Multi-modal Large Language Model

- Methodology
- Benchmark
- Platform
- Data
Multi-modal Large Language Model

PaLM 2

Meta

文心一言

商量 - 语言大模型

通义千问

360智脑 × 360搜索
Q: Can you explain this meme?

A: 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.
Adapt LLM to Multi-Modal LLM?

Balance Multi-Modal Conflict?

How to Construct Multi-Modal Instructions?

Decode Various Downstream Tasks?

Evaluate Multi-Modal LLM?
- Each modality is pre-trained **separately**
- **No** connection between different modalities

<Vision, Label> / <Vision, Vision>
CLIP goes 3D

- Link vision with language
- Need **explicit** connection between different modalities, including large-scale paired data and explicit alignment. -> Hard to extend

Input/Output

| ImageNet/Coco | Kinetics |

Network

- Encoder
- Decoder
- Tokenizer
- Explicit Align.
- Implicit Align.
- Flow
- Frozen
- Tune

<Vision, Label> / <Vision, Vision>

<Vision, Text>
- Separate models for each task
  - Separate decoders
  - Separate finetuned encoders
  - Limited task number
MLLM based on API

- LLM as controller
- N models, one framework
- Prompt engineering (e.g. flow design, prompt definition)
- No need to finetune models

**Request**

I. Task Planning

II. Model Selection

**Response**

IV. Response Generation

**Models/Tools**

- VisualChatGPT/HuggingGPT
- Dolphin/VideoChat-Text
- InternGPT
- MLLM based on **trainable models**
  - LLM as base
  - One model, one framework
  - Instruct tuning data (implicit alignment between different modalities)
  - Trainable, in-context learning (fewer data, parameter-efficient finetuning)
Dolphin: A General Video Interaction Platform Based on LLMs, 2023
Q: Is he singing and rapping in the video?
A: No, he is not singing or rapping in this video. He is playing basketball and dancing in some parts of the video.
Q：请为我生成一段手语，内容是“你们好，很高兴认识你们，我是一名手语老师”。

A：好的，这个视频被保存为video/68g4.mp4。

Q：能帮我将视频中的手语翻译为中文嘛？

A：好的。这段手语对应的中文是：进入五月后，各地水果货物等农副产品大量上市，丰富着全国市民的菜篮子，也给农户们带来了收入。
Q：请为我生成一段以“悠闲的河畔”为主题的音乐。

A：好的，这个视频被保存为video/pe7s.mp4。

Q：请为我生成一个小女孩的演说肖像，其演说内容为“大家好，我是一个演说肖像，我能够辅助直播，演讲，带货等各种工作，很高兴认识大家”。

A：好的，这个视频被保存为video/9qc7.mp4。
Dolphin VideoChat InternGPT

Video Chat 来啦

扫描二维码加入讨论群
https://github.com/OpenGVLab/Ask-Anything
InternGPT is a multimodal interaction system that first proposed the cursor instruction.

5 Modalities

It allows users to input visual signals to the language model through simple actions such as clicking and drawing lines.

9 Tasks

Image cutouts, replacements, style transfer, image-based Q&A, drag-and-drop editing, sound-to-image conversion …
Interactive Image Generation

Online Demo: https://igpt.opengvlab.com/
GitHub: https://github.com/OpenGVLab/InternGPT

Audio 🎧 Image

Online Demo: https://igpt.opengvlab.com/
GitHub: https://github.com/OpenGVLab/InternGPT

InternGPT + DragGAN

Online Demo: https://igpt.opengvlab.com/
GitHub: https://github.com/OpenGVLab/InternGPT
**LLAMA-Adapter** finetunes language models **efficiently** to support **multiple** modalities

### 4 Modalities
- Image/Video/Sound/PointCloud

### 4 LLMs
- StableLM/Pythia/INCITE/Falcon

### High Impact
- [Github] 4k Stars/[Twitter] 9M Views

---

**Instruct**

The president of Mexico in 2019?

---

**LLaMA**

7B/13B Parameters

1.2M Parameters

1 Hour Fine-tuning

Plug with Expertise

Multi-modal Instruct

---

**Response**

Andrés Manuel López Obrador ...

