

Topics in NLP (CS6803)

Show and Tell: Vision Language Models (VLMs)

Dr. Maunendra Sankar Desarkar

What are Vision-Language Models (VLMs)?

- VLMs aim to understand and interpret visual data (images, videos) along with natural language.
- Early models focussed on connecting visual perception with text-based tasks (e.g, image captioning, visual question answering, video captioning, image-text retrieval etc).
- Modern VLM architectures are transformer-based models for image and text processing because they efficiently capture long-range dependencies.

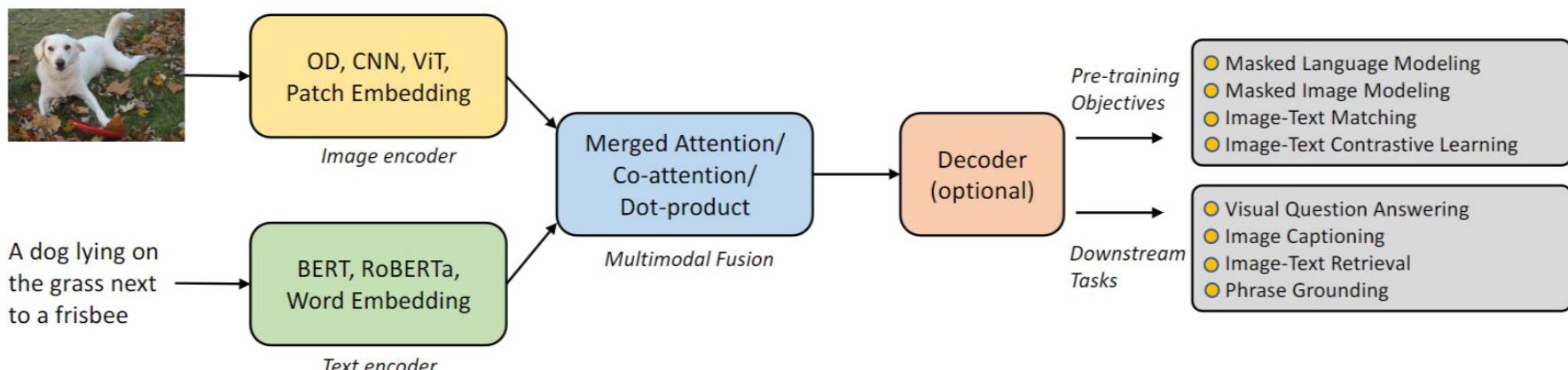


Fig: Illustration of a general framework for Transformer-based vision-language models.

Types of VL Tasks

VQA & Visual Reasoning

Q: What is the dog holding with its paws?
A: Frisbee.

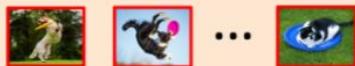
Image Captioning

Caption: A dog is lying on the grass next to a frisbee.

Text-to-Image Retrieval

Query: A dog is lying on the grass next to a frisbee.

Negative Images



Text-to-Video Retrieval

Query: A dog is lying on the grass next to a frisbee, *while shaking its tail*.

Negative Videos



Video Question Answering

Q: Is the dog perfectly still?
A: No.

Image Captioning

Caption: A dog is lying on the grass next to a frisbee.

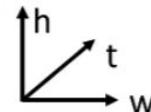


Image Classification

Labels: [dog, grass, frisbee]

Object Detection



dog, grass, frisbee

Segmentation



dog, grass, frisbee

Video Captioning

Caption: A dog is lying on the grass next to a frisbee, *while shaking its tail*.

Figure 1.2: Illustration of representative tasks from three categories of VL problems covered in this paper: **image-text tasks**, **vision tasks as VL problems**, and **video-text tasks**.

Evolution of VL models

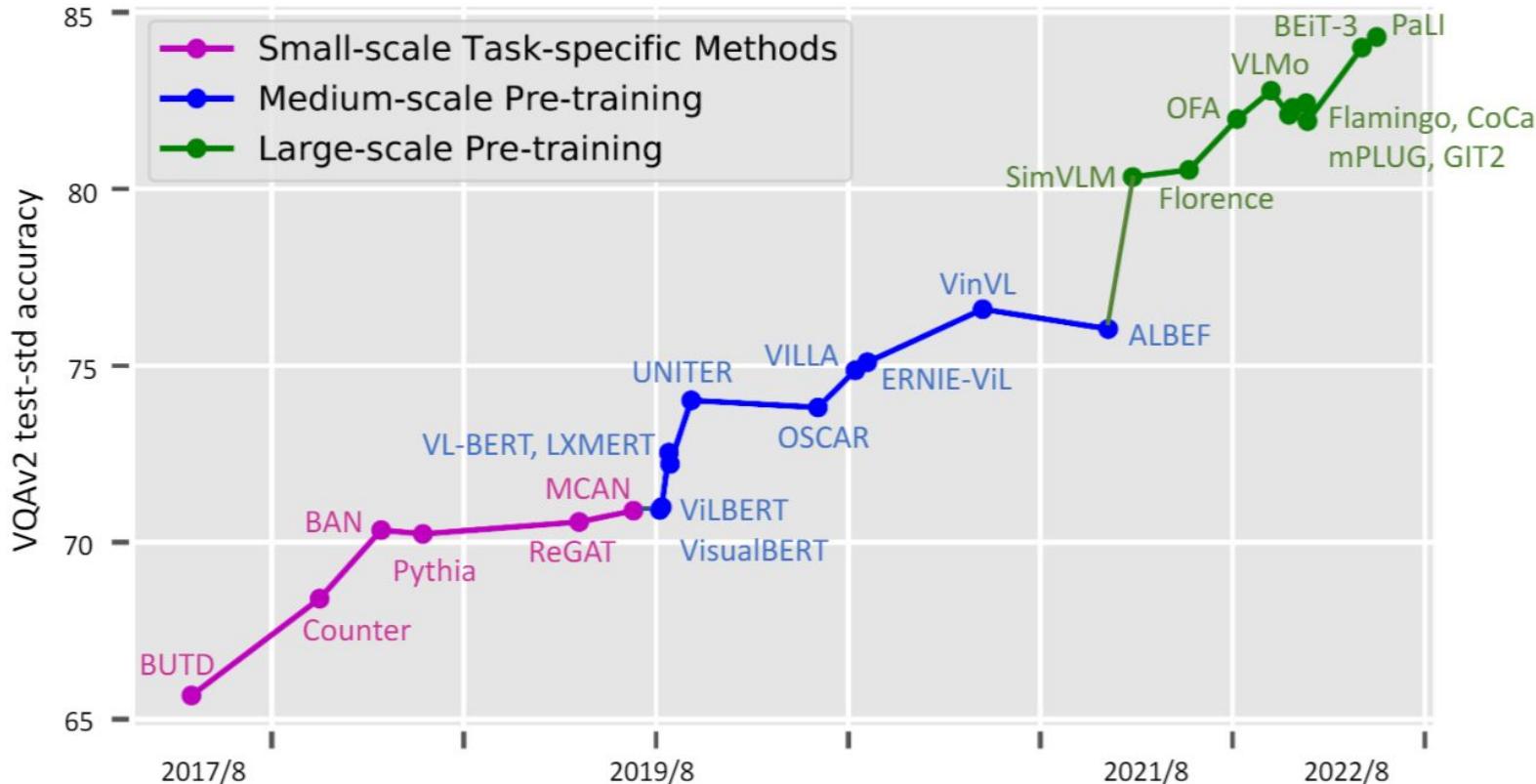


Fig: The transition from task-specific methods to large-scale pre-training, using the VQA task as a case study

Modeling Vision–Language Interaction

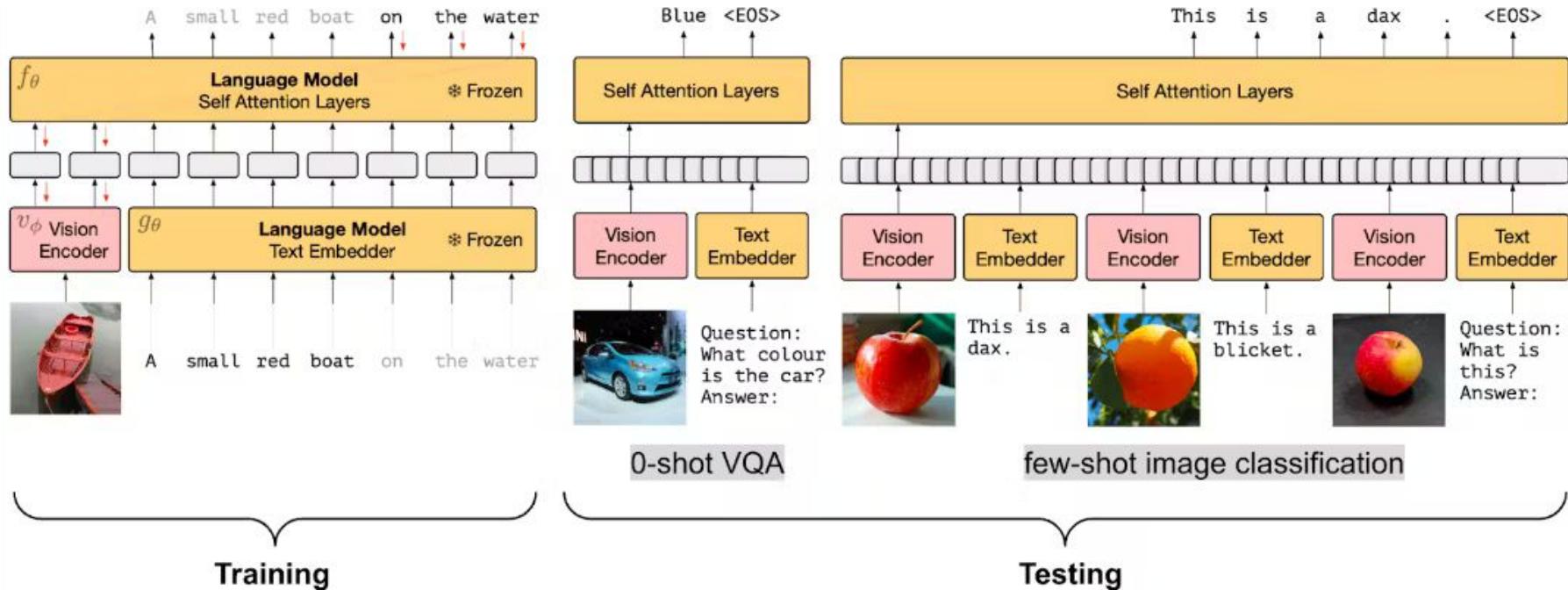
- **Fusion encoders:** directly combine image and text embeddings
- **Dual encoders:** process them separately before interaction
- **Hybrid methods** that leverage both strengths.

