ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240541

u-LLaVA: Unifying Multi-Modal Tasks via Large Language Model

Jinjin Xu[a], Liwu Xu[a], Yuzhe Yang[a], Xiang Li[a], Fanyi Wang[a], Yanchun Xie[a], Yi-Jie Huang[a] and Yaqian Li[a,*]

[a]OPPO AI Center

Abstract. Recent advancements in multi-modal large language models (MLLMs) have led to substantial improvements in visual understanding, primarily driven by sophisticated modality alignment strategies. However, predominant approaches prioritize global or regional comprehension, with less focus on fine-grained, pixel-level tasks. To address this gap, we introduce u-LLaVA, an innovative unifying multi-task framework that integrates pixel, regional, and global features to refine the perceptual faculties of MLLMs. We commence by leveraging an efficient modality alignment approach, harnessing both image and video datasets to bolster the model's foundational understanding across diverse visual contexts. Subsequently, a joint instruction tuning method with task-specific projectors and decoders for end-to-end downstream training is presented. Furthermore, this work contributes a novel mask-based multi-task dataset comprising 277K samples, crafted to challenge and assess the fine-grained perception capabilities of MLLMs. The overall framework is simple, effective, and achieves state-of-the-art performance across multiple benchmarks. We make model, data, and code publicly accessible at https://github.com/OPPOMKLab/u-LLaVA.

1 Introduction

Owing to the intrinsic difficulties associated with feature extraction in computer vision (CV) tasks, researchers have predominantly focused on perception rather than cognition over an extended duration. This emphasis bears a substantial impact on the development and understanding of various CV methodologies [29]. Although the development of deep neural networks and pre-training techniques has significantly reduced the difficulty of perception, it remains challenging to achieve homogeneity across downstream tasks due to substantial differences in their respective objectives. Recently, causal large language models such as GPT [35, 36, 5], Gemini [44] and LLaMA [46] have reached or come close to human-level performance on a variety of tasks. These advancements have also motivated researchers to incorporate LLMs as components [25, 66] or core elements [44, 50] in visual tasks, leading to the development of visual language models (VLMs), or multi-modal large language models (MLLMs). As a result, these methods have garnered increasing attention in recent times.

Typically, a multi-modal LLM consists of one or multiple encoders to extract features, paired with suitable mapping components (such as MLP [25], Q-Former[66], or cross-attention [2]), to align the other modalities with the textual domain. In comparison to the impressive performance of MLLMs on general-purpose understanding tasks, such as visual question answering (VQA), their capabilities in regional and pixel-level tasks are somewhat less remarkable [25, 66]. To achieve regional-level understanding, it is usual to convert target coordinates into tokens for causal language modeling, such as Shikra [7] and KOSMOS-2 [34]. To further realize pixel-level understanding, mask-level decoders or extractors are introduced, such as LISA [17], Osprey [60] and Next-Chat [61]. However, such region comprehension requires extensive data for training, entailing high costs. Pixel-level understanding methods offer more flexibility, but entail introducing or designing specific segmentation modules.

Table 1: Comparison of comprehension levels supported by existing MLLMs.

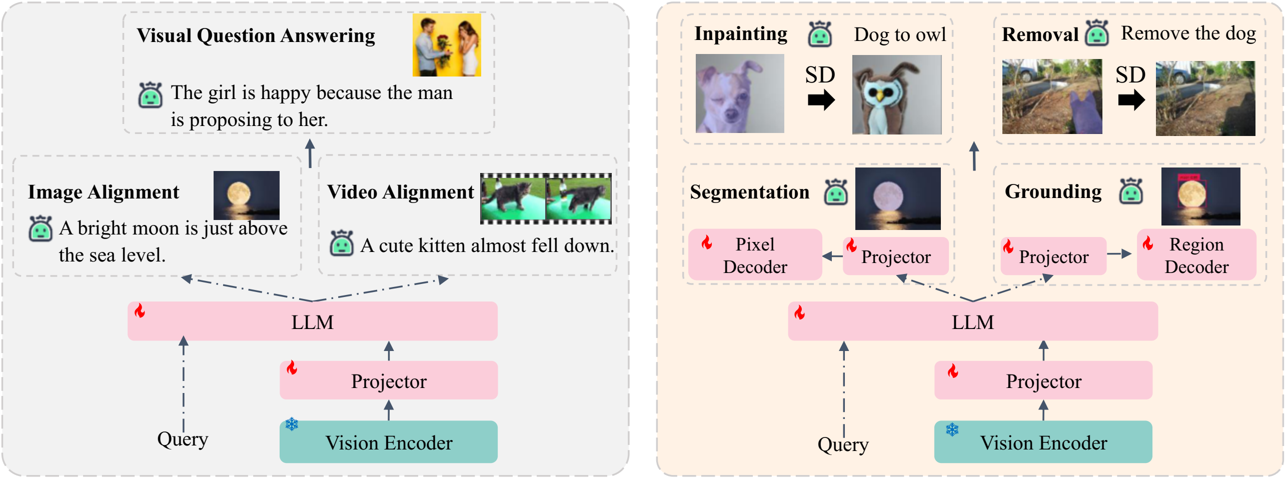
Methods	Image	Video	Region	Pixel
LLaVA [25]	✓	✗	✗	✗
MiniGPT-4 [66]	✓	✗	✗	✗
Video-LLaMA [62]	✓	✓	✗	✗
Video-ChatGPT [31]	✗	✓	✗	✗
Shikra [7]	✓	✗	✓	✗
CogVLM [50]	✓	✗	✓	✗
LISA [17]	✓	✗	✗	✓
u-LLaVA (ours)	✓	✓	✓	✓