---

**Multi-modal Reasoning of LLaMA-Adapter**

---

**Transformer Layers x L-1**

---

**Visual Context:**

- Frozen
- Fine-tune

---

**Visual Encoder**

---

**Adapter**

---

**Zero-ini Attention**

---

**Answer:** The answer is (A).

---

**Textual Context:**

Which force from the baby’s hand opens the door?

---

**Question:** (A) pull (B) push

---

**Options:**
**Answer:**

**Describe the 3D object in detail.**

The 3D object is a black car, specifically a minivan, parked in a parking lot. The car is positioned in the foreground of the image, with its front facing the viewer. The car appears to be a popular model, as there are several other cars parked in the lot, including a few cars in the background. The parking lot is located near a building, which could be a school or another type of establishment. The scene captures a typical day at a parking lot, with people going about their daily activities or running errands.

**What is the name of this game? What are the main characteristics?**

The name of this game is “League of Legends,” and it is a multiplayer online battle arena (MOBA) game. The main characteristics of League of Legends include a team-based format, where two teams of five players each compete to destroy the opposing team's Nexus, a central structure that houses the game's boss character, the “Nefarian.” Players use various champions, each with unique abilities and roles, to gain advantages and contribute to their team's success. The game is played on a map with multiple objectives, such as destroying enemy structures, capturing and defending areas, and engaging in teamfights. League of Legends is known for its strategic gameplay, skill-based mechanics, and competitive nature, making it a popular and engaging game for bot.
LAMM-InstructData

LAMM-Benchmark

LAMM-Framework

14 Datasets

70k <Image/PC, Instruction, Response>

12+ Tasks

6 4*A100 GPU Hours

180k <Image, Instruction, Response>

10k <PC, Instruction, Response>

LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark, 2023 June
**Instruction:** How many animals are depicted in the image?
**Response:** In the image, there are several animals depicted, including a large elephant and a group of antelope.

**Instruction:** What are the different animal species visible in the image?
**Response:** The image shows two distinct animal species - an elephant and several antelope.

**Instruction:** How can we identify the elephant in the image?
**Response:** The elephant in the image is identifiable by its large size, gray skin, long tusks, big ears, and a long trunk. These are the distinctive features of an elephant.

**Instruction:** How are the elephant and the antelope interacting in the image?
**Response:** The elephant appears to be walking peacefully among the antelope, suggesting a peaceful coexistence between these different species in their natural habitat.

**Instruction:** What can we infer from the image about the behavior of these animals?
**Response:** The image shows an elephant walking calmly among a group of antelope. This suggests that these animals are used to sharing the same habitat and living peacefully together. The absence of any visible signs of distress or aggression also indicates that there isn't any immediate threat present, and the animals feel safe in each other's presence.
Instruction: <image> Write a detailed description of the given image.

Response: The image shows a man standing on a street corner with a cart, selling unique umbrella-style hats. There are colorful umbrellas attached to poles sticking out of a basket next to the man, drawing the attention of passersby. Five umbrellas of varying sizes can be seen displayed, with one large umbrella placed in the foreground. Behind the man, two cars are parked along the side of the street, and another person is visible walking near the center of the scene. The man selling umbrella-style hats appears to be attracting interest or potentially waiting for customers to approach his makeshift street shop.

Instruction: <image> Portray the image with a rich, descriptive narrative.

Response: The image depicts an intriguing outdoor scene on a city sidewalk. A man is seated comfortably on a chair, overseeing his eclectic array of merchandise for sale. A conspicuous part of his offerings include umbrella-style hats and sunglasses, displayed attractively on mannequin heads to attract potential buyers. The man’s temporary storefront is situated near a corner of the street, where he seems to be ready to engage with passers-by. A collection of green-leaved trees offers a pleasant shade and contributes to the street's overall ambiance. The surroundings reveal an urban setting with multiple buildings constructed of red bricks and adorned with various windows. One building, in particular, has a sign indicating that it's a dental office, adding a touch of everyday life to the picture. A manhole cover can be seen on the sidewalk, suggesting a well-maintained urban infrastructure. Despite being an unusual sight, the man selling umbrella-style hats and sunglasses provides a touch of local color and personality to the city street corner. His presence adds a unique charm to the overall cityscape and hints at the diverse street culture in the area.
**Instruction:** What type of building is shown in the image?
**Response:** The building shown in the image is a monastery, which is the residence of a religious community.

