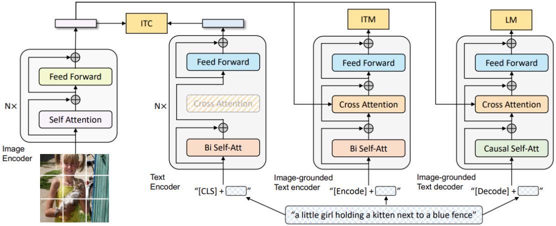


Figure: BLIP Architecture

BLIP is Pre-trained by :

- **Image-Text Contrastive (ITC) loss.**
- **Image-Text Matching (ITM) loss:** it performs fine-grained understanding by determining whether a given image-text pair is semantically matched.
- **Language Modeling (LM) loss.**

BLIP Architecture: Multimodal Mixture of Encoder-Decoder (MED)

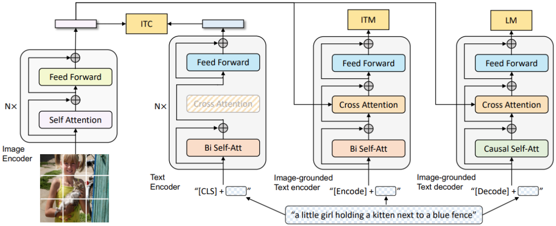


Figure: BLIP Architecture

BLIP is Pre-trained by :

- **Image-Text Contrastive (ITC) loss.**
- **Image-Text Matching (ITM) loss:**
- **Language Modeling (LM) loss:** it optimizes a cross entropy losswhich trains the model to maximize the likelihood of the text in an autoregressive manner.

Why BLIP Works

- CapFilt improves dataset quality by filtering noise and increasing the dataset size.
- Unified architecture enables multitasking.
- Diverse objectives enhance generalization.

Key Contribution

Combines understanding (e.g., retrieval) and generation (e.g., captioning) in one model.

Applications of BLIP

- Image-text retrieval (e.g., "Find an image for this caption").
- Image captioning.
- Visual Question Answering (VQA).
- Zero-shot video-language tasks.

Applications of BLIP

- Image-text retrieval (e.g., "Find an image for this caption").
- Image captioning.
- Visual Question Answering (VQA).
- Zero-shot video-language tasks.

Training Details

- CLIP:
  - Trained on 400M web-sourced pairs.
  - Optimizes contrastive loss.
- BLIP:
  - 14M curated and bootstrapped pairs.
  - Pre-training with ITC, ITM, and LM losses.

Training Procedure of BLIP: Overview

- Dataset Preparation:
  - Curated datasets (e.g., COCO, Visual Genome).
  - Web-sourced noisy datasets cleaned using CapFilt.
- Model Initialization:
  - Visual Transformer initialized from ImageNet pre-training.
  - Text Transformer initialized from BERT-base.
- Multi-task Pre-training:
  - ITC, ITM, and LM losses jointly optimized.

CapFilt Framework: Captioning and Filtering

- Captioner:
  - Generates synthetic captions for web images.
  - Fine-tuned with Language Modeling (LM) loss.
- Filter:
  - Removes noisy captions from original and synthetic texts.
  - Fine-tuned with Image-Text Matching (ITM) loss.
- Result:
  - High-quality bootstrapped dataset for pre-training.

Introduction to LLaVA

- LLaVA integrates vision and language models for general-purpose understanding.
- First attempt at instruction-tuning in the multimodal space.
- Combines a visual encoder with a large language model for multimodal tasks.

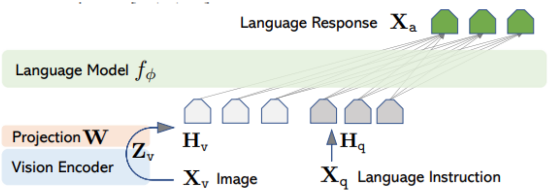
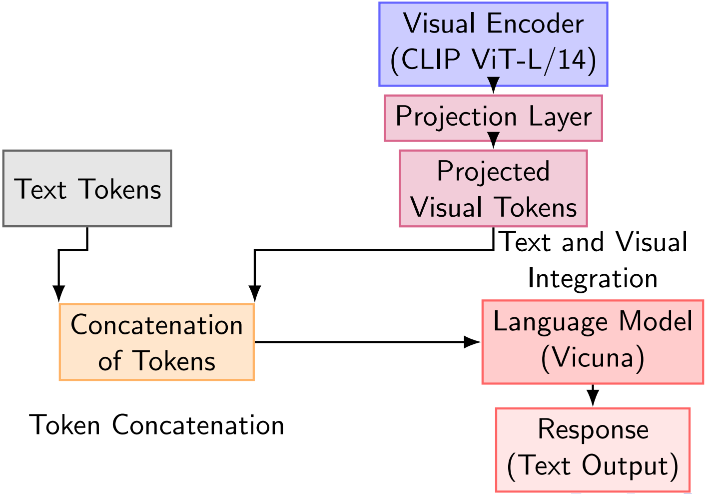


Figure: LLaVA Architecture

LLaVA's Architecture

- Visual encoder: CLIP ViT-L/14 processes images into features + Lightweight projection layer
- Language model: Vicuna, a LLM fine-tuned for chatting.



