Libra: Building Decoupled Vision System on Large Language Models

Yifan Xu [1 2 3] **Xiaoshan Yang** [1 2 3] **Yaguang Song** [2] **Changsheng Xu** [1 2 3]

Abstract

In this work, we introduce **Libra**, a prototype model with a decoupled vision system on a large language model (LLM). The decoupled vision system decouples inner-modal modeling and cross-modal interaction, yielding unique visual information modeling and effective cross-modal comprehension. Libra is trained through discrete auto-regressive modeling on both vision and language inputs. Specifically, we incorporate a routed visual expert with a cross-modal bridge module into a pretrained LLM to route the vision and language flows during attention computing to enable different attention patterns in inner-modal modeling and cross-modal interaction scenarios. Experimental results demonstrate that the dedicated design of Libra achieves a strong MLLM baseline that rivals existing works in the image-to-text scenario with merely 50 million training data, providing a new perspective for future multimodal foundation models. Code is available at https://github.com/YifanXu74/Libra.

1. Introduction

The integration of vision and language plays a vital role in machine perception and understanding of the world. Language serves as the basis for cognitive processing, while vision provides essential sensory information. In this context, the field of multimodal large language models (MLLMs) (Yin et al., 2023) has made remarkable progress, yielding impressive results across various domains, including multimodal conversation (Google, 2023b; OpenAI, 2023), interactive agents (Hong et al., 2023), and even autonomous driving (Cui et al., 2024).

A line of recent works (Lu et al., 2022a; 2023; Huang et al., 2023; Peng et al., 2023; Wang et al., 2022a) jointly trains multimodal models from scratch, naturally aligning vision and language under the unified structure design and modeling paradigm. However, these approaches often compromise on unified but not general models due to an information imbalance: general intelligence in the era of foundation models demands a large scale of language knowledge, but unfortunately, visual data falls short in matching the scale of language. For instance, Unified-IO (Lu et al., 2022a) achieves cross-modal comprehension by making sacrifices in certain language capabilities, especially the wide range of world knowledge and chat ability.

In light of this, another line of works (Alayrac et al., 2022; Liu et al., 2023b; Bai et al., 2023; Sun et al., 2023b), LLM-based approaches, follows a staged training paradigm: first training a large language model (LLM) to acquire a wide range of general knowledge, then integrating visual perception into the pretrained LLM. This paradigm is reasonable because it can efficiently transfer the general knowledge learned by the language model to the MLLMs. To this end, on the basis of well-built language systems like LLaMA (Touvron et al., 2023), building an effective vision system for essential visual sensory and a reasonable cross-modal interaction strategy for cross-modal comprehension upon LLMs becomes a natural idea.

A straightforward approach in most recent LLM-based works is to employ a pretrained vision encoder like CLIP (Radford et al., 2021) as the vision system, integrating its features into a pretrained LLM to facilitate cross-modal interaction, *e.g.*, through a trainable Q-Former (Li et al., 2023d) or a simple projection layer (Liu et al., 2023b). This integration is achieved through an image-captioning loss, where the supervision is only performed on the language part. However, this pipeline leads to a weak vision system because its visual understanding ability is limited by the pretrained vision encoder. To address this, several works (Dong et al., 2023), such as Emu (Sun et al., 2023b), attempt to directly build more sophisticated vision systems on LLMs. They perform contiguous auto-regressive image modeling, where each input visual feature predicts the input feature of the next position. Despite impressive image generation results, these works provide limited benefits to downstream tasks because 1) the unified architecture makes coupled vision and language systems, thereby losing unique visual

[1]MAIS, Institute of Automation, Chinese Academy of Sciences [2]Peng Cheng Laboratory [3]School of Artificial Intelligence, University of the Chinese Academy of Sciences. Correspondence to: Changsheng Xu <csxu@nlpr.ia.ac.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

information; 2) the contiguous vision supervision raises an infinite label space that increases the learning difficulty.

In this work, we aim to build a more reasonable vision system upon LLMs. From a biological perspective (Thiebaut de Schotten & Forkel, 2022), vision and language systems can exist independently, while vision-language comprehension requires further cross-modal interaction. This inspires us to consider *what is an ideal vision system on LLMs*. We believe that the following two aspects are equally important. 1) To retain an extensive and in-depth visual understanding ability, the vision system should be relatively independent from the language model due to the information imbalance. 2) To facilitate cross-modal comprehension, vision systems should be altruistic in aligning the vision and language features.

Based on the above inspiration, we propose to learn a decoupled vision system on LLMs, and build up a new prototype MLLM model **Libra**. We found that Libra is a strong MLLM baseline with limited training data (50M in this work vs. 1B in previous works (Li et al., 2023c)). The decoupled vision system of Libra can simultaneously retain unique visual information and support the cross-modal interaction, which is achieved by the following designs.

Routed visual expert. The core of the decoupled vision system is a routed visual expert module relied on LLMs, which comprises a simple visual expert and a cross-modal bridge module. Firstly, the visual expert has its own vision-specific parameters. It resembles a mixture of experts (MoE) (Jacobs et al., 1991; Fedus et al., 2022) structure, featuring an additional attention layer and a feed-forward network (FFN) for vision features alongside the existing frozen layers in the LLM for language features. Secondly, the cross-modal bridge module enables cross-modal interaction, routing the vision and language flows during attention computing to enable different attention patterns in inner-modal modeling and cross-modal interaction scenarios.

Discrete auto-regressive modeling. The vision system of Libra is learned through a discrete next-token-prediction paradigm on vision inputs, enabling a finite label space for stable learning of the vision system compared to previous contiguous image modeling approaches (Sun et al., 2023b;a). We focus on the image-to-text scenario, where the vision system (routed visual expert) learns unconditional image modeling, and the language system (LLM) learns vision-conditioned language modeling.

Hybrid image tokenization. A side effect of discrete auto-regressive modeling is the information loss brought by image discretization. To mitigate this, we propose a hybrid tokenization strategy that combines contiguous visual signals from the vision encoder with discrete modeling using tokenized ids. To leverage the pretrained knowledge of well-established vision encoders like CLIP (Radford et al., 2021), we construct a CLIP-based image tokenizer using lookup-free quantization (LFQ) (Yu et al., 2023a). This is the first time that a highly reconstructive image tokenizer can be constructed upon a frozen vision encoder like CLIP, which has not even been investigated in the work of LFQ.

With the dedicated designs, we demonstrate some noteworthy behaviors:

- We provide a new perspective for the design of MLLMs by modeling a decoupled vision system that decouples inner-modal modeling and cross-modal interaction.
- The decoupled vision system enhances the attention diversity across layers, reducing the learning redundancy and improving vision-language comprehension.
- Libra rivals modern MLLMs across more than 15 multimodal benchmarks, despite limited training data.

2. Related Work

Rapid developments have been witnessed in multimodal large language models (MLLMs) (Yin et al., 2023) that enable human interaction with both words and visual content. One line of works (Wu et al., 2023; Gupta & Kembhavi, 2023; Shen et al., 2023; Surís et al., 2023; Yang et al., 2023) utilizes LLMs as central controllers, integrating them with various functional agents, with language serving as a general interface. This plugin-style framework achieves remarkable success with very low training cost. Another line of works explores directly training MLLMs, including scratch training with unified architectures (Wang et al., 2022a; Lu et al., 2022a; Google, 2023b), integrating pretrained vision encoders with pretrained LLMs through simple projections (Liu et al., 2023b; Wang et al., 2023b; Bai et al., 2023; Li et al., 2023b; Dong et al., 2023; Sun et al., 2023b) or cross-attention (Alayrac et al., 2022). Several training strategies are proposed to reduce the training burden, including instruction tuning (Xu et al., 2022d; Liu et al., 2023b) and parameter-efficient tuning (Hu et al., 2021; Dettmers et al., 2023; Zhang et al., 2023).