In this paper, we propose u-LLaVA, a novel approach to enhance the general, region and even pixel-level comprehension abilities of MLLMs. To this end, we first design a efficient visual alignment strategy with image and spatio-temporal representations, and task-specific projectors and decoders are integrated for joint instruction tuning. The overall pipeline is illustrated in Figure 1.

To enable pixel-level understanding, we employ a projector to connect MLLMs and SAM [17], achieving two goals: a) imparting semantic understanding capacity to SAM by leveraging the world knowledge inherent in LLM; and b) enhancing the pixel-level understanding ability of LLM by harnessing SAM. To enhance the performance of regional-level comprehension, we introduced a independent location decoder to decode target coordinates from the hidden state or output of MLLMs, which greatly reduces the amount of data required for training. To accommodate the training of the aforementioned models, we have carefully designed a series of task-related prompt pools, and introduced a mask-based, region-specific dataset, namely ullava-277K. Most of the data was collected from publicly available datasets, with missing annotations carefully supplemented by the GPT-3.5.

Contributions can be summarized in three folds:

[*] Corresponding Author. Email: liyaqian@oppo.com.



(a) Stage I

(b) Stage II

Figure 1: Overview of u-LLaVA. In stage I, image and spatio-temporal features are used to efficiently boost the general-purpose modality alignment. In Stage II, task-specific projectors and decoders are jointly trained for region and pixel-level understanding. Further, additional modules such as stable diffusion [39] can be easily patched for downstream tasks.

- We propose a efficient visual alignment method for multi-modal pre-training, which leverages image (spatio features) and video (spatio-temporal features) to enhance the perceptual faculties of MLLMs.
- We first-time introduce joint instruction tuning approach in the same stage to enable multi-level understanding with task-specific projectors and decoders, see Table 1 for details.
- We release the joint instruction tuning dataset, ullava-277K, the model, and the code publicly available. Additionally, we conduct comprehensive experiments and demonstrate the effectiveness of the proposed method.

2 Related Work

2.1 MLLMs

Surprised by the remarkable abilities of large language models, researchers have shown great interest in transferring the capabilities of LLM to other domains [58, 53]. In recent months, remarkable progress has been made in this field, such as LLaVA [25], MiniGPT-4 [66], Otter [19], KOSMOS-1/2 [12, 34], mPLUG-owl [57] and Flamingo [1], etc. While having demonstrated impressive performance in image-level understanding, these methods show limited capabilities on pixel or region level tasks.

2.2 Region-Level Comprehension

Referring expression comprehension (REC) is one of the most typical region-Level comprehension tasks, and RefCOCO [59], RefCOCO+ [59] and RefCOCOg [32], RefCLEF [15] are popular datasets for REC. Recently, some methods have employed the pix2seq approach to achieve regional understanding [7, 34]. Some strategies further incorporate regional encoding-decoding [61], while others utilize external modules to complete the task [63].

2.3 Pixel-Level Understanding

The advent of MLLMs has reduced the difficulty of subjective visual tasks, but progress on mask-aware tasks, such as referring expression segmentation (RES) and salient object segmentation, has been relatively slow due to the difficulty in designing pixel-level tokens. The prevalent methods currently involve using the output of grounding as the input for SAM [61], or utilizing a specific decoder for end-to-end training [17, 60].

3 Methods

The overall framework of u-LLaVA is presented in Figure 1. As we can see, u-LLaVA is a multi-modal multitask chatbot that takes text, images, and videos as inputs. It achieves this by unifying the representation space of visual and textual elements at stage I, and understanding region and pixel representations jointly at stage II. In this part, we will first introduce the model architecture and modality alignment strategy in Section 3.1, followed by a discussion on the joint instruction tuning process in Section 3.2. Lastly, we will present dataset construction methods.

3.1 Efficient Visual Alignment

To align representations among different modalities, the projector-based structure is adopted in this work: the pre-trained CLIP ViT-L/14 [37] and a visual projector are combined to encode image inputs, while the Vicuna [8] is employed as the cognitive module. In addition, u-LLaVA supports video modality by concatenating spatial and temporal representations, requiring only the addition of two special video tokens and a minimal amount of trainable parameters.

Table 2: Special tokens for modality and task expressions, where T denotes the number of frames and is set 8 in this work.