**Instruction:** Can you describe the architectural features of the monastery in the image?
**Response:** While specific architectural features can vary widely between different monasteries, many share common elements such as a central courtyard or cloister, a chapel or church, dormitory-style living quarters, and communal spaces like a refectory for communal meals. The image shows a building that has a sturdy, typically stone-built structure designed for communal living and religious activities.

**Instruction:** What is the purpose of a monastery and how does its design reflect this?
**Response:** A monastery is a building or complex of buildings where monks or nuns live in a community. The primary purpose of a monastery is to support the religious work of its residents, whether that be prayer, meditation, or other forms of spiritual practice. As such, the design of a monastery often reflects these needs, with spaces dedicated to worship, study, work, and communal living. It's intended to be self-sufficient and promote a peaceful, contemplative environment.

**Instruction:** Can you accurately report the names and positions of all the objects visible in the image?
**Response:** The microphone object is present at the [0.49, 0.38, 0.56, 0.43] coordinate in the image. An object that can be classified as footwear is located at the [0.5, 0.77, 0.55, 0.84] position of the image. Upon close inspection of the image, it can be observed that there is an object positioned at [0.56, 0.78, 0.61, 0.85] that belongs to the footwear category. The image contains an object that can be classified as person and is located at the [0.48, 0.31, 0.78, 0.84] position.

**Instruction:** Please identify the number of objects in the image that are labeled as a certain category.
**Response:** The system has detected 30 Ice cream present in the image.

**Instruction:** what large name is on the framed picture on the shelf?
**Response:** A careful examination of the picture indicates that the correct answer is kasper.
**n-round Daily Dialogue**

**Instruction:** <point cloud> Is there a TV cabinet in the room?
**Response:** Yes. The point cloud describes a full-sized corner cabinet with a large opening in the center for a television.

**Instruction:** <point cloud> Where is the desk located in the room?
**Response:** The desk is located in front of a black office chair with wheels.

**1-round Detailed Description**

**Instruction:** <point cloud> What is this point cloud about?
**Response:** The point cloud shows a living space with a large gray L-shaped couch positioned in the middle of the room. The couch is adjacent to the kitchen area and across from the bedroom area, providing a cozy and functional gathering space....

**1-round Visual Task Dialogue**

**Instruction:** <point cloud> Provide a detailed account of the objects observed in the point cloud, including their precise locations?
**Response:** An object within the table category can be observed at the [0.070, 3.632, 0.624, 1.110, 1.976, 1.083] position in the point cloud. Positioned at the [-1.705, 0.211, 0.417, 0.611, 9.417, 1.632] location, an object within the window category...
You are an AI visual assistant...
C1. You need to generate a conversation based on Bamboo label and Wikipedia description...
C2. You need to generate multiple rounds of conversations based on Bamboo label and Wikipedia description...
C3. You need to describe the scene in detail...
C4. Give me 20 similar instruction-response prompts...

Human-labeled <instruction, response> pairs

C1. <Caption> + <Boxes> + <Relationships>
C2. <Bamboo labels> + <Wikipedia description>
C3. <Caption> + <Boxes> + <Relationships>
C4. N/A

Instructions: [What are the names of the objects present in the image, and where are they positioned? How many items in the image belong to a particular category? ...]
Responses: [There is an object at the {P} position of the image, and its category is {C}. There are {N} {C} visible in image ...]

Object Detection
Image Classification
Keypoint Detection
Object Counting
OCR
3D indoor Detection

C1: n-round Daily Dialogue
C2: n-round Factual Knowledge Dialogue
C3: 1-round Detailed Description
C4: 1-round Visual Task Dialogue

Dataset:
Bamboo CLS
Bamboo DET
COCO Captions
COCO DET
COCO Keypoint
Visual Genome
Locount
Text VQA
3RScan
CLEVR3D
3DSSG
ShapeNet
I. Traditional Metrics

<table>
<thead>
<tr>
<th>Task Definition</th>
<th>GT</th>
<th>Eval-Instruct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Accuracy</td>
<td>1. Model A</td>
</tr>
<tr>
<td>Detection</td>
<td>mAP</td>
<td>3. Model C</td>
</tr>
<tr>
<td>Caption</td>
<td>BLEU</td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td>MAE</td>
<td></td>
</tr>
<tr>
<td>OCR</td>
<td>Word ACC</td>
<td></td>
</tr>
</tbody>
</table>

II. Binary-Loc. Metric

Inside GT bbox
- Positive

Outside GT bbox
- Negative

Response: It is located at (0.790, 1.477, 0.219, 0.434, 0.687, 0.3667).

