Arcana: Improving Multi-modal Large Language Model through Boosting Vision Capabilities

Yanpeng Sun[1,2*], **Huaxin Zhang**[2,3*], **Qiang Chen**[2], **Xinyu Zhang**[2], **Nong Sang**[3]
Jingdong Wang[2], **Zechao Li**[1†]

[1]Nanjing University of Science and Technology,
[2]Baidu VIS, [3]Huazhong University of Science and Technology

{yanpeng_sun, zechao.li}@njust.edu.cn

Abstract

We are interested in improving the visual understanding capability for boosting the vision-language models. Therefore, we propose **Arcana**, a multimodal language model, introducing two crucial techniques. Firstly, we introduce multimodal LoRA (MM-LoRA) to construct a multimodal decoder. Unlike traditional language-driven decoders, multimodal LoRA employs separate parameters for visual and language modalities, avoiding information confusion while preserving the uniqueness of each modality. Secondly, we propose a Query Ladder adapter (QLadder) for the visual encoder. By retaining the pre-trained image encoder's capabilities and introducing a small number of visual tokens, QLadder significantly enhances the model's learning and representation abilities for visual information. Inspired by traditional query mechanisms, QLadder adopts a "ladder" structure to progressively refine and strengthen the extraction and fusion of visual features. Extensive experiments demonstrate the effectiveness and generalization ability of Arcana. Furthermore, we propose a data engine focused on the diversity and richness of visual perception annotations, ensuring comprehensive visual perception information in annotations through diverse image annotation models. We annotate images in the open-source datasets VG and COCO to obtain more detailed image descriptions, crucial for the understanding of images by multimodal models. The code and re-annotated data are available at `https://github.com/syp2ysy/Arcana`.

1 Introduction

In recent years, multimodal large language models (MLLMs) [52, 4, 32, 56] have made significant advancements. These models amalgamate images into large-scale language models (LLMs) [49, 60], leveraging their powerful language capabilities to excel in various multimodal tasks. In MLLMs' decoder (LLMs), visual information is usually directly integrated with language information, forming a unified multimodal representation space. However, the visual component typically depends only on the instance-level contrastive language-image pre-training model(e.g., CLIP [41]).

While these models showcasing remarkable proficiency in cross-modal tasks, they still face challenges in visual preception. As shown in Fig. 1(a), we observe deficiencies in current MLLMs regarding low-level visual perception, such as color and quantity, as well as high-level visual perception, such as small object detection and localization. This suggests that insufficient visual perception capabilities

*equal contribution
†Corresponding author.

Preprint. Under review.

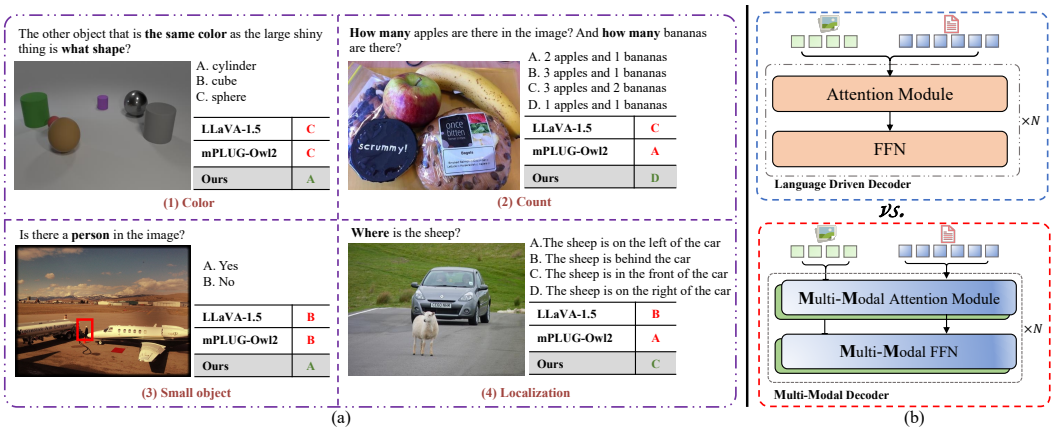

Figure 1: **(a)** Sampled some VQA examples involving color, quantity, small objects, and localization tasks, showcasing the importance of visual recognition capabilities for multimodal language models (MLLMs). **(b)** Contrasting Arcana's multimodal decoder with mainstream methods' language driven decoder. The language-driven decoder employs a language decoder (LLMs) directly to handle tokens from different modalities, which may lead to modality interference and performance degradation. In contrast, the multimodal decoder independently processes different token types to avoid modality interference.

may affect the model's performance, applicability, and reliability, thereby limiting the breadth and depth of multimodal applications. Therefore, enhancing the visual perception capabilities of MLLMs has become a pressing issue awaiting resolution in VL community.

The insufficient visual perception capabilities of MLLMs can mainly be attributed to two factors: decoder and visual encoder. As depicted in Fig. 1(b), existing language driven decoder structures directly couple visual and language modalities. This approach not only disregards their unique characteristics but also may lead to information confusion or blurring, thus impairing the model's accurate understanding and processing of visual information. Furthermore, the operation of freezing visual encoder directly limits the model's ability to learn and represent visual information. Therefore, improving the model's visual perception requires rethinking the decoder design and optimizing the use of the visual encoder to better capture and process visual features.

To this end, we propose **Arcana**, a model that aims to enhance visual perception capabilities from both structural and data perspectives. Specifically, we design a multimodal LoRA (MM-LoRA) to construct a multimodal decoder as show in Fig. 1(b). This decoder provides independent learning spaces for each modality, ensuring the decoupling of different modalities, avoiding information confusion, and preserving the uniqueness of each modality. Additionally, we introduce a new design called the Query Ladder Adapter (QLadder). QLadder significantly enhances the model's ability to learn and represent visual information with the introduction of a small number of visual tokens. Inspired by traditional query mechanisms, QLadder incorporates a "ladder" structure to progressively refine and strengthen the extraction and fusion of visual features. On the data side, we propose a new data engine to obtain detailed descriptions of images. Unlike previous work, we focus on achieving diversity and richness in visual perception annotations. By utilizing diverse data annotation methods, we ensure comprehensive visual perception information in the annotations. This approach not only improves the quality of training data but also enhances the model's generalization ability and performance in practical applications. Finally, extensive experiments validated the effectiveness and generalization ability of our model. Notably, our model demonstrated exceptional performance on various benchmark vision-language tasks, proving its competitiveness. Additionally, our data engine has provided detailed annotations for COCO and VG datasets, which will be released soon.

2 Related Work

Multi-Modal Large Language Models. Fueled by the tremendous success of large language models (LLMs) [1, 49, 19], there is growing interest in developing end-to-end multi-modal large language models (MLLMs) [11, 56, 12]. These models aim to enhance the visual perceptual capabilities of LLMs by integrating additional modalities, allowing for unified handling of multi-modal tasks.

