Development Roadmap from LLaVA to LLaVA-OneVision to LLaVA-3D

Chenming Zhu

Fall 2024

LLaVA Roadmap

- LLaVA (Visual Instruction Tuning)
- LLaVA-NeXT Blog:

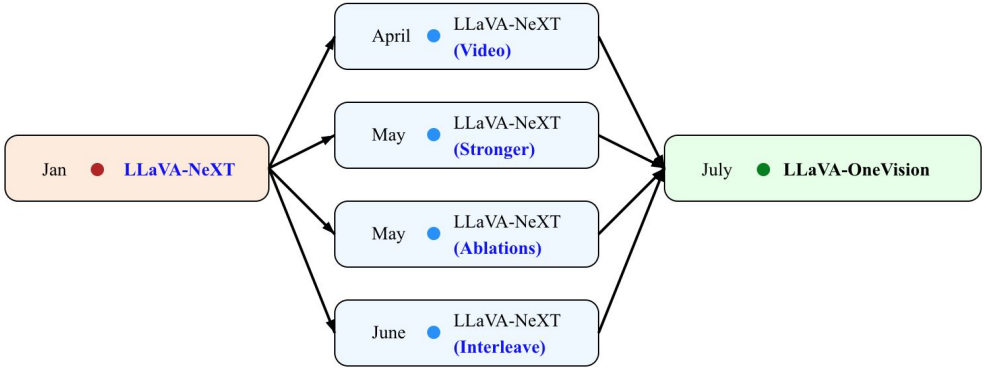
  Improved reasoning, OCR, and world knowledge

  A Strong Zero-shot Video Understanding Model

  Stronger LLMs Supercharge Multimodal Capabilities in the Wild

  What Else Influences Visual Instruction Tuning Beyond Data?

- LLaVA-OneVision
- LLaVA-3D



LLaVA: Visual Instruction Tuning

Model Architecture

- Vision Encoder + Projection Layer + LLM

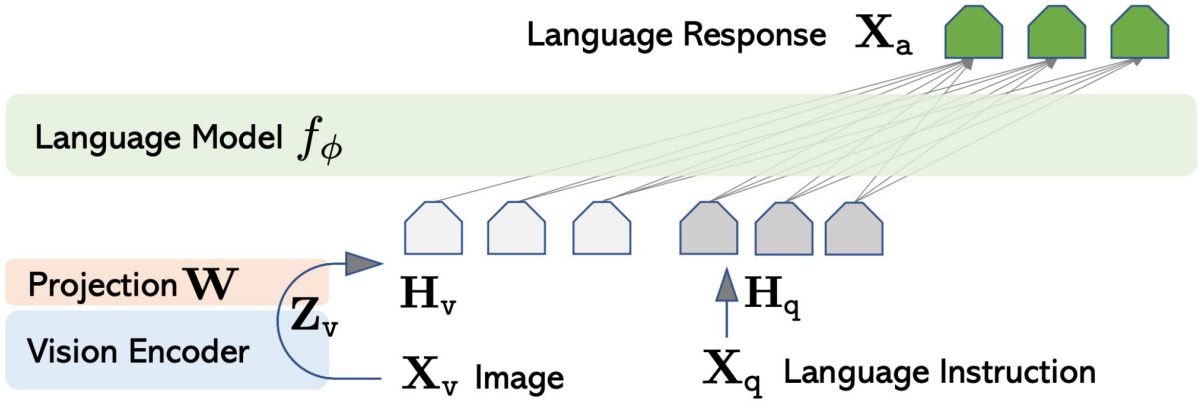
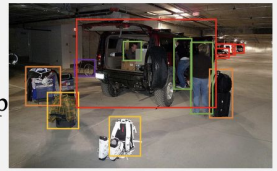


Figure 1: LLaVA network architecture.

Visual Instruction Tuning Data Construction

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>

Use GPT-4 to convert the COCO dataset with Caption and Bounding Boxes information to data:

Conversation: Dialogue data, totaling 58K samples.

Detailed description: Rich and comprehensive descriptions of images, totaling 23K samples.

Complex reasoning: Complex reasoning data, totaling 77K samples.

Training Recipe

Stage 1: Pre-training for Feature Alignment [projection layer]

Dataset: filter CC3M to 595K image-text pairs

Stage 2: Fine-tuning End-to-End for Instruction Following [projection layer and LLM]

Dataset: Instruction Tuning Dataset

Limitations

- The simple connector may limit the model's ability to deeply understand complex visual information.

- Limited training data scale and diversity

- Potential hallucination and misinformation

LLaVA-1.5: Improved Baselines with Visual Instruction Tuning

Model Architecture Modification:

- Replacing the original CLIP-ViT-L/14 visual encoder with the CLIP-ViT-L-336px visual encoder.
- Replacing the original single linear layer with an MLP layer (two linear layers).

Incorporating VQA data oriented towards academic tasks and specifying response format in prompts: This enhances LLaVA's performance on academic task benchmarks.

	Method	LLM	Res.	GQA	MME	MM-Vet
	InstructBLIP	14B	224	49.5	1212.8	25.6
	Only using a subset of InstructBLIP training data					
0	**LLaVA**	7B	224	–	809.6	25.5
1	+VQA-v2	7B	224	47.0	1197.0	27.7
2	+Format prompt	7B	224	46.8	1323.8	26.3
3	+MLP VL connector	7B	224	47.3	1355.2	27.8
4	+OKVQA/OCR	7B	224	50.0	1377.6	29.6
	Additional scaling					
5	+Region-level VQA	7B	224	50.3	1426.5	30.8
6	+Scale up resolution	7B	336	51.4	1450	30.3
7	+GQA	7B	336	62.0*	1469.2	30.7
8	+ShareGPT	7B	336	62.0*	1510.7	31.1
9	+Scale up LLM	13B	336	**63.3***	**1531.3**	**36.1**

Table 2. **Scaling results** on data, model, and resolution. We choose to conduct experiments on GQA [20], MME [16], and MM-Vet [52] to examine the representative capabilities of VQA with short answers, VQA with output formatting, and natural visual conversations, respectively. *Training images of GQA were observed during training.