Learning Strategies and Pre-training Objectives

- **Contrastive Learning:** Aligning images and texts to a joint feature space by image-text matching.
- **PrefixLM:** Jointly learning image and text embeddings by using images as a prefix to a language model.
- **Multi-modal Fusing with Cross Attention:** Fusing visual information into layers of a language model with a cross-attention mechanism.
- **MLM / ITM:** Aligning parts of images with text with masked-language modeling and image-text matching objectives.
- **No Training:** Using stand-alone vision and language models via iterative optimization.

Note: There are various other approaches, as well as hybrid strategies such as [Unified-IO](#).

PrefixLM



For Details on VL pre-training and models

1. Gan, Zhe, et al. "Vision-language pre-training: Basics, recent advances, and future trends." Foundations and Trends® in Computer Graphics and Vision 14.3–4 (2022). [link](#)
2. Du, Yifan, et al. "A survey of vision-language pre-trained models." arXiv preprint arXiv:2202.10936 (2022). [Link](#)
3. HuggingFace [blog](#)

Connecting Text and Images

CLIP: Contrastive Language Image Pre-training

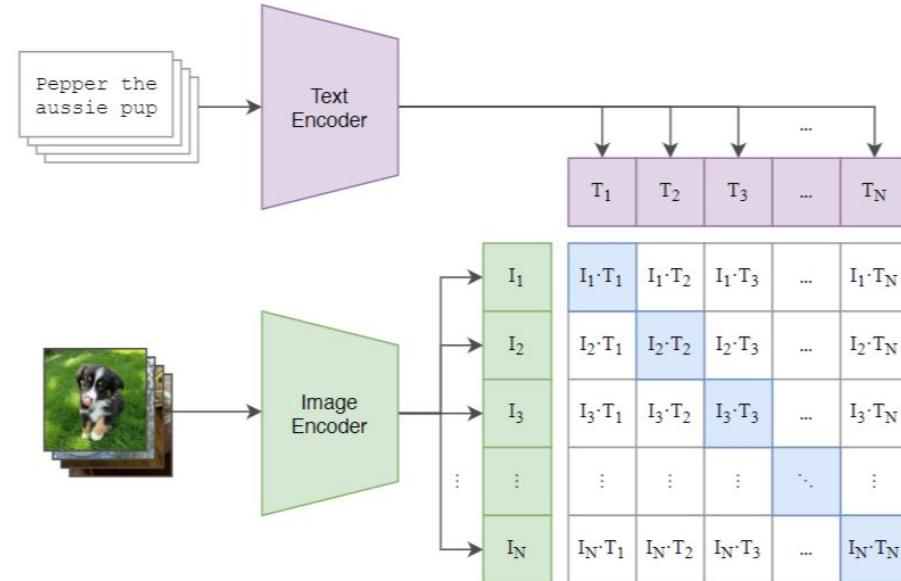
OpenAI

What is CLIP?

- CLIP → Contrastive Language Image Pre-training
- Vision-Language model pre-trained using (image, text) pairs for the task of predicting which caption goes with which image.
- Learn image representations from scratch on a dataset of 400M (image, text) pairs collected from the internet.
- Enables zero-shot transfer of the model to downstream tasks.
- Tested on 30+ CV datasets on tasks including OCR, action recognition in videos, object classification, etc.
- Zero-shot CLIP is often competitive with a fully supervised baseline.
 - E.g, Accuracy of 0-shot CLIP = original ResNet50 on ImageNet 1.28M training set

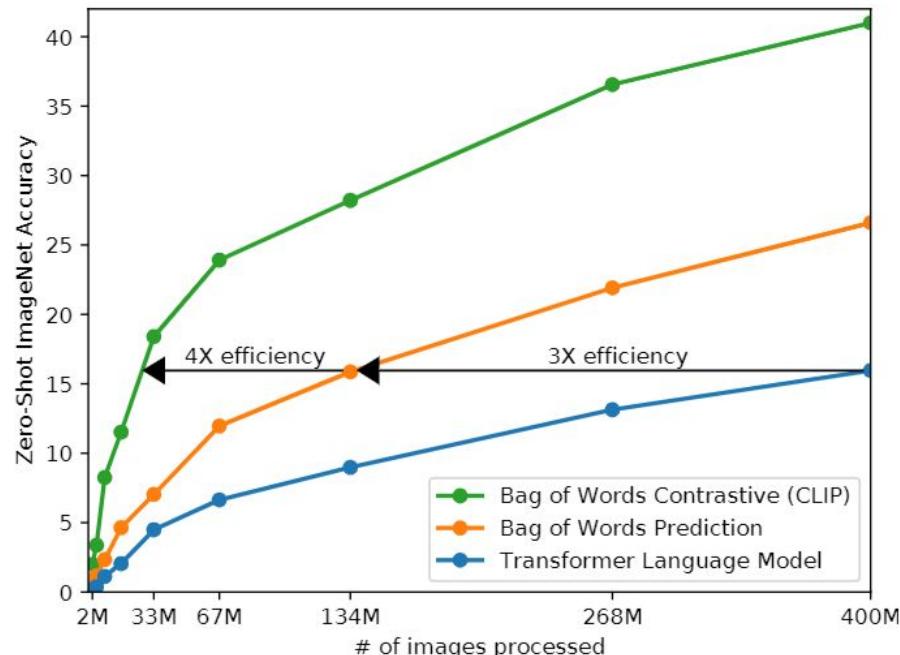
Contrastive Learning

- Standard image models jointly train an image feature extractor and a linear classifier to predict some label.
- Given a batch of N (image, text) pairs, CLIP is trained to predict which of the $N \times N$ possible (image, text) pairings across a batch actually occurred.
- CLIP learns a multi-modal embedding space by jointly training an image encoder and a text encoder to maximize the cosine similarity of the image and text embeddings of the N real pairs in the batch while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairings.
- CLIP optimizes a symmetric cross entropy loss over these similarity scores.



Why Contrastive pre-training?

- CLIP is more efficient at zero-shot transfer than their image caption baseline.
- Transformer-based caption generation LMs are relatively weak at zero-shot ImageNet classification.
- It learns 3x slower than a baseline which predicts a bag-of-words (BoW) text encoding.
- Swapping the prediction objective for the contrastive objective of CLIP improves efficiency by another 4x.