The most related studies to our work are Emu (Sun et al., 2023b) and CogVLM (Wang et al., 2023b). Emu performs contiguous auto-regressive image modeling using a CLIP vision encoder and a diffusion (Rombach et al., 2022) image decoder. CogVLM proposes a visual expert module on frozen LLMs to achieve deeper alignment between vision and language. Both works demonstrate that the contiguous modeling paradigm does not provide evident benefits for image-to-text vision-language comprehension, despite remarkable text-to-image generation capability. Instead, we show the importance of stable discrete image modeling with a reasonable cross-modal interaction strategy, which enables

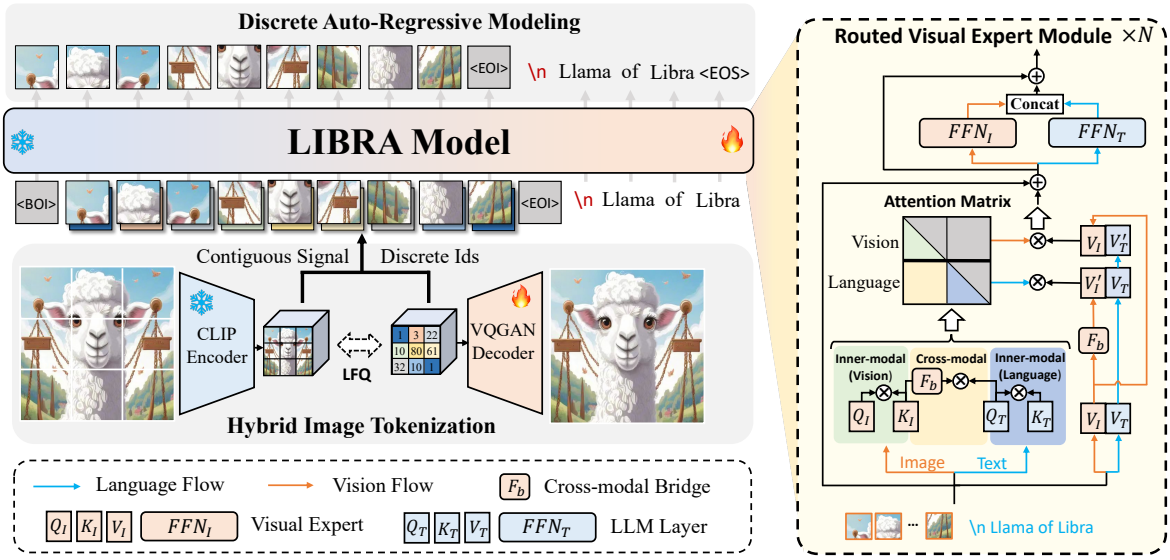


Figure 1. Libra investigates a decoupled vision system on the pretrained LLM. The vision system is built with a routed visual expert design. We train Libra through discrete auto-regressive modeling. The vision inputs consist of a hybrid of contiguous signals from the vision encoder and discrete "word" embeddings constructed based on the tokenized ids. <EOS> is the end-of-sequence token. In practice, the discrete ids are used to construct discrete vision embeddings from a codebook learned by auto-regressive image modeling of Libra.

an effective vision system on LLMs, ultimately enhancing vision-language comprehension.

3. Approach

3.1. Architecture

Libra comprises three fundamental components: a unified input tokenizer, a pretrained large language model, and a routed visual expert with a cross-modal bridge module. Fig. 1 illustrates an overview of Libra. The implementation details are introduced as follows.

Unified Input Tokenizer. Libra unifies both vision and language modeling into a *discrete* next-token-prediction paradigm. Given an input sequence with both image and corresponding language parts, we separately tokenize the image and the language parts into discrete tokens through a VQGAN (Esser et al., 2021) and a SentencePiece (Kudo & Richardson, 2018) tokenizer. We respectively prefix and suffix the image sequence with a ⟨BOI⟩ (beginning of image) token and a ⟨EOI⟩ (end of image) token. We use a newline token "\n" to separate images and texts. All token embeddings except the separation newline token "\n" are updated through a cross-entropy classification loss. Meanwhile, integrating images into completely discrete tokens results in severe information loss, as verified in Sec. 4.4. Therefore, we propose a hybrid image tokenization process (Sec. 3.2) to enable both stable discrete sequential modeling and contiguous visual comprehension.

Pretrained Large Language Model. Libra's model design is compatible with any off-the-shell GPT-style pretrained LLMs. We adopt the commonly used LLaMA2-7B-Chat (Touvron et al., 2023) for further training. We freeze the LLM during pretraining and unfreeze it during instruction tuning (see Sec. 3.3).

Routed Visual Expert. As described in Sec. 1, the LLM-based approaches are built upon a wide range of general knowledge brought by the LLM. A decoupled vision system can preserve unique visual information without distorting the inherent knowledge within the LLM. Therefore, we propose a routed visual expert for vision-specific encoding and decoupled cross-modal interaction. We add the routed visual expert to each layer of the LLM and freeze the LLM during pretraining to preserve its language knowledge.

The routed visual expert features: 1) additional attention and FFN layers for vision features alongside the original LLM layers for language features, and 2) a cross-modal bridge for cross-modal interaction. Formally, given the input hidden states $X \in \mathbb{R}^{B\times H\times(L_I+L_T)\times D}$ with the image part X_I of length L_I and the text part X_T of length L_T, where B is the batch size, H is the number of attention heads, and D is

the hidden size. The attention is computed as:

$$\begin{aligned} X_I^a, X_T^a &= \text{Attn}\left(X\right) = \\ &\text{softmax}\left(\frac{\text{Tril}\left(Q \cdot F_b(K)^T\right)}{\sqrt{D}}\right) F_b(V), \\ Q &= \text{concat}\left(X_I W_I^Q, X_T W_T^Q\right), \\ K &= \text{concat}\left(X_I W_I^K, X_T W_T^K\right), \\ V &= \text{concat}\left(X_I W_I^V, X_T W_T^V\right), \end{aligned} \tag{1}$$

where X_I^a, X_T^a is the attention outputs of the vision and language parts, respectively. F_b refers to the cross-modal bridge module introduced in the next part, W_I^*, W_T^* are the QKV matrices of the visual expert and original language model, and $\text{Tril}(\cdot)$ denotes the causal lower-triangular mask. For parameter-efficiency, we represent each visual expert matrix $W_I^* \in \mathbb{R}^{D \times D'}$ as the product of two low-rank matrices, namely: $W_I^* = A_I^* \cdot B_I^*$, where $A_I^* \in \mathbb{R}^{D \times D/4}$ and $B_I^* \in \mathbb{R}^{D/4 \times D'}$. Similarly, the visual expert in feed-forward network (FFN) layers performs as:

$$\text{FFN}(X) = \text{concat}\left(\text{FFN}_I\left(X_I\right), \text{FFN}_T\left(X_T\right)\right), \tag{2}$$

where FFN_I, FFN_T are the FFNs of the visual expert and the original language model.

It is worth noting that the visual expert design here is similar to the one proposed in CogVLM (Wang et al., 2023b). The differences lie in various aspects. 1) *Approach*: we further introduce a cross-modal bridge module to decouple inner-modal modeling and cross-modal interaction. 2) *Goal*: Libra uses the visual expert design as one reasonable path to achieve a decoupled vision system upon frozen LLMs, while CogVLM only uses it for better vision-language alignment. 3) *Insight*: we demonstrate that effective image modeling of the vision system significantly enhances vision-language comprehension under the visual expert design (see Sec. 4.4), in contrast to the findings of CogVLM described in Sec. 2.