LLaVA-NeXT: Improved reasoning, OCR, and world knowledge

Dynamic High Resolution

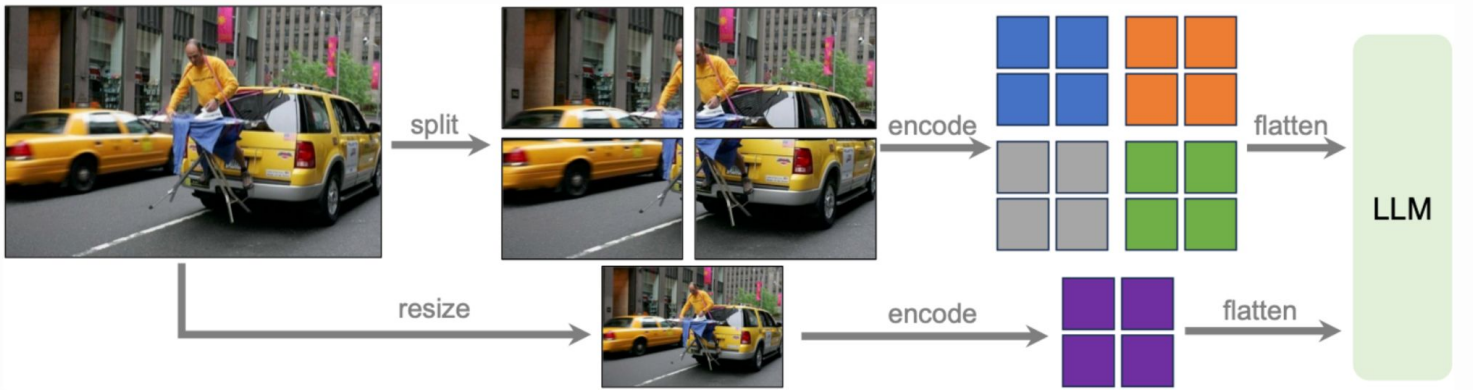


Illustration of dynamic high resolution scheme: a grid configuration of 2 × 2

AnyRes technique is designed to accommodate images of various high resolutions. It employs a grid configuration of {2×2,1×{2,3,4},{2,3,4}×1}, balancing performance efficiency with operational costs for the high-resolution image.

Data Mixture

- **High-quality User Instruct Data**.

First, the diversity of task instructions, ensuring adequately represent a broad spectrum of user intents that are likely to be encountered in real-world scenarios, particularly during the model's deployment phase. Second, the superiority of responses is critical, with the objective of soliciting favorable user feedback. To achieve this, it considers two data sources:

(1) Existing GPT-V data. LAION-GPT-V and ShareGPT-4V.

(2) To further facilitate better visual conversation for more scenarios, it collects a small 15K visual instruction tuning dataset covering different applications. The instructions and images come from LLaVA demo, which are real-world users requests. They carefully filter samples that may have privacy concerns or are potentially harmful, and generate the response with GPT-4V.

- **Multimodal Document/Chart Data.**

LLaVA-NeXT: Improved reasoning, OCR, and world knowledge

Compared with LLaVA-1.5, LLaVA-NeXT has several improvements:

1. Increasing the input image resolution to 4x more pixels. This allows it to grasp more visual details. It supports three aspect ratios, up to 672x672, 336x1344, 1344x336 resolution.
2. Better visual reasoning and OCR capability with an improved visual instruction tuning data mixture.
3. Better visual conversation for more scenarios, covering different applications. Better world knowledge and logical reasoning.