Pseudocode for CLIP implementation

```
# image_encoder - ResNet or Vision Transformer
# text_encoder - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]         - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t              - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Dataset for contrastive pre-training

- Popular datasets: MS-COCO, Visual Genome, YFCC100M.
 - Small
 - YFCC100M is **noisy** → After filtering, it shrunk by a factor of 6 i.e., 15M only
- The authors create **WebImageText (WIT)** dataset:
 - 400 million (image, text) pairs collected from the Internet
 - Text includes one of a set of 500,000 queries.
 - Approximate class balanced by including up to 20,000 (image, text) pairs per query

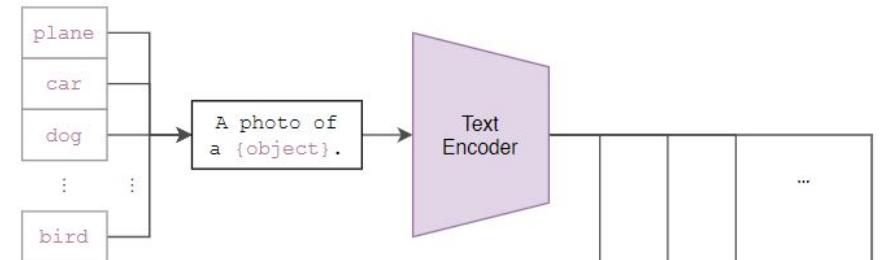
Architecture

- 5 ResNets:
 - ResNet-50, ResNet-101
 - 3 more EfficientNet-style model scaling and use ~ 4x, 16x, and 64x the compute of ResNet-50, denoted as RN50x4, RN50x16, and RN50x64.
- 3 Vision Transformers (ViT): ViT-B/32, ViT-B/16, and ViT-L/14
- Text encoder: Transformer architecture of GPT2: 12 layers, embed_dim = 512, 8 attention heads.
- Tokenizer: lower-cased byte pair encoding (BPE)
- The largest ResNet model – RN50x64, took 18 days to train on 592 V100 GPUs.
- The largest Vision Transformer took 12 days on 256 V100 GPUs.
- ViT-L/14 is also pre-trained at a higher 336 pixel resolution for one additional epoch to boost performance, denoted as ViT-L/14@336px → best performing model
- Image and text encoders are trained from scratch, without initializing from any pre-trained weights.

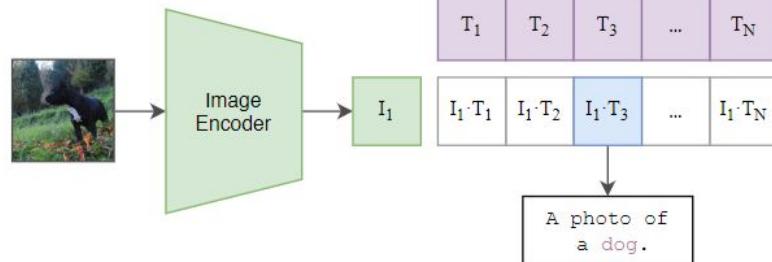
CLIP for Inference

- For each dataset, all the class names are used as the set of potential text pairings and the most probable (image, text) pair is predicted.
- Class names are used as captions. For example,
 - “A photo of a {label}.”
 - “A photo of a {label}, a type of pet.”
 - “a satellite photo of a {label}.”.
 - “A photo of a big {label}”
 - “A photo of a small {label}”.

(2) Create dataset classifier from label text

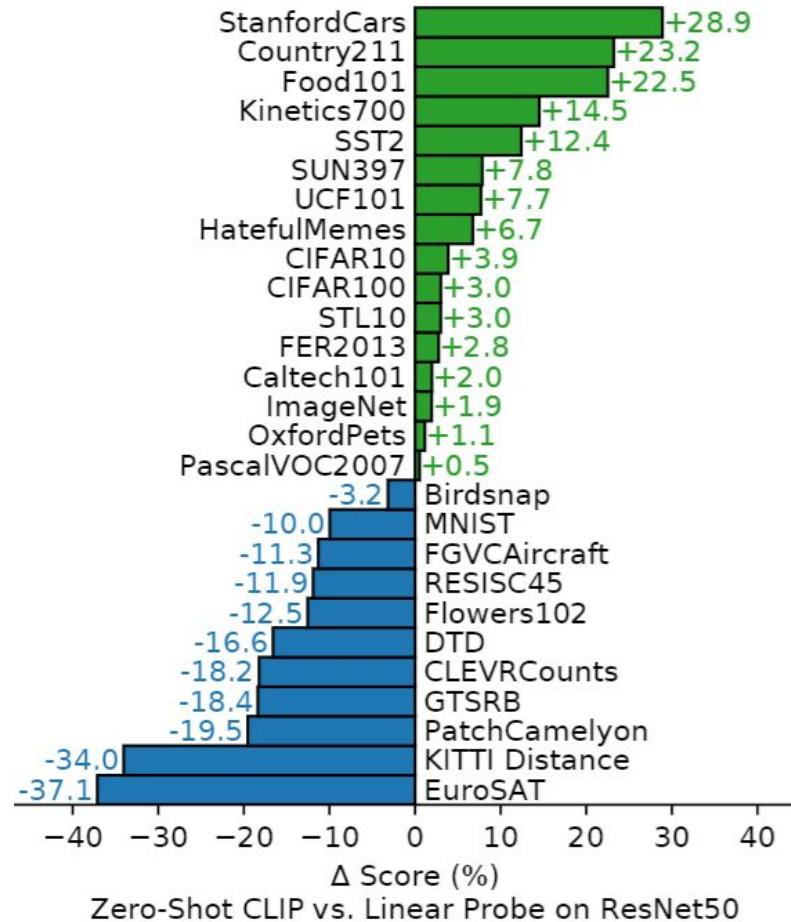


(3) Use for zero-shot prediction

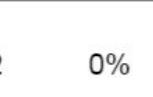
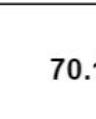
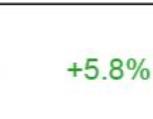
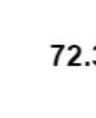
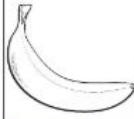
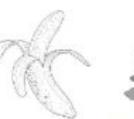
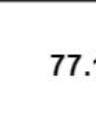


Performance on Classification

- Best CLIP model has 76.2% top-1 accuracy on ImageNet
= ResNet-50 trained on 1.28M train set.
- Top-5 accuracy is 95% and matches supervised Inception-V4.
- Zero-shot CLIP wins on 16 of the 27 datasets.
- Zero-shot CLIP is quite weak on several specialized, complex, or abstract tasks such as satellite image classification (EuroSAT and RESISC45), lymph node tumor detection (PatchCamelyon), counting objects in synthetic scenes (CLEVRCounts), self-driving related tasks such as German traffic sign recognition (GTSRB), recognizing distance to the nearest car (KITTI distance).



Zero-shot CLIP is much more robust to distribution shift

	Dataset Examples						ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet									
ImageNetV2									
ImageNet-R									
ObjectNet									
ImageNet Sketch									
ImageNet-A									

Examples

Food101

guacamole (90.1%) Ranked 1 out of 101 labels



- ✓ a photo of **guacamole**, a type of food.
- ✗ a photo of **ceviche**, a type of food.
- ✗ a photo of **edamame**, a type of food.
- ✗ a photo of **tuna tartare**, a type of food.
- ✗ a photo of **hummus**, a type of food.

SUN397

television studio (90.2%) Ranked 1 out of 397 labels



- ✓ a photo of a **television studio**.
- ✗ a photo of a **podium indoor**.
- ✗ a photo of a **conference room**.
- ✗ a photo of a **lecture room**.
- ✗ a photo of a **control room**.

Youtube-BB

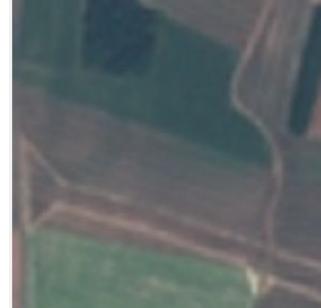
airplane, person (89.0%) Ranked 1 out of 23 labels



- ✓ a photo of a **airplane**.
- ✗ a photo of a **bird**.
- ✗ a photo of a **bear**.
- ✗ a photo of a **giraffe**.
- ✗ a photo of a **car**.

EuroSAT

annual crop land (46.5%) Ranked 4 out of 10 labels



- ✗ a centered satellite photo of **permanent crop land**.
- ✗ a centered satellite photo of **pasture land**.
- ✗ a centered satellite photo of **highway or road**.
- ✓ a centered satellite photo of **annual crop land**.
- ✗ a centered satellite photo of **brushland or shrubland**.

Limitations

- The performance of zero-shot CLIP is often just competitive with the supervised baseline of a linear classifier on ResNet-50 features. This **baseline** is well **below the overall SOTA**.
- Around 1000x increase in compute is required for zero-shot CLIP to reach overall SOTA performance across our evaluation suite. This is infeasible to train with current hardware. Improving the **computational and data efficiency** of CLIP is necessary.
- Despite stating zero-shot transfer, evaluations on validation sets were done to guide development. This is **unrealistic for true zero-shot** scenarios.
- Many complex tasks and visual concepts can be difficult to specify just through text. Actual training examples are undeniably useful but **CLIP does not optimize for few-shot** performance directly.
- **Social Bias:** CLIP is trained on text paired with images on the internet. These image-text pairs are unfiltered and uncurated.

More Details and References:

For more examples and performance comparisons of CLIP, please refer to:

CLIP's published paper: <https://proceedings.mlr.press/v139/radford21a/radford21a.pdf>

CLIP's arXived paper (additional details): <https://arxiv.org/pdf/2103.00020>

OpenAI CLIP blog: <https://openai.com/index/clip/>

For code, refer to:

<https://github.com/openai/CLIP>

https://huggingface.co/docs/transformers/model_doc/clip

Issues with early VLMs

- **Lack of Flexibility:** Early VLMs were not general-purpose; they were task-specific (e.g., image captioning).
- **Task-Specific Architectures:** Early models had rigid architectures, making them less adaptable to new tasks without significant finetuning.
- **Large scale Pretraining:** Required huge datasets and large-scale training
- **Poor Performance Across Domains:** Many VL models performed well on specific datasets but struggled with real-world data or images outside of their training set.
- **Lack of Human-Like Reasoning**

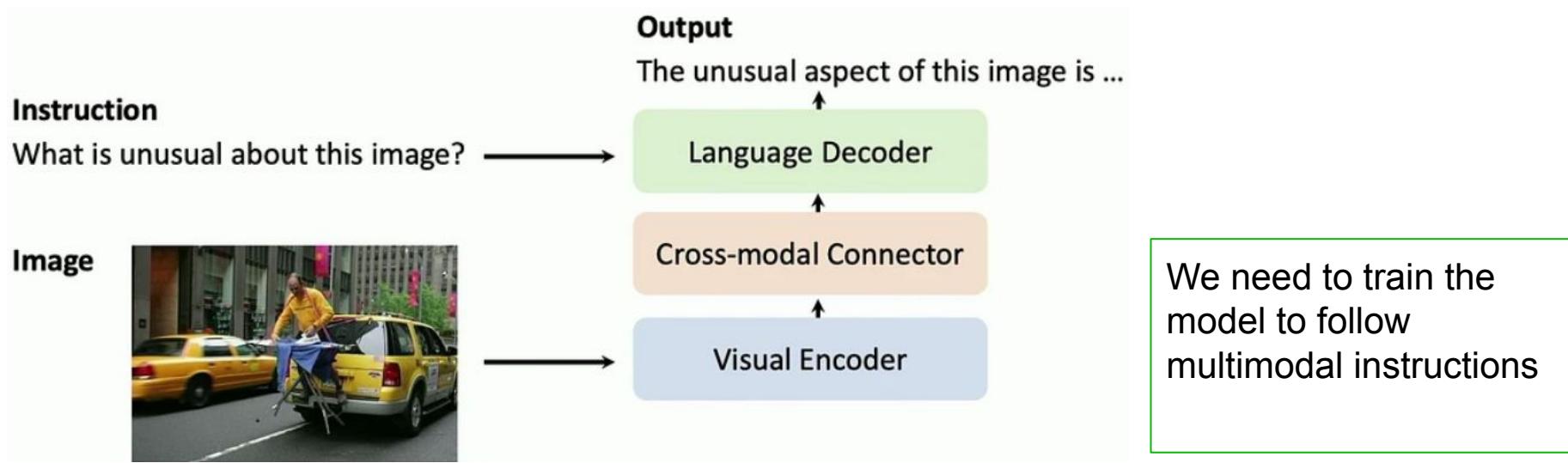
The new generation models : Instruction-tuned Multimodal Models

Instruction tuning large language models (LLMs) using machine-generated instruction-following data has been shown to improve zero-shot capabilities on new tasks, but the idea is less explored in the multimodal field.

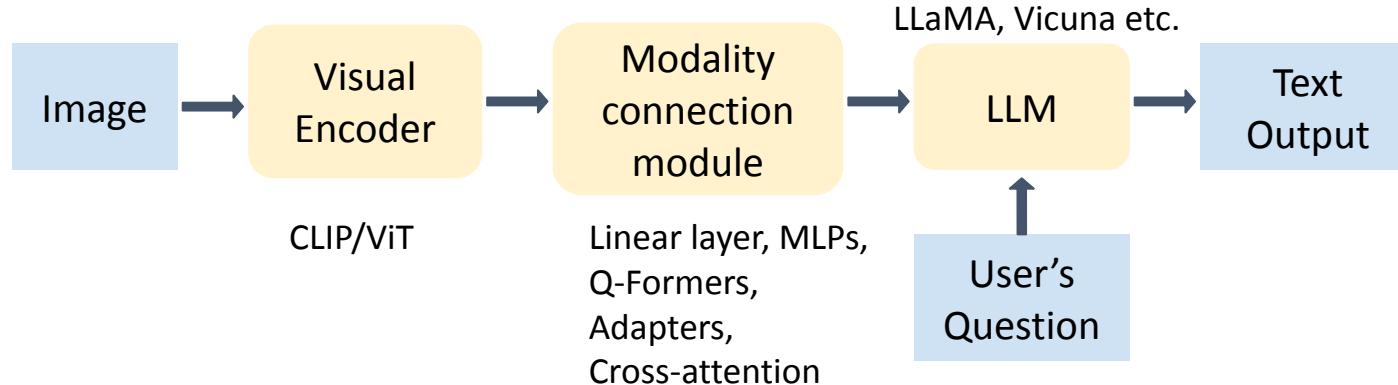
What are instruction-tuned multimodal models?

Able to follow human instructions in natural language, acts as an assistant.

How to build instruction-following models?



Large Vision Language Models (LVLMs)



Two key stages of Training:

- **Pre-training** – feature alignment/ LVLMs acquire vision-language knowledge from image-text pairs
- **Instruction-tuning** – to follow human instructions using a varied task dataset.

Can perform multiple tasks like Visual Question Answering (VQA), Visual Reasoning etc.

LLaVA
Visual Instruction Tuning

NeurIPS 2023

LLaVA



User

What are the things I should be cautious about when I visit here?



LLaVA

When visiting this peaceful location with a pier extending over a serene lake, surrounded by mountains, there are several things to be cautious about. First, pay attention to the weather conditions, as sudden changes in weather, especially strong winds or storms, can cause the water to become rough and pose risks to the stability of the pier. Second, be mindful of the structural integrity of the pier; avoid overloading it with too many people or excessive weight, as this could potentially damage the pier or cause it to collapse. Additionally, be aware of the surrounding wildlife and respect their space; do not harm or disturb them in any way. Finally, if you decide to go fishing, ensure that you adhere to any local fishing regulations and use appropriate fishing techniques to keep the lake ecosystem healthy and maintain the balance of the food chain.

Compose an engaging travel blog post about my trip to this place, highlighting the beautiful sceneries, joyful experiences, and include possible anecdotes.



LLaVA

Traveling to this picturesque location with a serene lake and mountains in the background was a refreshing and rejuvenating experience. The pier extending over the calm water provided an excellent vantage point from which to appreciate the breathtaking scenery. As I stood on the pier, I couldn't help but feel a deep connection with nature, taking a moment to soak in the peaceful atmosphere.

One of the highlights of my trip was undoubtedly the fishing experience. The area is teeming with life, and I was thrilled to land a sizeable catch, despite the modest size of the lake. The joy of fishing in such a beautiful setting was truly rewarding, and it allowed me to take a break from my busy life and truly immerse myself in the tranquility of nature.

While exploring the surrounding area, I came across a family of birds that had built a nest close to the pier, adding to the charm of the location. Interacting with the wildlife and observing their behavior was an enriching experience that I'll cherish for years to come.

In conclusion, this idyllic destination provided me with an opportunity to unwind and recharge, surrounded by mesmerizing natural beauty and the gentle sounds of water and wildlife. If you're looking for a serene escape from the hustle and bustle of everyday life, this pier overlooking the lake is the perfect place to find solace and reconnect with nature.



User



User



Do you know who drew this painting?



The painting depicts a woman, commonly believed to be Mona Lisa, the famous artwork by Leonardo da Vinci. It is a portrait painting that showcases the woman's enigmatic smile and has become one of the most famous and iconic art pieces in the world. The original work is displayed in the Louvre Museum in Paris, and it is known for its intricate details, use of oil paint, and the artist's innovative techniques that contributed to its enduring appeal and mystery.

Start a new conversation, and the history is cleared.



User



Do you know who drew this painting?



The painting depicts a dog in a humorous situation, where it appears to be wearing a woman's dress and standing in front of a landscape. Although the artist of the painting is not explicitly identified, it could be a reference to the famous artist Da Vinci, known for his artistic works, including the Mona Lisa. The painting is likely a creative and comical take on the traditional portrait style, replacing a human figure with a dog dressed in a woman's clothing, making it an interesting and amusing visual piece.

LLaVA: Large Language and Vision Assistant

LLaVA is an end-to-end trained large multimodal model that connects a vision encoder and an LLM for general purpose visual and language understanding and assistance.

Contributions:

- Multimodal instruction-following data: generated using GPT-4
- Instruction-tuned large multimodal models
- Multimodal instruction-following benchmark: LLaVA-Bench with two challenging benchmarks, with a diverse selection of paired images, instructions and detailed annotations, for inference.

LLaVA: Creating data to train the model

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.

Luggage surrounds a vehicle in an underground parking area

People try to fit all of their luggage in an SUV.

The sport utility vehicle is parked in the public garage, being packed for a trip

Some people with luggage near a van that is transporting it.

Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>



Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

“... we leverage language-only GPT-4 or ChatGPT as the strong teacher (both accept only text as input), to create instruction-following data involving visual content. Specifically, in order to encode an image into its visual features to prompt a text-only GPT, we use two types of symbolic representations: (i) **Captions** typically describe the visual scene from various perspectives; (ii) **Bounding boxes** usually localize the objects in the scene, and each box encodes the object concept and its spatial location.”

From the LLaVA paper

LLaVA: Creating data to train the model

