The (R)Evolution of Multimodal Large Language Models: A Survey

Conor McCarthy, Shuai Wang

The (R)Evolution of Multimodal Large Language Models: A Survey

- Multimodal Large Language Models (MLLMs)
  - Integrate visual and textual modalities, both as input and output
  - Dialogue-based interface and instruction-following capabilities
- Review of
  - Architectural Choices
  - Multimodal Alignment Strategies
  - Training Techniques
- Analysis of
  - Visual Grounding
  - Image Generation
  - Image generation and editing
  - Visual Understanding
  - Domain-specific applications

Contents

1 Introduction

- The introduction of the attention operator and the Transformer architecture [1] has enabled the creation of models capable of handling various modalities on an increasingly large scale
  - Led to LLMs
- MLLMs:
  - Merging single modality architectures for vision and language through vision-to-language adapters
  - Innovative training approaches

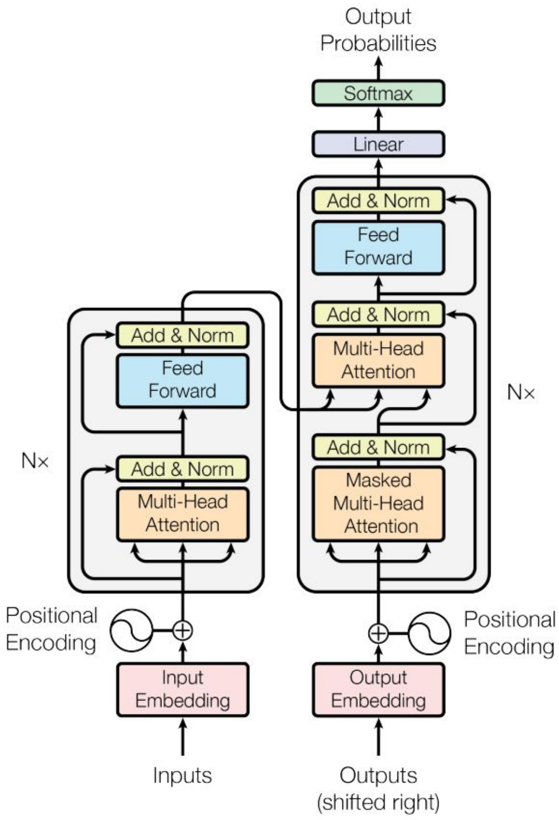


Figure 1: The Transformer - model architecture.

[1]

2 Empowering LLMs with Multimodal Capabilities

2.1 Preliminaries

- LLMs
  - In-context learning improves performance
  - Instruction-tuning: providing the LLM with the natural language description of the desired task for each training sample.

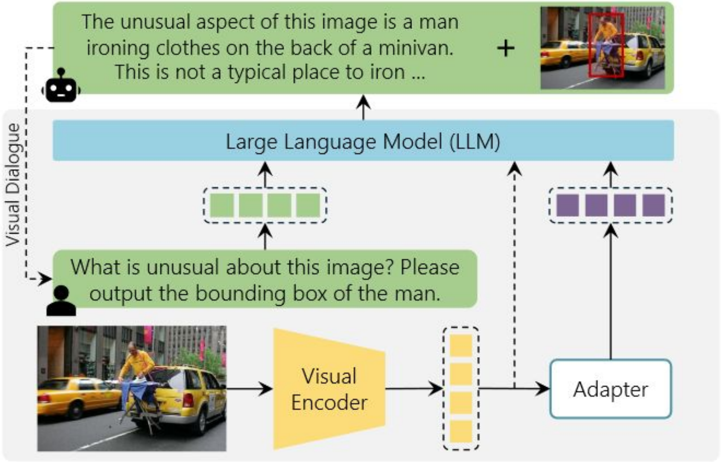


Figure 1: General architecture of Multimodal Large Language Models (MLLMs), composed of a visual encoder, a language model, and an adapter module that connects visual inputs to the textual space.

[7]

2 Empowering LLMs with Multimodal Capabilities

2.1 Preliminaries

- Parameter-Efficient Fine-Tuning (PEFT)
  - Adapt a pre-trained LLM to a specific domain/application
    - Prompt-tuning [2] - small set of vectors to be fed to the model as soft prompts before the input text
    - LoRA [3] - constrains the number of new weights by learning low-rank matrices
    - QLoRA [4] - further decreases the memory footprint of the LLM compared to the usual half-precision weights

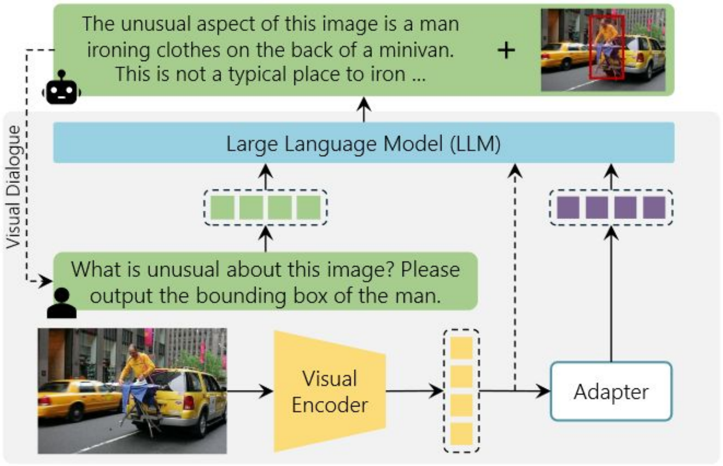


Figure 1: General architecture of Multimodal Large Language Models (MLLMs), composed of a visual encoder, a language model, and an adapter module that connects visual inputs to the textual space.

