MM-LAMA: Probe and Locate Knowledge in Large Vision-Language Models

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) have exhibited remarkable prowess in tackling multimodal tasks, yet their internal knowledge storage mechanisms remain opaque, impeding practical deployment. This paper undertakes a thorough examination of knowledge storage and activation within LVLMs, introducing the Multimodal Language Model Analysis (MM-LAMA) dataset and novel analytical tasks: multimodal knowledge probing and localization. To surmount analytical hurdles, we propose Multimodal Integrated Gradients (MIG), a gradient-based method tailored for LVLM analysis. Our empirical investigation unveils two critical neuron classes within LVLMs: Synesthesia Neurons, elucidating shared knowledge across modalities for noise filtration, and Complementarity Neurons, augmenting multimodal knowledge localization accuracy. These findings contribute significantly to advancing the understanding and application of multimodal models in various domains.

1 Introduction

Large Vision-Language Models (LVLMs) (Li et al., 2023; Liu et al., 2023; Touvron et al., 2023) have demonstrated exceptional performance across various knowledge-intensive multimodal tasks, such as visual question answering (Antol et al., 2015; Goyal et al., 2017; Marino et al., 2019), multimodal dialogue (Wang et al., 2021; Zheng et al., 2022), novel caption generation (Agrawal et al., 2019), and more. However, despite their success, the knowledge storage mechanism within the parameters of LVLMs remains unclear. The lack of transparency in this aspect poses a significant challenge, limiting the practical application of LVLMs, such as parameter editing (Meng et al., 2023, 2022a; Mitchell et al., 2022) or model pruning (Li et al., 2017; Han et al., 2016). While some work (Dai et al., 2022; Meng et al., 2022b; Li et al., 2022) has been proposed to analyze the knowledge storage mechanisms of textual LLMs, there is a notable absence of similar multimodal knowledge analysis for LVLMs. In multimodal knowledge graphs (Liu et al., 2019; Wang et al., 2020), the same knowledge is expressed in different representations in different modality data, which is defined as multimodal knowledge. For example, on Figure 1 Left, the head entity of the same knowledge triplet (e.g., (*Peter Morgan*,*Nationality*, *England*)) can be represented as the text "*Peter Morgan*" or the corresponding image.

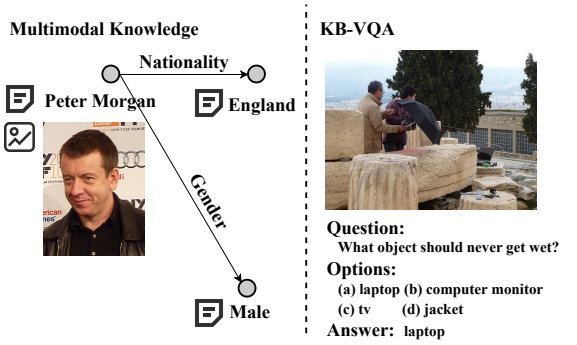


Figure 1: **Left**: Multimodal knowledge in multimodal knowledge graph FB15k-237-IMG (Liu et al., 2019). **Right**: An example from A-OKVQA (Schwenk et al., 2022), given an image, asking knowledge questions about the image and getting answers.

- What are the similarities and differences in the knowledge areas excited by different modality data in LVLMs?
- Are there any valuable unique knowledge areas in LVLMs?

We believe that answering these questions will help design better LVLMs. In light of this, our paper aims to gain insights into LVLMs by probing whether multimodal knowledge exists in LVLMs (i.e., **multimodal knowledge probe**), then locating and analyzing the knowledge storage area of

LVLMs (i.e., **multimodal knowledge localization**).