	Image	Video	Tag	Region	Pixel
Begin	<img_beg>	<vid_beg>	<tag>	<LOC>	<SEG>
Patches	<img_patch>	<vid_patch>	/	/	/
End	</img_end>	</vid_end>	</tag>	/	/
Special token length	256	256+T	1	1	1

Generally, maximizing the likelihood function below to align the representation spaces of image/video and text is a widely-used approach for pre-training [25]. For a given image or video embeddings $\boldsymbol{x}_e$, and a conversation list of L tokens $\boldsymbol{x}_t = \{x_t^1, x_t^2, ..., x_t^L\}$, we have the following training objectives, called ***coarse-grained loss***:

$$L_{cgl} = \sum_i \log P(x_i | \boldsymbol{x}_e, x_{i-k}, ..., x_{i-1}; \boldsymbol{\theta}), \tag{1}$$

where, in accordance with [35], k, P, and $\boldsymbol{\theta}$ are the size of context window, the conditional probability, and network parameters, respectively.

3.2 Joint Instruction Tuning

Visual instruction tuning is a common strategy for MLLM fine-tuning, but most methods only include one or two out of general, region, and pixel-level aspects during the training phase. In this work, we first-time jointly involve general, region, and pixel features in the same tuning stage.

General-aware tuning: In this part, there are no adjustments made to the model structure. However, unlike the first stage, we emphasize the use of multi-turn dialogues and complex reasoning datasets to further enhance the model's understanding capabilities. The special tokens used in this work are listed in Table 2.

Mask-aware tuning: Inspired by LISA [17], we employ a projector to map the hidden states of the mask-related special tokens and then incorporate them into SAM decoder as the text embeddings to facilitate pixel-level understanding. We use a projector to connect SAM and MLLM, and map the mask-related hidden states as the text embedding for SAM. This endows SAM with semantic perception capabilities while achieving pixel-level perception for MLLM.

Region-aware tuning: Similar to pixel perception, we utilize a projector and a location decoder, mapping the hidden states of location-related special tokens directly to the target coordinates, in which the decoder consists of a randomly initialized MLP. To enhance the data volume and improve the decoder's performance, we convert the segmentation annotations into bounding boxes for samples that lack detection annotations.

In general, we have the following training objective, namely ***fine-grained loss***:

$$L_{fgl} = L_{cgl} + \begin{cases} L_{pixel}, & \text{if mask exists} \\ L_{region}, & \text{if bbox exists} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where the term $L_{pixel} = \alpha_1 L_{bce} + \alpha_2 L_{dice}$ represents the mask prediction loss, and $L_{region} = \beta_1 L_1 + \beta_2 L_{giou}$ is the prediction loss for the target bounding box. The values of α_1, α_2, β_1, and β_2 are set to 2.0, 0.5, 1.0, and 1.0, respectively.

3.3 Dataset Construction

To accommodate the training of the aforementioned models, we reorganize or rebuild various types of public datasets, details are summarized in Table 3.

Table 3: Construction of the training datasets. The color blue indicates that the dataset is utilized in Stage I, while yellow signifies its usage in Stage II. Where annotations

	Dataset	Images/Videos	Annotations
Visual Captioning	LLaVA CC3M [25]	595,375	595,375
	TGIF [21]	125,782	125,782
VQA	LLaVA-Instruction-mix [25]	349,034	624,610
	ALLaVA-4V-Instruction [6]	643,315	692,097
	DVQA [14]	10,000	10,000
	DocVQA [45]	9,911	9,911
	AI2D [25]	4,060	4,060
RES	RefCOCO [59]	16,994	120,624
	RefCOCO+ [59]	16,992	120,191
	RefCOCOg [32]	21,899	80,512
	RefCLEF [15]	17,978	108,652
Semantic Segmentation	COCO-Stuff [37]	118,205	742,787
	VOC2010 [9]	4,366	81,139
	PACO-LVIS [38]	45,790	612,188
	ADE20K [64]	20,196	165,120
Salient-15K	MSRA-10K [10]	10,000	10,000
	MSRA-B [47]	5,000	5,000

As for the referring and semantic segmentation datasets, all references or semantic labels are extracted and then formed with the given templates. However, salient object detection/segmentation datasets usually lack descriptions of the target objects. To address this issue, we employ mask information to extract the primary objects from images within MSRA-10K [10] and MSRA-B [47]. The extracted objects are then fed into BLIP2 [20] to generate descriptions solely for the objects. Lastly, GPT-4o is used to phrase the object tags from the generated description, followed by the integration of predefined templates to complete the reconstruction process. We refer to the reconstructed salient instruction dataset as Salient-15K for short. The template examples and construction process of Salient-15K are summarized in Appendix.