[7]

2 Empowering LLMs with Multimodal Capabilities

2.1 Preliminaries

- Towards Multimodal LLMs
  - Any MLLM contains at least three components:
    - an LLM backbone serving as an interface with the user
    - one (or more) visual encoders
    - one or more vision-to-language adapter modules
  - LLM backbone typically in LLaMA family

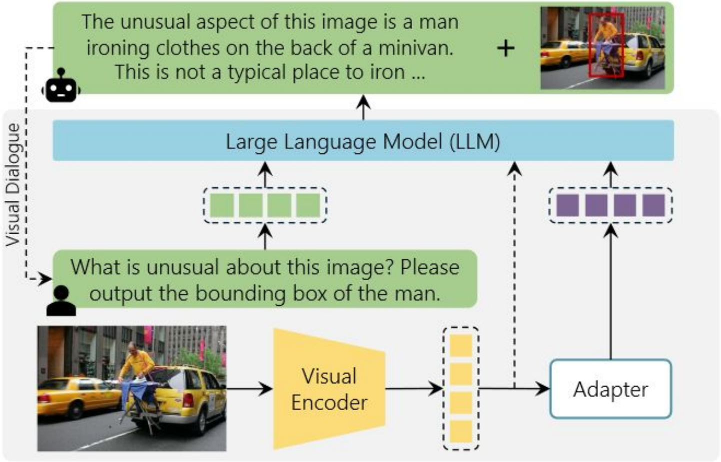


Figure 1: General architecture of Multimodal Large Language Models (MLLMs), composed of a visual encoder, a language model, and an adapter module that connects visual inputs to the textual space.

[7]

Model	LLM	Visual Encoder	V2L Adapter	VInstr. Tuning	Main Tasks & Capabilities
BLIP-2 (Li et al., 2023f)	FlanT5-XXL-11B★	EVA ViT-g	Q-Former	✗	Visual Dialogue, VQA, Captioning, Retrieval
FROMAGe (Koh et al., 2023b)	OPT-6.7B★	CLIP ViT-L	Linear	✗	Visual Dialogue, Captioning, Retrieval
Kosmos-1 (Huang et al., 2023a)	Magneto-1.3B♢	CLIP ViT-L	Q-Former*	✗	Visual Dialogue, VQA, Captioning
LLaMA-Adapter V2 (Gao et al., 2023)	LLaMA-7B▲	CLIP ViT-L	Linear	✗	VQA, Captioning
OpenFlamingo (Awadalla et al., 2023)	MPT-7B★	CLIP ViT-L	XAttn LLM	✗	VQA, Captioning
Flamingo (Alayrac et al., 2022)	Chinchilla-70B★	NFNet-F6	XAttn LLM	✗	Visual Dialogue, VQA, Captioning
PaLI (Chen et al., 2023i)	mT5-XXL-13B♦	ViT-e	XAttn LLM	✗	Multilingual, VQA, Captioning, Retrieval
PaLI-X (Chen et al., 2023g)	UL2-32B♦	ViT-22B	XAttn LLM	✗	Multilingual, VQA, Captioning
LLaVA (Liu et al., 2023e)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
MiniGPT-4 (Zhu et al., 2023a)	Vicuna-13B★	EVA ViT-g	Linear	✓	VQA, Captioning
mPLUG-Owl (Ye et al., 2023c)	LLaMA-7B▲	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA
InstructBLIP (Dai et al., 2023)	Vicuna-13B★	EVA ViT-g	Q-Former	✓	Visual Dialogue, VQA, Captioning
MultiModal-GPT (Gong et al., 2023)	LLaMA-7B▲	CLIP ViT-L	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
LaVIN (Luo et al., 2023)	LLaMA-13B▲	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
Otter (Li et al., 2023a)	LLaMA-7B★	CLIP ViT-L	XAttn LLM	✓	VQA, Captioning
Kosmos-2 (Peng et al., 2023)	Magneto-1.3B♢	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA, Captioning, Referring, REC
Shikra (Chen et al., 2023f)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Clever Flamingo (Chen et al., 2023b)	LLaMA-7B▲	CLIP ViT-L	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
SVIT (Zhao et al., 2023a)	Vicuna-13B♦	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
BLIVA (Hu et al., 2024)	Vicuna-7B★	EVA ViT-g	Q-Former+Linear	✓	Visual Dialogue, VQA, Captioning
IDEFICS (Laurençon et al., 2023)	LLaMA-65B★	OpenCLIP ViT-H	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
Qwen-VL (Bai et al., 2023b)	Qwen-7B♦	OpenCLIP ViT-bigG	Q-Former*	✓	Visual Dialogue, Multilingual, VQA, Captioning, REC
StableLLaVA (Li et al., 2023h)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
Ferret (You et al., 2023)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, Captioning, Referring, REC, GroundCap
LLaVA-1.5 (Liu et al., 2023d)	Vicuna-13B♦	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
MiniGPT-v2 (Chen et al., 2023e)	LLaMA-2-7B▲	EVA ViT-g	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Pink (Xuan et al., 2023)	Vicuna-7B▲	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
CogVLM (Wang et al., 2023c)	Vicuna-7B♦	EVA ViT-E	MLP	✓	Visual Dialogue, VQA, Captioning, REC
DRESS (Chen et al., 2023j)	Vicuna-13B▲	EVA ViT-g	Linear	✓	Visual Dialogue, VQA, Captioning
LION (Chen et al., 2023d)	FlanT5-XXL-11B★	EVA ViT-g	Q-Former+MLP	✓	Visual Dialogue, VQA, Captioning, REC
mPLUG-Owl2 (Ye et al., 2023d)	LLaMA-2-7B♦	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA, Captioning
SPHINX (Lin et al., 2023b)	LLaMA-2-13B♦	Mixture	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Honeybee (Cha et al., 2023)	Vicuna-13B♦	CLIP ViT-L	ResNet blocks	✓	Visual Dialogue, VQA, Captioning
VILA (Lin et al., 2023a)	LLaMA-2-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
SPHINX-X (Gao et al., 2024)	Mixtral-8×7B♦	Mixture	Linear	✓	Visual Dialogue, Multilingual, VQA, Captioning, Referring, REC

Table 1: Summary of generalist MLLMs for vision-to-language tasks. For each model, we indicate the LLM used in its best configuration as shown in the original paper (♢: LLM training from scratch; ♦: LLM fine-tuning; ▲: LLM fine-tuning with PEFT techniques; ★: frozen LLM). The * marker indicates variants to the reported vision-to-language adapter, while gray color indicates models not publicly available.