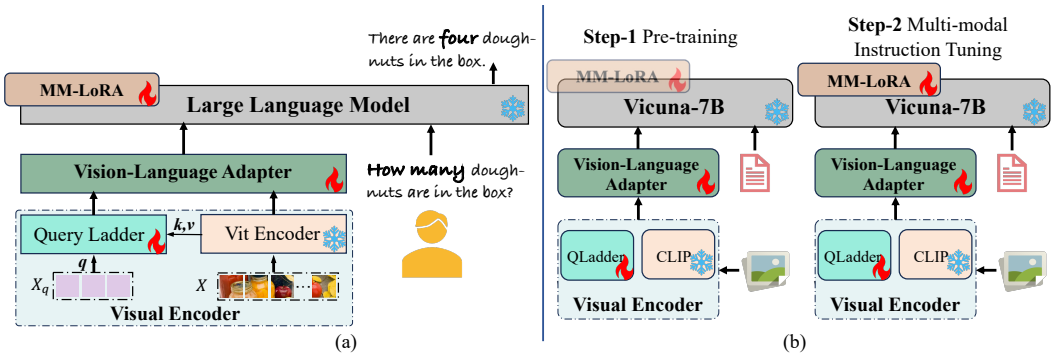

Figure 2: **(a)** The architecture of the Arcana. **(b)** The training pipeline of Arcana. MM-LoRA is optional during the pre-training phase.

Currently, there are three primary approaches to building Multi-Modal foundational models, each demonstrating strong potential for zero-shot generalization in the visual-language domain.

The first approach, exemplified by Flamingo [2], uses cross-attention to align visual models with large language models across modalities. The second approach, used by models like PaLM-E [13], directly integrates extracted visual features into a pre-trained PaLM [3] model via a linear layer, achieving robust performance. This method is widely adopted by mainstream models such as LLaVA [32], CogVLM [52] and Internlm-Xcomposer [59] but incurs high inference costs due to the lengthy visual tokens. To address this, the third approach, inspired by DETR [39, 64] and represented by BLIP-2 [28], employs a Q-former to effectively reduce the sequence length of visual features. Similar designs are used by mPLUG-OWL2 [56], and MiniGPT-4 [63]. However, these methods [3–5] couple visual and language modalities in the same space using language-guided decoders, overlooking the uniqueness of different modalities. This oversight may result in interference between modalities, potentially affecting performance. To this end, we employ MM-LoRA to implement a multimodal decoder, aiming to mitigate the impact of modality interference on the model.

Improve visual perception for MLLMs. Currently, MLLMs are the most popular approach in VL community [2, 28], and enhancing their visual recognition capabilities has become a prominent research trend. Integrating visual features into large language models (LLMs) via a linear layer has become the mainstream approach [32, 52]. However, this approach often relies on frozen vision encoders to provide visual features, which limits the visual recognition capabilities of multimodal large language models (MLLMs). To address this issue, existing methods enhance visual recognition in two ways. The first method [36, 48, 54] introduces new high-resolution vision encoders, significantly improving visual recognition by increasing the number of visual tokens. For example, LLaVA-HR [36] achieves this by incorporating ConvNeXt [34] to handle high-resolution images. However, these methods significantly increases the number of visual tokens. Therefore, we propose QLadder, which can significantly enhance the model's visual perception capability with the introduction of a small number of visual tokens. The second method [52, 12, 56] expands the learning space for visual tokens within the large language model to accelerate visual-language alignment, thereby enhancing visual perception. For instance, Internlm-Xcomposer2 [12] introduces Partial-LoRA, adding a LoRA to visual tokens to strengthen their representation. However, experiments with MM-LoRA have shown that directly increasing the learning space for visual tokens in the decoder does not improve the model's performance.

3 Method

3.1 Overview

We propose a new model, named Arcana as shown in Fig 2, designed to enhance visual perception in multimodal language models. Like most existing models [32, 5], it includes a visual encoder, a vision-language adapter, and a large language model. The key difference is that we use MM-LoRA to implement a multimodal decoder. Unlike traditional fine-tuning where visual and language modalities share LoRA parameters, MM-Lora assigns different LoRA parameters to each modality. Additionally, we introduce QLadder in the visual encoder, which significantly enhances the model's ability to learn and represent visual information with the introduction of a small number of visual tokens. We first briefly introduce Arcana's architecture in Section 3.2. Additionally, in Section 3.3, we

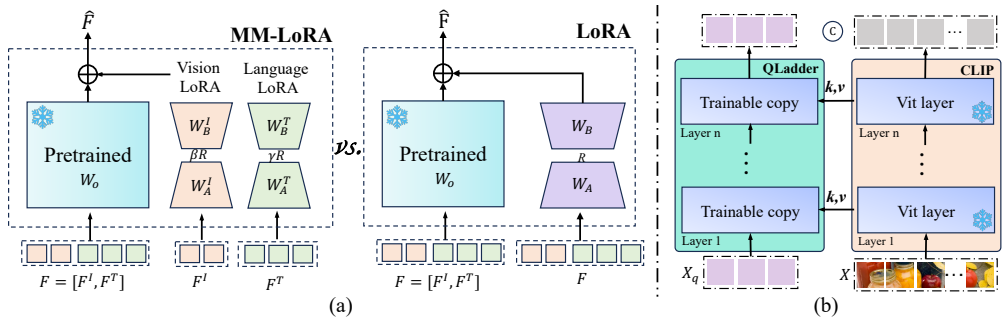

Figure 3: **(a)** The farmework of MM-LoRA *vs.* LoRA. MM-LoRA introduces two new hyperparameters, β and γ, to control the ranks of the visual and language LoRAs, respectively. Notably, we set $\beta + \gamma = 1$ to ensure that MM-LoRA has the same number of parameters as LoRA. **(b)** The architecture of the visual encoder includes the QLadder adapter and CLIP. The QLadder adapter consists of cross-attention and FFN layers, with weights initialized from those of CLIP.

detail MM-LoRA, which aims to decouple the learning spaces of different modalities to achieve a multimodal decoder. Lastly, we introduce the training paradigm of Arcana in Section 3.4.

3.2 Architecture

Arcana, as illustrated in Fig 2(a), consists of four components: **visual encoder**, **vision-language adapter** and **large language model**. Next, we will provide a detailed overview of the design and implementation details of each component.

Visual Encoder. The visual encoder's primary objective is to extract high-level features from images. To approximate the language modality, the CLIP [41] visual model has emerged as the optimal choice for the visual encoder. However, inherent drawbacks in CLIP result in the loss of fine-grained visual information, thereby limiting the performance of MLLMs on specific multimodal tasks. Toward this end, we introduce the Query Ladder adapter (QLadder) as shown in Fig 3(b). This adapter enhances the visual feature representation of the visual encoder by adding a small number of query visual tokens while retaining the pretrained image encoder. It improves Arcana's visual perception capability.

For input image X, we utilize CLIP to extract visual features $F_c \in \mathbb{R}^{N_I \times C_v}$, where C_v represents the channel of visual feature, N_I indicates the number of image patch. Simultaneously, a set of learnable vectors X_q is fed into QLadder to acquire additional visual features $F_q \in \mathbb{R}^{N_q \times C_v}$, where $N_q << N_I$. Ultimately, the visual features are concatenated from the outputs of CLIP and the query ladder to obtain $F_v = concat(F_c, F_q)$. As shown in Fig 2(b), QLadder comprises multiple layers composed of cross-attention and feed-forward networks (FFNs). In the cross-attention component, X_q serves as the Query, while the Key and Value are derived from CLIP's image features.

Vision-Language Adapter. To map the output of visual encoder to same space as the word embeddings of language features, we employ an Vision-Language adapter, denoted as g, consisting of two MLP layers. Thus, the visual features are processed through the vision-language tokens to obtain visual tokens, denoted as $F^I = g(F_v)$.