```
messages = [ {"role": "system", "content": f"""You are an AI visual assistant, and you are seeing a single image. What you see are provided with five sentences, describing the same image you are looking at. Answer all questions as you are seeing the image.
```

Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the **object types, counting the objects, object actions, object locations, relative positions between objects**, etc. Only include questions that have definite answers:

- (1) one can see the content in the image that the question asks about and can answer confidently;
- (2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently.

Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary.""}
]

Architecture

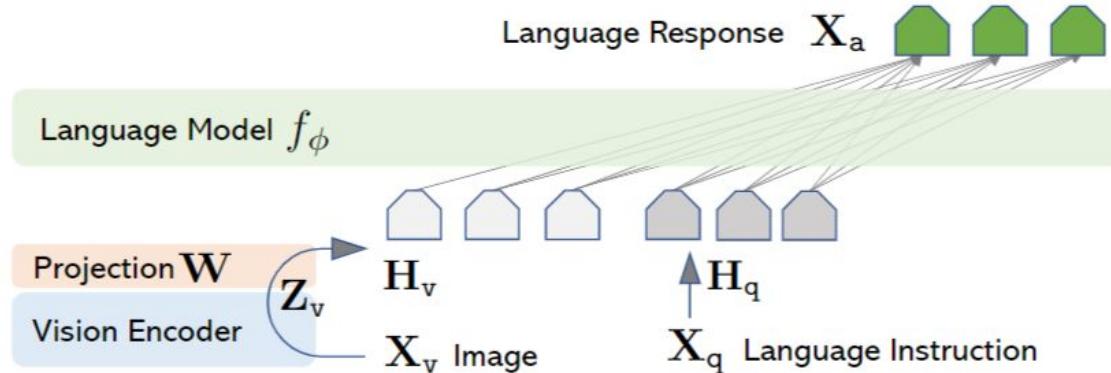


Figure 1: LLaVA network architecture.

- Image encoder: pre-trained [CLIP visual encoder ViT-L/14](#)
- LLM used: [Vicuna](#), [LLaMA](#)
- A simple [linear layer](#) is used to connect image features into the word embedding space, i.e., a trainable projection matrix W converts Z_v into language embedding tokens H_v , which have the same dimensionality as the word embedding space in the language model
- $H_v = W \cdot Z_v$, with $Z_v = g(X_v)$

Training

- For each image X_v , multi-turn conversation data is generated $(X_q^1, X_a^1, \dots, X_q^T, X_a^T)$, where T is the total number of turns.
- Organized as a sequence, by treating all answers as the assistant's response, and the instruction X_{instruct}^t at the t -th turn as:

$$X_{\text{instruct}}^t = \begin{cases} \text{Randomly choose } [X_q^1, X_v] \text{ or } [X_v, X_q^1], & \text{the first turn } t = 1 \\ X_q^t, & \text{the remaining turns } t > 1 \end{cases}$$

- The input sequence used to train the model:

```
Xsystem-message <STOP>  
Human : Xinstruct1 <STOP> Assistant: Xa1 <STOP>  
Human : Xinstruct2 <STOP> Assistant: Xa2 <STOP> ...
```

The model is trained to predict the assistant answers and where to stop (only green tokens are used to compute the loss in the autoregressive model).

Objective Function

For a sequence of length L, the probability of the target answers X_a is computed as:

$$p(\mathbf{X}_a | \mathbf{X}_v, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^L p_{\theta}(\mathbf{x}_i | \mathbf{X}_v, \mathbf{X}_{\text{instruct}, < i}, \mathbf{X}_{a, < i})$$

where θ is the trainable parameters, $\mathbf{X}_{\text{instruct}, < i}$ and $\mathbf{X}_{a, < i}$ are the instruction and answer tokens in all turns before the current prediction token x_i , respectively.

$\mathbf{X}_v \rightarrow$ the image is grounded for all answers

Two-stage instruction-tuning

→ Stage 1: Pre-training for Feature Alignment:

- ◆ Data: Conceptual Captions (CC)-3M is filtered to 595K image-text pairs.
- ◆ Converted to the instruction-following data using a naive expansion method
- ◆ Each sample can be treated as a single-turn conversation
- ◆ The Visual encoder and LLM weights are frozen
- ◆ Trainable parameters are: $\theta = W$ (the projection matrix) only
- ◆ This stage can be understood as training a compatible visual tokenizer for the frozen LLM.

- Stage 2: **Fine-tuning End-to-End:**
 - ◆ Visual encoder weights **frozen**.
 - ◆ The pre-trained weights of the **projection layer** and **LLM** are updated i.e., the trainable parameters are $\theta = \{W, \varphi\}$, where φ are the parameters of the LLM
 - ◆ Two Datasets:
 - **158K** language-image instruction-following synthetic data (proposed by the authors)
 - **ScienceQA** benchmark

Note that the visual encoder weights frozen in both stages of training.

Naive expansion of image-caption pairs to instruction-following data

- For an image X_v and its caption X_c , a set of questions X_q is created to instruct the assistant to describe the image content.
- GPT-4 is prompted to curate such a list of questions.
- So, a simple way to expand an image-text pair to its instruction-following version is:

Human : $X_q X_v <\text{STOP}>$ Assistant : $X_c <\text{STOP}>$

- This is used for Stage-1 pre-training on CC.
- **Limitation: lacks diversity and in-depth reasoning** in both the instructions and responses.

- "Describe the image concisely."
- "Provide a brief description of the given image."
- "Offer a succinct explanation of the picture presented."
- "Summarize the visual content of the image."
- "Give a short and clear explanation of the subsequent image."
- "Share a concise interpretation of the image provided."
- "Present a compact description of the photo's key features."
- "Relay a brief, clear account of the picture shown."
- "Render a clear and concise summary of the photo."
- "Write a terse but informative summary of the picture."
- "Create a compact narrative representing the image presented."

Fig: The list of instructions for brief image description

- "Describe the following image in detail"
- "Provide a detailed description of the given image"
- "Give an elaborate explanation of the image you see"
- "Share a comprehensive rundown of the presented image"
- "Offer a thorough analysis of the image"
- "Explain the various aspects of the image before you"
- "Clarify the contents of the displayed image with great detail"

Fig: Some examples of instructions for detailed image description. The complete list has more instructions.

GPT-assisted Visual Instruction Data Generation

- COCO **caption** and **bbox annotations**, and language-only GPT-4.
- Examples of conversations, detailed descriptions and complex reasoning using **few-shot prompts** is given to GPT-4.
- 158k image-language instruction following samples:
 - 58k conversations
 - 23k detailed descriptions
 - 77k complex reasoning
- This data is used for fine-tuning in Stage-2.

Context type 1: **Captions**

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.



The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: **Boxes**

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: **conversation**

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: **detailed description**

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: **complex reasoning**

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Fig: Example to illustrate the instruction-following data.

LLaVA's performance

85.1% relative score compared with GPT-4.

	Conversation	Detail description	Complex reasoning	All
Full data	83.1	75.3	96.5	85.1
Detail + Complex	81.5 (-1.6)	73.3 (-2.0)	90.8 (-5.7)	81.9 (-3.2)
Conv + 5% Detail + 10% Complex	81.0 (-2.1)	68.4 (-7.1)	91.5 (-5.0)	80.5 (-4.4)
Conversation	76.5 (-6.6)	59.8 (-16.2)	84.9 (-12.4)	73.8 (-11.3)
No Instruction Tuning	22.0 (-61.1)	24.0 (-51.3)	18.5 (-78.0)	21.5 (-63.6)

	Conversation	Detail description	Complex reasoning	All
OpenFlamingo [5]	19.3 ± 0.5	19.0 ± 0.5	19.1 ± 0.7	19.1 ± 0.4
BLIP-2 [27]	54.6 ± 1.4	29.1 ± 1.2	32.9 ± 0.7	38.1 ± 1.0
LLaVA	57.3 ± 1.9	52.5 ± 6.3	81.7 ± 1.8	67.3 ± 2.0
LLaVA [†]	58.8 ± 0.6	49.2 ± 0.8	81.4 ± 0.3	66.7 ± 0.3

Examples



In the image, a young woman is leaning on a man in front of a beautiful sunset, with both of them standing on a boat. They appear to be enjoying a romantic moment together. The scene suggests that they could be depicted as the characters Jack and Rose from the popular movie Titanic. In the movie, their romance blossoms aboard the fictional ocean liner Titanic, with the pair ultimately sharing a passionate kiss as the ship sinks into the ocean. The ending of the movie is known for its emotional and heartfelt scenes, where the characters are forced to say goodbye to each other as they are separated by the sinking ship. The image evokes a sense of romance and nostalgia, paying tribute to the iconic love story portrayed in the movie Titanic.



User

Figure 4: LLaVA relates the movie scenes to the textual knowledge from the pretrained LLM.



LLaVA

The name of the man in the photo is Elon Musk.



User

What is the name of the man in the photo?

Start a new conversation, and the history is cleared.



LLaVA

The name of the man in the photo is Elon Musk.



User

What is the name of the man in the photo?

Figure 6: An interesting emergent behavior of LLaVA is its ability to recognize Elon Musk both in a headshot and in a humorous meme where he is dressed as a doge. This implies that the pre-trained CLIP vision encoder may have seen images of Elon Musk. However, it is still surprising because Elon Musk *never* appears in the training data for either the visual feature alignment or visual instruction tuning stages of LLaVA, which indicates that the base language model generalizes to unseen visual concepts.

Visual input example, Extreme Ironing:



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

User
LLaVA

What is unusual about this image?

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
LLaVA

What's happening in the scene?

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
GPT-4 [35]

What is unusual about this image?

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
BLIP-2

What is unusual about this image?

a man is sitting on the back of a yellow cab