Investigating the multimodal knowledge probe or localization of LVLMs faces two challenges: **First, the lack of analytical data.** If we directly use existing multimodal task datasets (Marino et al., 2019; Schwenk et al., 2022; Wang et al., 2018) to analyze the knowledge storage mechanism of LVLMs, there are two issues: (1) *The mixture of diverse knowledge*. Many existing datasets involve multiple factual knowledge within a single sample. As shown in Figure 1 right, a sample in A-OKVQA dataset (Schwenk et al., 2022) contains at least two distinct knowledge: (i) "*The laptop will be damaged when it encounters water.*" (ii) "*The jacket will not be damaged when it encounters water.*" Using such a sample to locate knowledge storage areas about "*laptop*" may inadvertently lead the model to locate the area where knowledge about "*jackets*" is stored. (2) *The biases in aligning different modalities.* Alignment between different modalities may be not obvious. As illustrated in Figure 1 left, this image contains a lot of things: two people, a laptop, rain, and other information. The "*laptop*" is not the central focus of the image. Locating knowledge about '*laptop*" in LVLMs with this image as input becomes challenging. These challenges impede the analysis of neurons where data under different modality settings evoke the same knowledge. Consequently, the construction of a high-quality analytical dataset emerges as a crucial problem.

Second, Lack of analysis methods. Current analysis methods (Dai et al., 2022; Meng et al., 2022b) predominantly focus on textual unimodal language models, and adapting them for multimodal analysis proves challenging due to the dissimilarity between multimodal data and the complexity of multimodal models. Specifically, if knowledge from different modalities is located consistently, performance under each single modality can not reach the best. In addition, existing analysis methods are often tailored for a single type of Transformer (auto-regressive or auto-encoding models) (Vaswani et al., 2017). However, it is very common for LVLMs that leverage both autoregression and autoencoding structures, such as BLIP2 (Li et al., 2023) based on T5 (Chung et al., 2022), cannot be analyzed by existing methods. Addressing the challenge of conducting knowledge mechanism analysis on LVLMs is therefore an urgent challenge.

To address the first challenge, we construct

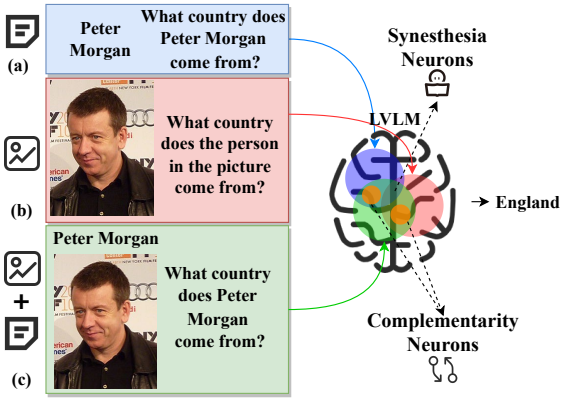


Figure 2: Examples under three modality settings of **MM-LAMA** with the relation "*nationality*". (a) Language setting. (b) Visual setting. (c) Multimodal setting. Synesthesia Neurons often exist in the intersection of different modal knowledge neurons, and Complementary Neurons exist in the relative complement of different modal knowledge neurons.

the **MM-LAMA** (**M**ulti**M**odal **LA**nguage **M**odel **A**nalysis) dataset to support studying multimodal knowledge probe and localization. This dataset consists of 5 types of relations, 1564 samples, 2 types of probing questions, and 2 types of modal triples, all derived from factual triples in a multimodal knowledge graph (Liu et al., 2019). As depicted in Figure 1 left, each triplet head entity is represented in two modalities (language or visual), and the tail entity is a simple short text. To tackle the issue "*the mixture of diverse knowledge*", each sample in our dataset pertains to only one fact triplet. To address the issue "*the biases in aligning different modalities.*", we manually filter image triples, prioritizing subjects with minimal ambiguity. Additionally, to ensure clarity in questions, we select relations as one-to-one as much as possible. To overcome the second challenge, we develop a novel gradient-based method (Ancona et al., 2019) named **MIG** (**M**ultimodal **I**ntegrated **G**radients). In order to address the differences between different modality data, we adopt different threshold settings for the localization of knowledge neurons in different modalities. Besides, we model the baseline vectors in different modes to reflect the characteristics of different modality settings. To solve the problem of being unable to handle complex multi-modal models, we adopt the adaptive gradient integration for the LVLMs to obtain the analysis. Notably, in our experiments, we identify two crucial classes of neurons: Synesthesia Neurons and Complementarity Neurons.