2.2 Visual Encoder

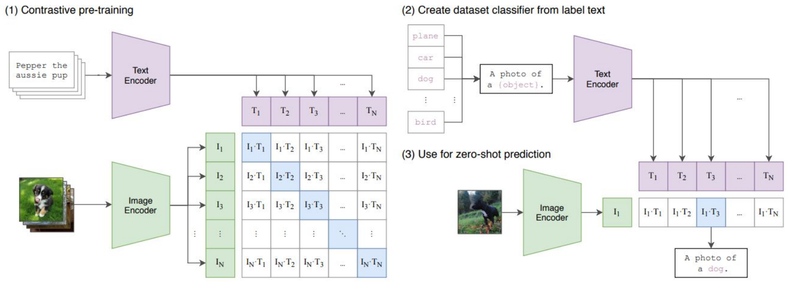


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

[5]

- Provides the LLM with the visual extracted features
  - It is common to employ a frozen pre-trained visual encoder while training only a learnable interface that connects visual features with the underlying LLM
- Common choices:
  - Pre-trained Vision Transformer(ViT) with a CLIP-based objective
  - Exploits inherent alignment of CLIP embeddings
    - ViT-L [5]
    - ViT-H
    - ViT-g

2.2 Visual Encoder

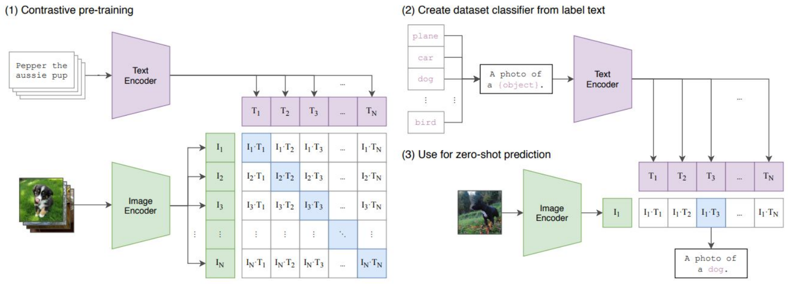


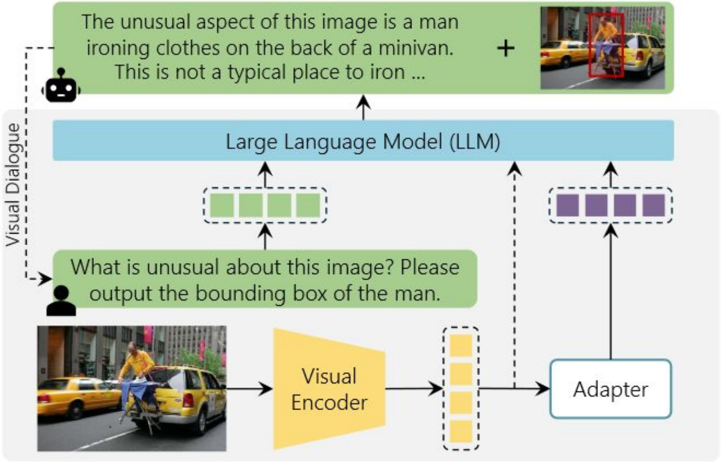
Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

[5]

- Stronger image encoder => Better Performance
- PaLI models propose scaling visual backbone to account for imbalance with LLM
- Larger Vision Encoders kept frozen during training
  - Can lead to inadequate alignment
  - May fragment the fine-grained image information and bring large computation due to the lengthy sequence when fed into the language model
- Drawbacks mitigated with two-stage training
  - First stage: LLM remains frozen
  - Better for visual Q&A or description
  - Leads to forgetting

2.3 Vision-to-Language Adapters

- Facilitate interoperability between the visual and textual domains
- Choices range in complexity

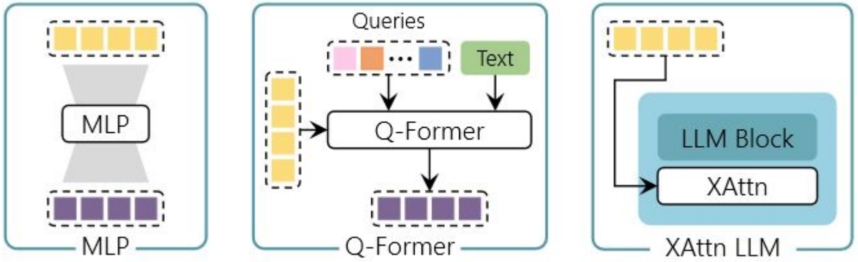


[7]

2.3 Vision-to-Language Adapters

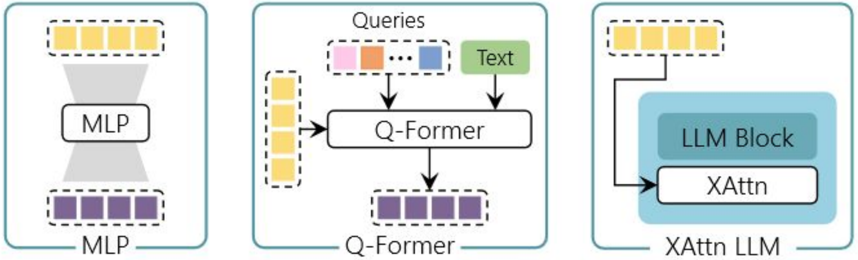
Linear and MLP Projections

- Most Straightforward
- Linear mapping, which translates visual features to the same dimensionality as the textual counterpart
- Single or two layer
- Simple yet effective
- Still effective in recent methods (more advanced understanding of visual input
- Can replace linear layers with convolutional ones



[7]

2.3 Vision-to-Language Adapters



[7]