III. GPT Metric

Model A Response
- GPT Ranking
1. Model C
2. Model A
3. Model B
**Instruction**

What's the type of room in the <point cloud>?

**Response**

The room shown can be a bathroom with a separate toilet...

**Instruction**

How many fruits are there in the <image> and what are they?

**Response**

There’re 15 fruits. There are 3 apples, 2 pears and...
<table>
<thead>
<tr>
<th>IDs</th>
<th>Task</th>
<th>Dataset</th>
<th>Metric</th>
<th>SOTA</th>
<th>LLaVA</th>
<th>MiniGPT4</th>
<th>mPLUG</th>
<th>LAMM</th>
<th>LAMM-FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Classification</td>
<td>CIFAR10</td>
<td>Acc ↑</td>
<td>99.5</td>
<td>60.83</td>
<td>46.22</td>
<td>42.5</td>
<td>34.5</td>
<td>91.2</td>
</tr>
<tr>
<td>2</td>
<td>Detection</td>
<td>VOC2012</td>
<td>mAP ↑</td>
<td>97.2</td>
<td>1.42</td>
<td>0.92</td>
<td>0.158</td>
<td>4.82</td>
<td>13.48</td>
</tr>
<tr>
<td>3</td>
<td>VQA</td>
<td>SQAImage</td>
<td>Acc ↑</td>
<td>92.53</td>
<td>40.5</td>
<td>43.43</td>
<td>36.39</td>
<td>47.15</td>
<td>74.27</td>
</tr>
<tr>
<td>4</td>
<td>Image Caption</td>
<td>flickr30k</td>
<td>BLEU4 ↑</td>
<td>30.1</td>
<td>6.65</td>
<td>5.1</td>
<td>2.74</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>F-g classification</td>
<td>UCMerced</td>
<td>Acc ↑</td>
<td>100</td>
<td>47</td>
<td>33.6</td>
<td>32.5</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Counting</td>
<td>FSC147</td>
<td>MAE ↑</td>
<td>10.79</td>
<td>56.2</td>
<td>Failed</td>
<td>60.67</td>
<td>53.97</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>OCR</td>
<td>SVT</td>
<td>Word Acc↑</td>
<td>97.9</td>
<td>37.78</td>
<td>16.97</td>
<td>30.39</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Facial Classification</td>
<td>CelebA(Smile)</td>
<td>Acc ↑</td>
<td>N/A</td>
<td>Failed</td>
<td>66.36</td>
<td>Failed</td>
<td>51.3</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Keypoints Detection</td>
<td>LSP</td>
<td>PCK ↑</td>
<td>99.5</td>
<td>46.42</td>
<td>43.47</td>
<td>40.93</td>
<td>30.48</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Binary Detection</td>
<td>LAMM</td>
<td>Acc ↑</td>
<td>N/A</td>
<td>14.73</td>
<td>13.12</td>
<td>4.42</td>
<td>31.2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>GPT Evaluation</td>
<td>LAMM</td>
<td>Score ↑</td>
<td>N/A</td>
<td>11</td>
<td>N/A</td>
<td>N/A</td>
<td>89</td>
<td></td>
</tr>
</tbody>
</table>