Synesthesia Neurons: As depicted in Figure 2, some neurons exhibit consistent activation for the same knowledge across all different modality settings, and we term them as synesthesia neurons. This phenomenon mirrors the way humans may experience similar sensations when exposed to different types of perceptions about the same thing. For instance, when humans view images or videos depicting ice and snow, they might feel a sense of coldness. The more realistic the image, the stronger this feeling becomes. Through our experimental verification, we observed a parallel phenomenon in LVLMs. In multimodal tasks, synesthesia neurons constitute a larger proportion of activated neurons when images and texts align, and a smaller proportion when they do not align. This observation can be leveraged for filtering noise samples from multimodal datasets where different modalities exhibit poor alignment.

Complementarity Neurons: As illustrated in Figure 2, for different modal inputs of the same knowledge, although some neurons are not activated by the input in a specific modal setting, experiments have proven that adding these knowledge neurons will be beneficial to the prediction results in this modal setting, and we term them complementarity neurons. For instance, when determining a person's nationality based on the name "*Peter Morgan*", the name may correspond to multiple individuals from different countries. However, knowledge inspired by the image modality can help eliminate this ambiguity and yield the correct answer. Locating such complementarity neurons enhances the accuracy of multimodal knowledge localization, contributing to improved knowledge editing and model explanations.

Our contributions can be summarized as follows:

- We propose two new tasks, the multimodal knowledge probe and localization, and first construct a high-quality new dataset Multimodal Language Model Analysis (**MM-LAMA**) to support the deep analysis of multimodal knowledge in LVLMs.
- We design a novel method Multimodal Integrated Gradients (**MIG**), which models different modality data separately and can handle complex LVLMs.
- We conduct extensive experiments and analyze under three settings. We uncover the existence of synesthesia neurons in LVLMs, which can be utilized to filter noise image samples, and complementarity neurons in LVLMs, enhancing the accuracy of multimodal knowledge localization.

2 Methodology

Our primary objectives encompass two main facets: (1) Probing multimodal factual knowledge mastered by different LVLMs. (2) Analyzing the similarities and differences of activated knowledge neurons under various modality inputs using knowledge localization methods.

2.1 Problem Formulation

Multimodal Knowledge Probe: Our approach aligns with the methodology introduced by LAMA (Petroni et al., 2019) in which a triplet (h, r, t) within a knowledge graph represents an indivisible unit of factual knowledge. In the context of knowledge probe tasks, the head entity information and relationships are commonly fed to the model through prompts to predict the tail entity information. This process can be expressed as $\hat{t} = f(p(h, r))$, where $p(\cdot, \cdot)$ denotes the prompt function combining the information of the head entity and relation, $f(\cdot)$ represents the model prediction function, and $\hat{t}$ is the predicted tail entity. A correct prediction implies that the model has successfully grasped the underlying knowledge (Petroni et al., 2019). For example, as illustrated in Figure 2 (a), given a triplet (*Peter Morgan*, *Nationality*, *England*), the information about the head entity and relation is presented in the form of a question such as "*What country does Peter Morgan come from?*". Feeding with this question, the model is expected to accurately predict the tail entity as "*England*", called the knowledge probe task. We extend the scope of the knowledge probe task into the multimodal domain to investigate the knowledge mastery of LVLMs. Concretely, in the multimodal knowledge probe task, the head entity $h = \{h^t, h^v\}$ can be represented in two modalities: textual modality representation h^t and visual modality representation h^v. In addition, we also propose three input settings to probe the knowledge in LVLMs:

(1) *Language setting*: As shown in Figure 2 (a), the language setting is consistent with traditional knowledge probe tasks about language models.

(2) *Visual setting*: As shown in Figure 2 (b), under the visual setting, the model needs to predict tail entities based on images containing head entity information and questions containing relation information.

(3) *Multimodal setting*: As shown in Figure 2 (c), under the multimodal setting, the model needs to predict tail entities based on images containing head entity information and questions containing information of relation and head entity.