LLaVA-NeXT: A Strong Zero-shot Video Understanding Model

Zero-shot video representation capabilities with AnyRes

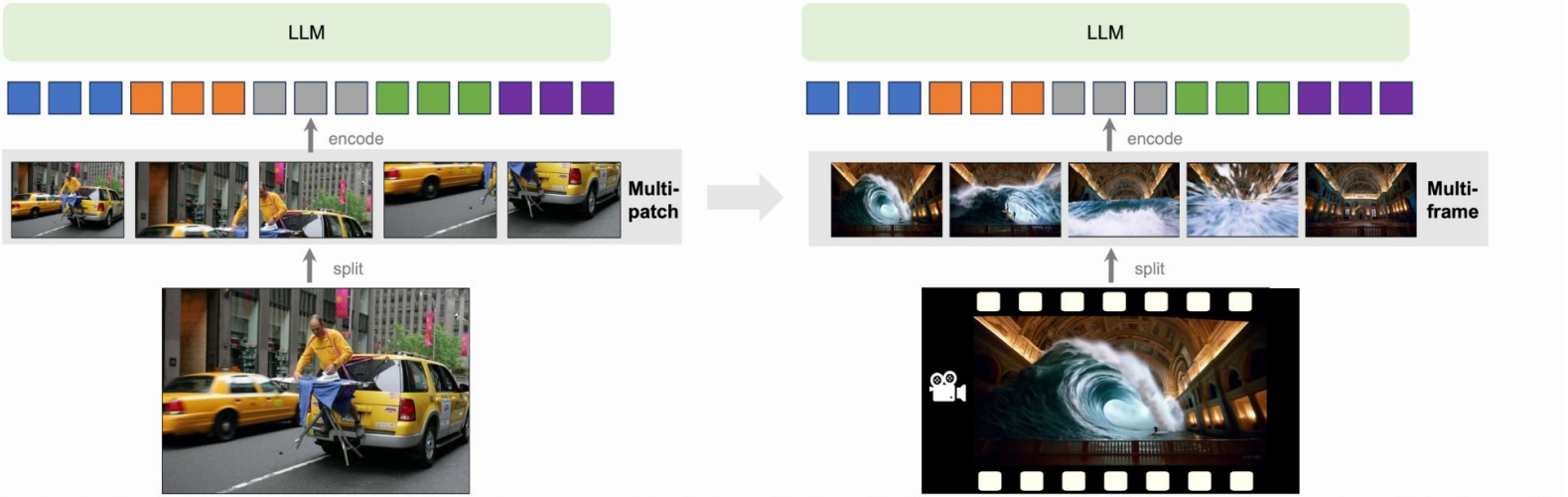
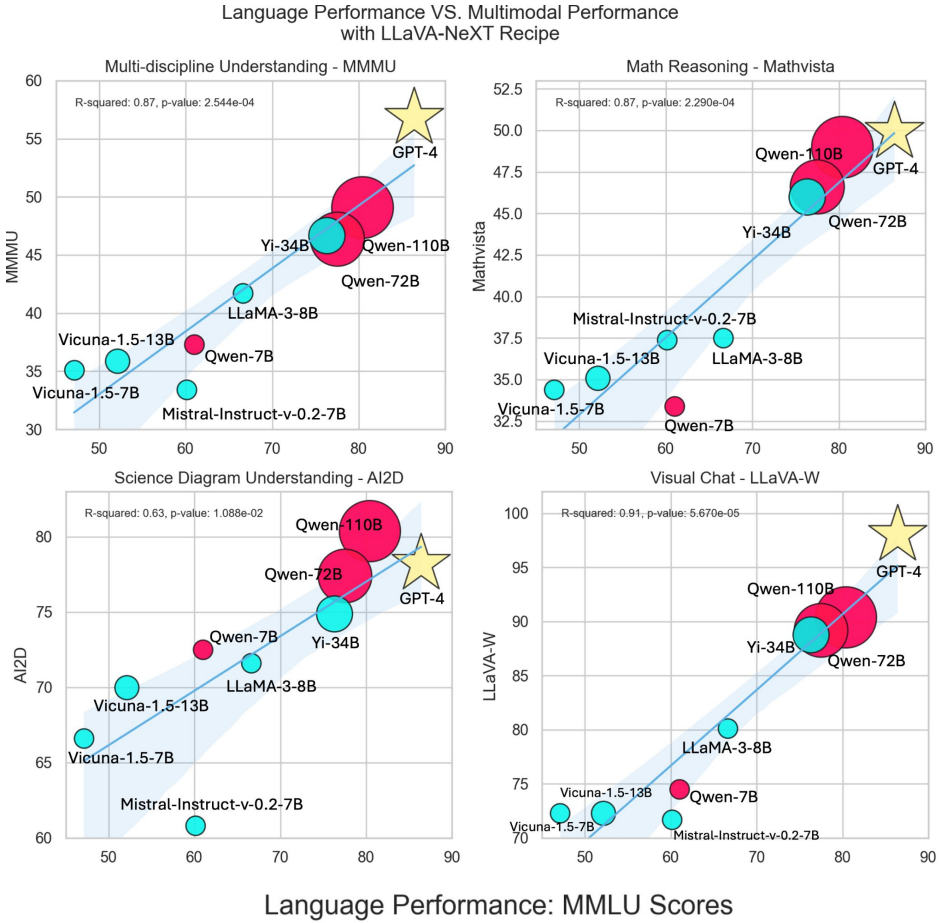


Illustration that AnyRes digests a set of image as a sequence of concatenated visual tokens, allowing unified image and video input, which natually suppots the evolution from multi-image to multi-frame

With minor code adjustments, LLaVA-NeXT can process N video frames arranged in a {1xN} grid. Assuming each frame comprises 24x24 tokens, the total token count for a video would be 24x24xN. However, considering the "max_token_length" limit of 4096 for the LLM, it is crucial to ensure that 24x24xN + the number of text tokens < 4096 to avoid nonsensical outputs.

How effectively can the language capabilities of LLMs be transferred to multimodal settings?

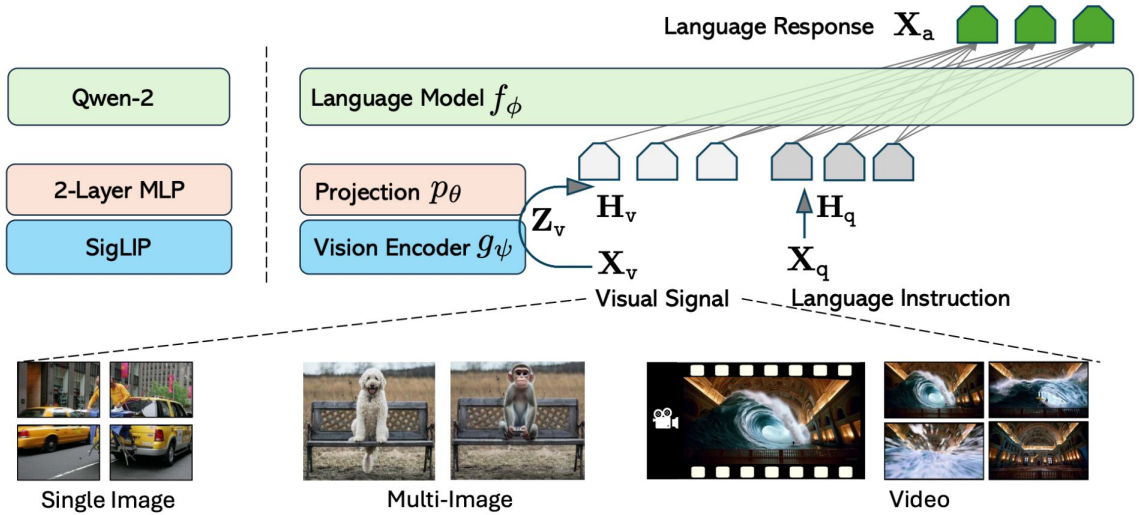


Improved Language Capability:
Across LLMs of comparable sizes (e.g., 7B Mistral/Vicuna, 7B Qwen, 8B LLaMa3), there exists a consistent pattern where higher language proficiency, as measured by MMMU scores, corresponds to improved multimodal capabilities.