**Instruction:** What is the category for this image?

**Ground Truth:** car

**LAMM Response:** Fiat 500L 2012
**LLaVA Response:** Car

=CIFAR10-Car Fiat 500L 2012

Fine-Grained Recognition Capability
<table>
<thead>
<tr>
<th>IDs</th>
<th>Task</th>
<th>Dataset</th>
<th>Metric</th>
<th>SOTA</th>
<th>LLaVA</th>
<th>MiniGPT4</th>
<th>mPLUG</th>
<th>LAMM</th>
<th>LAMM-FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Classification</td>
<td>CIFAR10</td>
<td>Acc ↑</td>
<td>99.5</td>
<td>60.83</td>
<td>46.22</td>
<td>42.5</td>
<td>34.5</td>
<td>91.2</td>
</tr>
<tr>
<td>2</td>
<td>Detection</td>
<td>VOC2012</td>
<td>mAP ↑</td>
<td>97.2</td>
<td>1.42</td>
<td>0.92</td>
<td>0.156</td>
<td>4.82</td>
<td>13.48</td>
</tr>
<tr>
<td>3</td>
<td>VQA</td>
<td>SQAIimage</td>
<td>Acc ↑</td>
<td>92.53</td>
<td>40.5</td>
<td>43.43</td>
<td>36.39</td>
<td>47.15</td>
<td>74.27</td>
</tr>
<tr>
<td>4</td>
<td>Image Caption</td>
<td>flickr30k</td>
<td>BLEU4 ↑</td>
<td>30.1</td>
<td>6.65</td>
<td>5.1</td>
<td>2.74</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>F-g classification</td>
<td>UCMerced</td>
<td>Acc ↑</td>
<td>100</td>
<td>47</td>
<td>33.6</td>
<td>32.5</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Counting</td>
<td>FSC147</td>
<td>MAE ↑</td>
<td>10.79</td>
<td>56.2</td>
<td>Failed</td>
<td>60.67</td>
<td>53.97</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>OCR</td>
<td>SVT</td>
<td>Word Acc ↑</td>
<td>97.9</td>
<td>37.78</td>
<td>16.97</td>
<td>30.39</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Facial Classification</td>
<td>CelebA(Smile)</td>
<td>Acc ↑</td>
<td>N/A</td>
<td>Failed</td>
<td>66.36</td>
<td>Failed</td>
<td>51.3</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Keypoints Detection</td>
<td>LSP</td>
<td>PCK ↑</td>
<td>99.5</td>
<td>46.42</td>
<td>43.47</td>
<td>40.93</td>
<td>30.48</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Binary Localization</td>
<td>LAMM</td>
<td>Acc ↑</td>
<td>N/A</td>
<td>14.73</td>
<td>13.12</td>
<td>4.42</td>
<td>31.2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>GPT Evaluation</td>
<td>LAMM</td>
<td>Score ↑</td>
<td>N/A</td>
<td>11</td>
<td>N/A</td>
<td>N/A</td>
<td>89</td>
<td></td>
</tr>
</tbody>
</table>

**Instruction:** Give the bounding box of all the objects in the image.
**LLaVA Response:** There is a white horse grazing in a fenced-in area with grass. The horse is positioned near the left side of the image. To provide the precise coordinates, the horse's bounding box would be (0.156, 0.477, 0.767, 0.523).