Q-Former

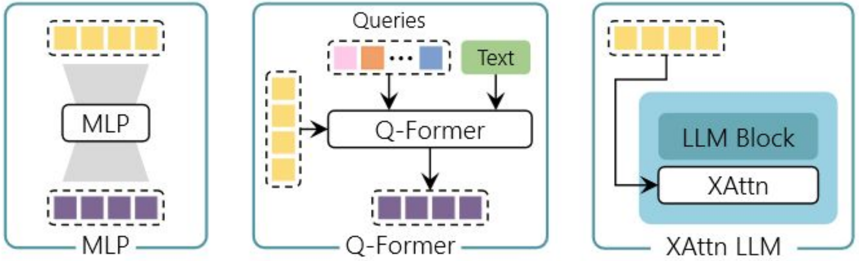
- Transformer-based model proposed in BLIP-2 [6]
- Adaptable architecture
- 2 Transformer blocks sharing mutual self-attention layers =>aligning visual and textual representations
- Learnable queries, interact with visual features via cross-attention
- Textual & Visual: shared self-attention
- Modifications have also been proposed

2.3 Vision-to-Language Adapters

Additional Cross-Attention Layers

- Integration of dense cross-attention blocks among the existing pre-trained layers of the LLM
- Additional layers must be trained from scratch
- Number of visual tokens reduced using a Perceiver based component [8]
- Shown to give enhanced training stability and improved performance

[7]



2.3 Multimodal Training

Training: Single-stage or Two-stage

- Standard cross-entropy loss is utilized for predicting the next token - serving as an auto-regressive objective

2.3 Multimodal Training

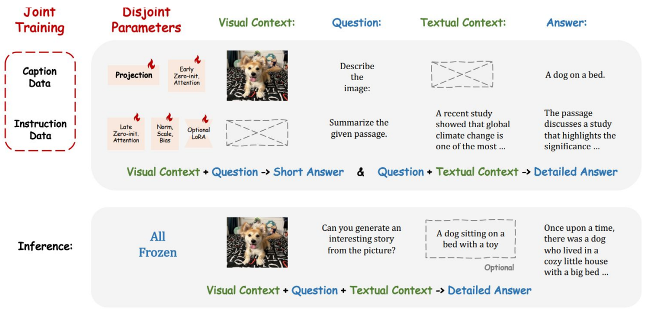


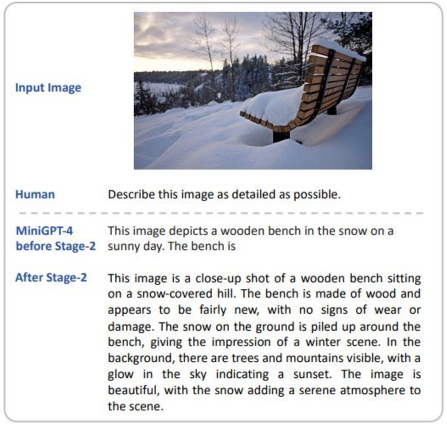
Figure 2. **Joint Training Paradigm in LLaMA-Adapter V2.** We utilize both image-text caption and language-only instruction data to jointly train LLaMA-Adapter V2, optimizing disjoint groups of learnable parameters.

[9]

Single-Stage Training

- Joint training using image-text pairs and instructions
- Two contrastive losses for image-text retrieval - 3 linear layers updated
- Frozen visual backbone - trains language model from scratch
- Train cross-attention layers and Perceiver based component
- All-in-one training stage

2.3 Multimodal Training



[10] Figure 5: MiniGPT-4 before second-stage finetuning fails to output completed texts. The generation is improved after the finetuning.

Two-Stage Training

- In the first stage: objective is to align the image features with the text embedding space
- In the second step: improve multimodal conversational capabilities
- Visual instruction-following training scheme, which is performed as a second training stage updating the parameters of both the multimodal adapter and LLM. During the first stage, instead, only the multimodal adapter is trainable
- Training solely the linear layer responsible for multimodal alignment across both stages - using filtered data in the second stage [10]
- Freezing of the visual encoder and LLM. In both training stages, only the Q-Former and the connection module are trainable
- Updates all weights in both stages, with the only exception of the visual backbone which is kept frozen

2.3 Multimodal Training

Training Data

- During the first (or single) training stage, image-text pairs from different sources are usually employed
  - LAION-2B, LAION400M, Conceptual Captions, COYO-700M, and DataComp
- Used in combination with one or more datasets characterized by text interleaved with images usually scraped from the web
  - WebLI, MMC4, MMDialog , and OBELICS
- Can introduce new data to address noise and bias (in first stage)
  - Uses ChatGPT to generate image-generating prompts and content based dialogues
  - Uses Stable Diffusion to generate the corresponding images.
- Datasets for visual instruction tuning
  - LLaVA-Instruct
  - Manual and generated data
- Multi-turn Dialogue datasets
  - LRV-Instruction (more robust instructions)
  - LLaVAR (text-rich images)

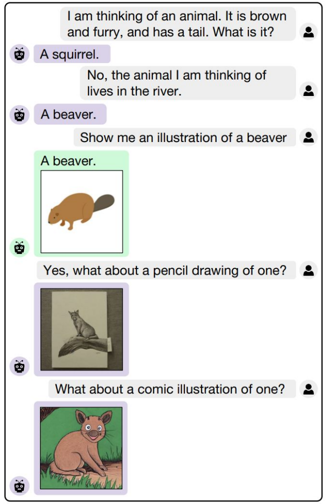
Tackling Visual Tasks with MLLMs

Tackling Visual Tasks with MLLMs

- Standard Visual understanding task:
  - Visual Question Answering
  - Image Captioning
  - Multi-turn conversation
- More fine-grained tasks:
  - Visual grounding
  - image generation

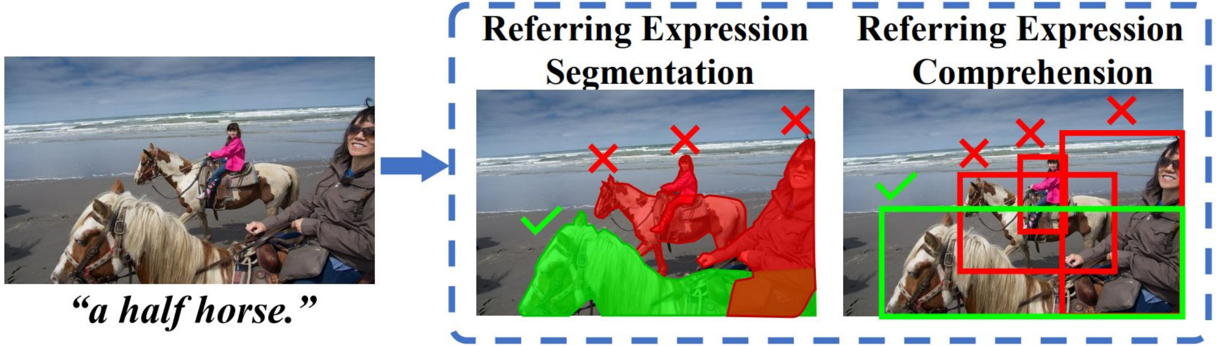
The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.