Influence of Model Size:
Within the same LLM family (e.g., Qwen LLM: 7B, 72B, 110B), larger models consistently demonstrate superior performance on multimodal benchmarks.

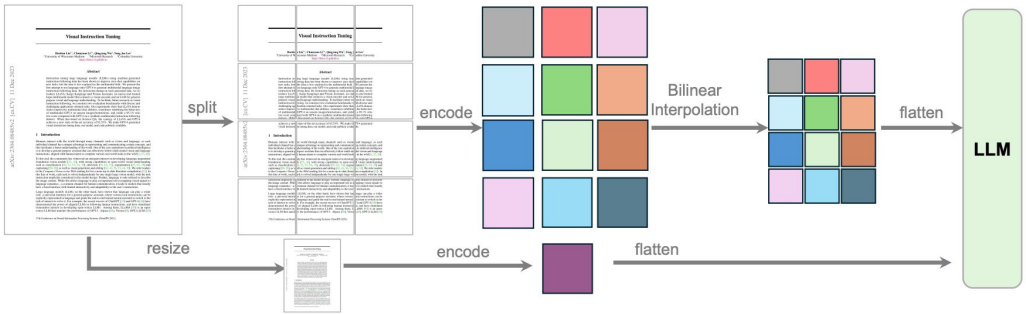
LLaVA-OneVision: Easy Visual Task Transfer

LLaVA-OneVision is the first single model that can simultaneously push the performance boundaries of open LMMs in three important computer vision scenarios: single-image, multi-image, video scenarios.

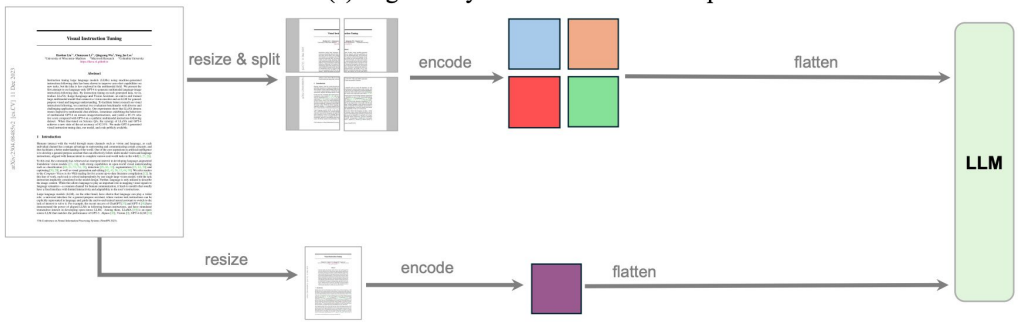


Visual Representation

To strike a balance of performance and cost, the author observe that the scaling of resolution is more effective than the scaling of token numbers, and recommend an AnyRes strategy with pooling.



(a) Higher AnyRes with Bilinear Interpolation



(b) The original AnyRes

Visual Representation

Single-Image: consider a large maximum spatial configuration (a, b) for single-image representation to maintain the original image resolution without resizing.

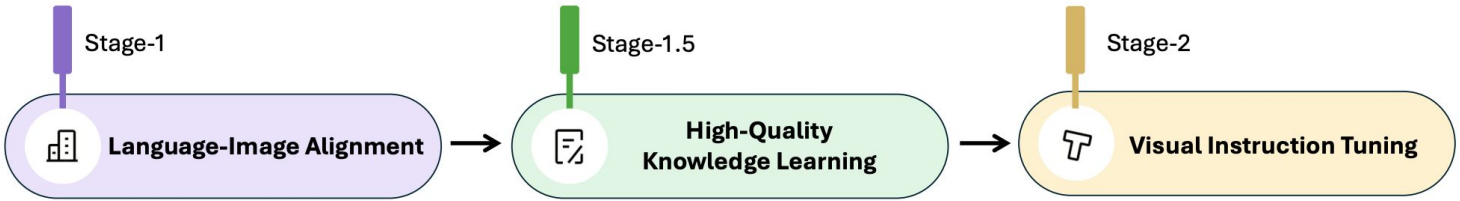
Multi-image: Only the base image resolution is considered.

Video: Each frame of the video is resized to the base image resolution and processed by the vision encoder to generate feature maps. Bilinear interpolation is employed to reduce the number of tokens

Insights on Training Strategies

Prior LLaVA models mainly explore Stage-2 for new scenarios and improved performance. However, the first two functionalities are less frequently investigated and therefore constitute the primary focus of this section.

- Stage-1: Language-Image Alignment.
- Stage-1.5: High-Quality Knowledge Learning.
- Stage-2:Visual Instruction Tuning.



High-Quality Knowledge

To illustrate high-quality knowledge, we consider data from three major categories:

- **Re-Captioned Detailed Description Data:** LLaVA-NeXT-34B is known for its strong detailed caption ability among open-source LMMs. We used the model to generate new captions for the images from the following datasets: COCO118K, BLIP558K, and CC3M.

- **Document / OCR Data:** We utilized the Text Reading subset from the UReader dataset, totaling 100K, which is easily accessible through PDF rendering. We used this text reading data along with the SynDOG EN/CN 1M datasets.

- **ShareGPT4V Chinese Detailed Caption:** We used the original ShareGPT4V[3] images and utilized GPT-4V provided by the Azure API to generate detailed Chinese caption data, aiming to improve the model's capability in Chinese.