Data Generation for LLaVA training

- Uses GPT-4 to generate multimodal instruction-following data.
- Three types of responses:
  - Conversation: Q&A about objects, actions, and relations.
  - Detailed description: Comprehensive descriptions of scenes.
  - Complex reasoning: In-depth reasoning about visual content.
- Collected 158k samples of language-image pairs for instruction tuning.

LLaVA's Training Details

- **Stage 1: Feature Alignment**
  - Trained on 595k image-text pairs (filtered from CC3M).
  - Freezes visual encoder and language model.
  - Optimizes projection matrix to align visual and textual embeddings.
- **Stage 2: Fine-tuning**
  - Trained on 158k multimodal instruction data.
  - Updates projection layer and language model weights.
- Training on $8\times$ A100 GPUs, learning rates of 2e-3 (pre-training) and 2e-5 (fine-tuning).

LLaVA's Key Features

- End-to-end training enables rich multimodal interactions.
- Instruction-tuned for tasks like:
  - Visual question answering.
  - Visual chat with reasoning and detailed descriptions.
- Leverages high-quality data generated by GPT-4.

LLaVA's Evaluation and Performance

Visual input example, Extreme Ironing:

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [36]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

LLaVA's Evaluation and Performance

- Evaluated on:
  - Multimodal chatbot tasks.
  - ScienceQA dataset: New state-of-the-art accuracy (92.53

Challenging examples from LLaVA-Bench (In-the-Wild):		
	 ICHIRAN Ramen [source]	 Filled fridge [source]
Annotation	A close-up photo of a meal at ICHIRAN. The chashu ramen bowl with a spoon is placed in the center. The ramen is seasoned with chili sauce, chopped scallions, and served with two pieces of chashu. Chopsticks are placed to the right of the bowl, still in their paper wrap, not yet opened. The ramen is also served with nori on the left. On top, from left to right, the following sides are served: a bowl of orange spice (possibly garlic sauce), a plate of smoke-flavored stewed pork with chopped scallions, and a cup of matcha green tea.	An open refrigerator filled with a variety of food items. In the left part of the compartment, towards the front, there is a plastic box of strawberries with a small bag of baby carrots on top. Towards the back, there is a stack of sauce containers. In the middle part of the compartment, towards the front, there is a green plastic box, and there is an unidentified plastic bag placed on it. Towards the back, there is a carton of milk. In the right part of the compartment, towards the front, there is a box of blueberries with three yogurts stacked on top. The large bottle of yogurt is Fage non-fat yogurt, and one of the smaller cups is Fage blueberry yogurt. The brand and flavor of the other smaller cup are unknown. Towards the back, there is a container with an unknown content.
Question 1	What's the name of the restaurant?	What is the brand of the blueberry-flavored yogurt?
Question 2	Describe this photo in detail.	Is there strawberry-flavored yogurt in the fridge?

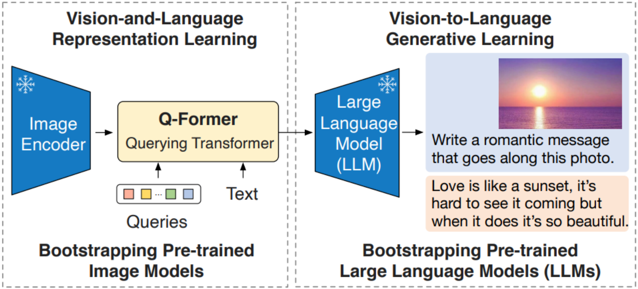
LLaVA's Applications

- Visual tutoring and education tools.
- Human-AI collaboration for complex problem-solving.
- Context-aware visual chatbots for customer support.
- Multimodal research and creative tools.

Introduction to BLIP-2

- BLIP-2 introduces a new strategy for vision-language pre-training.
- **Goal**: Efficiently bridge the gap between frozen image encoders and large language models (LLMs).
- **Key Idea**: Use a lightweight Querying Transformer (Q-Former) to enable interaction between modalities.
- Advantages:
  - Reduced computation cost.
  - Strong zero-shot performance.
  - Fewer trainable parameters than state-of-the-art methods.