3.1 Visual Grounding

- Visual Grounding (VG) aims to link textual descriptions with specific parts of an image.
- Corresponds to tasks such as referring expression comprehension (REC), referring expression segmentation (RES)



(a) Illustration of Referring Expression Comprehension (REC) and Segmentation (RES).

3.1 Visual Grounding

Two main components are required to equip MLLMs with these capabilities

- A region-to-sequence method
    - Process input regions
- A sequence-to-region method
    - Ground nouns and phrases

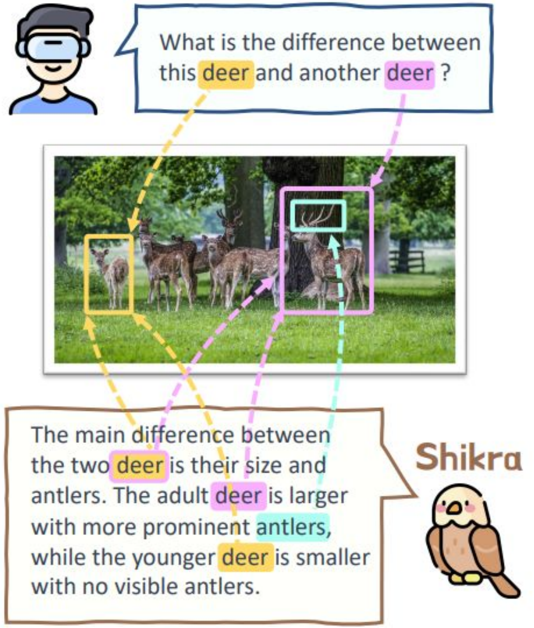


Figure 1: **Demo of Referential Dialogue (RD).** Users can point to specific areas and ask questions. In turn, Shikra will indicate the specific regions when replying, if necessary. More interesting dialogues can be found in Figure 2 and Appendix C.

Visual Grounding

Model	LLM	Visual Encoder	Supporting Model	Main Tasks & Capabilities
ContextDET (Zang et al., 2023)	OPT-6.7B★	Swin-B	-	Visual Dialogue, VQA, Captioning, Detection, REC, RES
DetGPT (Pi et al., 2023)	Vicuna-13B★	EVA ViT-g	G-DINO★	Visual Dialogue, Detection
VisionLLM (Wang et al., 2023e)	Alpaca-7B▲	Intern-H	Deformable-DETR▲	VQA, Captioning, Detection, Segmentation, REC
BuboGPT (Zhao et al., 2023c)	Vicuna-7B★	EVA ViT-g	RAM, G-DINO, SAM★	Visual Dialogue, Audio Understanding, Captioning, GroundCap
ChatSpot (Zhao et al., 2023b)	Vicuna-7B♦	CLIP ViT-L	-	Visual Dialogue, VQA, Captioning, Referring
GPT4RoI (Zhang et al., 2023f)	LLaVA-7B♦	OpenCLIP ViT-H	-	Visual Dialogue, VQA, Captioning, Referring
ASM (Wang et al., 2023d)	Husky-7B▲	EVA ViT-g	-	VQA, Captioning, Referring
LISA (Lai et al., 2023)	LLaVA-13B▲	CLIP ViT-L	SAM♦	Visual Dialogue, Captioning, RES
PVIT (Chen et al., 2023a)	LLaVA-7B♦	CLIP ViT-L	RegionCLIP★	Visual Dialogue, VQA, Captioning, Referring
GLaMM (Rasheed et al., 2023)	Vicuna-7B▲	OpenCLIP ViT-H	SAM♦	Visual Dialogue, Captioning, Referring, REC, RES, GroundCap
Griffon (Zhan et al., 2023)	LLaVA-13B♦	CLIP ViT-L	-	REC, Detection, Phrase Grounding
LLaFS (Zhu et al., 2023c)	CodeLLaMA-7B▲	CLIP RN50	-	Few-Shot Segmentation
NExT-Chat (Zhang et al., 2023a)	Vicuna-7B♦	CLIP ViT-L	SAM♦	Visual Dialogue, Captioning, Referring, REC, RES, GroundCap
GSVA (Xia et al., 2023b)	LLaVA-13B▲	CLIP ViT-L	SAM♦	VQA, Segmentation, REC, RES
Lenna (Wei et al., 2023)	LLaVA-7B▲	CLIP ViT-L	G-DINO♦	VQA, Captioning, REC
LISA++ (Yang et al., 2023b)	LLaVA-13B▲	CLIP ViT-L	SAM♦	Visual Dialogue, Captioning, RES
LLaVA-G (Zhang et al., 2023d)	Vicuna-13B♦	CLIP ViT-L	OpenSeeD, S-SAM♦	Visual Dialogue, REC, RES, Grounding
PixelLLM (Xu et al., 2023a)	FlanT5-XL-3B▲	EVA ViT-L	SAM★	Referring, REC, RES, GroundCap
PixelLM (Ren et al., 2023b)	LLaVA-7B▲	CLIP ViT-L	-	Visual Dialogue, RES
VistaLLM (Pramanick et al., 2023)	Vicuna-13B♦	EVA	-	Visual Dialogue, VQA, Referring, REC, RES, GroundCap
ChatterBox (Tian et al., 2024b)	LLaVA-13B▲	CLIP ViT-L	iTPN-B★, DINO♦	Visual Dialogue, Referring, REC, GroundCap
GELLA (Qi et al., 2024)	LLaVA-13B▲	CLIP ViT-L	Mask2Former♦	Segmentation, RES, GroundCap
PaLI-3 (Chen et al., 2023h)	UL2-3B♦	SigLIP ViT-g	VQ-VAE♦	VQA, Captioning, Retrieval, RES