Multimodal Knowledge Localization: The knowledge localization task aims to use factual knowledge triples (h, r, t) to locate the parameters that are activated in the model, considered as knowledge neurons $\mathcal{N}$ that express this knowledge. Similarly, multimodal knowledge localization aims to locate the knowledge neurons representing multimodal factual knowledge triples. As illustrated in Figure 2, the blue part is the knowledge neurons $\mathcal{N}_T$ obtained under the language setting, the red part is the knowledge neurons $\mathcal{N}_V$ obtained under the visual setting, and the green part is the knowledge neurons $\mathcal{N}_M$ obtained under the multimodal setting. We build the method based on the integral gradient (Ancona et al., 2019) to locate knowledge neurons, which will be detailed in Section 2.2. We then analyze the similarities and differences of knowledge neurons under different settings, unveiling interesting phenomena in the process.

2.2 Multimodal Intergrated Gradients

This Section will introduce how to localize knowledge neurons under different modality settings. Inspired by the gradient-based knowledge localization method (Sundararajan et al., 2017; Dai et al., 2022; Ancona et al., 2019), we adopt the intermediate parameters in the FFN of each layer in the transformer(Vaswani et al., 2017) with the large attribution score as the knowledge neurons.

$$P(\hat{w}_i^{(l)}) = p(\bar{t}|x, w_i^{(l)} = \hat{w}_i^{(l)}), \tag{1}$$

where $\bar{t}$ is the ground truth tail entity of the sample, $w_i^{(l)}$ is the i-th parameter of l-th layer FFN of the transformer, $\hat{w}_i^{(l)}$ is the value assign to $w_i^{(l)}$. The attribution score $S(w_i^{(l)})$ of the neuron $w_i^{(l)}$ is calculated as

$$S(w_i^{(l)}) = \bar{w}_i^{(l)} \int_{\alpha=0}^{1} \frac{\partial P(\alpha \bar{w}_i^{(l)})}{\partial w_i^{(l)}} d\alpha, \tag{2}$$

where the value of $w_i^{(l)}$ assigned from 0 to it original value $\bar{w}_i^{(l)}$, $\frac{\partial P(\alpha \bar{w}_i^{(l)})}{\partial w_i^{(l)}}$ is the gradient from the model output. Intuitively, with the value changing of neuron $w_i^{(l)}$ from 0 to the original value, if this neuron is important for the model output, the model should have a relatively large change in its gradient change. It is complex to calculate the integrals directly, thus we use the Riemann approximation to calculate this value, which is

$$\tilde{S}(w_i^{(l)}) = \frac{\bar{w}_i^{(l)}}{n} \sum_{k=1}^{n} \frac{\partial P(\frac{k}{n} \bar{w}_i^{(l)})}{\partial w_i^{(l)}}, \tag{3}$$

where n is the number of approximation steps. Under different modal settings, the detected neurons can be defined as

$$\mathcal{N}_m = \{w_i^{(l)} | \tilde{S}(w_i^{(l)}) > \lambda_m \max_{i,l} \tilde{S}(w_i^{(l)})\}, \tag{4}$$

where $m \in \{L, V, M\}$ represent three modality settings, λ_m is the threshold hyperparameter, which determines how much activation value belongs to the knowledge neuron.

3 Experiment Settings

3.1 Evaluation tasks

In addition to the **MM-LAMA** dataset introduced in Appendix A, to assess the nature of Synesthesia Neurons, we generate negative examples $(\hat{h} = \{h^t, \hat{h}^v\}, r, t)$ for each triplet $(h = \{h^t, h^v\}, r, t)$ in the dataset. Specifically, we randomly assign an inconsistent image $\hat{h}^v$ to each head entity $h = \{h^t, h^v\}$. To ensure the difficulty of distinguishing negative examples from positive examples, we randomly sample from the set of head entities of triples with the same relation. The negative example data and the positive example data together constitute the Fake-MMLAMA. To summarize, the tasks we test include the multimodal knowledge probe task, the multimodal knowledge localization task, and the noise image filtering task.

3.2 Implementation Details

The completion of all tasks does not involve parameter training. Our primary method is based on BLIP2-FlanT5XXL and BLIP2-FlanT5XL (Li et al., 2023). For tasks involving only textual input, we directly utilize the probe results of the textual language model on which the multimodal model is based as the probe results under the text modal setting. The hyperparameters λ_L and λ_M

are set to 0.05, while λ_V is set to 0.04. All experiments are conducted using NVIDIA GeForce RTX 3090 GPUs. On average, it takes 57.6 seconds to identify knowledge neurons for one sample in BLIP2-FLANT5XL.