Large Language Model. For multimodal tasks [15, 18], leveraging pre-trained large language models (LLMs) can provide valuable linguistic priors. Through multimodal instruction tuning, LLMs learn to comprehend visual features within images, enabling comprehensive understanding and processing of multimodal data. Typically, this process is accomplished through full fine-tuning or LoRA [17]. However, these methods overlook the unique characteristics of modalities, leading to modality confusion. This not only damages MLLMs' accurate understanding and processing of visual information but also affects natural language understanding. Therefore, a multimodal decoder that provides separate learning spaces for each modality is a better choice for MLLMs.

3.3 Multimodal LoRA

To implement a multimodal decoder based on a large language model, we propose a multimodal LoRA. This approach projects visual and language features into separate semantic spaces to decouple their representations, thereby avoiding modality interference. This allows Arcana to retain the unique

characteristics of each modality, enhancing its visual perception without compromising natural language understanding. Next, we detail the MM-LoRA process.

MM-LoRA, as illustrated in Fig. 3, consists of visual LoRA and language LoRA. In comparison to LoRA, we introduces two parameters, β and γ, to control the rank size of (R) visual LoRA and language LoRA. It's worth noting that $\beta+\gamma = 1$ to ensure that no additional parameters are introduced compared to LoRA. Specifically, given a sequence of visual-language features $F \in \mathbb{R}^{(N_v+N_t)\times C}$ and a multimodal mask $M \in \{0,1\}^{(N_v+N_t)}$, where C represents the hidden dimension in LLMs, N_v and N_t indicates the number of visual and language tokens, respectively. We define a modality separation function Θ to separate the tokens of different modalities within F.

$$\Theta(F, M, m) = F \odot (M == m) \tag{1}$$

where $m \in \{0, 1\}$ is used to select between visual tokens ($m = 0$) and language tokens ($m = 1$). Therefore, based on multimodal mask M , we can obtain F^I and F^T.

$$F^I = \Theta(F, M, 0) \qquad F^T = \Theta(F, M, 1) \tag{2}$$

Then, F^I and F^T are separately inputted into the visual part and language part of MM-LoRA. In Visual LoRA, the weights are denoted as $W_A^I \in \mathbb{R}^{C\times\beta R}$ and $W_B^I \in \mathbb{R}^{\beta R\times C}$, while in Language LoRA, the weights are denoted as $W_A^T \in \mathbb{R}^{C\times\gamma R}$ and $W_B^T \in \mathbb{R}^{\gamma R\times C}$.

Similarly to LoRA [17], F is inserted into the LLM layer to obtain $\hat{F}$. Finally, the output results of MM-LoRA are added to the output of LLM according to the mask M_I.

$$\begin{gathered} \hat{F} = W_o \times F \\ \Theta(\hat{F}, M, 0) += W_B^I \times W_A^I \times F^I \qquad \Theta(\hat{F}, M, 1) += W_B^T \times W_A^T \times F^T \end{gathered} \tag{3}$$

In Arcana, MM-LoRA is applied to all linear layers of the large language model, thereby achieving an optimal multimodal decoder.

3.4 Training Paradigm

Following prior work [32, 52], we adopt a two-stage approach involving pretraining and multimodal instruction fine-tuning to train Arcana, as illustrated in Fig. 2(b). The purpose of the pretraining stage is to align the visual encoder with the language model, while multimodal instruction fine-tuning aims to adapt the model better to specific tasks through fine-tuning. We found that freezing the visual encoder limits the MLLM's ability to capture complex visual information, such as scene text and visual knowledge. To address this issue, we introduce Qladder and enable it to be trained in both the pretraining and instruction fine-tuning stages. This strategy allows the model to more effectively capture both low-level and high-level semantic visual information. Additionally, we introduce MM-LoRA fine-tuning as an alternative to full fine-tuning and LoRA fine-tuning, enabling a multimodal decoder that minimizes modality interference. Specifically, in the pretraining stage, we train Qladder and the vision-language adapter, while in the instruction fine-tuning stage, we train Qladder, the vision-language adapter, and MM-LoRA. Furthermore, to ensure the linguistic capabilities of Arcana, we employ joint training, adjusting the entire model during instruction fine-tuning, integrating textual and multimodal instructions.

4 Data Engine

In the pretraining stage of multimodal large language models (MLLMs), using image captions to help decoders understand image content is crucial. However, existing datasets in the open-source community, such as LLaVA [32] and ShareGPT4V [6], lack detailed descriptions rich in various visual information. These open-source image descriptions typically cover only the primary content, missing many important visual details like color, quantity, relationships, and textures, which are vital for enhancing the generalization capabilities of MLLMs.

To address this issue, we designed a data engine that employs diverse annotation models to provide various types of visual information annotations for images. We then use large language models (LLMs) like GPT [1] to integrate these annotations into high-quality image descriptions. The detailed process is provided in the **Appendix**. Notably, we used this data engine to generate richly detailed

Table 1: Performance on six General Visual Question Answering benchmarks. Specialist models, indicated in gray, are fine-tuned on each individual dataset. The red and blue colors respectively represent the optimal and suboptimal results on each benchmark. $*$ indicates that MM-LoRA is trained during the pretrain stage.

Type	Model	LLM	In-domain VQA Tasks			Zero-shot VQA Tasks		
			VQAv2	OKVQA	GQA	TextVQA	ScienceQA	Ai2d
Generalists	BLIP2 [28]	Flan-T5	65.0	45.9	41.0	42.5	61.0	-
	InstructBLIP [11]	Vicuna (7B)	-	-	49.2	50.1	60.5	40.6
	InstructBLIP [11]	Vicuna (13B)	-	-	49.5	50.7	63.1	-
	Shikra [5]	Vicuna (7B)	77.4	47.2	-	-	-	-
	IDEFICS-Instruct [25]	LLaMA (65B)	37.4	36.9	-	28.3	61.8	54.8
	LLaVA-v1.5 [31]	Vicuna (7B)	78.5	-	**62.0**	58.2	66.8	55.5
	Qwen-VL-Chat [4]	Qwen (7B)	78.2	56.6	57.5	**61.5**	68.2	-
	mPLUG-Owl2 [56]	LLaMA (7B)	**79.4**	57.7	56.1	58.2	68.7	55.7
	Arcana	Vicuna (7B)	79.2	**57.9**	61.6	**59.5**	**71.2**	**56.8**
	Arcana*	Vicuna (7B)	**79.5**	**58.9**	**61.8**	58.7	69.5	**56.9**
Specialists	GIT2 [51]	-	81.7	-	-	59.8	-	-
	PaLI-17B [8]	-	84.3	64.5	-	58.8	-	-

descriptions for the Visual Genome (VG) [23] and COCO [30] datasets, and this annotated data will soon be open-sourced. Arcana has utilized only a subset of this data.

5 Experiments

5.1 Implementation Details

Model. In the visual encoder, we utilize the CLIP-L [41] model with an input resolution of 336 and a patch size of 14×14. Furthermore, the QLadder adapter adopts the same structure as CLIP-L, replacing self-attention with cross-attention. Notably, QLadder utilizes pre-trained CLIP weights as its initial weights. For the LLMs, we employ the pre-trained Vicuna-7B [9] model. The Vision-Language adapter comprises two layer MLP. MM-LoRA, used for fully supervised multimodal instruction tuning, consists of a visual LoRA with a rank of $\beta \times R$ and a language LoRA with a rank of $\gamma \times R$.