**Instruction:** Give the position of all the objects in the image.
**LLaVA Response:** In the image, there is a white horse grazing on grass near a wall. The horse's position is approximately [0.5, 0.5].
<table>
<thead>
<tr>
<th>IDs</th>
<th>Task</th>
<th>Dataset</th>
<th>Metric</th>
<th>SOTA</th>
<th>LLaVA</th>
<th>MiniGPT4</th>
<th>mPLUG</th>
<th>LAMM</th>
<th>LAMM-FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Classification</td>
<td>CIFAR10</td>
<td>Acc</td>
<td>99.5</td>
<td>60.83</td>
<td>46.22</td>
<td>42.5</td>
<td>34.5</td>
<td>91.2</td>
</tr>
<tr>
<td>2</td>
<td>Detection</td>
<td>VOC2012</td>
<td>mAP</td>
<td>97.2</td>
<td>1.42</td>
<td>0.92</td>
<td>0.158</td>
<td>4.82</td>
<td>13.48</td>
</tr>
<tr>
<td>3</td>
<td>VQA</td>
<td>SQAImage</td>
<td>Acc</td>
<td>92.53</td>
<td>40.5</td>
<td>43.43</td>
<td>36.39</td>
<td>47.15</td>
<td>74.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AI2D</td>
<td>Acc</td>
<td>N/A</td>
<td>18.13</td>
<td>Failed</td>
<td></td>
<td>19.31</td>
<td>19.5</td>
</tr>
<tr>
<td>4</td>
<td>Image Caption</td>
<td>flickr30k</td>
<td>BLEU4</td>
<td>30.1</td>
<td>6.65</td>
<td>5.1</td>
<td>2.74</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>F-g classification</td>
<td>UCMerced</td>
<td>Acc</td>
<td>100</td>
<td>47</td>
<td>33.6</td>
<td>32.5</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Counting</td>
<td>FSC147</td>
<td>MAE</td>
<td>10.79</td>
<td>56.2</td>
<td>Failed</td>
<td>60.67</td>
<td>53.97</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SVT</td>
<td>Word Acc</td>
<td>97.9</td>
<td>37.78</td>
<td>16.97</td>
<td>30.39</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>OCR</td>
<td>CelebA(Smile)</td>
<td>Acc</td>
<td>N/A</td>
<td>Failed</td>
<td>66.36</td>
<td>Failed</td>
<td>51.3</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Facial Classification</td>
<td>CelebA(Hair)</td>
<td>Acc</td>
<td>N/A</td>
<td>Failed</td>
<td>46.42</td>
<td>43.47</td>
<td>40.93</td>
<td>30.48</td>
</tr>
<tr>
<td>9</td>
<td>Keypoints Detection</td>
<td>LSP</td>
<td>PCK</td>
<td>99.5</td>
<td>Failed</td>
<td>Failed</td>
<td>Failed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Binary Localization</td>
<td>LAMM</td>
<td>Acc</td>
<td>N/A</td>
<td>14.73</td>
<td>13.12</td>
<td>4.42</td>
<td>31.2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>GPT Evaluation</td>
<td>LAMM</td>
<td>Score</td>
<td>N/A</td>
<td>11</td>
<td>N/A</td>
<td>N/A</td>
<td>89</td>
<td></td>
</tr>
</tbody>
</table>

**LLaVA Response:** The image shows a group of people playing a game of volleyball on a beach. They are actively engaged in the game, with one man jumping in the air to hit the ball. The scene is lively and energetic, capturing the spirit of outdoor sports and beach activities.

**LAMM Response:** The image captures an engaging scene from a professional sand volleyball game. Two men can be seen in the center of the court, fully immersed in the game. Both players are dressed in casual sportswear, with one wearing a white shirt and the other in a shirt of the same color. The sand court stretches across the image, providing ample space for the players to move around. There are multiple flags...
Insights for Future

Search Engine

**Instruction:** `<action>` Based on my previous trip to Hawaii, how should I arrange my trip to Phuket?
**Response:** I recommend you to book a Marriott and then...

Autonomous Driving

**Instruction:** `<image> <pcloud>` Drive me efficiently to work and park underground.
**Response:** OK. We will take Rd. 66 and then...

Human-Centric Applications

**Instruction:** `<image> <mesh>` Generate a motion of 3D digital humans while keeping the identification of the person in `<image>` but changing his clothes by `<image>` and acting as shooting the basketball.
**Response:** OK. Shooting the basketball should first jump and raise his right arm `<mesh>`....

Interactive Image Synthesis

**Instruction:** `<action> <image>` Draw a herd of running house...
**Response:** Here are a few images you might like... `<image>`

Video Applications

**Instruction:** Here is a real `<video>`, transfer its style to make an animation of 2mins, 30FPS...
**Response:** Here is the animation `<video>`, it tells a funny story that...

Low Level Applications

**Instruction:** `<image>` Denoise the image, remove the raindrop in the background, super resolution to 8K.
**Response:** Here are a few images... `<image>`
Hi3D
"From Geometry to Texture: A Hierarchical Framework for Efficient Text-to-3D Generation"

- **Efficient**: $10 \times 7 \times 4$ speed up
  - DreamFusion/Magic3D/Fantasia3D

- **Geometric Stable**
  - DreamFusion

- **Controllable**
  - DreamFusion

UniG3D
"UniG3D: A Unified 3D Object Generation Dataset"

- **Comprehensive data format**
  - <Text,3D-PCL,3D-Mesh,2D>

- **Unified Pipeline**
  - Adapt to any 3D dataset

- **Large-scale**
  - <550K,550K,5.5M,11M>

Otter+MIMIC-IT
"MIMIC-IT: Multi-Modal In-Context Instruction Tuning"
Thank You