4 Experiments

4.1 Implementation Details

All experiments are conducted with 8 NVIDIA Tesla A100 80G GPUs and Pytorch framework [33]. Vicuna v1.1 [8] and CLIP ViT-L/14 [37] are set to the foundational language model and image encoder. For the region understanding task, the projector and decoder are implemented using two MLPs. Specifically, the projector is configured with layers of [4096->4096, 4096->256], while the decoder comprises layers of [256->256, 256->128, 128->4]. For pixel understanding, a two-layer MLP is used as the projector with layers of [4096->4096, 4096->256]. The decoder is implemented using the off-the-shelf SAM ViT-H [16]. For alignment and instruction training, AdamW is utilized as the optimizer with a weight decay of 0. The learning rate is set to 2e-3 and 2e-5 (2e-4 if LoRA [11] is employed). The batch size per device is configured to 48 and 16 (32 if LoRA), with a gradient accumulation step of 1. Additionally, the token length is set to 1024 and 512. Under the above settings, each training step requires approximately 7s and 5s (9.5s if LoRA), with BF16 and DeepSpeed ZeRO-2 enabled.

4.2 Evaluation Metrics

We follow the previous works [23, 17] to validate the quantitative performance of the proposed algorithm, with details as follows:

Pixel Segmentation: Cumulative-IoU (cIoU) is a widely-used performance indicator in segmentation tasks, which calculates the total intersection pixels over the total union pixels. In some works, it is also referred to as the overall-IoU (oIoU), as seen in [56, 54].

Region Grounding: The percentage of samples with IoU higher than a threshold X is a commonly used metric in visual grounding

Table 4: RES results with cIoU indicator. Specialists represent models that are specifically designed for CV tasks. Where * in MLLMs denotes using LoRA [11] for parameter efficient training. The top 2 results are outlined in **bold** and with underline.

Type	Method	Segmentation Masks	RefCOCO			RefCOCO+			RefCOCOg	
			val	test A	test B	val	test A	test B	val	test
Specialists	LAVT [56]	0.03M	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09
	X-Decoder(L) [67]	0.12M	-	-	-	-	-	-	64.60	-
	ReLA [23]	-	73.82	76.48	70.18	66.04	71.02	57.65	65.00	65.97
	SEEM(B) [68]	0.12M	-	-	-	-	-	-	65.00	-
	SEEM(L) [68]	0.12M	-	-	-	-	-	-	65.60	-
	PolyFormer(B) [26]	0.16M	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05
	PolyFormer(L) [26]	0.16M	75.96	78.29	73.25	69.33	74.56	61.87	69.20	70.19
	UNINEXT(L) [54]	3M	80.32	82.61	77.76	70.04	74.91	62.57	73.41	73.68
	UNINEXT(H) [54]	3M	**82.19**	**83.44**	**81.33**	**72.47**	**76.42**	**66.22**	**74.67**	**76.37**
MLLMs	LISA-7B* [17]	∼ 0.80M	74.10	76.50	71.10	62.40	67.40	56.50	66.40	68.50
	LISA-7B* (ft) [17]	-	74.90	79.10	72.30	65.10	70.80	58.10	67.90	70.60
	NExT-Chat-7B [61]	0.15M	74.70	78.90	69.50	65.10	71.90	56.70	67.00	67.00
	u-LLaVA-7B* (Ours)	∼ 0.66M	81.11	82.98	77.63	71.36	76.88	65.66	73.45	74.65
	u-LLaVA-7B (Ours)	∼ 0.66M	**83.00**	**85.09**	**80.51**	**77.10**	**81.68**	**70.56**	**77.14**	**77.97**

tasks, denoted as Precision@X (Prec@X). In this work, we set the threshold to 0.5 according to [7].

4.3 Pixel-Level Understanding Performance

To demonstrate the performance of the proposed method on pixel-level understanding, we conduct experiments on widely-used RES benchmarks, RefCOCO, RefCOCO+, and RefCOCOg. The comparison is made between existing state-of-the-art (SOTA) specialist models and MLLMs with cIoU indicator, as presented in Table 4.

As can be seen from the table, even with LoRA, our method still achieves the best results among the MLLMs methods. More notably, u-LLaVA-7B surpasses the performance of the prevailing state-of-the-art MLLM method, LISA-7B*(ft), achieving an average improvement of 9.28 in the cIoU indicator. It is also noteworthy that u-LLaVA surpasses the performance of the current leading expert model, UNINEXT(H) [54], on the three benchmarks, all while utilizing merely a tenth of the training data. These findings serve as a testament to the efficacy of LLM in tasks that necessitate comprehension-based capabilities.

4.4 Pixel-level Intent Understanding Performance

We further examine the zero-shot performance of u-LLaVA in widely recognized salient segmentation datasets to clarify the superiority of MLLMs in comprehending human subjective intentions.

Here, DUT-OMRON [55] (5,168 test images), DUTS-TE [48] (5,019 test images), and ECSSD [40] (1000 test images) datasets are selected for validation. To ensure fairness, we draw parallels between our method and a range of other previously conducted unsupervised algorithms. As summarized in Table 5, u-LLaVA outperforms the rest, achieving SOTA performance across all three benchmarks, further solidifying the effectiveness and superiority of our method.