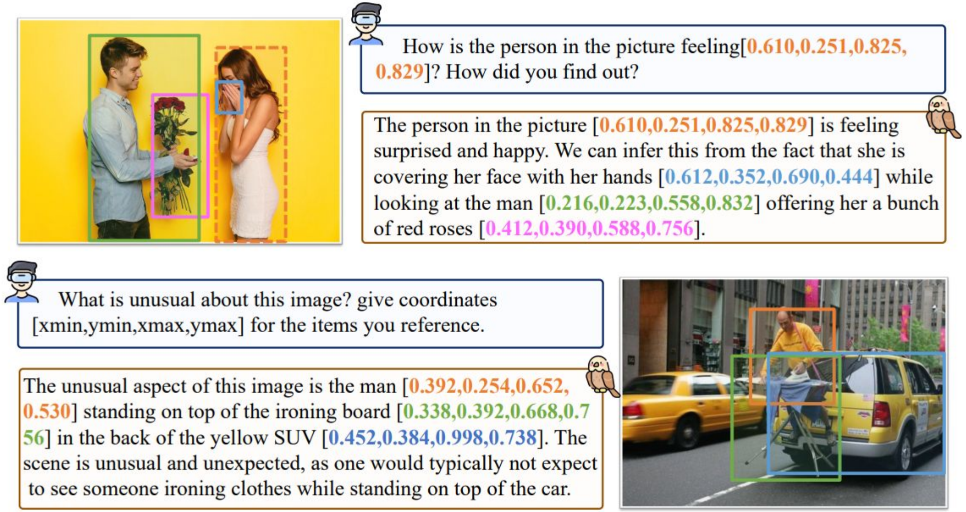
Table 2: Summary of MLLMs with components specifically designed for visual grounding and region-level understading. For each model, we indicate the LLM used in its best configuration, in some cases initialized with the weights of a pre-trained MLLM, and any supporting models used to perform the task (♦: fine-tuning; ▲: fine-tuning with PEFT techniques; ★: frozen). Gray color indicates models not publicly available.

Visual Grounding

Region-as-Text

- Generate text as a series of coordinates, represented as numbers or as special tokens dedicated to location bins

Embedding-as-Region



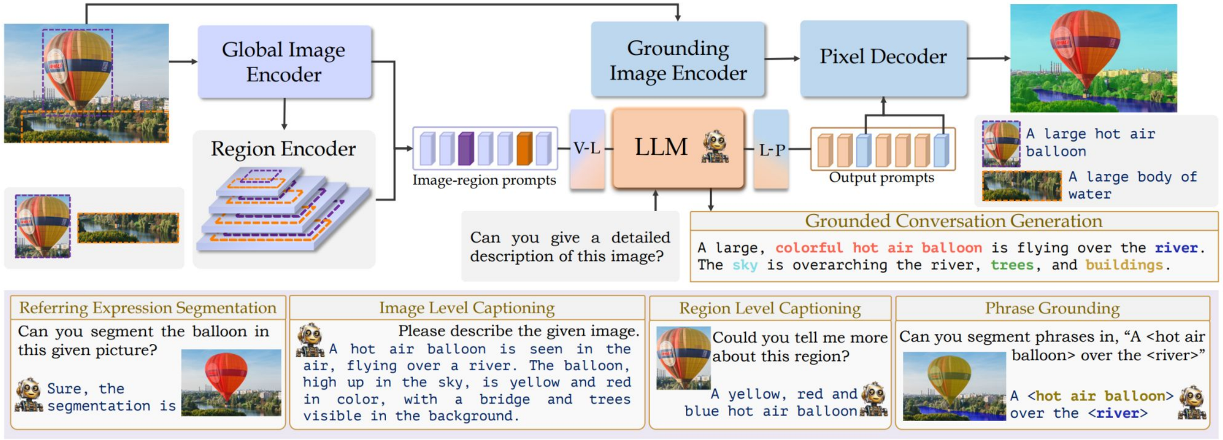
Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., & Zhao, R. (2023). Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. ArXiv, abs/2306.15195.

Visual Grounding

Region-as-Text

Embedding-as-Region

- Read input regions through region encoders and provide output regions as embedding to decoder



Rasheed, H.A., Maaz, M., Mullappilly, S.S., Shaker, A.M., Khan, S.H., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M., & Khan, F.S. (2023). GLaMM: Pixel Grounding Large Multimodal Model. ArXiv, abs/2311.03356.