Cross-modal Bridge. In addition to the modality-specific modeling brought by the visual expert and the LLM, we found that a decoupled cross-modal interaction strategy plays a vital role in cross-modal comprehension. We observed that image modeling fails with a simple visual expert, indicating an invalid vision system (see Fig. 3(a)). This is because a simple visual expert does not really build a decoupled vision system. In image-to-text scenarios, language predictions are based on the image condition. The frozen LLM places the entire learning burden of cross-modal interaction on the visual expert module. Consequently, the vision modeling of the visual expert tends to align with language, resulting in an inability to learn meaningful visual representations.

Therefore, we design a cross-modal bridge module to decouple the inner-modal modeling and cross-modal interaction. The bridge adds an additional learnable projection upon the keys and values when computing cross-modal attention. Formally, given the input hidden states $X = [X_I, X_T]$, the attention keys of the vision part are computed as:

$$\begin{aligned} F_b\left(K_I \mid Q_*, X_I\right) &= \begin{cases} K_I & \text{, if inner-modal,} \\ K_I' & \text{, if cross-modal,} \end{cases} \\ K_I' &= K_I + X_I W_I^{K\prime}, \end{aligned} \tag{3}$$

where $K_I = X_I W_I^K$ in Eqn. (1), $W_I^{K\prime}$ is the learnable transformation projection of the bridge module. Eqn. (3) denotes that: we transform the original keys to new values if Q_* and K_I are from different modalities (cross-modal); if Q_* and K_I are from the same modality (inner-modal), we keep the original keys. Note that the condition suffix of $F_b(\cdot)$ in Eqn. (3) is omitted in the other parts of the paper for concision. Similarly, we can get the attention keys $F_b(K_T)$ of the text part, which is unused in the image-to-text scenario. Finally, the attention matrix can be computed as:

$$Q \cdot F_b(K) = \begin{bmatrix} Q_I K_I^\top & Q_I K_T'^\top \\ Q_T K_I'^\top & Q_T K_T^\top \end{bmatrix}. \tag{4}$$

Similarly, the computing of the bridge module on attention values can be illustrated through Fig. 1, where we transform and keep the original values under cross-modal and inner-modal scenarios, respectively. Formally,

$$\begin{aligned} X_I^a &= \sigma\left(\begin{bmatrix} Q_I K_I^\top & Q_I K_T'^\top \end{bmatrix}\right) \cdot \begin{bmatrix} V_I \\ V_T' \end{bmatrix}, \\ X_T^a &= \sigma\left(\begin{bmatrix} Q_T K_I'^\top & Q_T K_T^\top \end{bmatrix}\right) \cdot \begin{bmatrix} V_I' \\ V_T \end{bmatrix}, \\ V_I' &= V_I + X_I W_I^{V\prime}, \\ V_T' &= V_T + X_T W_T^{V\prime}, \end{aligned} \tag{5}$$

where $V_I = X_I W_I^V$ and $V_T = X_T W_T^V$ in Eqn. (1). $W_I^{V\prime}$ and $W_T^{V\prime}$ are the learnable transformation projections of the bridge module. σ denotes a softmax function. We omit the normalization factor and the causal mask for concision. In practice, V_T' takes no effect during pretraining due to the causality of auto-regression, as the data are always formulated as `<Image>\n <Text>`, with visuals preceding the text.

Last but not least, the transformation brought by the bridge module should not be too large, in order to leverage the learned knowledge in the original keys and values. Thus, we apply a low-rank strategy to the design of the transformation projection $W_I^{*\prime}, W_T^{*\prime} \in \mathbb{R}^{D \times D'}$. Take the vision part as example: $W_I^{*\prime} = A_I^{*\prime} \cdot B_I^{*\prime}$, where $A_I^{*\prime} \in \mathbb{R}^{D \times 8}$ and $B_I^{*\prime} \in \mathbb{R}^{8 \times D'}$.

3.2. Hybrid Image Tokenization

The unified discrete next-token-prediction paradigm raises two obstacles for effective vision-language comprehension.

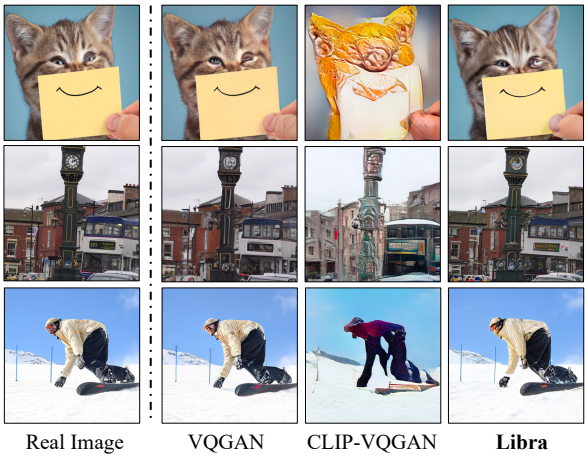


Figure 2. Image reconstruction results. Directly replacing the image encoder of VQGAN with CLIP distorts the visual information. Libra largely alleviates this problem via lookup-free quantization.

1) The discretization process of VQGAN can cause severe visual information loss, leading to low perception on visual details. 2) Naive discrete sequential modeling hardly benefits from the pretrained knowledge of the vision encoder, since the model receives newly-constructed embeddings based on the input ids instead of the features of the vision encoder. To this end, as shown in Fig. 1, we propose a hybrid image tokenization process from two aspects: contiguous visual signals and pretrained visual knowledge.

Contiguous Visual Signal vs. Discrete Modeling. We leverage a hybrid tokenization strategy with a combination of contiguous visual signals from the vision encoder and discrete modeling using tokenized ids. Specifically, as illustrated in Fig. 1, given an input image, we first feed it into the vision encoder and obtain the output features as contiguous visual signals. Then, a quantization/discretization process is performed to tokenize the contiguous visual signals into discrete token ids based on a vision codebook. The token ids are used to construct vision "word" embeddings similar to the ones in LLMs. Finally, we concatenate the contiguous visual signals and the discrete vision embeddings in the channel dimension as the final vision inputs of the Libra model. Sec. A.1 provides more details of the tokenization process. Meanwhile, a discrete auto-regressive image modeling is performed on the output features of Libra, *i.e.*, each vision input is used to predict the token id of the next position. This simple design enables both contiguous visual comprehension and stable discrete sequential modeling.

Pretrained Visual Knowledge. To leverage the pretrained knowledge in existing well-established vision encoders like CLIP (Radford et al., 2021), we replace the vision encoder of VQGAN with a *frozen* CLIP-ViT-L-336px. However, training a CLIP-based VQGAN is non-trivial. The features in CLIP are highly semantic with less low-level visual information. Directly emulating such features in the quantization process of the original VQGAN results in poor reconstruction performance, as demonstrated in Fig. 2 (CLIP-VQGAN). Instead, we find that the lookup-free quantization (LFQ) (Yu et al., 2023a), which does not need to emulate the input features, can largely address this problem. To this end, we use a CLIP-based VQGAN with LFQ as Libra's image tokenizer (Libra in Fig. 2). This is the first time that a highly reconstructive image tokenizer can be constructed based on a frozen vision encoder like CLIP, which has not even been investigated in the work of LFQ.

We train our image tokenizer using 10M images collected by (Kirillov et al., 2023). The vision vocabulary size is enlarged to 2^{18} thanks to LFQ. For computational efficiency, we predict in two concatenated codebooks, each of size 2^9. More details can be found in Sec. A.1.