Data Sets. Arcana uses open-source data and data obtained from our data engine (soon to be open-sourced). During pre-training, we used approximately 1.2M image-text pairs from ShareGPT4V [6]. In the multimodal instruction tuning stage, we utilize six types of supervised data totaling 934k, namely: (1) text-only instruction data (ShareGPT [43]); (2) vision question-answering data (VQAv2 [15], GQA [18], A-OKVQA [42], OK-VQA [38]); (3) OCR QA (OCRVQA [40], TextCaps [45]); (4) Region-aware QA (RefCOCO [20, 37], VG [24]); (5) multi-modal instruction data (LLaVA-instruct [32]); and (6) image captions (VG-COCO [30, 23], shareGPT4V). Types (1)-(5) of the data constitute the instruction data in LLaVA-v1.5. In type (6), most of the data comes from our data engineering efforts, with some contributions from shareGPT4V.

Training Setting. During the pretraining step, we use language modeling loss with a batch size of 256 for 1 epoch. The learning rates are set to $1e-3$ for the vision-language adapter and $2e-5$ for Qladder. In the multimodal instruction tuning step, we integrated MM-LoRA into the LLM to create a multimodal decoder, thus preventing information interference between modalities. We set the learning rate for MM-LoRA to $1e-4$, and for both Qladder and the vision-language adapter, to $2e-5$. MM-LoRA is configured with a default rank R of 256, β set to 0.25, and γ set to 0.75. All experiments are conducted on 8 NVIDIA A100 GPUs.

5.2 Main Results

General Visual Question Answering Benchmarks. In Table 1, we compare with both SOTA MLLMs model on six General VQA benchmarks, including VQAv2 [15], OKVQA [42], GQA [18], TextVQA [46], ScienceQA [35] and Ai2d [21]. We found that Arcana achieved competitive results

Table 2: Performance on five Large Vision-Language Models (LVLM) benchmarks.The red and blue colors respectively represent the optimal and suboptimal results on each benchmark. $*$ indicates that MM LoRA is trained during the pretrain stage.

Method	Vision Encoder	Language Model	MME	MMBench	MM-Vet	SEED-Bench	LLaVA^W	POPE
BLIP-2 [28]	ViT-g (1.3B)	Vicuna (7B)	1293.84	-	22.4	46.4	38.1	85.3
MiniGPT-4 [63]	ViT-g (1.3B)	Vicuna (7B)	581.67	23.0	22.1	42.8	45.1	-
LLaVA [32]	ViT-L (0.3B)	Vicuna (7B)	502.82	36.2	28.1	33.5	63.0	80.2
mPLUG-Owl [56]	ViT-L (0.3B)	LLaMA (7B)	967.34	46.6	-	34.0	-	-
InstructBLIP [11]	ViT-g (1.3B)	Vicuna (7B)	1212.82	36.0	26.2	53.4	60.9	78.9
LLaMA-Adapter-v2 [14]	ViT-L (0.3B)	LLaMA (7B)	1328.40	39.5	31.4	32.7	-	-
Otter [27]	ViT-L (0.3B)	LLaMA (7B)	1292.26	48.3	24.6	32.9	-	-
Qwen-VL-Chat [4]	ViT-G (1.9B)	Qwen (7B)	1487.58	60.6	-	58.2	-	-
LLaVA-v1.5 [31]	ViT-L (0.3B)	Vicuna (7B)	**1510.70**	64.3	30.5	58.6	63.4	85.9
mPLUG-Owl2 [57]	ViT-L (0.3B)	LLaMA (7B)	1450.19	64.5	**36.2**	57.8	-	86.2
Arcana	ViT-L (0.3B)	Vicuna (7B)	1476.48	**66.9**	**34.8**	**62.6**	**67.3**	**86.5**
Arcana*	ViT-L (0.3B)	Vicuna (7B)	**1520.93**	**67.4**	34.4	**63.2**	**72.7**	**87.1**

on six VQA benchmarks. Notably, it achieved accuracies of 57.9 on OKVQA, 71.2 on ScienceQA, and 56.8 on Ai2d , surpassing most recently proposed MLLMs methods. Additionally, Arcana* with MM-LoRA used during the pre-training stage achieved better performance, indicating the importance of preserving the uniqueness of different modalities during pre-training. The superior performance on zero-shot VQA tasks particularly highlights strong generalization ability and potential across different domains of our model.

Large Vision-Language Model Benchmarks. Table 2 presents our comparative results on five different LVLM benchmarks: MMBench [33], MM-Vet [58], SEED-Bench [26], LLava^W [32], and POPE [29]. It is evident that Arcana achieves highly competitive performance across these benchmarks. Compared to mPLUG-OWL2 [57], Arcana scores 2.4 and 4.8 points higher on MMBench and SEED-Bench, respectively. Additionally, Arcana achieves a score of 86.5 on the hallucination evaluation dataset POPE, indicating significant advancements in visual recognition capabilities. These impressive results not only demonstrate its strong reasoning and multi-task generalization abilities but also clearly show that Arcana significantly outperforms others in these areas. Notably, we achieved this using a 0.3B visual encoder, with MM-LoRA and QLadder significantly enhancing the model's visual perception and generalization.

Table 3: **Performance on language benchmarks of our model** compared to LLaMA-2 0-shot for BBH, AGIEval, ARC.

Method	BBH	AGIEval	ARC-c	ARC-e
LLaMA-2 [50]	38.2	21.8	40.3	56.1
WizardLM [53]	34.7	23.2	47.5	59.6
LLaMA-2-Chat [50]	35.6	28.5	54.9	71.6
Vicuna-v1.5 [9]	41.2	21.2	56.6	72.8
Arcana	**42.1**	**29.3**	**61.4**	**78.3**

Natural Language Understanding. Although MLLMs excel in various multimodal downstream tasks, existing work [32, 12] often overlooks their natural language understanding capabilities. To address this, we also evaluated our model's language understanding performance on BIG-Bench Hard (BBH) [47], AGIEval [62], and ARC [10], as shown in Table 3. Compared to LLaMA-like [49] language models, Arcana achieved competitive results across multiple benchmarks. This demonstrates that our model not only performs well in multimodal tasks but also excels in language understanding, further highlighting the superiority of our approach.

5.3 Ablation Study

To validate the effectiveness of QLadder and MM-LoRA, we designed a series of experiments. Additionally, to ensure fairness, we used only LLAVA-v1.5 [31] data for these experiments.

Multimodal LoRA (MM-LoRA).

To validate the effectiveness of the multimodal decoder, we compared the performance of MM-LoRA and LoRA. Additionally, to investigate the importance of visual tokens and language tokens in the multimodal instruction tuning process within the decoder, we compared different ratios of β and γ

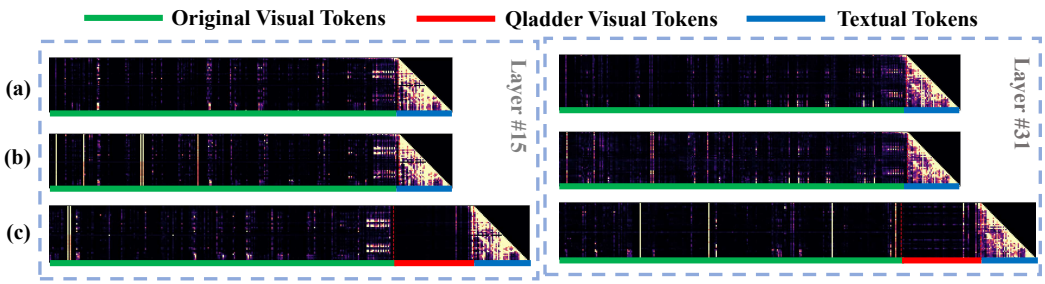

Figure 4: **Visualization of attention maps**. We compare the attention maps in different layer of LLM between different composition, include **(a)** Baseline, **(b)**Baseline+MM-LoRA, and **(c)** Baseline+MM LoRA+QLadder. Higher brightness indicates higher attention values, with the x-axis representing all tokens, and the y-axis containing only the generated text tokens.

parameters. In all experiments, the RANK of MM-LoRA and LoRA was set to 256. The results are shown in Table 4. It clearly indicate that MM-LoRA achieves optimal performance when $\beta = 0.25$ and $\gamma = 0.75$. When β is set to 1, performance significantly drops, indicating that aligning language distribution using only visual tokens is challenging for

Table 4: Ablation of β and γ in MM-LoRA. The default rank is set to 256, while β and γ are used to control the rank values in visual and language LoRA components, respectively.