3.3 Evaluation Metrics

For the multimodal knowledge probe task and the noise image filtering task, we adopt accuracy (ACC.) as the evaluation criterion. That is, for multimodal knowledge detection tasks, it is judged whether the tail entity information predicted by the model is consistent with the real tail entity information. For the problem of noise image filtering, it is judged whether the model can separate positive examples and negative examples. Besides, for the multimodal knowledge localization task, following Dai et al. (Dai et al., 2022), we adopt suppressing score and enhancing score as the criterion. Given a set of knowledge neurons $\mathcal{N}$, the suppressing score refers to the difference between the predicted probability of the correct answer under all knowledge neurons activation being set to 0 and the predicted probability of the answer in the original state. On the contrary, the enhancing score refers to the difference between the predicted probability of the correct answer under all knowledge neuron activation being set to 2 times its original activation and the predicted probability of the answer in the original state. This can be formulated as

$$S_{suppress} = p(\hat{t}|x, w_i^{(l)} = 0) - p(\hat{t}|x), w_i^{(l)} \in \mathcal{N},$$
$$S_{enhance} = p(\hat{t}|x, w_i^{(l)} = 2\bar{w}_i^{(l)}) - p(\hat{t}|x), w_i^{(l)} \in \mathcal{N}, \tag{5}$$

where $\hat{t}$ is the ground truth answer of the sample, $w_i^{(l)}$ is the i-th parameter of l-th layer FFN of the transformer, $\bar{w}_i^{(l)}$ is the original value of $w_i^{(l)}$. In order to unify the evaluation indicators, we define the localization score (LS) as the difference between the two scores, which is:

$$S_L = S_{enhance} - S_{suppress}. \tag{6}$$

3.4 Baselines

We evaluate our method mainly based on BLIP2 (Li et al., 2023) which consists of a language model (i.e., FlanT5 (Chung et al., 2022)), a vision encoder (i.e., eva-clip (Sun et al., 2023)) and a multimodal interface (i.e., Q-former). In addition, we also used LLaVA2 (Liu et al., 2023) with a similar structure, which uses Llama2 (Touvron et al., 2023) as the language model, as the baseline. In the task of removing noise images, we prompt the model "*Do you think the image is a companion image of entity h^t? (Yes or No)*" as the baseline. In the knowledge localization task, we randomly select the same number of neurons as the set to be compared as the baseline.

4 Results and Analysis

4.1 Multimodal Knowledge Probe Results

The primary results of the multimodal knowledge probe task in MM-LAMA are summarized in Table 1. Key observations include:

(1) Results under the multimodal probe setting consistently outperform those of any single-modal probe setting. For instance, LLAVA results increased by 6.1 compared to the best language settings, while BLIP2-xl increased by 3.8, and BLIP2-xxl increased by 4.0. This highlights the advantage of integrating information from different modalities, indicating a complementary effect between them. However, in sports prediction, BLIP2-xl's multimodal performance is less effective than pure text, suggesting that introducing images may introduce unnecessary influencing factors, leading to a decline in performance.

(2) Text modalities generally yield better results than image modalities. Visual modalities' probe settings are challenging, such as predicting nationality based on images often requires prior knowledge, making it more challenging than predicting nationality based on the person's name. However, there are exceptions. in predicting gender, BLIP2 demonstrates that visual modality characteristics can be more effective.

4.2 Multimodal Knowledge Neurons Coincidence Analysis

Through multimodal knowledge localization, we found that knowledge neurons under different settings always have different parts as well as overlapping parts, so we conducted an in-depth discussion on this phenomenon. As illustrated in Figure 3, We divided the activated knowledge neurons under the three settings into 7 regions of neuron sets based on the overlapping parts, which were recorded as $\mathcal{N}_A, \mathcal{N}_B, \mathcal{N}_C, \mathcal{N}_D, \mathcal{N}_E, \mathcal{N}_F$ and $\mathcal{N}_G$