High-Quality Knowledge

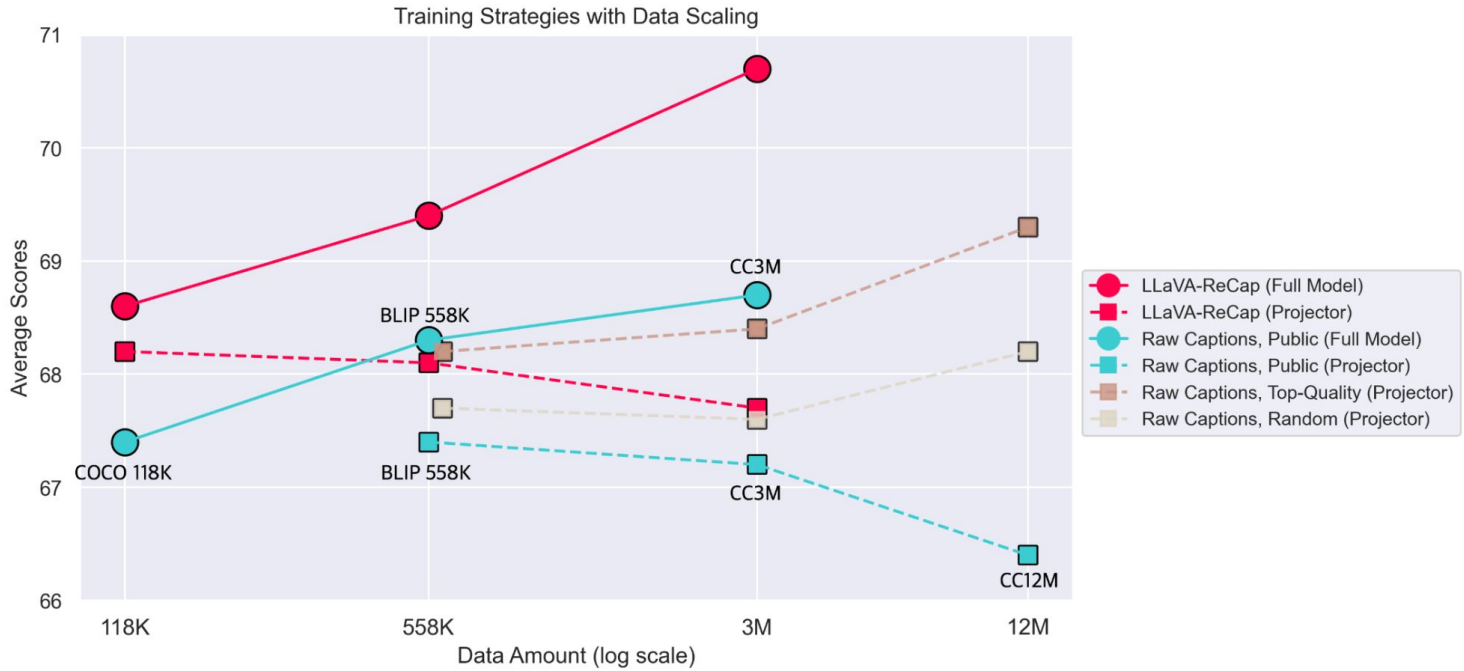
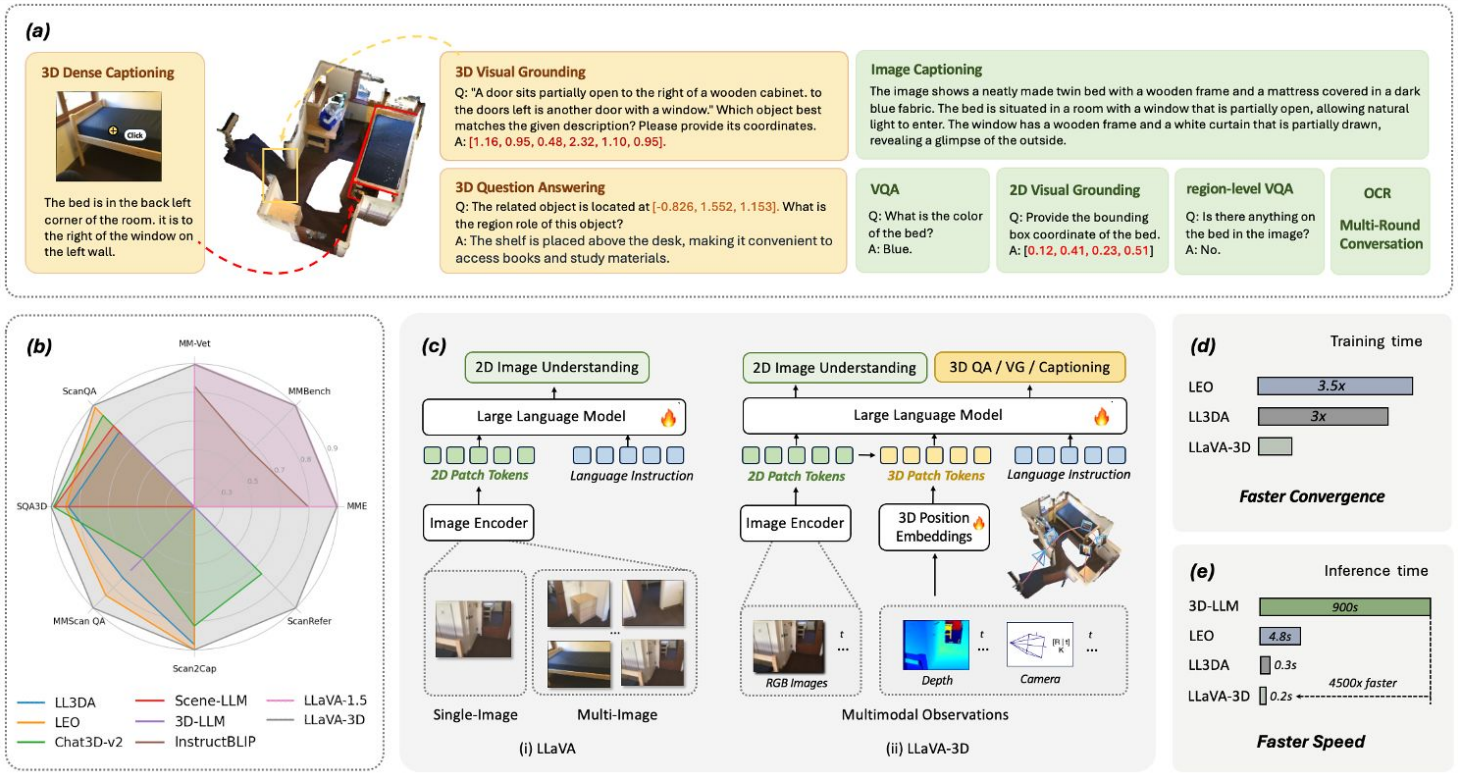