Method	RANK		TextVQA	ScienceQA	MMBench	MME
	β	γ				
LoRA	-	-	58.1	69.1	63.8	1460
MMLoRA	1	0	51.2$_{(-6.9)}$	65.8$_{(-3.3)}$	56.4$_{(-7.4)}$	1356$_{(-104)}$
	0.75	0.25	**58.7**$_{(+0.6)}$	68.6$_{(-0.5)}$	63.3$_{(-0.5)}$	1465$_{(+5.0)}$
	0.5	0.5	58.5$_{(+0.4)}$	70.1$_{(+1.0)}$	64.4$_{(+0.6)}$	1483$_{(+23)}$
	0.25	0.75	**58.7**$_{(+0.6)}$	**71.2**$_{(+2.1)}$	64.8$_{(+1.0)}$	**1500**$_{(+40)}$
	0	1	57.9$_{(-0.2)}$	70.1$_{(+1.0)}$	**65.4**$_{(+1.6)}$	1480$_{(+20)}$

Table 5: Ablation of query number in QLadder. N_q represents the number of learnable query.

Method	N_q	ScienceQA	MMBench	MME
baseline	-	69.1	63.8	1460
Arcana	16	70.4$_{(+1.3)}$	63.9$_{(+0.1)}$	1481$_{(+21)}$
	32	70.6$_{(+1.5)}$	64.6$_{(+0.8)}$	1493$_{(+33)}$
	64	**71.2**$_{(+2.1)}$	**64.8**$_{(+1.0)}$	**1500**$_{(+40)}$
	128	69.7$_{(+0.6)}$	64.2$_{(+0.4)}$	1473$_{(+13)}$

MLLMs. However, introducing γ greatly improves performance, demonstrating that learning both vision and language simultaneously accelerates modality alignment. When γ is set to 1, there is a slight performance decline, but MM-LoRA still matches LoRA's performance, suggesting that visual token learning is less critical than language token learning in LLMs. This indicates that during the instruction tuning phase of MLLM training, more emphasis should be placed on learning language tokens. Furthermore, when both β and γ are set to 0.5, the performance of MM-LoRA significantly outperforms LoRA. This intuitively demonstrates that the multimodal decoder can avoid interference between modalities by separating them, thus significantly enhancing the performance of MLLMs.

QLadder in Vision Encoder. To validate the effectiveness of QLadder and determine the optimal number of queries, we conducted experiments with QLadder. The results, shown in Table 5, indicate that the inclusion of QLadder significantly enhances our model's performance. This demonstrates that even with a slight increase in visual tokens, without introducing a new visual encoder, the model's visual recognition capabilities can be improved. As the number of queries increased, our model's performance gradually improved, reaching its best performance with 64 queries. However, further increasing the number of queries led to a performance decline, indicating that too many queries can negatively impact the model's performance.

Impact of MM-LoRA and QLadder in MLLMs. To investigate the impact of MM-LoRA and QLadder in multimodal scenarios, we visualized the attention maps of Arcana with and without these modules in MM-Vet benchmark [58]. The visualization results, shown in Fig. 4, display the attention scores of generated tokens over the input sequence during the generation process. It can be seen that MLLM decoder initially focuses more on text tokens and gradually increases attention to visual tokens in the middle and subsequent layers. This indicates that visual and language information play different roles in MLLMs. The discussion about shallow-level attention maps, which also reflects this point, is provided in the Appendix.

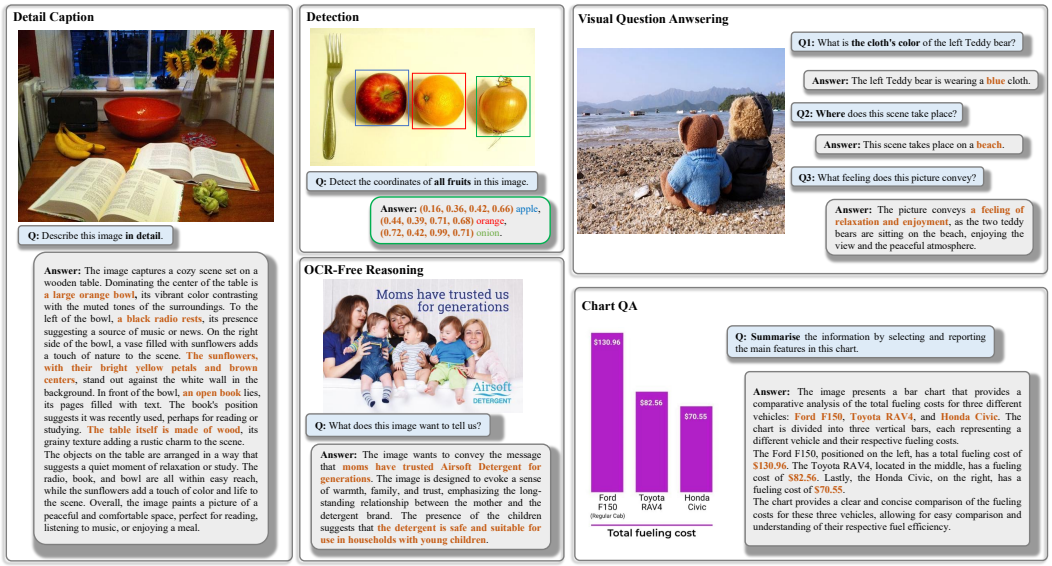

Figure 5: Examples of results generated by Arcana were sampled, focusing on tasks that test visual perception capabilities, such as detailed captions, detection, and OCR-reasoning. In the answers, all visual recognition-related responses are highlighted in orange.

Additionally, with MM-LoRA, we observe a significant increase in attention to visual tokens in the middle and subsequent layers, indicating that MM-LoRA helps prevent information confusion and promotes cooperation between different modalities. With the introduction of QLadder, the MLLM decoder shows increased attention to visual tokens across all layers. The highlighted regions of visual tokens further indicate that QLadder not only enhances the model's focus on visual tokens but also enriches the visual information, achieving optimal performance in multimodal tasks.

Visualization results. To showcase Arcana's outstanding performance in visual perception, we visualized its performance across various types of multimodal tasks. As illustrated in Fig. 5, visual perception information is highlighted in orange. In detailed description tasks, our model not only accurately identifies and describes low-level visual information such as colors and textures for each object in the image but also precisely recognizes and describes high-level visual information such as positions and relationships of each object. Moreover, detection tasks further demonstrate our model's effectiveness in visual recognition and localization. OCR-Free inference and chart-based question answering tasks not only exhibit our model's OCR recognition capabilities but also demonstrate its reasoning prowess. Visual question answering tasks showcase our model's excellent multi-turn dialogue capabilities on the foundation of precise identification.