3.3. Training

Pretraining. Libra is pretrained under unified sequential modeling, where a next-token-prediction objective is performed on all input tokens, as:

$$p(X) = \prod_{\ell=1}^{L} p\left(X_\ell \mid X_{<\ell}\right), \tag{6}$$

where $X = [X_I, X_T]$ is the input multimodal sequence and L is the sequence length. In practice, the objective is computed through a discrete cross-entropy classification loss. We use image-text pairs for training. We freeze the LLM and only update the routed visual expert during pretraining.

Multimodal Instruction Tuning. Language instruction tuning has helped LLMs to align with user intentions (Ouyang et al., 2022; Wang et al., 2022b) and generalize to unseen tasks (Wei et al., 2021; Chung et al., 2022). Similarly, we apply multimodal instruction tuning on the pretrained Libra model. We train the whole model during tuning. All instruction-tuning data are arranged based on this template:

```
<System Message>
[USER]: <Image> <Instruction>
[ASSISTANT]: <Answer>
```
(7)

where only `<Answer>` is accounted for computing loss, as:

$$p(X_a \mid X_v, X_{instruct}) = \prod_{\ell=1}^{L} p\left(x_\ell \mid X_v, X_{instruct}, x_{<\ell}\right), \tag{8}$$

where $X_a = \{x_\ell\}_{\ell=1}^{L}$, X_v, $X_{instruct}$ are the answers, images, and instructions.

Data. In this work, we only build Libra as a prototype model, thereby using much less pretraining data than most

Table 1. Performance comparison on visual question answering (VQA) and image captioning. Specialists perform dataset-specific finetuning, while generalists commonly perform zero-shot evaluation. The pretraining data sizes are reported. *The training images of the datasets are observed during training. †Includes in-house data that is not publicly accessible.

Method	#Params	#Data	General VQA					Image Caption		
			VQAv2	OKVQA	GQA	VizWiz	SQA	NoCaps	Flickr	COCO
Specialists										
BEiT-3 (Wang et al., 2023a)	1.9B	50M	84.0	-	-	-	-	-	-	147.6
PaLI-X (Chen et al., 2023b)	55B	-	86.1	66.1	-	70.9	-	124.3	-	149.2
OFA (Wang et al., 2022a)	930M	60M	82.0	-	-	-	-	-	-	154.9
CogVLM (Wang et al., 2023b)	17B	1.5B†	84.7	64.7	65.2	75.8	92.7	126.4	94.9	144.9
Generalists										
BLIP-2 (Li et al., 2023d)	12.1B	129M	65.0	45.9	44.7	-	-	121.6	74.9	**144.5***
Flamingo (Alayrac et al., 2022)	80B	2.1B†	56.3	50.6	-	31.6	-	-	67.2	84.3
Unified-IO$_{XL}$ (Lu et al., 2022a)	2.9B	-	77.9*	54.0*	-	57.4*	-	100.0	-	122.3*
PaLM-E (Driess et al., 2023)	12B	70M†	76.2*	55.5*	-	-	-	-	-	135.0*
InstructBLIP (Dai et al., 2023)	14.2B	129M	-	-	49.5	33.4	63.1	121.9	82.8	104.2*
Emu (Sun et al., 2023b)	14B	4B	40.0*	34.7	-	35.4	-	-	-	117.7*
Qwen-VL (Bai et al., 2023)	9.6B	1.4B†	78.2*	56.6	57.5*	38.9	68.2	120.2	81.0	-
Shikra (Chen et al., 2023a)	13.3B	600K	77.4*	53.8	-	-	-	-	73.9*	117.5*
IDEFICS (IDEFICS, 2023)	80B	353M	60.0	-	45.2	36.0	-	-	-	-
LLaVA1.5 (Liu et al., 2023a)	13.4B	558K	**80.0***	-	63.3*	53.6	71.6	-	-	129.8*
Libra (ours)	11.3B	50M	77.3*	**59.7**	**63.8***	**59.5**	**73.5**	**123.8**	**86.6**	135.2*

of previous works. For pretraining, we use 50M image-text pairs randomly sampled from COYO-700M (Byeon et al., 2022) and CC12M (Changpinyo et al., 2021). We use additional 500K image-text pairs from COCO (Chen et al., 2015) training split to standardize the caption outputs. For instruction tuning, we leverage the 665K high-quality supervised data from LLaVA-Instruct (Liu et al., 2023b). More training details can be found in Sec A.2.

4. Experiments

4.1. Implementation

Libra consists of 11.3 billion parameters, with 7 billion from the LLM, 4 billion from the routed visual expert, and 0.3 billion from the CLIP vision encoder. We conduct comprehensive evaluation on various tasks, including visual question answering (VQA), image captioning, and MLLM-oriented multimodal benchmarks. We refer to Sec. B.1 for more details on the evaluation benchmarks and metrics. All evaluations are performed based on greedy search for replication.

4.2. Vision-Language Comprehension

Visual Question Answering and Image Captioning. We evaluate Libra on a wide range of academic benchmarks, including 5 popular general VQA benchmarks and 3 image captioning benchmarks. Tab. 1 shows the results. Libra exhibits strong generalization capabilities in zero-shot captioning and question answering tasks, surpassing previous generalist models with more parameters or larger pretraining data sizes, *e.g.*, it achieves a notable improvement of +20.6% on the VizWiz dataset compared to Qwen-VL, despite using only 4% of the pretraining data. Moreover, when using the same instruction tuning data, Libra outperforms LLaVA1.5 on zero-shot tasks, indicating the effectiveness of the vision system in Libra.

MLLM-oriented Multimodal Benchmarks. Recent studies (Fu et al., 2023; Liu et al., 2023c) found that traditional academic benchmarks often fall short in providing a comprehensive ability assessment. To fully evaluate the generality of MLLMs, research communities have introduced a series of benchmarks. We evaluate Libra on 8 MLLM-oriented multimodal benchmarks in Tab. 2. We highlight the best and second-best results in bold and underlined, respectively. The results confirm that Libra rivals existing modern MLLMs.

4.3. Visual Sequential Modeling

Despite promising results in vision-language comprehension tasks in Sec. 4.2, the metrics cannot directly reflect the effectiveness of the vision system. A vision system with good visual representation should at least learn the basic image distribution. Therefore, we examine the vision system of Libra from the perspectives of image completion and text-to-image generation. We disable the contiguous visual signal (see Sec. 3.2) by replacing it with zero values to enable image generation. For text-to-image generation, we

Table 2. The zero-shot evaluation on MLLM-oriented multimodal benchmarks.

Method	LLM	#Data	POPE	MME	MME^C	MMB	MMB^{CN}	SEED	MM-Vet	MMVP
BLIP-2 (Li et al., 2023d)	FlanT5-11B	129M	85.3	1293.8	290.0	-	-	46.4	22.4	-
InstructBLIP (Dai et al., 2023)	FlanT5-11B	129M	78.9	1212.8	291.8	-	-	-	-	16.7
Shikra (Chen et al., 2023a)	Vicuna-13B	600K	-	-	-	58.8	-	-	-	-
MiniGPT-4 (Zhu et al., 2023)	Vicuna-13B	3.5K	-	581.6	144.2	23.0	-	42.8	22.1	12.7
Otter (Li et al., 2023b)	LLaMA-7B	2.1B	-	1292.2	-	48.3	-	32.9	24.6	-
IDEFICS (IDEFICS, 2023)	LLaMA-65B	353M	-	-	54.5	38.1	-	-	-	-
mPLUG-Owl (Ye et al., 2023)	LLaMA-7B	1.2B	-	967.3	276.0	46.6	-	34.0	-	-
Qwen-VL (Bai et al., 2023)	Qwen-7B	1.4B	-	1487.5	**360.7**	60.6	56.7	58.2	-	-
LLava1.5 (Liu et al., 2023a)	Vicuna-7B	558K	85.9	**1510.7**	274.3	64.3	58.3	58.6	30.5	24.7
Libra (ours)	LLaMA2-7B	50M	**88.2**	1494.7	281.1	**65.2**	**58.8**	**62.7**	**31.8**	**30.0**
Commercial Chatbots										
GPT-4V (OpenAI, 2023)	-	-	-	1409.4	517.1	77.0	74.4	71.6	-	38.7
Gemini-Pro (Google, 2023b)	-	-	-	1496.5	436.7	73.6	74.3	70.7	-	40.7
Bard (Google, 2023a)	-	-	-	-	-	-	-	-	-	19.0