User
OpenFlamingo

What is unusual about this image?

The man is drying his clothes on the hood of his car.

Visual input example, Chicken Nugget Map:

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.



User
LLaVA

Can you explain this meme in detail?

The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.

User
GPT-4 [35]

Can you explain this meme?

This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

User
BLIP-2

Can you explain this meme in detail?

sometimes i just look at pictures of the earth from space and marvel how beautiful it is

User
OpenFlamingo

Can you explain this meme in detail?

It's a picture of a chicken nugget on the International Space Station.



User



What are the meals that I can cook with these?



With the variety of food items stored in the refrigerator, you can prepare several meals. For example, you can create a fruit salad using the strawberries, blueberries, and carrots. Additionally, you can make a delicious smoothie using the fruits and vegetables, which would be a healthy and nutritious snack or breakfast option. The milk and eggs can be used to create various dishes, such as a milk-based smoothie, a baked casserole, or scrambled eggs. The juice can also be used to make a fresh fruit and juice mix or as a refreshing homemade popsicle. The possibilities are vast, and the meals you can create will depend on your preferences and dietary needs.

I would like to do the fruit salad. Show me the recipe.



User



Certainly! Here's a simple fruit salad recipe using the items in your refrigerator:

Ingredients:

- 2 cups fresh strawberries, sliced
- 1 cup fresh blueberries
- 1 large carrot, grated
- Juice of half a lemon
- A handful of chopped fresh parsley or mint
- A few ice cubes
- Optional: honey or maple syrup (adjust to taste)

Instructions:

1. In a large bowl, combine the sliced strawberries, blueberries, and carrot. Gently toss them together to mix.
2. Drizzle the juice of half a lemon over the fruit and vegetables to prevent them from oxidizing and add a refreshing citrus flavor.
3. Sprinkle the chopped fresh parsley or mint over the fruit and gently toss once more.
4. If desired, add a drizzle of honey or maple syrup for a touch of sweetness. Adjust the amount according to your taste preferences.
5. Cover the bowl and refrigerate the fruit salad for at least 30 minutes or up to a few hours. This allows the flavors to meld and creates a more refreshing salad.
6. Serve the fruit salad in cups or bowls, garnished with a sprinkle of fresh parsley or mint, if desired.

This fruit salad recipe is perfect for a healthy snack or side dish, and with the variety of ingredients available in your refrigerator, it should be easy to put together quickly.

Challenging examples from LLaVA's mistakes

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?

Limitations

- **Hallucination:** Similar to LLMs, LLaVA might generate outputs that aren't grounded in facts or input data. Cannot be used in critical applications (e.g., medical).
- **Biases:** Bias can be transferred from the base models to LLaVA, both from the vision encoder (CLIP) and the language decoder (LLaMA/Vicuna).
- **Energy consumption:** may become a concern when scaling up the pretraining dataset or increasing the model size, e.g., to a larger LLaMA version like the 65B model.
- **Evaluation:** Challenging as it involves both language and vision.
 - Degree of visual content hallucination and fine-grained understanding of visual content is not considered in evaluation.
 - Text-only GPT-4 based multimodal evaluation is used: Its robustness and reliability in different situations and capability to evaluate other unexplored aspects are doubtful.

More Details and References

Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* (2023).

LLaVA [Paper](#)

Code: <https://github.com/haotian-liu/LLaVA>

HuggingFace: https://huggingface.co/docs/transformers/en/model_doc/llava

Dataset: <https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K>

LLaVA website: <https://llava-vl.github.io/>

Paper presentation at Neurips'23: <https://nips.cc/virtual/2023/oral/73817>

InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning

NeurIPS 2023



What could have happened based on the current scene?

Based on the current scene in the image, it is possible that a hurricane or severe weather event caused significant damage to the buildings and infrastructure in the area. The person walking through the debris-covered area suggests that they are inspecting the damage caused by the disaster. The presence of palm trees indicates that the location may be in a tropical or sub-tropical region, which can be more susceptible to hurricanes and other natural disasters.



Introduce me this painting in detail.

The painting depicts a young girl with a pearl earring, also known as "The Girl with the Pearl Earring" or "Girl with a Pearl Earring". She is wearing a blue dress and has a pearl earring on her left ear. Her hair is pulled back into a bun, and she is looking directly at the viewer. This painting is a famous work of art by Dutch painter Johannes Vermeer, created in 1665. It is considered one of his masterpieces and is widely recognized for its intricate details and realistic portrayal of the young girl's facial expression.



Can you tell me about this image in detail?

The image depicts a collection of various vegetables including carrots, cucumbers, tomatoes, and nuts arranged on a stone surface. There are several jars filled with different types of ingredients, such as peanuts, cashews, sunflower seeds, and pumpkin seeds. These ingredients are likely to be part of a healthy meal or snack.

Can you give me an example of such meal?

An example of a healthy meal or snack that incorporates some of the ingredients in the image could be a salad made with mixed greens, carrots, cucumbers, tomatoes, and nuts. The salad can be dressed with a homemade vinaigrette made with olive oil, lemon juice, salt, and pepper. Additionally, the salad can be topped with pumpkin seeds or sunflower seeds for added crunch and nutritional value.

Sounds good! Show me steps of making this salad.

Sure! Here are the steps to make this healthy salad:

1. Wash and chop the mixed greens, carrots, cucumbers, and tomatoes into bite-sized pieces.
2. In a small bowl, whisk together olive oil, lemon juice, salt, and pepper to create a homemade vinaigrette dressing.
3. Place the chopped vegetables in a large bowl and toss them with the vinaigrette dressing.
4. Sprinkle pumpkin seeds or sunflower seeds on top of the salad for added crunch and nutritional value.
5. Serve the salad immediately or store it in the refrigerator until ready to enjoy.



Can you describe this image in detail?

The image depicts a person standing in front of an open doorway, looking out into space. They are wearing a jacket and appears to be contemplating the vastness of the universe beyond the doorway. In the background, there are multiple stars visible, creating a stunning celestial scene. The person's gaze is directed towards the center of the image, where a bright star can be seen shining brightly. There are also smaller stars scattered throughout the scene, adding to the overall cosmic atmosphere.