In summary, Arcana utilizes a multimodal decoder to avoid information interference between different modalities. QLadder offers an innovative strategy for enhancing visual representations with limited data. By adding a small number of visual tokens, it significantly improves the performance of large multimodal language models. This finding is significant for the future of multimodal model, as it presents an effective approach to achieving notable performance improvements even with limited data resources. By combining these techniques, future multimodal models will handle complex tasks with greater flexibility and efficiency.

6 Conclusion

In this paper, we introduce a new multimodal large language model, **Arcana**, which incorporates two novel techniques. Unlike current mainstream methods, Arcana employs MM-LoRA for a multimodal decoder, enabling more efficient information processing and integration across different modalities. MM-LoRA effectively combines data from various modalities without significantly increasing computational complexity, reducing information interference between modalities. Secondly, we present the QLadder structure, which demonstrates for the first time that with limited multimodal training data, retaining the capabilities of a pre-trained model and adding a small number of visual encoders can still enhance the performance of multimodal language models. This hierarchical structure progressively

refines and enhances the expression of visual information, resulting in improved adaptability and generalization in multimodal tasks. With these two key techniques, Arcana not only excels in multimodal tasks but also shows potential for performance improvement even in data-constrained environments. Additionally, the severe lack of visual information in the image captions of open-source data limits the visual perception capabilities of multimodal large language models. To address this, we designed a data engine that uses diverse visual annotation models and large language models to generate captions rich in visual information.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in neural information processing systems*, pages 23716–23736, 2022.

[3] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023.

[5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[6] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[7] Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6633–6642, 2023.

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2022.

[9] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[10] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

[13] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, pages 8469–8488. PMLR, 2023.

[14] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[16] Jiale Han, Bo Cheng, and Wei Lu. Exploring task difficulty for few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2605–2616, 2021.

[17] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

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

[19] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[20] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara 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.

[21] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pages 235–251, 2016.

[22] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[25] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[26] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[27] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[29] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[35] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, pages 2507–2521, 2022.

[36] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024.

[37] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[38] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[39] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660, 2021.

[40] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[42] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

[43] ShareGPT. `https://sharegpt.com/`, 2023.

[44] Tao Sheng, Jie Chen, and Zhouhui Lian. Centripetaltext: An efficient text instance representation for scene text detection. In *Advances in Neural Information Processing Systems*, pages 335–346, 2021.

[45] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*. Springer, 2020.

[46] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[47] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, 2023.

[48] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024.

[49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[50] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[51] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*, 2022.

[52] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[53] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

[54] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024.

[55] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, pages 38571–38584, 2022.

[56] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[57] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023.

[58] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[59] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[60] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2024.

[61] Huang Zhizhong, Dai Mingliang, Zhang Yi, Zhang Junping, and Shan Hongming. Point, segment and count: A generalized framework for object counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

[62] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

[63] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[64] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020.

[65] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. Fourier contour embedding for arbitrary-shaped text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3123–3131, 2021.

A Appendix

A.1 Broader Impact

This paper present Arcana, which target at improving the visual understanding capability for boosting the vision-language models. To achieve this goal, Arcana conducts a series of explorations into visual learning within the model structure. On one hand, Arcana demonstrates that decoupling the learning of visual and language representation within the LLM is beneficial for avoiding information confusion while preserving the uniqueness of each modality, and based on this, proposes MM-LoRA. On the other hand, Arcana asserts that under limited training data, it is important to retain the pre-trained image encoder's capabilities and introduces QLadder, which incorporates a small number of visual tokens to enhance the model's learning and representation abilities for visual information. Extensive experiments demonstrate the effectiveness and generalization ability of Arcana.

The positive societal impacts of the work include:

- **Improved Human-Machine Interaction**: Enhanced visual perception in multimodal models can lead to more intuitive and effective human-machine interactions. This could improve applications such as virtual assistants, customer service bots, and educational tools, making them more responsive and capable of understanding complex visual contexts.
- **Advancements in AI Research**: The Arcana model's innovative architecture and data handling approaches could stimulate further research in the AI community, leading to new breakthroughs and applications in various fields, from healthcare to autonomous vehicles, where precise visual perception is crucial.
- **Better Performance in Real-World Applications**: By addressing the deficiencies in low-level and high-level visual perception, Arcana can improve performance in practical applications like object detection in surveillance, quality control in manufacturing, and detailed image analysis in medical diagnostics.

The negative societal impacts may include:

- **Privacy Concerns**: Enhanced visual perception capabilities may lead to more invasive surveillance technologies. The ability to detect and interpret small objects and detailed visual information could be misused to infringe on individuals' privacy, leading to unauthorized tracking and monitoring.
- **Security Risks**: Advanced visual perception models could be exploited for malicious purposes, such as by enhancing the capabilities of autonomous weapons or by improving the precision of surveillance systems used by authoritarian regimes to suppress dissent.
- **Dependence on Technology**: Increasing reliance on advanced AI for visual tasks may lead to a decrease in human skills and awareness in certain fields. Over-dependence on such technology without proper human oversight could have negative implications for critical decision-making processes.

In summary, while the Arcana model holds promise for significant advancements and positive contributions to society, it is crucial to address the associated risks through responsible development, deployment, and regulation to mitigate potential negative impacts.

A.2 Limitations and Future Work

The previous experiments have demonstrated the effectiveness of Arcana. Although the multimodal decoder has proven effective, giving each modality its own learning space significantly increases the number of parameters. While MM-LoRA only adds a small number of parameters to achieve a multimodal decoder, the independent LoRA parameters for different modalities cannot be merged into the LLMs' weights, thereby increasing inference costs. The introduction of QLadder enhances visual representation capabilities, but it comes at the cost of adding visual tokens, which also increases inference costs. Additionally, compared to existing state-of-the-art methods, we used only about 2M training data, limiting Arcana's performance.

To further unlock Arcana's potential, we will design a more efficient multimodal decoder that improves performance while reducing inference costs. We will also focus on designing a more

Table 6: Object hallucination benchmark using POPE evaluation pipeline. "Yes" signifies the likelihood of the model producing a positive response.

Datasets	Metrics	**Arcana (Ours)**	mPLUG-Owl2 [57]	LLaVA-v1.5 [31]	Shikra [5]	InstructBLIP [11]	MiniGPT-4 [63]
Random	Accuracy (↑)	88.87	88.28	88.38	86.90	88.57	79.67
	Precision (↑)	96.59	94.34	96.56	94.40	84.09	78.24
	Recall (↑)	81.27	82.20	80.33	79.27	95.13	82.20
	F1-Score (↑)	88.27	87.85	87.70	86.19	**89.27**	80.17
	Yes (→ 50%)	43.37	44.91	42.89	43.26	56.57	52.53
Popular	Accuracy (↑)	88.07	86.20	87.67	83.97	82.77	69.73
	Precision (↑)	94.06	89.46	94.14	87.55	76.27	65.86
	Recall (↑)	81.27	82.06	80.33	79.20	95.13	81.93
	F1-Score (↑)	**87.20**	85.60	86.69	83.16	84.66	73.02
	Yes (→ 50%)	43.20	45.86	42.67	45.23	62.37	62.20
Adversarial	Accuracy (↑)	86.57	84.12	85.23	83.10	72.10	65.17
	Precision (↑)	90.90	85.54	89.06	85.60	65.13	61.19
	Recall (↑)	81.27	82.13	80.33	79.60	95.13	82.93
	F1-Score (↑)	**85.81**	83.80	84.47	82.49	77.32	70.42
	Yes (→ 50%)	44.70	48.00	45.10	46.50	73.03	67.77

efficient visual encoder that uses fewer visual tokens to represent visual features, enhancing training efficiency and reducing inference costs. Finally, we plan to leverage our data engine to annotate more high-quality caption data to fully unleash Arcana's potential.