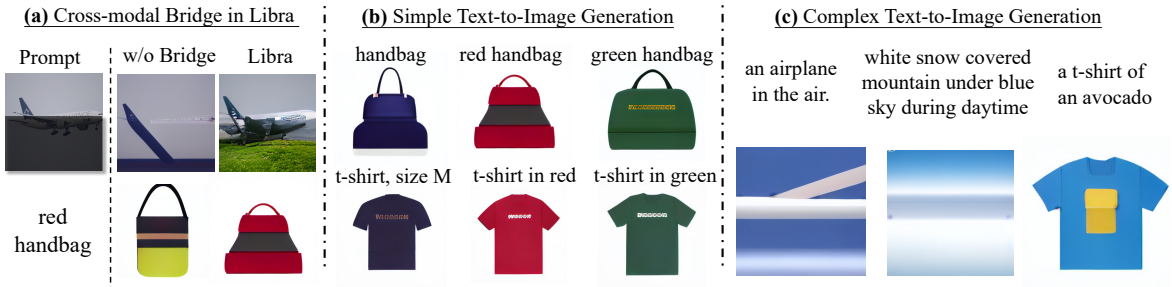


Figure 3. Results of visual sequential modeling.

further finetune Libra with additional 10 million text-image pairs (7B tokens in total) from the pretraining data. Note that our aim here is to validate the effectiveness rather than striving for state-of-the-art performance.

Cross-modal Interaction with Cross-modal Bridge. Fig. 3(a) shows the image completion and text-to-image generation results of Libra and its variant without the cross-modal bridge module. The results show that the variant without the cross-modal bridge module learns a coupled and weak vision system, which: 1) only learns repetitive patterns in image completion, and 2) hardly follows the language instruction during text-to-image generation. A reasonable cross-modal interaction strategy brought by the cross-modal bridge largely boosts effective vision system learning on an LLM. Tab. 3(e) also quantitatively shows the effectiveness of the cross-modal bridge module.

Naive Image Generation. The text-to-image generation results in Fig. 3(b)(c) indicate that Libra can learn basic structures and concepts (*e.g.*, colors). Fig. 3(c) shows that Libra fails under complex text-to-image generation. This might be due to the limited training data (training tokens: 7B in Libra vs. 400B in DALL-E (Ramesh et al., 2021)). Despite naive image generation performance, the results sufficiently prove that Libra learns the basic image distribution.

4.4. Discussion

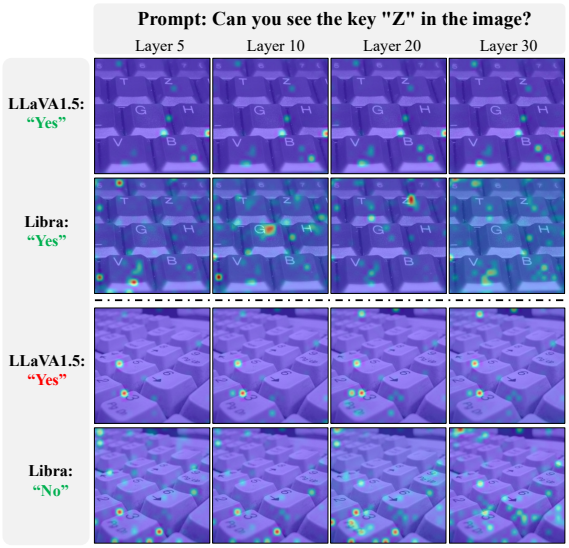
Impact of a Decoupled Vision System. We found that a decoupled vision system impacts the following aspects:

(1) *Attention diversity*. We analyze Libra's attention patterns in Fig. 4. In Fig. 4(a), LLaVA1.5 (Liu et al., 2023a) shows a consistent attention pattern across layers in VQA, while Libra exhibits diverse attention patterns across layers. The implementation details can be found in Sec. B.2. Quantitative results in Fig. 4(b)(c), averaged 100 VQA samples, reveal interesting findings: 1) In Fig. 4(b), LLaVA1.5 demonstrates distinct attention patterns only in few middle and deep layers, while Libra shows diverse attention patterns across all layers. 2) In Fig. 4(c), LLaVA1.5 exhibits low inner-layer attention differences in shallow layers, whereas Libra diversifies attention patterns across all layers. These results suggest that the decoupled vision system has lower learning redundancy in both cross-layer and inner-layer aspects, as observed through diverse attention patterns.

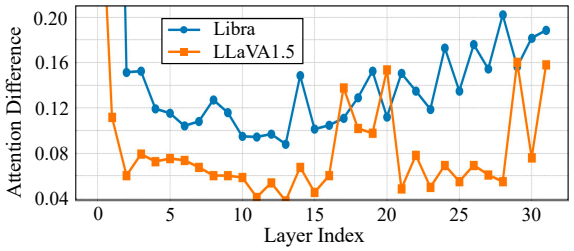
(2) *Learning bias*. The MMVP (Tong et al., 2024) benchmark in Tab. 2 is designed to detect the perception bias in

Table 3. Ablation results on VQA and MLLM benchmarks. *The training images of the datasets are observed during training.

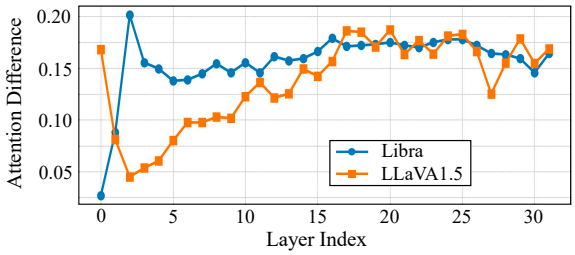
Ablated Setting		Ablated Details	Libra Original Value	→	Changed Value	#Params	General VQA VQAv2*	GQA*	VizWiz	MLLM Benchmark MME	POPE	SEED
Libra model						11.3B	77.3	63.8	59.5	1494.7	88.2	62.7
Training	**(a)**	Paradigm	Unified		Language	11.3B	77.0	60.8	52.9	1465.5	86.0	58.8
	(b)	Supervision	Discrete		Contiguous	11.3B	77.2	60.5	53.3	1473.4	85.2	58.2
	(c)	#Data	50M		3M	11.3B	67.8	54.9	48.8	1189.8	82.1	51.8
Architecture	**(d)**	Expert	✓		✗	7.5B	77.0	61.4	50.8	1450.2	85.9	58.4
	(e)	Bridge	✓		✗	11.2B	76.3	61.4	53.6	1458.4	86.0	59.6
	(f)	Input	Hybrid		Discrete	11.3B	65.0	53.1	38.6	1127.8	80.3	48.5
	(g)	Vision Encoder	CLIP		Scratch	11.3B	67.2	55.7	40.6	1148.4	80.7	50.8



(a) Attention visualization



(b) Cross-layer attention difference



(c) Inner-layer attention difference

Figure 4. Attention patterns across layers. (a) Attention activation of single-word answers on images. (b) Cross-layer attention difference: the difference between each layer's attention score (averaged across all heads) and the mean value of all layers, averaged along the spatial dimension. (c) Inner-layer attention difference: the difference between each head's attention score and the mean value of all heads in each layer, averaged along the spatial dimension. The implementation details can be found in Sec. B.2.