What if they jump over the doorway?

If the person were to jump over the doorway, they would be leaving the confines of their current location and venturing into the vastness of space. This could be a metaphorical representation of stepping out of one's comfort zone or embarking on a new adventure. However, it is important to note that such an action would be extremely dangerous and potentially life-threatening due to the harsh conditions of space.



What makes this dog special?

The dog in the image is wearing futuristic armor and has a glowing green eye, giving it a unique and futuristic appearance. This combination of elements makes the dog stand out from other dogs and adds an element of fantasy or sci-fi to the image. Additionally, the dog's posture and facial expression convey a sense of confidence and determination, further emphasizing its unique features.

InstructBLIP

Proposes a new vision-language instruction-tuning framework using BLIP-2 models.

Contributions:

- A comprehensive and systematic study on vision-language instruction tuning:
 - **26 datasets**: 13 held-in for instruction tuning and 13 held-out for zero-shot evaluation.
 - **11 task categories**
 - Four entire task categories for zero-shot evaluation at the task level.
- Propose **instruction-aware visual feature extraction**:
 - enables flexible and informative feature extraction according to the given instructions
→ the textual instruction is given not only to the frozen LLM, but also to the Q-Former, enabling it to extract instruction-aware visual features from the frozen image encoder.
 - Propose a balanced sampling strategy to synchronize learning progress across datasets.
- **State-of-the-art** zero-shot performance on a wide range of vision-language tasks. State-of-the-art finetuning performance when used as the model initialization on individual downstream tasks.

Task and Datasets

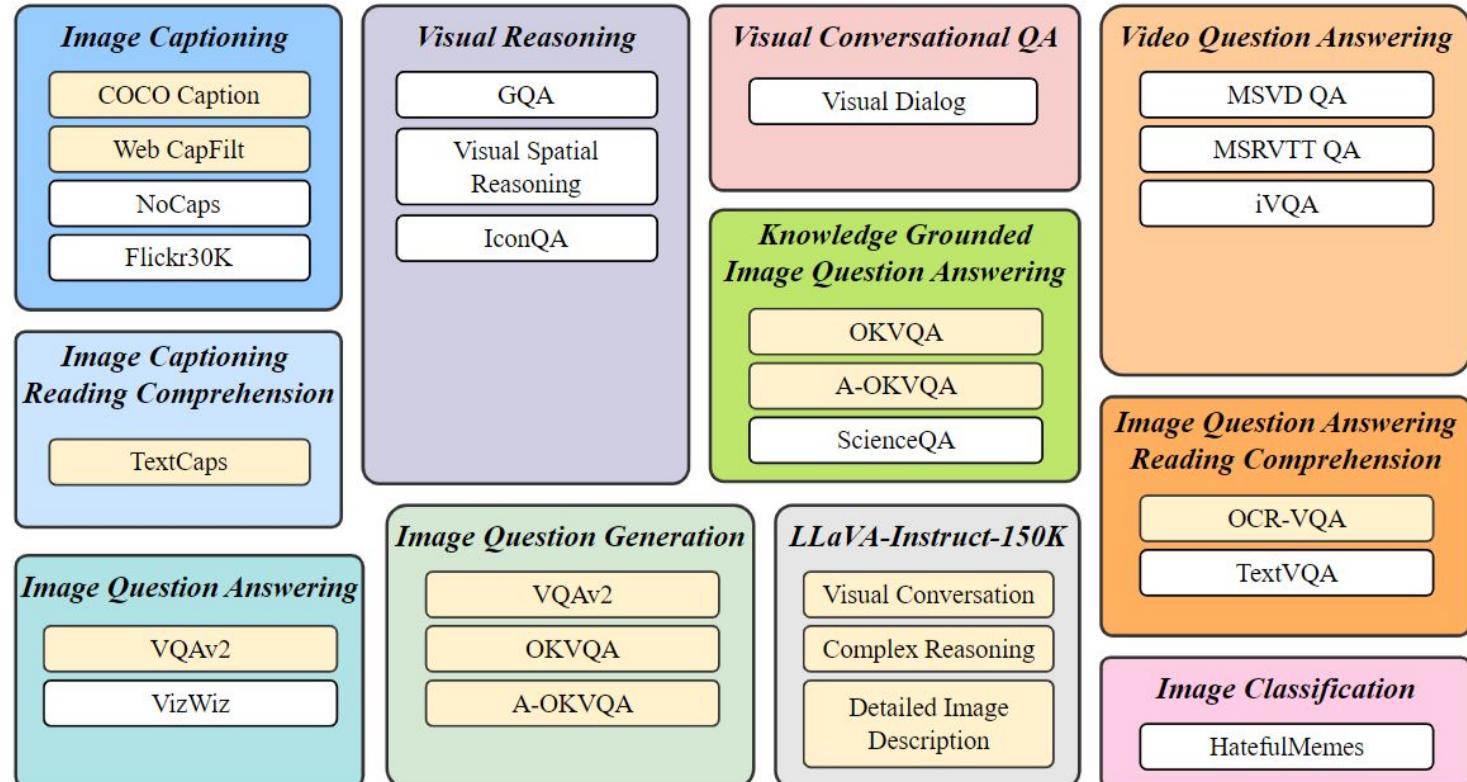


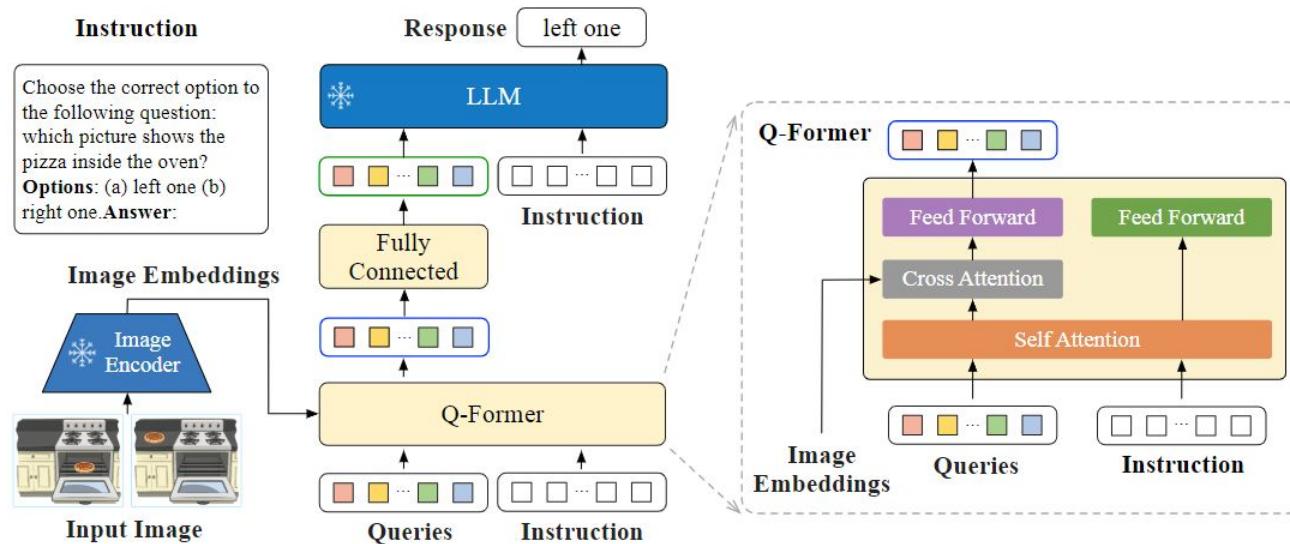
Figure 2: Tasks and their corresponding datasets used for vision-language instruction tuning. The held-in datasets are indicated by yellow and the held-out datasets by white.

Converting datasets to instruction-following format

- 10 to 15 distinct instruction templates is crafted, consisting of the task and the objective.
- For public datasets with short responses, terms like “short” and “briefly” is put into some of their corresponding instruction templates.
- For the LLaVA-Instruct-150K dataset, no change is made.

Task	Instruction Template
Image Captioning	<Image>A short image caption: <Image>A short image description: <Image>A photo of <Image>An image that shows <Image>Write a short description for the image. <Image>Write a description for the photo. <Image>Provide a description of what is presented in the photo. <Image>Briefly describe the content of the image. <Image>Can you briefly explain what you see in the image? <Image>Could you use a few words to describe what you perceive in the photo? <Image>Please provide a short depiction of the picture. <Image>Using language, provide a short account of the image. <Image>Use a few words to illustrate what is happening in the picture.
VQA	<Image>{Question} <Image>Question: {Question} <Image>{Question} A short answer to the question is <Image>Q: {Question} A: <Image>Question: {Question} Short answer: <Image>Given the image, answer the following question with no more than three words. {Question} <Image>Based on the image, respond to this question with a short answer: {Question}. Answer: <Image>Use the provided image to answer the question: {Question} Provide your answer as short as possible: <Image>What is the answer to the following question? "{Question}" <Image>The question "{Question}" can be answered using the image. A short answer is
VQG	<Image>Given the image, generate a question whose answer is: {Answer}. Question: <Image>Based on the image, provide a question with the answer: {Answer}. Question: <Image>Given the visual representation, create a question for which the answer is "{Answer}". <Image>From the image provided, craft a question that leads to the reply: {Answer}. Question: <Image>Considering the picture, come up with a question where the answer is: {Answer}. <Image>Taking the image into account, generate a question that has the answer: {Answer}. Question:

Architecture



- Like BLIP-2 model, InstructBLIP uses a **Query Transformer (Q-Former)**, to extract visual features from a frozen image encoder.
- Input to Q-Former:
 - **K learnable query embeddings**, which interact with the image encoder's output through cross-attention
 - **Instruction text tokens** → *instruction-aware* Q-former
- Output of Q-Former: **K encoded visual vectors**, one per query embedding.

Model Components and Details

- Architecture framework used: **BLIP-2**
- Image encoder (**frozen**): ViT-g/14
- LLMs used (**frozen**): FlanT5XL (3B), FlanT5-XXL (11B), Vicuna-7B and Vicuna-13B.
- FlanT5 is an instruction-tuned model based on the encoder-decoder Transformer T5.
- Vicuna is a decoder-only Transformer instruction-tuned from LLaMA.
- During vision-language instruction tuning, the model is initialized from pre-trained BLIP-2 checkpoints.
- Only the **parameters of Q-Former** are finetuned.
- **188M** trainable parameters
- No. query embeddings = 32

Training

- The Q-Former is **pre-trained in two stages** using image-caption data before instruction tuning:
 - Stage 1: pretrains the Q-Former with the frozen image encoder for vision-language representation learning.
 - Stage 2: adapts the output of Q-Former as soft visual prompts for text generation with a frozen LLM.
- After pretraining, the **Q-Former is finetuned with instruction tuning**, where the LLM receives as input the visual encodings from the Q-Former and the task instruction.
- The model is instruction-tuned with the **language modeling loss** to generate the response.

Balancing Training Datasets

Size imbalance: Due to the large number of training datasets and the significant differences in their size, mixing them uniformly could cause the model to overfit smaller datasets and underfit larger datasets.

Solution: Datasets are sampled with probabilities proportional to the square root of their sizes, or the numbers of training samples.

Given D datasets with sizes $\{S_1, S_2, \dots, S_D\}$, the probability of a data sample being selected from a dataset d during training is:

$$p_d = \frac{\sqrt{S_d}}{\sum_{i=1}^D \sqrt{S_i}}$$

Zero-shot Evaluation

	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA image	MSVD QA	MSRVTT QA	iVQA
Flamingo-3B [6]	-	60.6	-	-	-	30.1	-	53.7	28.9	-	27.5	11.0	32.7
Flamingo-9B [6]	-	61.5	-	-	-	31.8	-	57.0	28.8	-	30.2	13.7	35.2
Flamingo-80B [6]	-	67.2	-	-	-	35.0	-	46.4	31.6	-	35.6	17.4	40.7
BLIP-2 (FlanT5 _{XL}) [3]	104.5	76.1	44.0	60.5	45.5	43.1	45.7	53.0	29.8	54.9	33.7	16.2	40.4
BLIP-2 (FlanT5 _{XXL}) [3]	98.4	73.7	44.6	68.2	45.4	44.1	46.9	52.0	29.4	64.5	34.4	17.4	45.8
BLIP-2 (Vicuna-7B)	107.5	74.9	38.6	50.0	39.7	40.1	44.9	50.6	25.3	53.8	18.3	9.2	27.5
BLIP-2 (Vicuna-13B)	103.9	71.6	41.0	50.9	40.6	42.5	45.1	53.7	19.6	61.0	20.3	10.3	23.5
InstructBLIP (FlanT5 _{XL})	119.9	84.5	48.4	64.8	50.0	46.6	46.6	56.6	32.7	70.4	43.4	25.0	53.1