Relation	LLaVA2			BLIP2-xl			BLIP2-xxl		
	Lan.	Vis.	MM.	Lan.	Vis.	MM.	Lan.	Vis.	MM.
nationality	69.2	24.7	**75.9**	55.5	39.1	**56.9**	61.3	39.1	**66.4**
state	81.4	35.1	**83.3**	**70.3**	37.9	**70.3**	72.2	37.9	**73.1**
sport	71.2	68.3	**72.4**	**71.8**	46.9	70.8	73.8	46.9	**78.9**
gender	95.7	92.4	**97.1**	93.3	98.2	**98.5**	90.1	99.1	**99.5**
profession	13.4	15.4	**22.1**	7.2	14.1	**14.8**	14.2	17.6	**17.8**
Average	54.8	33.5	**60.9**	46.1	39.3	**49.9**	50.5	40.6	**54.5**

Table 1: Multimodal knowledge probe performance (%ACC.) on **MM-LAMA** of different multimodal language models in different modal settings. Lan. represents for language setting, Vis. represents for visual setting, MM. represents for multimodal setting. BLIP2-xl stands for BLIP2-FLANT5XL and BLIP2-xxl stands for BLIP2-FLANT5XXL.

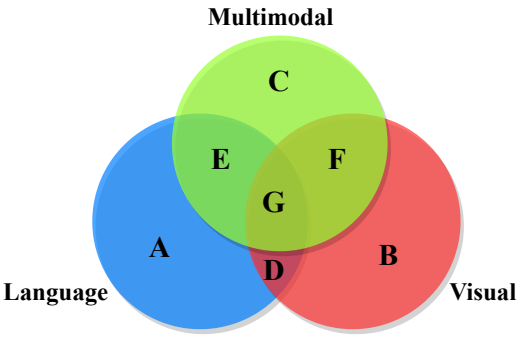


Figure 3: Schematic diagram of the overlap area of knowledge neurons under different settings.

respectively. In order to better understand the role of each region, we visualize the proportion of neurons in different regions that predicted correctly or the entire dataset, in the encoder or the entire Language model, as Figure 4. Among them, if and only if the predictions are correct under the three modal settings, we consider the model's answer to be correct. We found that when the answer is correct, that is, the model masters this knowledge, the proportion of $\mathcal{N}_G$ will increase. In detail, for the whole language model, $\mathcal{N}_G$ increases by 7.89%, and for the encoder it increases by 12.7%. This drastic change shows that there is a certain overlap between knowledge neurons in different modalities. When knowledge is mastered, no matter what setting is used, some specific neurons will be activated. We define this type of neuron as the Synesthesia Neurons, which will be introduced in Section 4.3 In addition, Section 4.1 pointed out that different modal data also complement each other. We define knowledge neurons that play complementary roles as the Complementary Neurons, which will be introduced in Section 4.4.

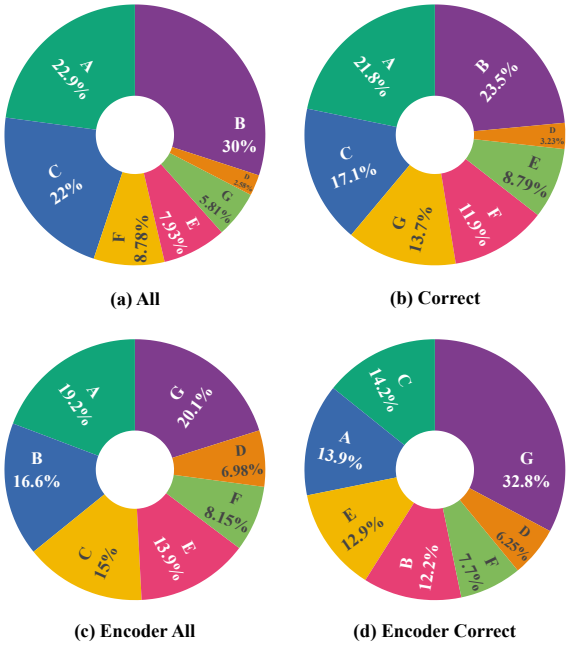


Figure 4: Proportion of activated knowledge neurons of BLIP2-FlanT5xl in different regions on **MM-LAMA**. (a) For all samples, all parameters. (b) For all correctly predicted samples, all parameters. (c) For all samples, only for some parameters of Encoder. (d) For all correctly predicted samples, only for some parameters of the Encoder.