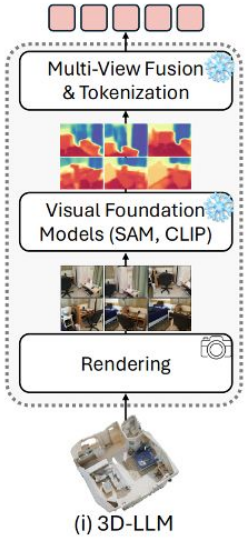
Figure 2. This figure shows that using LLaVA-ReCap Data in Stage 1.5 training yields the most significant improvements (red circles). Performance with raw captions data like COCO18K, BLIP558K, and CC3M is also strong (blue circles). We also include the results from Section 3.1 (squared shape), where only the projector was trained using raw captions data at various scales (e.g. from BLIP558K to Web 12M).

LLaVA-3D: A Simple yet Effective Pathway to Empowering LMMs with 3D Capabilities

https://zcmax.github.io/projects/LLaVA-3D/

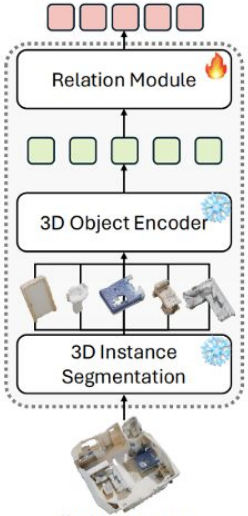


Motivation 1: Lack of "3D CLIP"



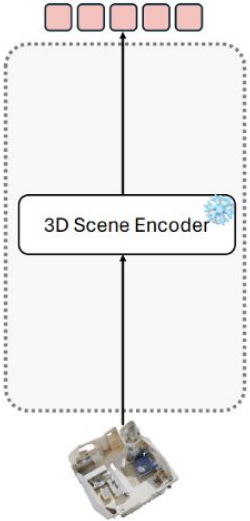
(i) 3D-LLM

Rely on 2D Feature, but time-consuming (600s)



(ii) Chat-3D v2

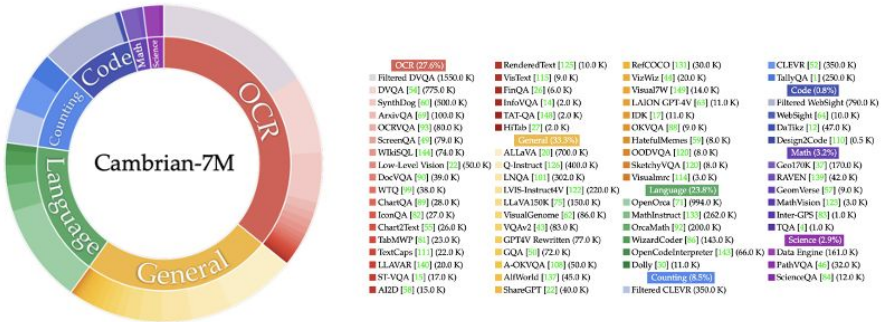
Rely on 3D segmentor and object point encoder, Hard to handle non-object related task



Pre-trained 3D point encoder (generalization problem)

Motivation 2 – Lack of High Quality 3D Vision-Language Data

2D Vision Language Data:



Diverse, Rich text description

3D Vision Language Data:

Limited

Simple Text Description

low-quality

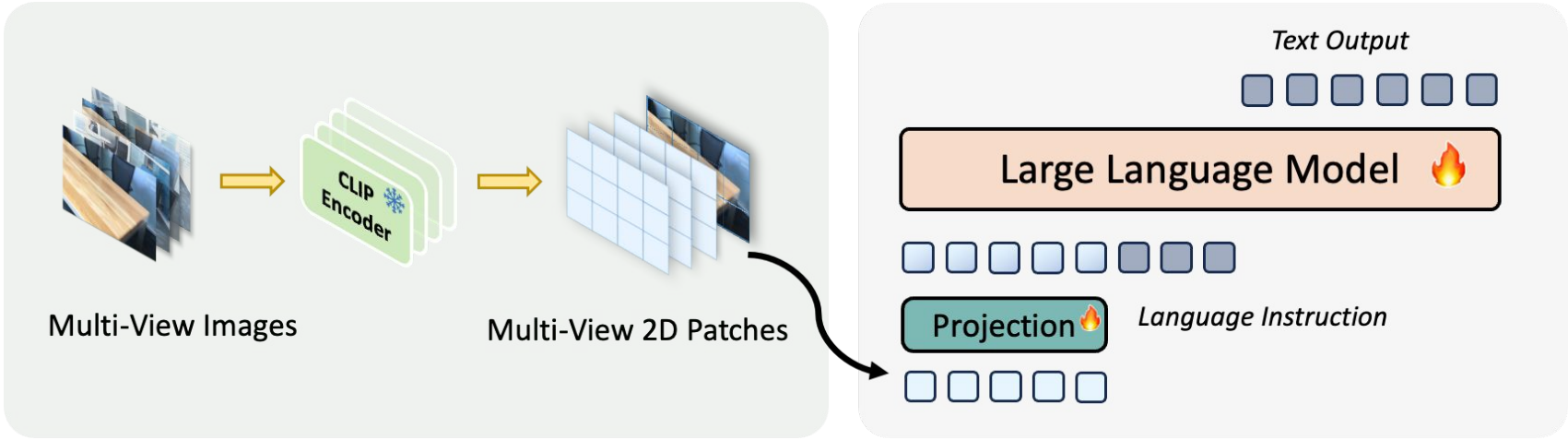
Can we unify 2D and 3D LMM Architecture?