Image generation and Editing

Model	LLM	Visual Encoder	Supporting Model	Main Tasks & Capabilities
GILL (Koh et al., 2023a)	OPT-6.7B★	CLIP ViT-L	SD v1.5★	Visual Dialogue, Retrieval, Image Generation
Emu (Sun et al., 2023b)	LLaMA-13B♦	EVA ViT-g	SD v1.5♦	Visual Dialogue, VQA, Captioning, Image Generation
SEED (Ge et al., 2023a)	OPT-2.7B▲	EVA ViT-g	SD v1.4★	VQA, Captioning, Image Generation
DreamLLM (Dong et al., 2023)	Vicuna-7B♦	CLIP ViT-L	SD v2.1★	Visual Dialogue, VQA, Captioning, Image Generation, Interleaved Generation
LaVIT (Jin et al., 2023)	LLaMA-7B♦	EVA ViT-g	SD v1.5♦	VQA, Captioning, Image Generation
MGIE (Fu et al., 2024)	LLaVA-7B★	CLIP ViT-L	SD v1.5♦	Image Editing
TextBind (Li et al., 2023e)	LLaMA-2-7B♦	EVA ViT-g	SD XL★	Visual Dialogue, VQA, Captioning, Image Generation
Kosmos-G (Pan et al., 2023)	Magneto-1.3B◊	CLIP ViT-L	SD v1.5★	Image Generation, Compositional Image Generation
MiniGPT-5 (Zheng et al., 2023)	Vicuna-7B▲	EVA ViT-g	SD v2.1★	Visual Dialogue, Image Generation, Interleaved Generation
SEED-LLaMA (Ge et al., 2023b)	LLaMA-2-13B♦	EVA ViT-g	SD unCLIP★	Visual Dialogue, VQA, Captioning, Image Generation, Interleaved Generation
CoDi-2 (Tang et al., 2023)	LLaMA-2-7B▲	ImageBind	SD unCLIP★	Visual Dialogue, Audio Understanding, Image Generation, Image Editing
Emu2 (Sun et al., 2023a)	LLaMA-33B♦	EVA ViT-E	SD XL♦	Visual Dialogue, VQA, Captioning, Image Generation, Image Editing
LLMGA (Xia et al., 2023a)	LLaVA-13B♦	CLIP ViT-L	SD XL♦	Visual Dialogue, VQA, Image Generation, Image Editing
SmartEdit (Huang et al., 2023b)	LLaVA-13B▲	CLIP ViT-L	SD♦	Image Editing
VL-GPT (Zhu et al., 2023b)	LLaMA-7B▲	CLIP ViT-L	SD v1.5★	Visual Dialogue, VQA, Captioning, Image Generation, Image Editing
MM-Interleaved (Tian et al., 2024a)	Vicuna-13B♦	CLIP ViT-L	SD v2.1♦	VQA, Captioning, REC, Image Generation, Interleaved Generation
JAM (Aiello et al., 2024)	LLaMA*-7B♦	-	CM3Leon♦	Image Generation, Interleaved Generation

Table 3: Summary of MLLMs with components specifically designed for image generation and editing. For each model, we indicate the LLM (✻: LLM variants) used in its best configuration, in some cases initialized with the weights of a pre-trained MLLM, and any supporting models used to perform the task (◊: training from scratch; ♦: fine-tuning; ▲: fine-tuning with PEFT techniques; ★: frozen). Gray color indicates models not publicly available.

Connecting MLLMs with Diffusion Models

- GILL (Koh et al., 2023a) maps the output embedding space of a frozen LLM to frozen diffusion model
- Inspired by Q-Former, a mapper component is trained by minimizing the L2 distance between image output representation of the language model and the expected conditioning embedding of SD

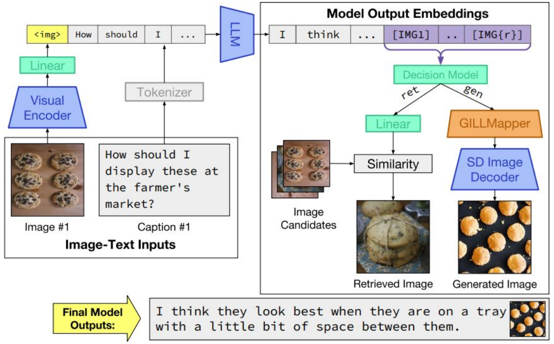


Figure 3: Inference time procedure for GILL. The model takes in image and text inputs, and produces text interleaved with image embeddings. After deciding whether to retrieve or generate for a particular set of tokens, it returns the appropriate image outputs.

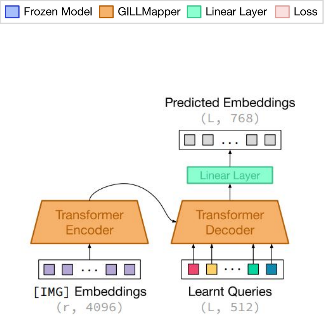
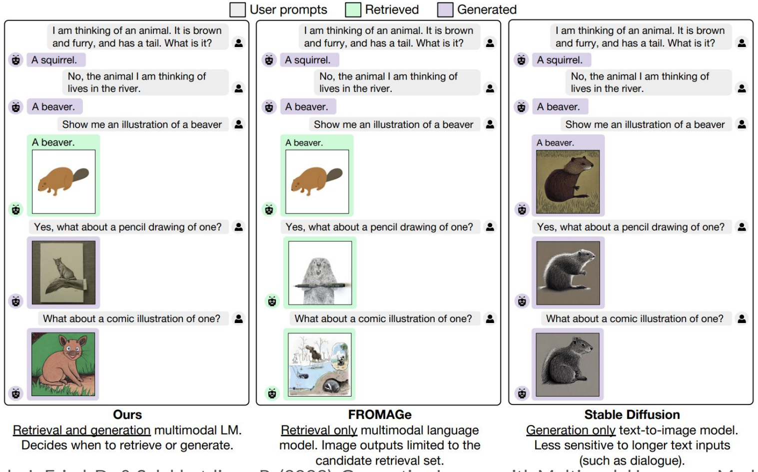


Figure 4: GILLMapper model architecture. It is conditioned on the hidden [IMG] representations and a sequence of learnt query embedding vectors.



Koh, J., Fried, D., & Salakhutdinov, R. (2023). Generating Images with Multimodal Language Models. ArXiv, abs/2305.17216.

Connecting MLLMs with Diffusion Models

- Kosmos-G (Pan et al., 2023) is developed through a training regime that integrates the output of the LLM with an encoder-decoder structure - AlignerNet.
- Leveraging a reconstruction loss and the minimization of the distance within a CLIP-text embedding.