Architecture Overview

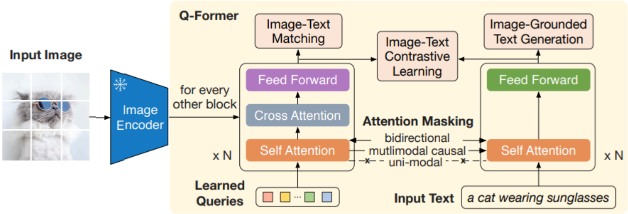


- **Components:**
  - **Image Encoder:** Frozen Vision Transformer (ViT).
  - **Q-Former:** Lightweight transformer with learnable queries.
  - **LLM:** Frozen language model (e.g., FlanT5, OPT).

Q-Former architecture

Q-Former consists of two transformer submodules that share the same self-attention layers:

- an image transformer that interacts with the frozen image encoder
- a text transformer that can function as both a text encoder and a text decoder.
- QFormer is initialized with the pre-trained weights of BERTbase



Two-Stage Pre-Training

- *Stage 1: Vision-Language Representation Learning*
  - Align image and text features using image-text contrastive (ITC) and matching (ITM) objectives.
  - Train Q-Former to extract text-relevant image features.
- *Stage 2: Vision-to-Language Generative Learning*
  - Connect Q-Former to a frozen LLM.
  - Train for tasks like image captioning and visual question answering (VQA).
- *Result:* Effective modality alignment with frozen models.

Differences Between BLIP and BLIP-2

- **Model Efficiency:**
  - BLIP requires end-to-end training; BLIP-2 uses frozen encoders.
- **Architecture:**
  - BLIP uses cross-attention layers for alignment.
  - BLIP-2 introduces Q-Former as an intermediary.
- **Training Objectives:**
  - BLIP uses image-text matching and captioning losses.
  - BLIP-2 adds generative training with frozen LLMs.
- **Parameter Efficiency:**
  - BLIP-2 achieves better performance with fewer trainable parameters.

Advantages of BLIP-2

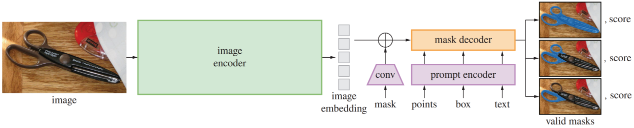
- **Compute Efficiency:**
  - No fine-tuning of image encoder or LLM.
- **Performance:**
  - Outperforms models like Flamingo with fewer parameters.
- **Zero-Shot Capabilities:**
  - Excels in tasks like VQA, captioning, and image-text retrieval.
- **Generality:**
  - Can integrate newer unimodal models for better results.

Applications of BLIP-2

- **Visual Question Answering (VQA):**
  - Answers questions based on image context.
- **Image Captioning:**
  - Generates textual descriptions for images.
- **Image-Text Retrieval:**
  - Finds images or text based on queries.
- **Zero-Shot Tasks:**
  - Handles novel tasks without additional training.

Introduction to SAM

- SAM is a foundation model for image segmentation.
- Designed for **promptable segmentation**, enabling zero-shot generalization.
- Components:
  - Image Encoder
  - Prompt Encoder
  - Mask Decoder
- Real-time segmentation with interactive use cases.



Architecture Overview

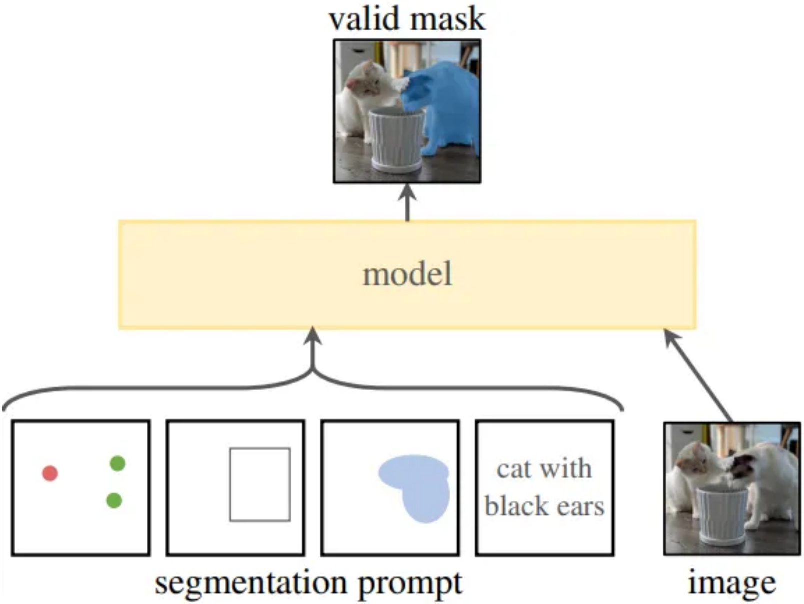
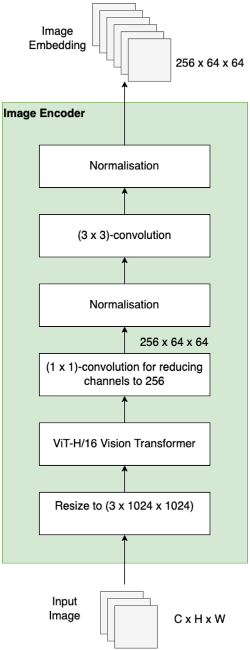