CLIP-based MLLMs, where most MLLMs perform even lower than random guess (25%). Libra achieves remarkable performance on this benchmark, clearly surpassing previous MLLMs (*e.g.*, +5.3% over LLaVA1.5). This indicates that a decoupled vision system preserves unique visual information through discrete auto-regressive image modeling. This resembles a regularization effect in previous self-supervised pretraining approaches (He et al., 2022; Caron et al., 2021), alleviating the learning bias in MLLMs.

(3) *General performance*. To verify the impact of image modeling on the vision system, we only supervise the language part of Libra, as shown in Tab. 3(a). The results show that Libra has an obvious performance degradation without the discrete image modeling. This might be because image modeling encourages unique visual information learning, enabling meaningful visual representation.

Discrete Modeling vs. Contiguous Modeling. Previous studies (Wang et al., 2023b; Sun et al., 2023b) show that a contiguous image modeling paradigm in pretraining makes limited benefits to downstream tasks, where each visual feature predicts the CLIP feature of the next position for visual self-supervision. In contrast, Libra's discrete image modeling effectively addresses these issues. To validate this, we convert Libra to contiguous image modeling in Tab. 3(b). We observed that contiguous image modeling achieves similar performance to the variant without any image modeling (Tab. 3(a)), with significant decrease compared to Libra with discrete image modeling (*e.g.*, -6.2% on VizWiz).

Impact of Data Scale. We reduce the pretraining data size to 3M. In Tab. 3(c), we found that larger trainable parameters

(4B in Libra) require more training data for convergence.

Impact of Routed Visual Expert. In Tab. 3(d), we remove the routed visual expert design in Libra, where Libra degenerates to LLaVA (Liu et al., 2023b). The results show that the coupled vision and language systems exhibit obvious performance degradation on zero-shot tasks.

We further investigate the impact of the cross-modal bridge in the routed visual expert, as shown in Tab.3(e). The results show that a simple visual expert alone yields limited performance benefits, *i.e.*, the variant with a simple visual expert (Tab. 3(e)) vs. the variant without any visual experts (Tab. 3(d)). This indicates the importance of a reasonable cross-modal interaction strategy. Fig. 3(a) provides further evidence of the effectiveness of the cross-modal bridge.

Impact of Hybrid Vision Inputs. The contiguous visual signal plays a crucial role in accurate visual perception. To validate this, we remove the contiguous visual signals in Libra's hybrid inputs and only retain the discrete embeddings. As shown in Tab. 3(f), a clear performance degradation rises when solely using discrete inputs (*e.g.*, -20.9% on VizWiz). Tab. 3(g) presents the variant without the CLIP encoder.

5. Conclusion

Through building up Libra, we found that a decoupled vision system can boost vision-language comprehension in the image-to-text scenario. Libra achieves this through the routed visual expert design, where a simple visual expert ensures separate parameter spaces for vision and language, and a cross-modal bridge module decouples inner-modal modeling and cross-modal interaction. Meanwhile, the hybrid image tokenization enables both contiguous visual comprehension and stable discrete modeling. We found that the design of Libra yields diverse attention patterns across layers, indicating potentially low learning redundancy. Vision and language should be integrated in a more reasonable manner beyond simple modality alignment. We hope our work could provoke more consideration in MLLM designs.

Acknowledgements

We'd like to thank Menghao Hu for data management, and Chaoyou Fu for early discussion. This work was supported by National Natural Science Foundation of China (No. 62036012, U23A20387, 62322212, 62072455).

Impact Statement

Libra presents a range of advantages along with potential risks. Its capability to quickly adapt to diverse tasks has the potential to empower non-expert users to achieve satisfactory performance even in data-scarce scenarios. This characteristic can lower the barriers for beneficial applications, but this flexibility also raises concerns regarding malicious and negative applications, necessitating careful consideration. Additionally, Libra shares similar risks with LLMs, such as generating offensive language, perpetuating social biases and stereotypes, and potential privacy breaches. Furthermore, Libra's promising capability to process visual inputs introduces specific risks, including gender and racial biases associated with input image content. To mitigate these risks, we take various measures, such as utilizing debiased pretraining datasets, blurring human faces in training data, and carefully validating instruction tuning data.

References

Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., and Anderson, P. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8948–8957, 2019.

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

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., and Kim, S. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

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

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C. R., Goodman, S., Wang, X., Tay, Y., et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023b.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.-D., et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 958–979, 2024.

Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2023.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Google. Bard. https://bard.google.com/, 2023a.

Google. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023b.

Gupta, T. and Kembhavi, A. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Dong, Y., Ding, M., et al. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2023.

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

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

Hudson, D. A. and Manning, C. D. 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*, pp. 6700–6709, 2019.

IDEFICS. Introducing idefics: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics, 2023.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

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

Kudo, T. and Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

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

Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., and Gao, J. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 1(2):2, 2023c.

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

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023e.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

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

Lu, J., Clark, C., Zellers, R., Mottaghi, R., and Kembhavi, A. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022a.

Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., and Kembhavi, A. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*, 2023.

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

Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204, 2019.

OpenAI. Gpt-4v(ision) system card. https://openai.com/research/gpt-4v-system-card, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

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

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.

Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023a.

Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023b.

Surís, D., Menon, S., and Vondrick, C. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.

Thiebaut de Schotten, M. and Forkel, S. J. The emergent properties of the connected brain. *Science*, 378(6619): 505–510, 2022.

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., and Xie, S. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19175–19186, 2023a.

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023b.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022b.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., and Duan, N. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Xu, Y., Sheng, K., Dong, W., Wu, B., Xu, C., and Hu, B.-G. Towards corruption-agnostic robust domain adaptation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(4):1–16, 2022a.

Xu, Y., Wei, H., Lin, M., Deng, Y., Sheng, K., Zhang, M., Tang, F., Dong, W., Huang, F., and Xu, C. Transformers in computational visual media: A survey. *Computational Visual Media*, 8:33–62, 2022b.

Xu, Y., Zhang, Z., Zhang, M., Sheng, K., Li, K., Dong, W., Zhang, L., Xu, C., and Sun, X. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2964–2972, 2022c.

Xu, Y., Zhang, M., Fu, C., Chen, P., Yang, X., Li, K., and Xu, C. Multi-modal queried object detection in the wild. *arXiv preprint arXiv:2305.18980*, 2023a.

Xu, Y., Zhang, M., Yang, X., and Xu, C. Exploring multi-modal contextual knowledge for open-vocabulary object detection. *arXiv preprint arXiv:2308.15846*, 2023b.

Xu, Z., Shen, Y., and Huang, L. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*, 2022d.

Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., and Wang, L. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

Yao, M., Hu, J., Hu, T., Xu, Y., Zhou, Z., Tian, Y., XU, B., and Li, G. Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=1SIBN5Xyw7.

Yao, M., Hu, J., Zhou, Z., Yuan, L., Tian, Y., Xu, B., and Li, G. Spike-driven transformer. *Advances in Neural Information Processing Systems*, 36, 2024b.

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

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

Yu, L., Lezama, J., Gundavarapu, N. B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A. G., et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023a.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023b.

Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., and Qiao, Y. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

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

A. Implementation Details

A.1. Image Tokenization

Tokenization Process. Thanks to lookup-free quantization (LFQ) (Yu et al., 2023a), we can enlarge the vision vocabulary size to 2^{18}. However, directly implementing such a large vocabulary size brings huge computing costs, *e.g.*, 1B parameters of a simple linear prediction head. Therefore, we predict in two concatenated codebooks, each of size 2^9.

Specifically, the image tokenization process can be illustrated as:

$$E_c = \Phi(I), \quad id_1, id_2 = LFQ(E_c), \quad E_d = \texttt{concat}(E_1(id_1), E_2(id_2)), \quad X_I = \texttt{concat}(E_c, E_d), \tag{9}$$

where I is an input image, Φ is the image encoder, E_c refers to the contiguous visual signals, E_1 and E_2 are two separate vision "word" embedding banks, E_d denotes the discrete vision embeddings, and X_I represents for the vision inputs of the Libra model. `concat` denotes concatenation in the channel dimension. During LFQ, we utilize two codebooks, each with a size of 2^9, to predict the tokenized ids id_1 and id_2. The sizes of embedding banks E_1 and E_2 are also set to 2^9. We maintain two prediction heads for vision outputs of Libra to separately predict id_1 and id_2. This design largely reduces the computational costs, *i.e.*, the prediction heads comprise: 2 (codebook number) $\times$ 2M (head parameters) = 4M parameters.

CLIP-based Image Tokenizer. Libra's image tokenizer is built on a CLIP-based VQGAN with LFQ. As far as we know, it is the first time that a highly reconstructive image tokenizer can be constructed based on a frozen vision encoder like CLIP (Radford et al., 2021). The most relevant approach to our image tokenizer is the one in DALL-E 2 (Ramesh et al., 2022), which tokenizes images into discrete embeddings using a frozen CLIP image encoder and decodes to original images with a diffusion (Rombach et al., 2022) decoder. We compare the reconstruction performance of the tokenizers in Libra and DALL-E 2 in Fig. 5. As the results show, the tokenizer of DALL-E 2 can catch basic visual concepts but largely distort the original visual information. This is helpful for diverse text-to-image generation but detrimental to accurate image-to-text comprehension. In contrast, the tokenizer of Libra effectively captures comprehensive visual information while preserving the pretrained CLIP knowledge.

Figure 5. Image reconstruction results of the image tokenizers in Libra and DALL-E 2 (Ramesh et al., 2022).

Brief Introduction of LFQ. Lookup-free quantization (LFQ) (Yu et al., 2023a) reduces the embedding dimension of the VQ codebook (Van Den Oord et al., 2017) to zero. Specifically, the codebook $\mathbf{C} \in \mathbb{R}^{K \times d}$, similar to the one in VQGAN (Esser et al., 2021), is replaced with an integer set $\mathbb{C}$ where $|\mathbb{C}| = K$, where K is the vision vocabulary size and d represents the embedding dimension. This approach eliminates the need for embedding lookup entirely. Unlike previous quantization methods (Esser et al., 2021; Van Den Oord et al., 2017) that require codebook embeddings to mimic input features for image reconstruction, LFQ does not require such emulation as it has no codebook embeddings. In light of this, we utilize LFQ to successfully quantize highly semantic CLIP features, which has not even been investigated in LFQ.

A.2. Training Details

Training Hyperparameter. The training process of Libra consists of 3 stages: language pretraining (already done in the pretrained LLM), multimodal pretraining, and instruction tuning/supervised finetuning (SFT). We present the hyperparameters of Libra during multimodal pretraining and instruction tuning stages in Tab. 4. The multimodal pretraining stage takes 8400 NVIDIA A100-40G GPU hours and the instruction tuning stage takes 380 NVIDIA A100-40G GPU hours.

Table 4. Training hyperparameters of Libra in different stages.

Configuration	Pretraining	SFT
Total steps	40000	7000
Warmup steps	2000	300
Batch size	1280	128
Learning rate	1e-4	2e-5
Learning rate decay	cosine decay	
Weight decay	0.01	
Dropout ratio	0.0	
Optimizer	AdamW	
Adam ϵ	1e-8	
Adam β	(0.9, 0.99)	
Gradient clipping	1.0	
Numerical precision	bfloat16	
LLM	LLaMA2-7B-Chat	
Vision encoder	CLIP-ViT-L-336px	
Image resolution	336^2	
Patch size	14×14	
Image token number	578	
Vision vocab size	$(2^9)^2 = 2^{18}$	
Language vocab size	32000	

Instruction Template. We arrange the instruction tuning data based on the template described in Sec 3.3. The `<System Message>` in Eqn. (7) is:

```
A chat between a curious user and an artificial intelligence assistant.
The assistant gives helpful, detailed, and polite answers to the user's questions.
```
(10)

B. Evaluation Details

B.1. Benchmarks and Metrics

We provide detailed information of the evaluation benchmarks used in this work in Tab. 5. We use different language prompts for each dataset according to corresponding data forms, as shown in Tab. 6.

B.2. Details on Attention Difference

We provide more details on the computing process of attention differences in Fig. 4. We present the pseudo code in Fig. 7.

B.3. Comparison in the Era of Foundation Models

In the era of foundation models, it is hard to achieve a completely fair performance comparison due to numerous variables such as model parameter size, model architecture, and training data. It is only possible to conduct relatively fair comparisons in scenarios where these variables are approximately equal. Meanwhile, smaller model parameters often imply easier training, which can lead to better performance particularly when dealing with limited data. For example, LLaVA1.5 (Liu et al., 2023b) achieves remarkable performance with merely 1.2M total training data, thanks to its small trainable parameter size (30M during pretraining). We perform the comparison between Libra and other MLLMs under similar model parameter

Table 5. Detailed information of the evaluation benchmarks.

Task	Dataset	Description	Split	Metric
General VQA	VQAv2	VQA on natural images.	test-dev	VQA Score (↑)
	OKVQA	VQA on natural images requiring outside knowledge.	val	VQA Score (↑)
	GQA	VQA on scene understanding and reasoning.	test-balanced	EM (↑)
	VizWiz	VQA on photos taken by people who are blind.	test-dev	VQA Score (↑)
	SQA	Multi-choice VQA on a diverse set of science topics.	Img-test	Accuracy (↑)
Image Caption	NoCaps	Captioning of natural images.	val	CIDEr (↑)
	Flickr	Captioning of natural images.	karpathy-test	CIDEr (↑)
	COCO	Captioning of natural images.	karpathy-test	CIDEr (↑)
MLLM Benchmark	POPE	Object existence by yes/no questions.	random/popular/adversarial	F1 Score(↑)
	MME	Visual perception by yes/no questions.	Perception	MME Score (↑)
	MME^C	Visual cognition by yes/no questions.	Cognition	MME Score (↑)
	MMB	Multi-choice VQA with circular evaluation.	test	Accuracy (↑)
	MMB^{CN}	Multi-choice VQA in Chinese with circular evaluation.	test	Accuracy (↑)
	SEED	Open-ended multi-choice VQA.	Image & Video	Accuracy (↑)
	MM-Vet	Open-ended VL benchmark with various abilities	test	GPT-4 Score (↑)
	MMVP	Detecting CLIP bias by multi-choice VQA.	test	GPT-4 Score (↑)

sizes. Last but not least, *Libra is not intended to achieve state-of-the-art performance; rather, it serves as a prototype model.* Our aim in developing Libra is to offer a promising perspective beyond simple modality alignment for the design of future MLLMs. The evaluations conducted on the Libra prototype have effectively showcased the potential of the decoupled vision systems within MLLMs.