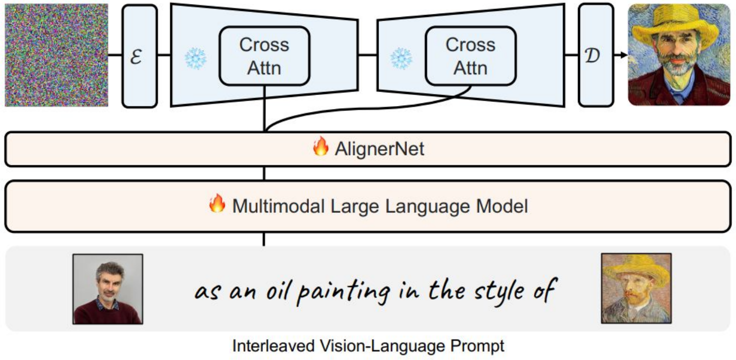
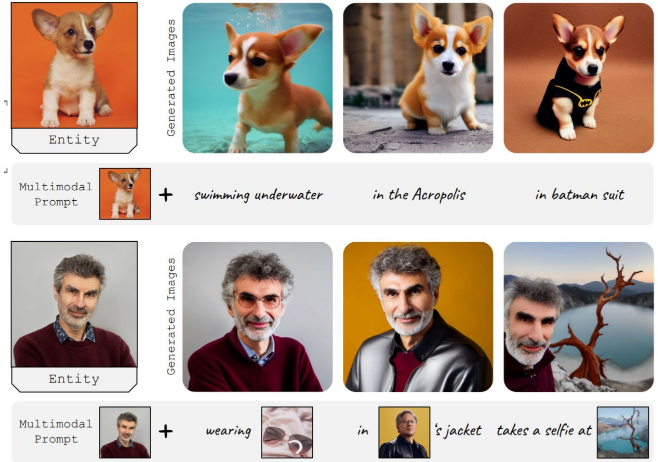


Figure 2: KOSMOS-G comprises an MLLM for multimodal perception, coupled with an AlignerNet that bridges the MLLM to the diffusion U-Net image decoder. KOSMOS-G can pass the fine concept-level guidance from interleaved input to image decoder, and offer a seamless alternative to CLIP. Orange denotes the trainable modules; Blue denotes the frozen ones.



Pan, X., Dong, L., Huang, S., Peng, Z., Chen, W., & Wei, F. (2023). Kosmos-G: Generating Images in Context with Multimodal Large Language Models. ArXiv, abs/2310.02992.

Other Modalities and Applications

Video Understanding

Any-Modality Models

Domain-Specific MLLMS

Conclusion and Future Directions

- Correction of Hallucinations.
- Prevent Harmful and Biased Generation.
- Reduce Computational Load

Input — *"Make them look like flight attendants"* — *"Make them look like doctors"*

Figure 14. Our method reflects biases from the data and models it is based upon, such as correlations between profession and gender.

Brooks, T., Holynski, A., & Efros, A.A. (2022). InstructPix2Pix: Learning to Follow Image Editing Instructions. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18392-18402.

Bibliography

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NeurIPS.

[2] Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. Warp: Word-level adversarial reprogramming. arXiv preprint arXiv:2101.00121.

[3] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In ICLR.

[4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient Finetuning of Quantized LLMs. arXiv preprint arXiv:2305.14314.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In ICML.

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023f. BLIP-2: Bootstrapping Language-Image Pretraining with Frozen Image Encoders and Large Language Models. arXiv preprint arXiv:2301.12597

[7]Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., ... & Cucchiara, R. (2024). The (R) Evolution of Multimodal Large Language Models: A Survey. arXiv preprint arXiv:2402.12451.

[8] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In ICML.

[9]Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. arXiv preprint arXiv:2304.15010

[10]Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv preprint arXiv:2304.10592.

Visual Grounding

Region-as-Text

Embedding-as-Region

Text-to-Grounding

- Based on open-vocabulary models that accept textual categories as input

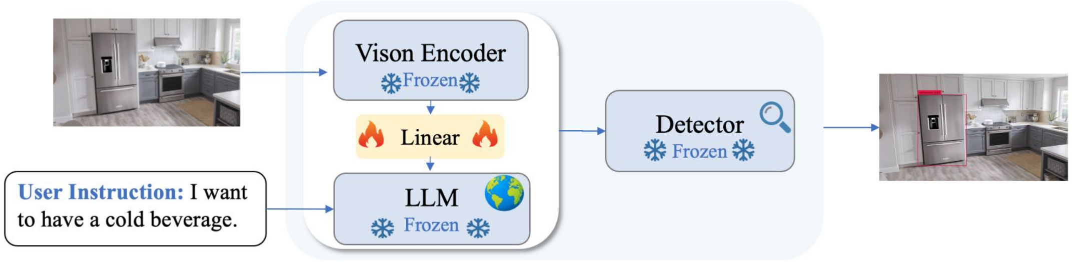


Figure 2: Framework of DetGPT. The multi-model model consisted of vision encoder and LLM interprets the user instruction, reasons over the visual scene, and finds objects matching the user instruction. Then, the object names/phrases are passed to the open-vocabulary detector for localization.

Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., & Zhang, L.K. (2023). DetGPT: Detect What You Need via Reasoning. ArXiv, abs/2305.14167.

End-to-End Training

- SD U-Net is directly fine-tuned with the continuous visual embeddings generated by the LLM
- employ a feature synchronizer, that intervenes in intermediate layers of the LLM and diffusion decoder to cross-attend multi-scale high-resolution image features

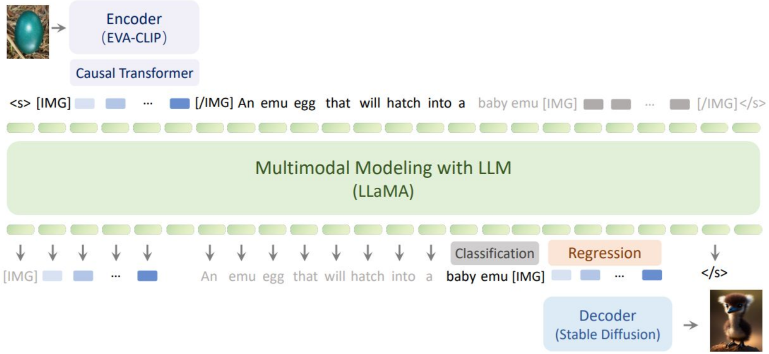


Figure 2: **Emu** unifies the modeling of different modalities in an auto-regressive manner. Visual signals are first encoded into embeddings, and together with text tokens form an interleaved sequence. The training objective is to either classify the next text token or regress the next visual embedding. In inference, regressed visual embeddings are decoded into a realistic image via a fine-tuned latent diffusion model.

Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., & Wang, X. (2023). Generative Pretraining in Multimodality. *ArXiv, abs/2307.05222*.