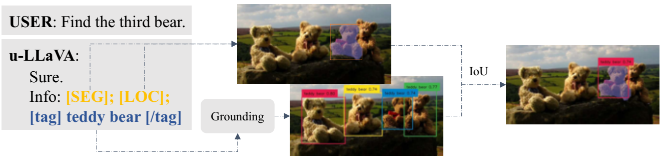
4.5 Region-Level Understanding Performance

In this section, we conduct a comparative analysis to evaluate u-LLaVA's performance against other 7B MLLM models in the context of region-level understanding tasks, using the REC task as the benchmark.

Table 5: Salient segmentation results on salient object detection benchmarks among different methods, where † denotes the method with Bilateral solver [3], and cIoU is adopted as the metric.

Type	Method	DUT-OMRON	DUTS-TE	ECSSD
Specialists	LOST† [42]	48.90	57.20	72.30
	TokenCut† [52]	61.80	62.40	77.20
	SELFMASK† [41]	**65.50**	66.60	**81.80**
	MOVE † [4]	63.60	**68.70**	80.10
MLLMs	u-LLaVA-7B* (Ours)	72.10	74.64	89.34
	u-LLaVA-7B (Ours)	**74.19**	**75.97**	**90.83**

It should be highlighted that we incorporate all intermediate results of our model, including the output of the region decoder and the generated mask, to generate the final regression box thus optimizing the performance. One can further utilize the off-the-shelf grounding model and the generated tags for redundancy, as encapsulated in Figure 2, and the corresponding experimental results are summarized in Table 6. Observably, u-LLaVA outperforms other MLLMs such as Shikra [7], while utilizing a mere one-tenth of the data. However, there exists a discernible performance gap when compared to expert models such as UNINEXT(H), it is essential to consider that this is influenced by multiple factors, including, but not limited to, input resolution and task interference.



USER: Find the third bear.

u-LLaVA:
Sure.
Info: [SEG]; [LOC];
[tag] teddy bear [/tag]

Figure 2: The inference workflow of u-LLaVA for region-level comprehension tasks.

4.6 General Benchmarks

In Table 7, we present a comparison of our model, u-LLaVA, with popular 7B MLLMs across several multi-modal benchmarks, including MMBench-Dev/Test [28], TextVQA [43], GQA [13], ScienceQA-IMG [30], and RefCOCO val. Notably, we enlarge the

Table 6: Comparative experiments of existing 7B MLLM models on REC tasks with Prec@0.5 indicator. The top 2 results are outlined in **bold** and with underline.

Type	Method	Grounding Boxes	RefCOCO			RefCOCO+			RefCOCOg	
			val	test A	test B	val	test A	test B	val	test
Specialists	SeqTR [65]	1.5M	87.00	90.15	83.59	78.69	84.51	71.87	82.69	83.37
	GroundingDINO(L) [27]	-	90.56	93.19	88.24	82.75	88.95	75.92	86.13	87.02
	OFA [49]	-	92.04	94.03	88.44	**87.86**	**91.70**	**80.71**	88.07	88.78
	UNINEXT(L) [54]	3M	91.43	93.73	88.93	83.09	87.90	76.15	86.91	87.48
	UNINEXT(H) [54]	3M	**92.64**	**94.33**	**91.46**	85.24	89.63	79.79	**88.73**	**89.37**
MLLMs	Shikra-7B [7]	∼ 4M	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19
	VisonLLM-H-7B [51]	0.15M	-	86.70	-	-	-	-	-	-
	NeXT-Chat-7B [61]	∼ 4M	85.50	90.00	77.90	77.20	84.50	68.00	80.10	79.80
	u-LLaVA-7B* (Ours)	0.66M	82.95	89.08	76.29	72.91	82.43	63.41	76.23	76.56
	u-LLaVA-7B (Ours)	0.66M	**91.20**	**94.29**	**87.22**	**85.48**	**90.76**	**78.11**	**86.54**	**87.25**