C. Qualitative Evaluation

Fig. 6 shows several conversations between users and Libra. We discovered that Libra demonstrates robust visual perception capabilities and inherits the cognitive abilities of LLMs. For example, it is capable of identifying objects within an image and performing further deducing (*e.g.*, the funny point of an image). Additionally, Libra can catch the relationship between objects, *e.g.*, locations. We also found that Libra, like many commercial chatbots, is capable of error correction based on the user feedback, as shown in the last case.

D. Further Discussion

D.1. Societal Impact

In terms of societal impact, Libra presents a range of advantages along with potential risks. Its capability to quickly adapt to diverse tasks has the potential to empower non-expert users to achieve satisfactory performance even in data-scarce scenarios. This characteristic can lower the barriers for beneficial applications, but this flexibility also raises concerns regarding malicious and negative applications, necessitating careful consideration. Additionally, Libra shares similar risks with LLMs, such as generating offensive language, perpetuating social biases and stereotypes, and potential privacy breaches. Furthermore, Libra's promising capability to process visual inputs introduces specific risks, including gender and racial biases associated with input image content. To mitigate these risks, we take various measures, such as utilizing debiased pretraining datasets, blurring human faces in training data, and carefully validating instruction tuning data.

D.2. Limitations

First, our model is built on pretrained LLMs, and as a side effect, directly inherit their weakness. For example, LLM priors generally provide helpful contextual information, but occasionally demonstrate hallucinations and ungrounded guesses. In addition, it is observed that LLMs exhibit poor generalization when faced with sequences longer than the ones they were trained on.

Second, the routed visual expert design introduces a novel attention computing mechanism, which has not been officially supported by existing acceleration frameworks (*e.g.*, FlashAttention (Dao et al., 2022)) yet. Addressing this issue can make Libra more efficient and more friendly to the downstream implementation.

Table 6. Language prompts for different datasets.

Task	Dataset	Language Prompt
General VQA	VQAv2 (Antol et al., 2015)	Answer the question using a single word or phrase.
	OKVQA (Marino et al., 2019)	Answer the question using a single word or phrase.
	GQA (Hudson & Manning, 2019)	Answer the question using a single word or phrase.
	VizWiz (Gurari et al., 2018)	When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase.
	SQA (Lu et al., 2022b)	Answer with the option's letter from the given choices directly.
Image Caption	NoCaps (Agrawal et al., 2019)	Provide a one-sentence caption for the provided image.
	Flickr (Plummer et al., 2015)	Provide a one-sentence caption for the provided image.
	COCO (Chen et al., 2015)	Provide a one-sentence caption for the provided image.
MLLM Benchmark	POPE (Li et al., 2023e)	Answer the question using a single word or phrase.
	MME (Fu et al., 2023)	Answer the question using a single word or phrase.
	MME^C (Fu et al., 2023)	Answer the question using a single word or phrase.
	MMB (Liu et al., 2023c)	Answer with the option's letter from the given choices directly.
	MMB^{CN} (Liu et al., 2023c)	请直接回答选项字母。
	SEED (Li et al., 2023a)	Answer with the option's letter from the given choices directly.
	MM-Vet (Yu et al., 2023b)	-
	MMVP (Tong et al., 2024)	Answer this question concisely.

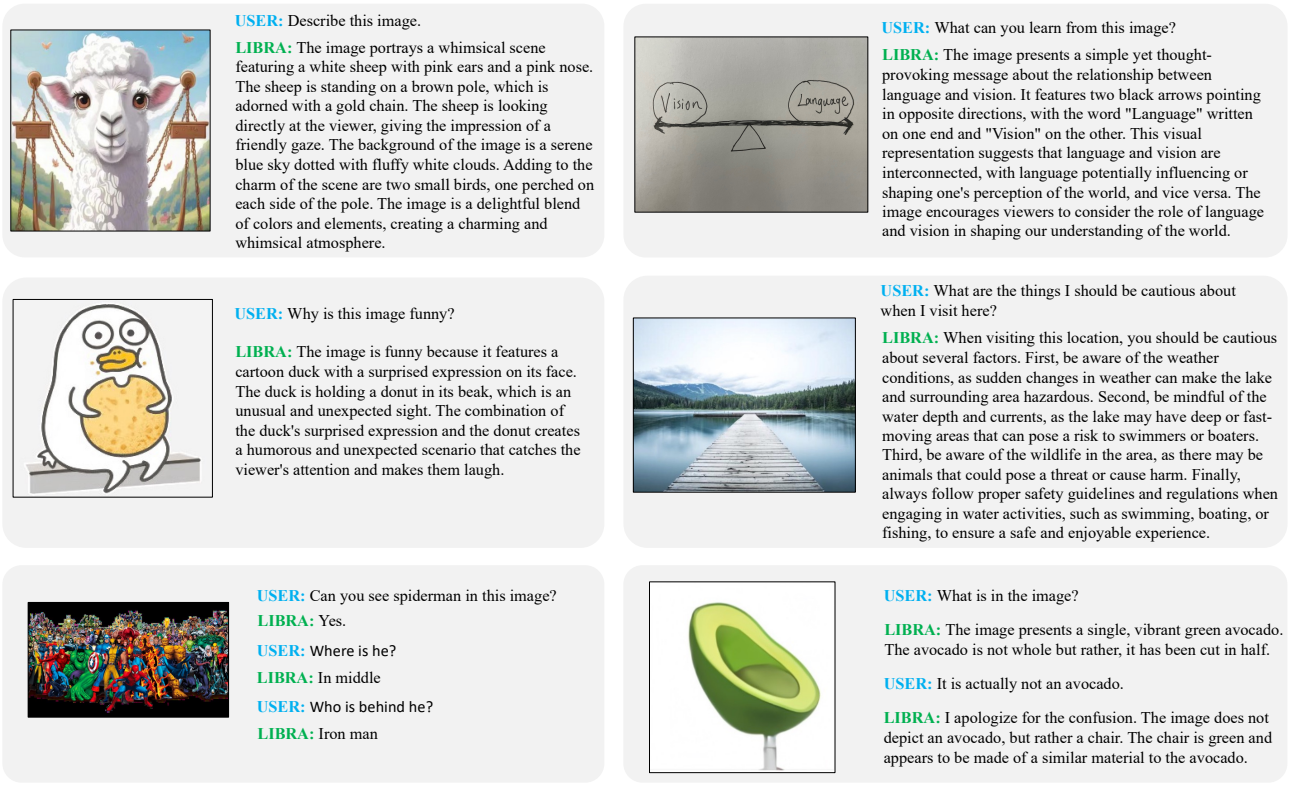


Figure 6. Conversations between users and Libra.

```
def attention_difference(
    attn_maps,
):
'''
attn_maps: (L,B,H,N) - attention activation of single-word answers on the images
L: layer number, B: batch size, H: head number, N: image token number
'''

# 1. Normalization
attn_maps = attn_maps / attn_maps.max()

# 2. cross-layer difference
layer_mean = attn_maps.mean(dim=0) # shape: (1,B,H,N)
cross_layer_difference = (attn_maps - layer_mean).abs() # shape: (L,B,H,N)
cross_layer_difference = cross_layer_difference.mean(dim=1).mean(dim=2).mean(dim=3) # shape: (L)

# 3. inner-layer difference
inner_layer_difference = []
for attn in attn_maps:
    # attn - shape: (B,H,N)
    head_mean = attn.mean(dim=1) # shape: (B,1,N)
    head_wise_difference = (attn - head_mean).abs() # shape: (B,H,N)
    head_wise_difference = head_wise_difference.mean(dim=0).mean(dim=1).mean(dim=2) # shape: (1)
    inner_layer_difference.append(head_wise_difference) # final shape: (L)
return cross_layer_difference, inner_layer_difference
```

Figure 7. Pseudo code for the computing process of attention differences in Fig. 4.