Input Vision Unify --> Multi-View Images --> Implicit 3D world understanding

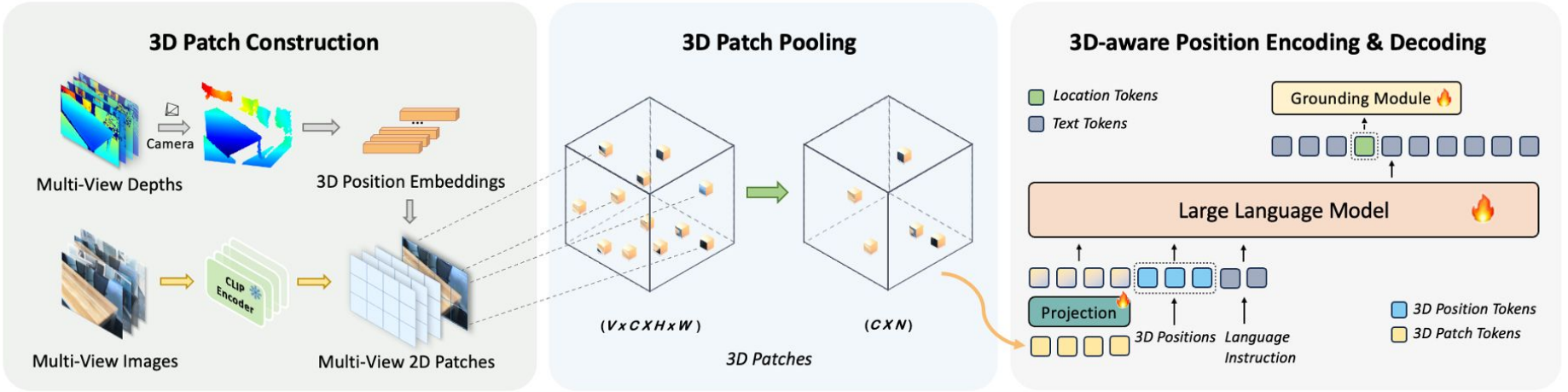
2D & 3D Differences --> 3D coordinates inputs and outputs --> Explicit 3D world understanding

How can we make LLaVA explicitly understand 3D world?

Multi-View LLaVA Pipeline

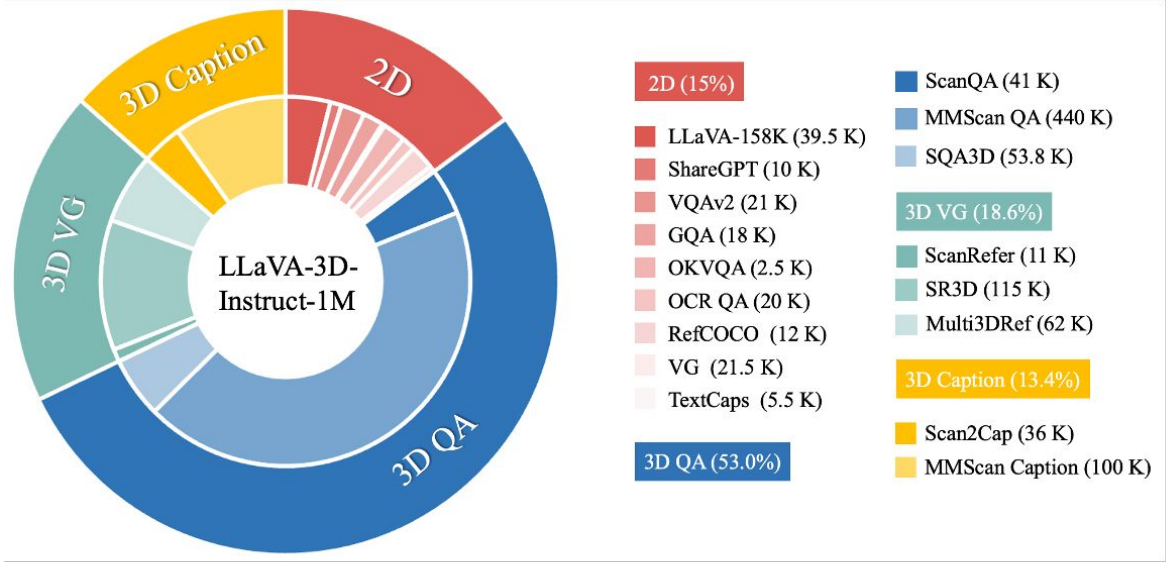


LLaVA-3D Architecture



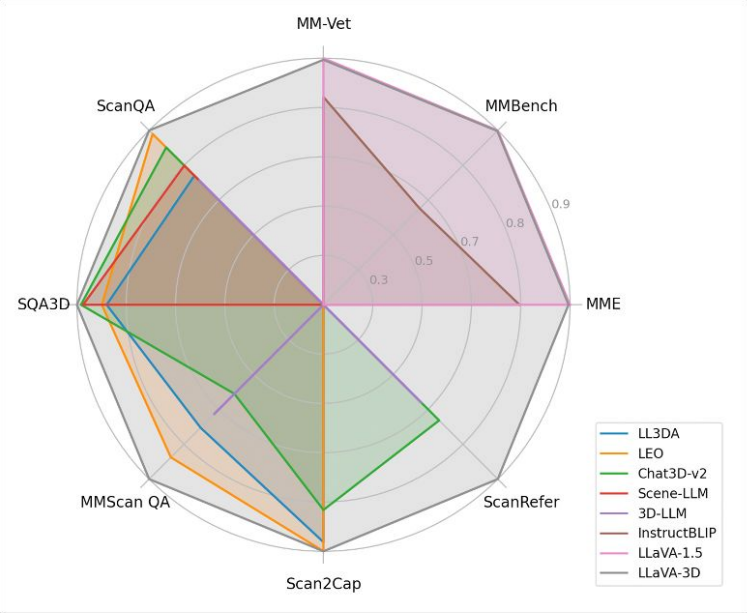
The core innovation in our approach is the introduction of 3D Patch, a new 3D representation that bridges 2D features within a 3D spatial context.