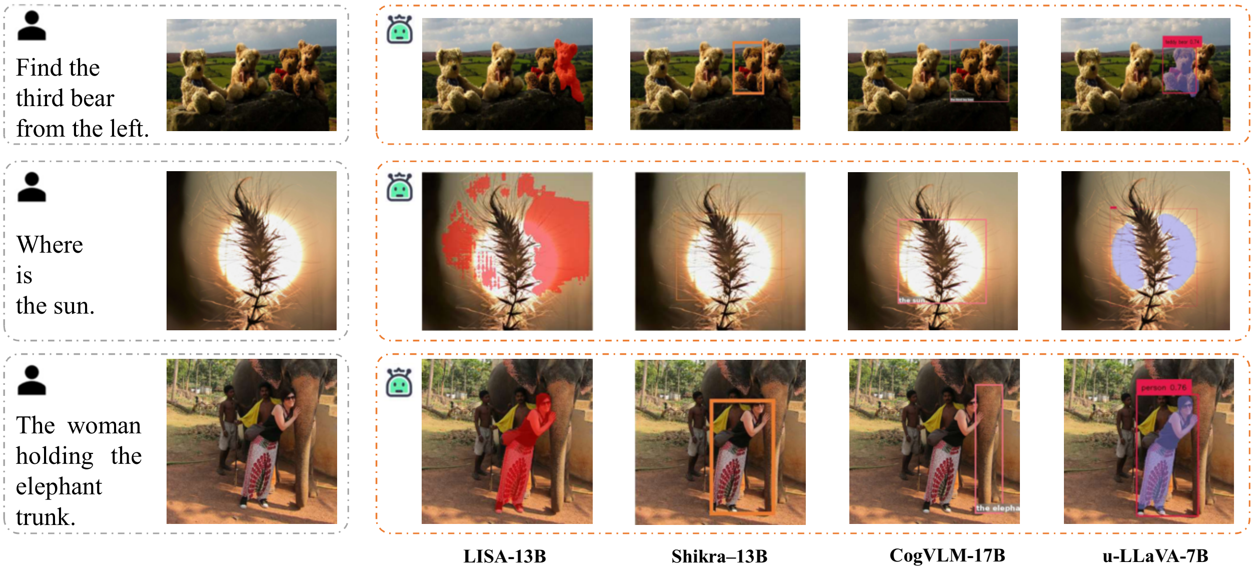

Find the third bear from the left.
Where is the sun.
The woman holding the elephant trunk.
person 0.76
LISA-13B
Shikra–13B
CogVLM-17B
u-LLaVA-7B

Figure 3: Qualitative examples of existing methods for regional and pixel-level understanding.

input resolution of u-LLaVA to 336, namely u-LLaVA-1.5, to enhance the performance of the model on such tasks. While these tasks are not the primary focus of the present study, our method demonstrates competitive performance relative to other 7B models. Specifically, u-LLaVA-1.5 achieves best results on the ScienceQA-IMG task using Vicuna-7B-v1.1 and ranks second only to LLaVA-1.5 on the MMBench-Test and GQA benchmarks.

4.7 *Dataset Ablation*

As shown in Table 8, we validate the impact of employing varied types of datasets during the second stage of the model's training on its overall performance. The results indicate that embracing diversity in dataset types fosters improved generalization of the algorithm, thereby circumventing the potential risk of overfitting on specific tasks. In essence, the algorithm's robustness is enhanced with an increased variety of dataset types.

4.8 *Qualitative examples*

Qualitative comparison with existing multi-task MLLM methods, LISA [17], Shikra [7] and CogVLM [50], on grounding and segmentation tasks are given in Figure 3. More conversation illustrations can be found in Figure 4 and Figure 5.

5 Conclusions

In this work, we introduce u-LLaVA, a multi-modal large language model that jointly tunes instructions at the global, regional, and pixel levels. Through innovative structural design and data configuration, we have achieved optimal performance in various comprehension-based tasks.

Currently, the pre-training and task adaptation of MLLMs remain an open area with many directions yet to be explored. This study represents an exploratory and experimental effort building upon previous works such as LLaVA and LISA. We believe that the open-sourcing of our work can provide valuable assistance to the development of this field.

6 Acknowledgement

This work is sponsored by Shanghai Pujiang Program (23PJ1421800).

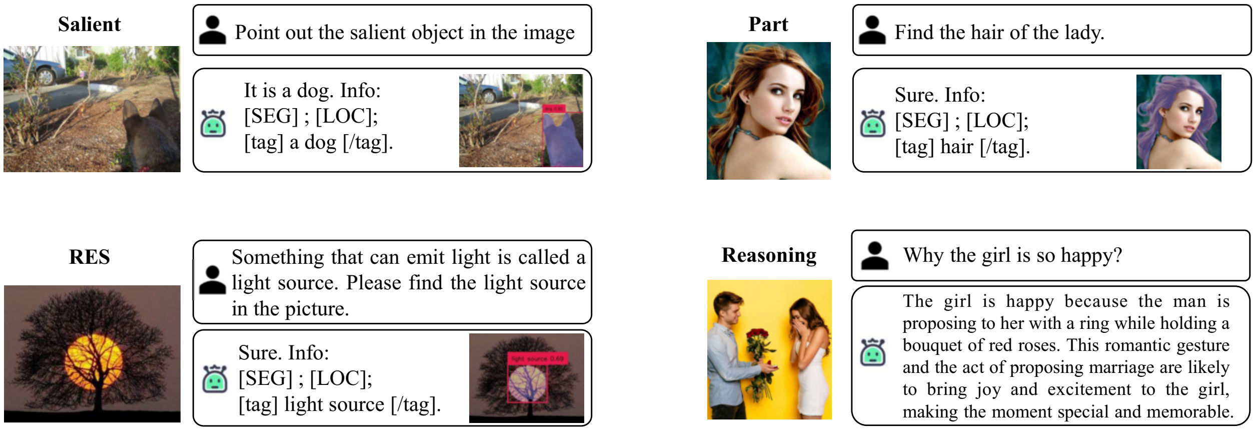
7 Appendix

7.1 *Templates*

Here, we present examples of task templates used by u-LLaVA on different type of training data.

Table 7: Experimental results with leading methods on popular multi-modal benchmarks, where the symbol † indicates that the model is trained for 1 epoch for fair comparison. The top 2 results are outlined in **bold** and with underline.