Image Encoder

- Uses a Vision Transformer (ViT) pre-trained with Masked Autoencoder (MAE).
- Processes high-resolution images to produce a compact embedding.
- Operates once per image and enables reuse for multiple prompts.

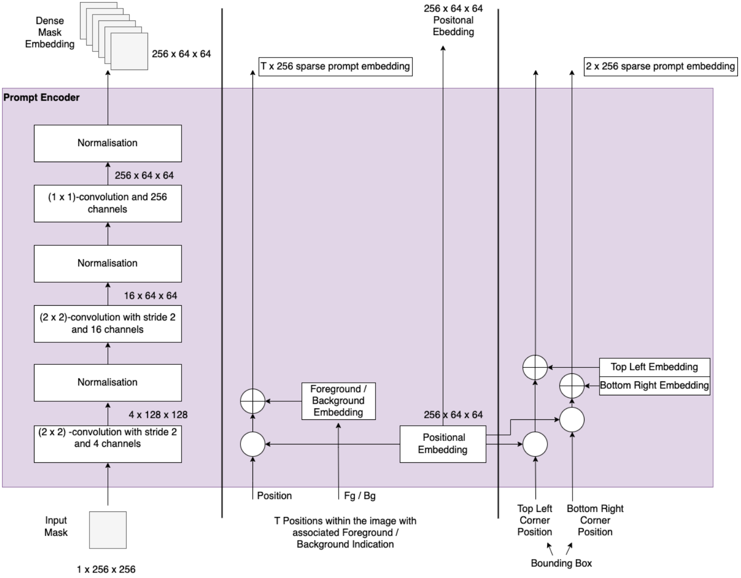
Image Encoder



Prompt Encoder

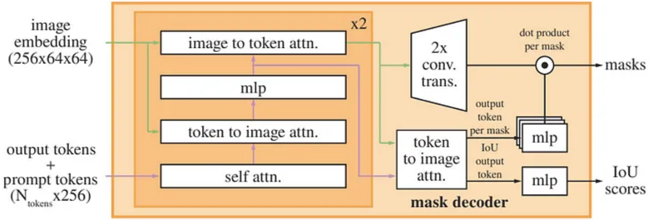
- Supports **sparse prompts** (points, boxes, and text) and **dense prompts** (masks).
- Sparse prompts:
  - Points and boxes use positional encodings with learned embeddings.
  - Text uses CLIP's text encoder.
- Dense prompts:
  - Masks are embedded using convolutional layers.
  - Combined with image embeddings via element-wise addition.

Prompt Encoder



Mask Decoder

- Maps embeddings to segmentation masks efficiently.
- Uses a modified Transformer decoder:
  - **Prompt Self-Attention:** Updates prompt embeddings.
  - **Cross-Attention:** Interacts between image and prompt embeddings.
- Outputs masks using a dynamic mask prediction head.



Key Features and Workflow

- Efficient: Mask decoding in 50ms for real-time interactive use.
- Ambiguity-aware: Predicts multiple masks for ambiguous prompts.
- Generalization: Supports a wide range of segmentation tasks using prompts.

Conclusion

- CLIP is highly effective for zero-shot understanding tasks due to its robust vision-language alignment.
- Models like BLIP, BLIP-2, and LLAVA excel at bridging understanding and generation, offering versatile capabilities across diverse tasks.
- SAM demonstrates adaptability, performing well across a wide range of segmentation and annotation tasks.

Remember that these models vary in size, and the most suitable choice depends on the application, not just model weight. Quantized versions can provide a lightweight yet powerful alternative for resource-constrained scenarios.

Bibliography:

1 Pan, F., Shin, I., Rameau, F., Lee, S., & Kweon, I. S. (2020). Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3764-3773).

2 Li, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." International conference on machine learning. PMLR, 2022.International Conference on Computer Vision. 2021.

3 Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." International conference on machine learning. PMLR, 2023.

4 Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

Bibliography:

5 Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2024).

6 Kirillov, Alexander, et al. "Segment anything." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.