InstructBLIP (FlanT5 _{XXL})	120.0	83.5	47.9	65.6	51.2	46.6	48.5	54.1	30.9	70.6	44.3	25.6	53.8
InstructBLIP (Vicuna-7B)	123.1	82.4	49.2	54.3	43.1	50.1	45.2	59.6	34.5	60.5	41.8	22.1	52.2
InstructBLIP (Vicuna-13B)	121.9	82.8	49.5	52.1	44.8	50.7	45.4	57.5	33.4	63.1	41.2	24.8	51.0

Table 1: Zero-shot results on the held-out datasets. Here, Visdial, HM and SciQA denote the Visual Dialog, HatefulMemes and ScienceQA datasets, respectively. For ScienceQA, we only evaluate on the set with image context. Following previous works [6, 25, 34], we report the CIDEr score [35] for NoCaps and Flickr30K, iVQA accuracy for iVQA, AUC score for HatefulMemes, and Mean Reciprocal Rank (MRR) for Visual Dialog. For all other datasets, we report the top-1 accuracy (%).

- New 0-shot on all datasets
- Smallest InstructBLIP FlanT5XL-4B> Flamingo-80B

Ablation

Model	Held-in Avg.	GQA	ScienceQA (image-context)	IconQA	VizWiz	iVQA
InstructBLIP (FlanT5 _{XL})	94.1	48.4	70.4	50.0	32.7	53.1
w/o Instruction-aware Visual Features	89.8	45.9 (\downarrow 2.5)	63.4 (\downarrow 7.0)	45.8 (\downarrow 4.2)	25.1 (\downarrow 7.6)	47.5 (\downarrow 5.6)
w/o Data Balancing	92.6	46.8 (\downarrow 1.6)	66.0 (\downarrow 4.4)	49.9 (\downarrow 0.1)	31.8 (\downarrow 0.9)	51.1 (\downarrow 2.0)
InstructBLIP (Vicuna-7B)	100.8	49.2	60.5	43.1	34.5	52.2
w/o Instruction-aware Visual Features	98.9	48.2 (\downarrow 1.0)	55.2 (\downarrow 5.3)	41.2 (\downarrow 1.9)	32.4 (\downarrow 2.1)	36.8 (\downarrow 15.4)
w/o Data Balancing	98.8	47.8 (\downarrow 1.4)	59.4 (\downarrow 1.1)	43.5 (\uparrow 0.4)	32.3 (\downarrow 2.2)	50.3 (\downarrow 1.9)

Table 2: Results of ablation studies that remove the instruction-aware Visual Features (Section 2.3) and the balanced data sampling strategy (Section 2.4). For held-in evaluation, we compute the average score of four datasets, including COCO Caption, OKVQA, A-OKVQA, and TextCaps. For held-out evaluation, we show five datasets from different tasks.

Fine-tuning InstructBLIP on Downstream Tasks

	ScienceQA (image-context)	OCR-VQA	OKVQA	A-OKVQA			
				Direct Val	Answer Test	Multi-choice Val	Test
Previous SOTA	LLaVA [28] 89.0	GIT [39] 70.3	PaLM-E (562B) [38] 66.1	[40] 56.3	[41] 61.6	[40] 73.2	[41] 73.6
BLIP-2 (FlanT5 _{XXL})	89.5	72.7	54.7	57.6	53.7	80.2	76.2
InstructBLIP (FlanT5 _{XXL})	90.7	73.3	55.5	57.1	54.8	81.0	76.7
BLIP-2 (Vicuna-7B)	77.3	69.1	59.3	60.0	58.7	72.1	69.0
InstructBLIP (Vicuna-7B)	79.5	72.8	62.1	64.0	62.1	75.7	73.4

Table 3: Results of finetuning BLIP-2 and InstructBLIP on downstream datasets. Compared to BLIP-2, InstructBLIP provides a better weight initialization model and achieves SOTA performance on three out of four datasets.

Example

[Input Image]



[Input Instruction] What is unusual about this image?

[Model Response]

InstructBLIP 

The unusual aspect of this image is that a man is ironing clothes on the back of a yellow taxi cab while it is driving down a bus city street. He is standing on top of the car's trunk, holding an iron and using it to press his clothing items. This situation is not only unconventional but also potentially hazardous, as the man could be at risk of falling off the vehicle or causing an accident due to his actions.

GPT-4 

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.

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.

MiniGPT-4

The image shows a man standing on the back of a yellow taxi, with a yellow shirt and black pants, and a blue backpack on his back. The taxi is driving on a city street with other cars and taxis in the background.

Example showing that long responses are not always preferable

[Input Image]



[Input Instruction] Do you know who drew this painting?

[Model Response]

InstructBLIP  Leonardo da Vinci.

LLaVA

The painting depicts a woman, commonly believed to be Mona Lisa, the famous artwork by Leonardo da Vinci. It is a portrait painting that showcases the woman's enigmatic smile and has become one of the most famous and iconic art pieces in the world. The original work is displayed in the Louvre Museum in Paris, and it is known for its intricate details, use of oil paint, and the artist's innovative techniques that contributed to its enduring appeal and mystery.

MiniGPT-4

The painting is the Mona Lisa, also known as the Mona Lisa, by Leonardo da Vinci. It is one of the most famous paintings in the world and is known for its enigmatic smile and the mysterious identity of the sitter. The painting is housed in the Louvre Museum in Paris, France.

Limitations

- **Hallucination:** can inherit from the original frozen LLMs. Not recommended for critical applications without prior assessment.
- **Bias:** can be inherited too from the frozen LLMs
- InstructBLIP **may not always precisely follow given instructions**, since it is not finetuned on the parameters of the LLM.

References and more details

Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems (2023).

InstructBLIP [Paper](#)

Code: <https://github.com/salesforce/LAVIS/tree/main/projects/instructblip>

Hugging Face: https://huggingface.co/docs/transformers/en/model_doc/instructblip

More instruction-tuned LLMs

- MiniGPT4
- CogVLM
- QWEN-VL
- VILA etc

Flamingo (2022)

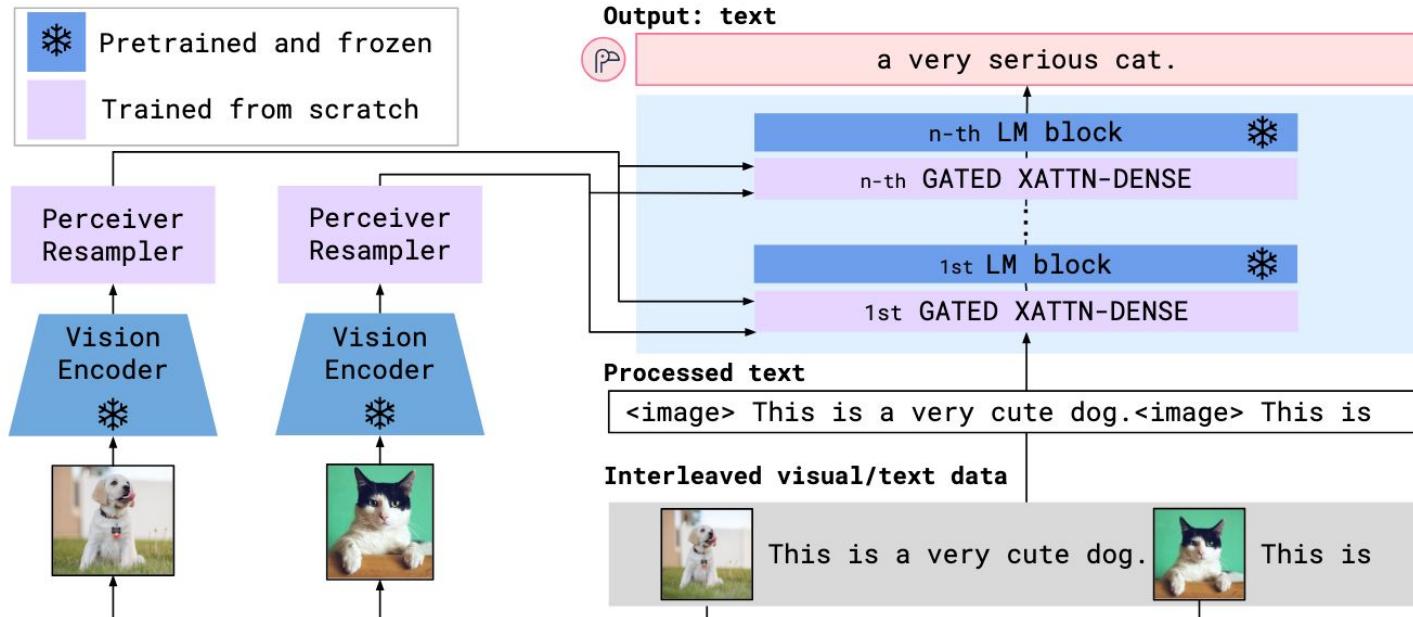


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

Cross-attention in Flamingo

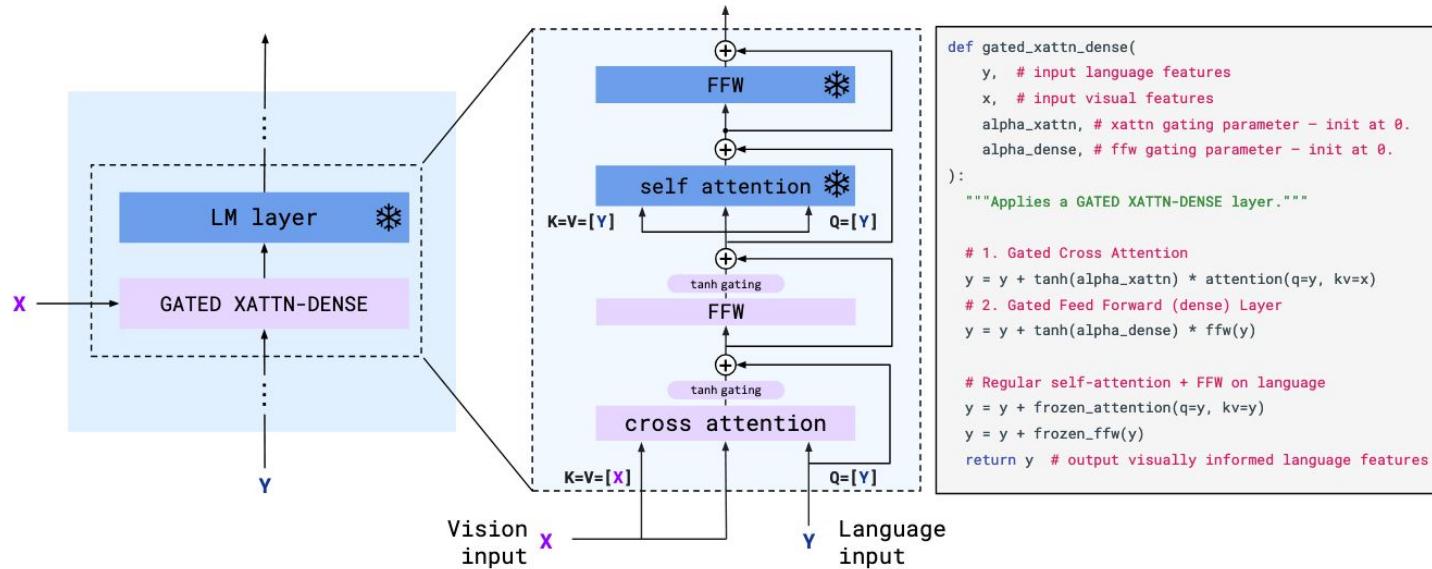


Figure 4: **GATED XATTN-DENSE layers.** To condition the LM on visual inputs, we insert new cross-attention layers between existing pretrained and frozen LM layers. The keys and values in these layers are obtained from the vision features while the queries are derived from the language inputs. They are followed by dense feed-forward layers. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.

QWEN-VL (2023)

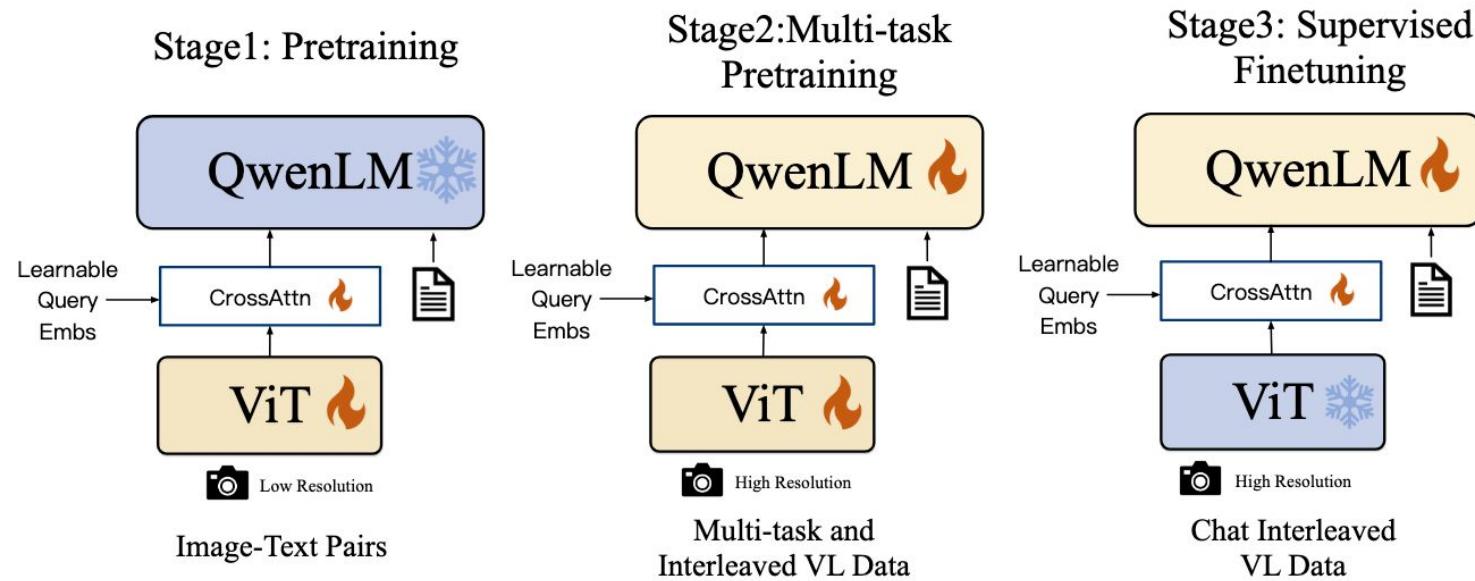


Figure 3: The training pipeline of the Qwen-VL series.

QWEN-VL Instructions

Image Captioning

cc3m/01581435.jpgGenerate the caption in English: the beautiful flowers for design.<eos>

Vision Question Answering

VG_100K_2/1.jpg Does the bandage have a different color than the wrist band?
Answer: No, both the bandage and the wrist band are white.<eos>

OCR VQA

ocr_vqa/1.jpg What is the title of this book? Answer: Asi Se Dice!, Volume 2: Workbook And Audio Activities (Glencoe Spanish) (Spanish Edition)<eos>

Caption with Grounding

coyo700m/1.jpgGenerate the caption in English with grounding: Beautiful shot of <ref>bees</ref><box>(661,612),(833,812)</box><box>(120,555),(265,770)</box> gathering nectars from <ref>an apricot flower</ref><box>(224,13),(399,313)</box><eos>

Referring Grounding

VG_100K_2/3.jpg<ref>the ear on a giraffe</ref><box>(176,106),(232,160)</box><eos>

Grounded Captioning

VG_100K_2/4.jpg<ref>This</ref><box>(360,542),(476,705)</box> is Yellow cross country ski racing gloves<eos>

OCR

synthdog/1.jpgOCR with grounding: <ref>It is managed</ref> <quad> (568,121), (625,131), (624,182), (567,172)</quad>...<eos>

The Dataset Format Example of ChatML

<im_start>user
Picture 1: vg/VG_100K_2/649.jpgWhat is the sign in the picture?<im_end>
<im_start>assistant
The sign is a road closure with an orange rhombus.<im_end>
<im_start>user
How is the weather in the picture?<im_end>
<im_start>assistant
The shape of the road closure sign is an orange rhombus.<im_end>

Thank you !!