Method	LLM	Epoch	Image	General		VQA			RES	REC
			Size	MMB^D	MMB^T	TextVQA	GQA	SciQA^{IMG}	RefCOCO	RefCOCO
InstructBLIP [25]	Vicuna-7B	-	224	-	36	-	-	-	-	-
Shikra [7]	Vicuna-7B	3	224	-	-	-	-	-	-	87.01
IDEFICS-9B [18]	Vicuna-7B	-	224	48.2	45.3	25.9	38.4	-	-	-
Qwen-VL [2]	Qwen-7B	-	448	38.2	32.2	63.8	-	67.1	-	89.4
Mini-Gemini [22]	Vicuna-7B-v.5	-	336(768)	**69.3**	-	**65.2**	-	-	-	-
LLaVA [25]	Vicuna-7B-v1.1	1	224	38.7	-	-	-	-	-	-
LLaVA-1.5 [24]	Vicuna-7B-v1.5	1	336	65.2	**66.5**	58.5	**62.0**	66.8	-	-
u-LLaVA †	Vicuna-7B-v1.1	1	224	56.2	57.0	47.4	55.8	64.5	80.01	83.96
u-LLaVA-1.5 †	Vicuna-7B-v1.1	1	336	61.6	62.3	55.6	57.8	**67.2**	**81.05**	86.81



Point out the salient object in the image

It is a dog. Info: [SEG] ; [LOC]; [tag] a dog [/tag].

Find the hair of the lady.

Sure. Info: [SEG] ; [LOC]; [tag] hair [/tag].

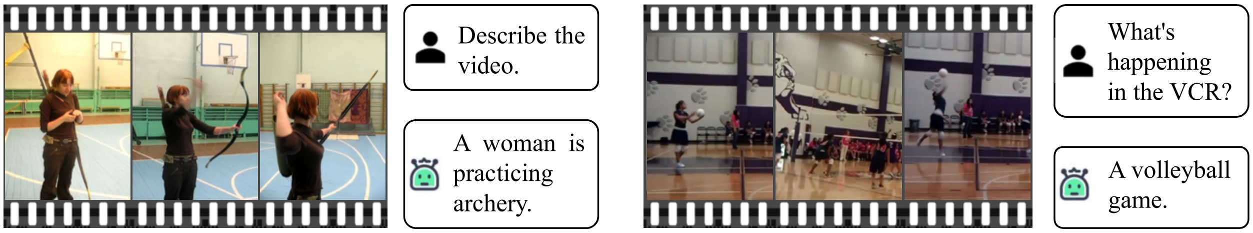
Something that can emit light is called a light source. Please find the light source in the picture.

Sure. Info: [SEG] ; [LOC]; [tag] light source [/tag].

Why the girl is so happy?

The girl is happy because the man is proposing to her with a ring while holding a bouquet of red roses. This romantic gesture and the act of proposing marriage are likely to bring joy and excitement to the girl, making the moment special and memorable.

Figure 4: Conversation examples with u-LLaVA-7B.



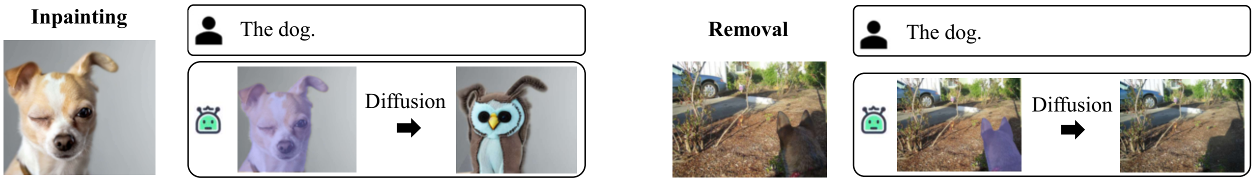
Describe the video.

A woman is practicing archery.

What's happening in the VCR?

A volleyball game.

Figure 5: Video captioning examples with u-LLaVA-7B.



The dog.

The dog.

Figure 6: Image inpainting and object removal examples with u-LLaVA-7B.

Template examples for salient segmentation task

```
<image> What makes the image stand out?
<image> What is salient one in this image?
<image> Look at the image, segment the main
    object in the picture and explain.
```

Template examples for video captioning task

```
<video> Describe the video concisely.
<video> What's happening in this video?
<video> Write a terse but informative summary of
    the VCR.
```

Template examples for RES task

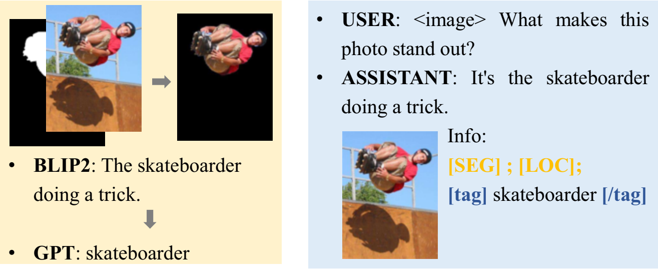
Table 8: Ablations on the Stage II training datasets, and cIoU is used as the performance indicator.