A.3 Detailed Evaluation Results.

POPE. We conduct the hallucination evaluation using POPE [29], the results are shown in Table. From the results in the Table 6, we can find Arcana achieves higher F1 scores on the popular and adversarial split, showing the robustness of our model in terms of object hallucination compared to other MLLMs.

MMBench. MMBench [33] is used to evaluate the model's ability of Perception and Reasoning. The detail results for various MLLMs are presented in Table 7.

Table 7: CircularEval multi-choice accuracy results on MMBench [33] dev set. We adopt the following abbreviations: LR for Logical Reasoning; AR for Attribute Reasoning; RR for Relation Reasoning; FP-C for Fine-grained Perception (Cross Instance); FP-S for Finegrained Perception (Single Instance); CP for Coarse Perception.

Method	Language Model	Vision Model	Overall	LR	AR	RR	FP-S	FP-C	CP
MiniGPT-4 [63]	Vicuna-7B	EVA-G	12.0	13.6	32.9	8.9	28.8	11.2	28.3
InstructBLIP [11]	Vicuna-7B	EVA-G	33.9	21.6	47.4	22.5	33.0	24.4	41.1
LLaMA-Adapter-v2 [14]	LLaMa-7B	CLIP ViT-L/14	38.9	7.4	45.3	19.2	45.0	32.0	54.0
LLaVA [32]	Vicuna-7B	CLIP ViT L/14	36.2	15.9	53.6	28.6	41.8	20.0	40.4
Shikra [5]	Vicuna-7B	CLIP ViT-L/14	60.2	33.5	69.6	53.1	61.8	50.4	71.7
LLaVA-v1.5 [31]	Vicuna-7B	CLIP ViT-L/14	64.3	33.1	69.3	57.4	68.9	54.5	76.4
mPLUG-Owl2 [57]	LLaMA2-7B	CLIP ViT-L/14	65.4	29.2	**69.7**	61.7	67.0	**60.0**	79.5
Arcana (Ours)	Vicuna-7B	CLIP ViT-L/14	**67.4**	**34.7**	69.3	**62.6**	**69.6**	58.7	**83.1**

A.4 More visualization results.

To demonstrate the effectiveness and generalization ability of Arcana, we provide more qualitative results in Fig. 6. We visualize its performance across various types of multimodal tasks, including Detail Caption, Detection, Knowledge, OCR-Free Reasoning, Visual Question Answering and ChartQA. To investigate the impact of MM-LoRA and QLadder in multimodal scenarios, we visualized the attention maps of different layers with and without these modules in Fig. 7.

Figure 6: More qualitative results. Main feature in answer is highlight in **orange**.

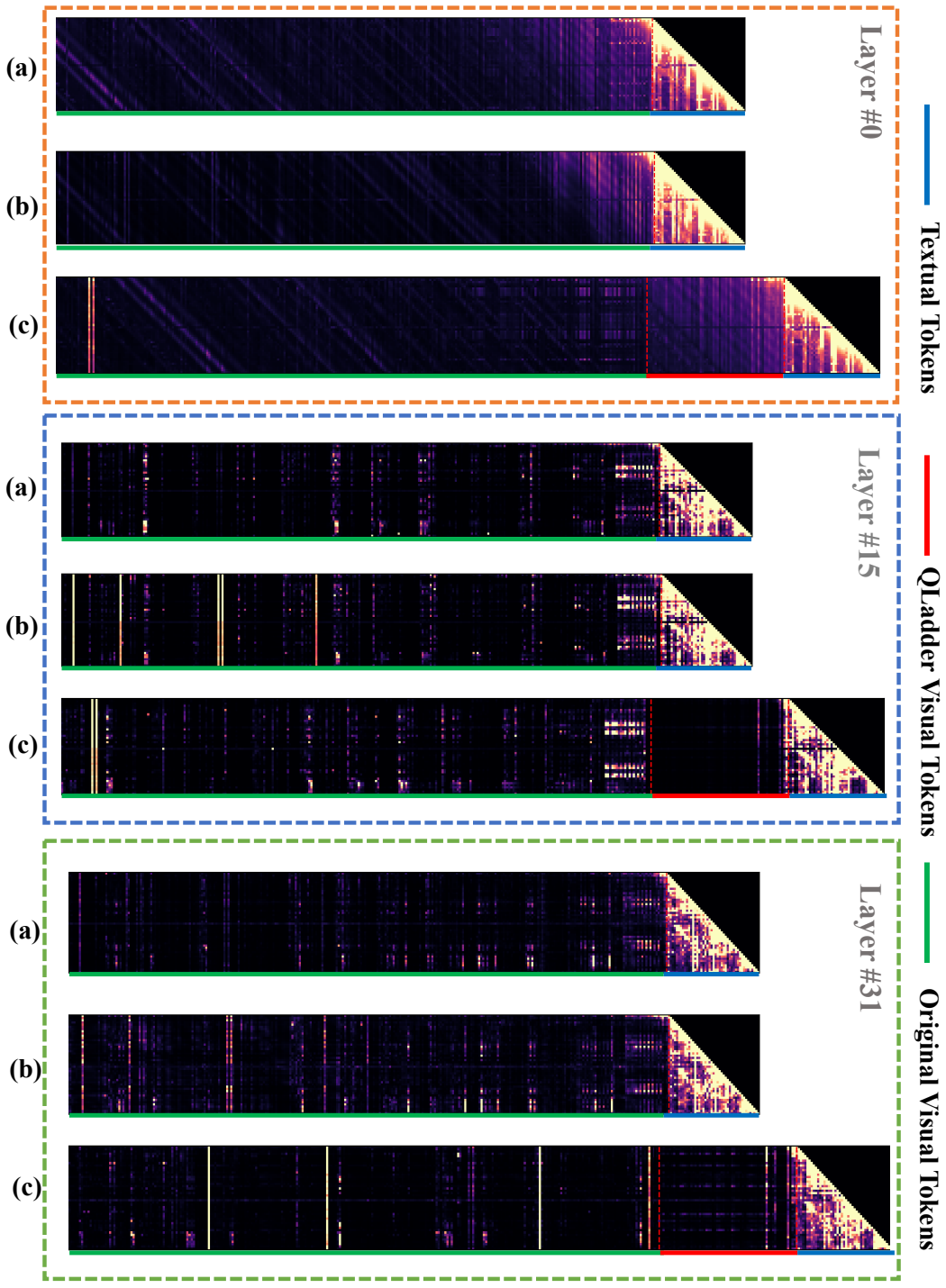

Figure 7: **Visualization of attention maps**. We compare the attention maps in different layer of LLM between different composition, include **(a)** Baseline, **(b)**Baseline+MM-LoRA, and **(c)** Baseline+MM LoRA+QLadder. Higher brightness indicates higher attention values, with the x-axis representing all tokens, and the y-axis containing only text tokens.