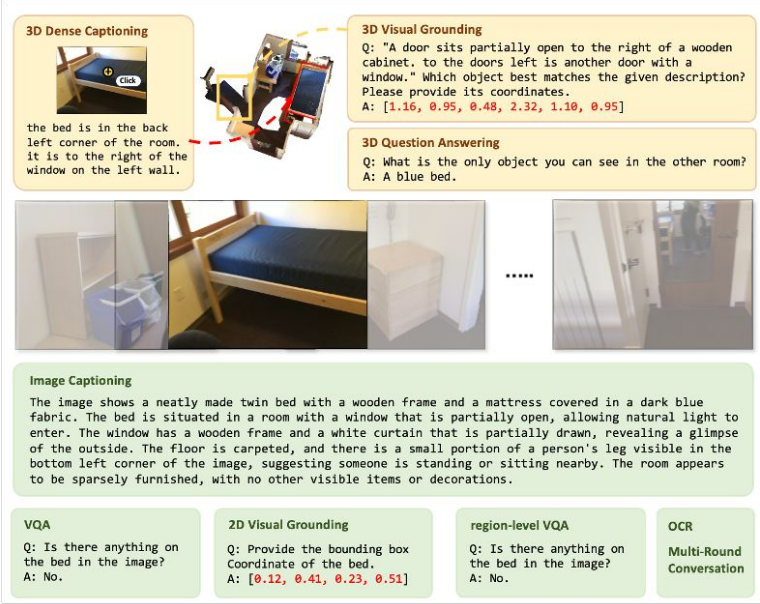
2D & 3D Joint Dataset



We collect the first joint 2D & 3D 1 Million instruction tuning dataset

2D & 3D Tasks

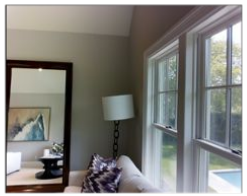


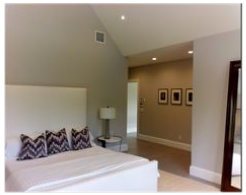


We achieve SOTA 3D V&L ability while maintaining the original 2D understanding and reasoning ability by joint tuning on 2D&3D Instruction Tuning datasets.

User-friendly 3D World Interaction

2D Click-based 3D Dense Captioning





User The related object is located at [Click]. Describe the object in detail and localize the object.

LLaVA-3D This is a single pillow placed on the sofa. The pillow is rectangular in shape, with a white background and black and white zigzag pattern. It is made of soft fabric, likely cotton or a similar material. The pillow is of medium size and is in good condition, showing no obvious signs of wear. It stands upright on the sofa, leaning against the backrest. The design of the pillow features a zigzag pattern, adding a touch of modernity to the overall look of the sofa. The placement of the pillow on the sofa indicates that it is used for decoration rather than providing support or comfort.

LLaVA-3D could perform 2D Click-based 3D dense captioning, generating the corresponding object caption and 3D bounding box.

User-friendly 3D World Interaction

2D Click-based 3D Question Answering

User The related objects are located at [Click1] [Click2]. Are these two objects similar in size?

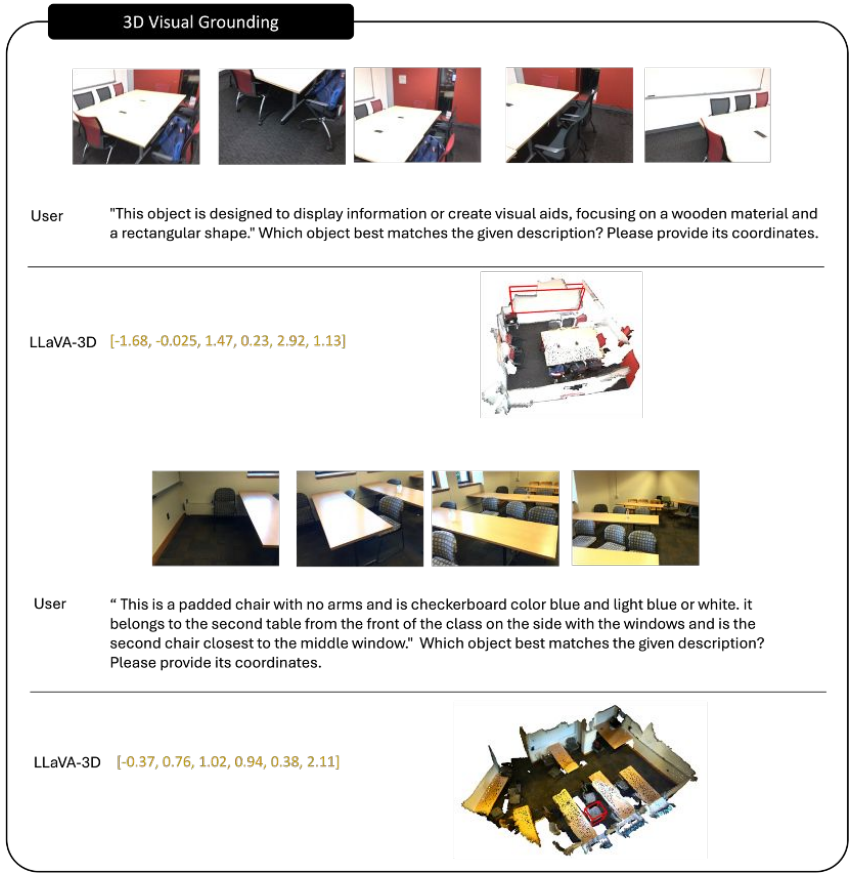
LLaVA-3D No.

User The related object is located at [Click]. Looking at the front of the object, select the cabinet that is to the left of it. What is the state of the cabinet?

LLaVA-3D The cabinet is full of ovens.

LLaVA-3D could perform 2D Click-based 3D question answering, now users could click on the 2D images and ask the question.

User-friendly 3D World Interaction



LLaVA-3D exhibits powerful 3D visual grounding capability, enabling accurate 3D bounding boxes output.

Thank you !