Exp.	Referring	Semantic	Salient	VQA	RefCOCOg test	DUT-OMRON
1	✓				72.83	52.04
2	✓	✓			75.04	46.70
3	✓	✓	✓		75.10	65.45
4	✓	✓	✓	✓	**77.97**	**74.19**

```
<image> Segment out the <class>.
<image> Output the mask of the <class>.
<image> Find the <class> in the picture.
```

7.2 Construction of Salient-15K

As shown in Figure 7, given that the MSRA-10K and MSRA-B datasets are devoid of label information and image descriptions pertaining to the principal subjects, we initially proceed by extracting the subjects from the images and subsequently inputting them into BLIP2 [20] for a rudimentary description. Following this, we employ GPT3.5 to parse the target labels emanating from the elementary description, thereby enabling an expansion of the description information. This approach facilitates a more comprehensive understanding of the subjects within the datasets while compensating for the initial lack of descriptive data.



- **BLIP2**: The skateboarder doing a trick.
- **GPT**: skateboarder

- **USER**: <image> What makes this photo stand out?
- **ASSISTANT**: It's the skateboarder doing a trick.

Info:
[SEG] ; [LOC];
[tag] skateboarder [/tag]

Figure 7: The process workflow of Salient-15K.

References

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[2] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023.

[3] J. T. Barron and B. Poole. The fast bilateral solver. In *European conference on computer vision*, pages 617–632. Springer, 2016.

[4] A. Bielski and P. Favaro. Move: Unsupervised movable object segmentation and detection. *Advances in Neural Information Processing Systems*, 35:33371–33386, 2022.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] G. H. Chen, S. Chen, R. Zhang, J. Chen, X. Wu, Z. Zhang, Z. Chen, J. Li, X. Wan, and B. Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024.

[7] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[8] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[9] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[10] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. *IEEE TPAMI*, 41(4):815–828, 2019. doi: 10.1109/TPAMI.2018.2815688.

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[12] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, Q. Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

[13] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[14] K. Kafle, B. Price, S. Cohen, and C. Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.

[15] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[16] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[17] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

[18] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. Rush, D. Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[20] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[21] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016.

[22] Y. Li, Y. Zhang, C. Wang, Z. Zhong, Y. Chen, R. Chu, S. Liu, and J. Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

[23] C. Liu, H. Ding, and X. Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23592–23601, June 2023.

[24] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[25] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[26] J. Liu, H. Ding, Z. Cai, Y. Zhang, R. K. Satzoda, V. Mahadevan, and R. Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18653–18663, June 2023.

[27] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[28] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[29] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[30] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[31] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[32] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[34] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[35] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.

[36] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1 (8):9, 2019.

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[38] V. Ramanathan, A. Kalia, V. Petrovic, Y. Wen, B. Zheng, B. Guo, R. Wang, A. Marquez, R. Kovvuri, A. Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023.

[39] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[40] J. Shi, Q. Yan, L. Xu, and J. Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2015.

[41] G. Shin, S. Albanie, and W. Xie. Unsupervised salient object detection with spectral cluster voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3971–3980, 2022.

[42] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, and J. Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

[43] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[44] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[45] R. Tito, D. Karatzas, and E. Valveny. Document collection visual question answering. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 778–792. Springer, 2021.

[46] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[47] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng. Salient object detection: A discriminative regional feature integration approach. *International Journal of Computer Vision*, 123(2):251–268, 2017. ISSN 1573-1405. doi: 10.1007/s11263-016-0977-3.

[48] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017.

[49] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[50] W. Wang, Q. Lv, W. Yu, W. Hong, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, X. Bin, H. Li, Y. Dong, M. Ding, and J. Tang. Cogvlm: Visual expert for large language models. *arXiv preprint*, 2023.

[51] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

[52] Y. Wang, X. Shen, S. X. Hu, Y. Yuan, J. L. Crowley, and D. Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022.

[53] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

[54] B. Yan, Y. Jiang, J. Wu, D. Wang, P. Luo, Z. Yuan, and H. Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15325–15336, June 2023.

[55] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013.

[56] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18155–18165, June 2022.

[57] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[58] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[59] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.

[60] Y. Yuan, W. Li, J. Liu, D. Tang, X. Luo, C. Qin, L. Zhang, and J. Zhu. Osprey: Pixel understanding with visual instruction tuning. *arXiv preprint arXiv:2312.10032*, 2023.

[61] A. Zhang, L. Zhao, C.-W. Xie, Y. Zheng, W. Ji, and T.-S. Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023.

[62] H. Zhang, X. Li, and L. Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[63] Y. Zhao, Z. Lin, D. Zhou, Z. Huang, J. Feng, and B. Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023.

[64] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[65] C. Zhu, Y. Zhou, Y. Shen, G. Luo, X. Pan, M. Lin, C. Chen, L. Cao, X. Sun, and R. Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022.

[66] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[67] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan, N. Peng, L. Wang, Y. J. Lee, and J. Gao. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15116–15127, June 2023.

[68] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Gao, and Y. J. Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023.