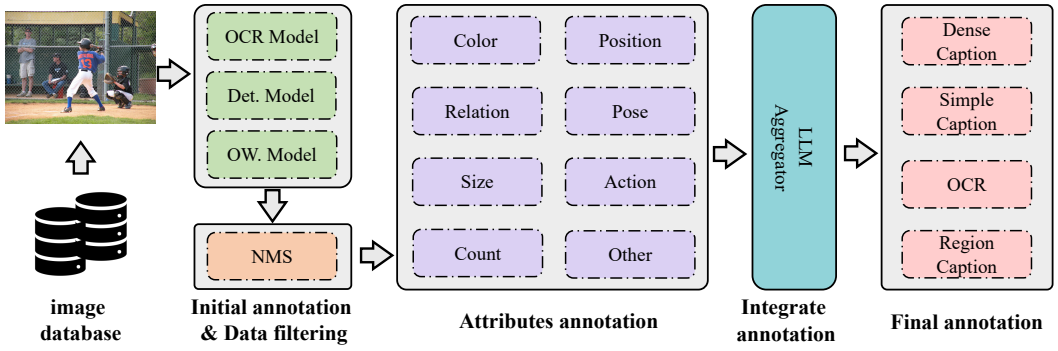

Figure 8: The **Arcana data engine** involves three crucial steps: **(1)** initializing and filtering image annotations, **(2)** obtaining diverse attribute annotations for each annotated region, and **(3)** using a large language model to integrate these visual annotations into different types of captions.

A.5 Data Engine

The LLM model lacks the ability to comprehend image information, hence introducing an image translation task during the SFT stage accelerates MLLM's understanding of image content. The primary objective of the image translation task is to comprehensively describe all information within the image. However, existing caption data is overly simplistic, overlooking numerous important details within images, which hampers MLLM's comprehension of images. Therefore, we devised a pipeline to acquire comprehensive annotations for images. The process, depicted in Fig. 8, consists of four stages: **initial annotation and data filtering**, **attribute annotation**, **integrate annotations** to obtain final annotation information.

Initial annotation and data filtering. To obtain initial annotation information, we utilize OCR models [44, 65], detection models [7, 39], and open-word classification models to extract text from images, pinpoint precise locations of objects, and identify all categories present within the images. Although initial annotations obtained from specialized models offer comprehensive information, they are susceptible to noise and inaccuracies. To address this challenge, we implement a multifaceted filtering process to refine and eliminate unnecessary annotations. Specifically, we aggregate results using NMS and apply thresholding to filter out noisy annotations.

Attribute annotation. To acquire various attributes of different objects in images, we introduce multiple existing models [61, 55, 22, 16] for obtaining attribute annotations of targets. Specifically, these annotations encompass information regarding attributes such as color, position, relationships, pose, size, actions, quantity, and more. These attribute annotations are aggregated in textual form.

Integrate annotation. To consolidate the discrete annotation results into a detailed caption for translating image content, we introduce a large language model. This model assists in integrating the aforementioned discrete annotation results to generate detailed captions for the images and annotated regions within them. Ultimately, we obtain dense captions and simple captions at the image level, region captions at the object level, and OCR annotations within the image.

Ultimately, each image ultimately receives five types of visual annotations: dense caption, simple caption, OCR, detection, and region caption. Here is a brief overview of each type:

- Dense Caption: This detailed description of object attributes such as color, behavior, and relationships enhances the model's understanding of image content, potentially improving its ability to generate accurate and informative captions.
- Simple Caption: These captions focus on significant events in the image, aiding the model in capturing the essence of the visual scene and generating concise and contextually relevant captions.
- OCR: Incorporating text information present in images through OCR enhances the model's ability to describe textual elements such as signs, labels, or captions within the image, enriching the generated captions with textual context.

- Detection: The positional information of objects obtained through detection enables the model to spatially ground its generated captions, ensuring that descriptions accurately correspond to the locations of objects in the image.
- Region Caption: By providing information about specific regions in the image, this data component helps the model localize objects and understand their spatial relationships, contributing to the precision and coherence of generated captions.

Table 8: Influence of the incorporation of the detailed caption data.

Detailed Caption	SQA_I	TextVQA	POPE	MME	MM-Vet	SEED
✗	66.8	58.2	85.9	1510.7	31.1	58.6
✓	**69.6**$_{\uparrow 2.8}$	**59.4**$_{\uparrow 1.2}$	**86.6**$_{\uparrow 0.7}$	**1519.1**$_{\uparrow 8.4}$	**31.4**$_{\uparrow 0.3}$	**62.1**$_{\uparrow 4.5}$

A.6 Influence of fine-grained caption.

In Fig. 9, we compare our enhanced caption data with the original caption data used in the pre-training stage. It is clearly that our enhanced annotations provide a more fine-grained description of the images, thereby improving the model's visual perception without increasing the data volume. To validate the effectiveness of enhanced annotations, we ensured the use of the same model structure and training method as LLaVA-v1.5-7B [31], and introduced our enhanced caption data. The results are presented in the Table. 8. We observe a significant improvement simply by incorporating our caption data into the training, demonstrating the benefits of fine-grained image descriptions for enhancing model visual perception.

Coarse-grained Caption

auckland skyline, with view of the famous landmark in the foreground

two house facades with skylights on and a few windows above them

maggie hurwitz and jennifer anderson in'the americans '

Fine-grained Caption

Image-level Caption

{'from': 'human', 'value': '<image>\nProvide a detailed description of the given image'}, {'from': 'gpt', 'value': '**A man in a wetsuit** is skillfully surfing on **a white surfboard** on a wave in the ocean. The surfboard, made of lightweight materials like **foam or fiberglass**, floats on the water as the man navigates the waves. The large blue-green wave is moving towards the shore, **where two surfers are waiting to catch it**. They are lying on their surfboards, ready to ride the approaching wave. The foam wave created by the surfer riding the wave is white and foamy, indicative of it breaking and forming a crest. The surfer, a man, actively engages in surfing, **catching and riding the waves on the surfboard**. The sunny day adds to the scenic ocean view as the surfers enjoy the waves'}

Region-level Caption

{'from': 'human', 'value': '<image>\nCould you elaborate on the characteristics of the area represented by (0.002, 0.183, 0.998, 0.384)?'}, {'from': 'gpt', 'value': 'It is **a large wave** in the ocean. The wave is blue-green in color, indicating that it is likely a sunny day. The wave is relatively small, but it is still powerful enough to create a good surfing experience. The wave is moving towards the shore, where two surfers are waiting to catch it. The surfers are lying on their surfboards, which are positioned horizontally in the water, ready to ride the wave as it approaches.'}, {'from': 'human', 'value': 'I would like a comprehensive depiction of the region bounded by (0.000, 0.111, 0.996, 0.997).'}, {'from': 'gpt', 'value': 'It is **a surfboard**. It is white and made of lightweight materials, such as foam or fiberglass. The surfboard is designed to float on the water and is used by the surfer to ride the waves. The surfboard is being ridden by **a man in a wetsuit**, who is skillfully navigating the waves in the ocean.'}, {'from': 'human', 'value': 'Provide a detailed description of the region defined by the coordinates (0.582, 0.345, 0.946, 0.664).'}, {'from': 'gpt', 'value': 'It is **a foam wave** created by the surfer riding the wave. The wave is white and foamy, indicating that it is breaking and forming a crest. The wave is relatively small, and the surfer is skillfully riding it on top of a surfboard. The surfer is a man, and he is actively engaged in the sport of surfing, which involves catching and riding waves on a surfboard}

Figure 9: LLaVA's coarse-grained caption v.s. Our fine-grained caption. Important visual recognition-related description are highlighted in orange.