4.3 Synesthesia Neurons Analysis

The results of knowledge localization for Synesthesia Neurons are presented in Table 2. Our findings indicate that synesthesia neurons (i.e., overlapping regions) are more crucial for the model in predicting the correct answer compared to non-synesthesia neurons (i.e., non-overlapping regions).

Neurons	Modality	$S_{sup.}$ ↓	$S_{en.}$ ↑	S_L ↑
$\mathcal{N}_G$	Vis.	-0.053	0.023	**0.076**
	Lan.	-0.112	0.058	**0.171**
$\mathcal{N}_B$	Vis.	-0.003	0.007	0.011
$\mathcal{N}_A$	Lan.	-0.014	0.017	0.031
Random	Vis.	-0.003	-0.001	0.002
	Lan.	-0.001	0.002	0.003

Table 2: The influence of Synesthesia Neurons on **MM-LAMA** prediction of correct samples.

Method	Accuracy
Random	50.1
BLIP2-xl (Li et al., 2023)	54.1
BLIP2-xxl (Li et al., 2023)	57.1
Synesthesia Neurons (Ours)	**74.6**

Table 3: The influence of Synesthesia neurons on **MM-LAMA** prediction of correct samples.

Specifically, neurons activated by $\mathcal{N}_G$ improve the localization score S_L by 0.065 in the visual modality and 0.14 in the language modality relative to areas activated by $\mathcal{N}_B$ or $\mathcal{N}_A$. Moreover, all results surpass those of Random by a large margin, affirming the accuracy of the located knowledge neurons.

Additionally, we observe that the number of synesthesia neurons reflects the correlation between images and text to a certain extent. Based on this, we conduct experiments on Fake-MMLAMA, which includes noise images in samples. When the proportion of $\mathcal{N}_G$ in the BLIP2-XL encoder is greater than 0.22 of $\mathcal{N}_L, \mathcal{N}_V$, it is considered that images and text are related, and vice versa. The results are shown in Table 3, leading us to the following conclusions:

(1) The results of prompting BLIP2-xl and BLIP2-xxl are not as good as the method utilizing synesthesia neurons. Directly using the prompt method makes it challenging for the model to distinguish relevant images from irrelevant ones. In contrast, the size of the overlapping area can continuously and directly reflect the degree of correlation of information between different modalities.

(2) Simultaneously, if we prompt the multimodal model directly to inquire whether the image and text are related, the model is likely to predict "*No*". This method of obtaining prediction through prompts is not as stable as judging the correlation through the proportion of synesthesia neurons. The proportion of synesthesia neurons may more directly express the model's perception of different modal data.

4.4 Complementarity Neurons Analysis

In the experiments outlined in Section 4.1, we identified complementarity between knowledge of different modalities. To further pinpoint the areas where these neurons exist, we conduct analyses using different sets of neurons under different modal settings. The results are presented in Table 4, and the following conclusions can be drawn:

(1) Comparing the results of (i) and (ii), under the language setting, using all neurons of $\mathcal{N}_L$ and $\mathcal{N}_V$ improve the localization score S_L, but under the visual setting, the score decreased. This suggests that overall knowledge activated by vision is often useful for solving language problems. However, for images, activating all language knowledge may lead to a certain degree of misleading effects.

(2) Comparing the results of (iii), (ii), and (i), adding the neurons of $\mathcal{N}_E$ and $\mathcal{N}_F$ to the originally activated neurons maximizes the localization score S_L. This indicates that the knowledge neurons in $\mathcal{N}_E$ play complementary roles for the language modality, and the activated neurons in $\mathcal{N}_F$ for the visual modality.

(3) Results (iv), (iii), and (i) show that adding the neurons of $\mathcal{N}_A$ and $\mathcal{N}_B$ to the originally activated neurons actually decreased the performance. This suggests that the knowledge of areas A and B misled the model. It may be analogous to when we focus on eliminating text ambiguity through images; we don't pay attention to the important information of the image but rather focus on useless background information, distracting our thinking.

(4) Based on (v), using non-overlapping neurons from other modalities is counterproductive, indicating misleading knowledge from other modalities in solving problems in the current modality.

(5) A comparison of (v) and (vi) reveals that non-overlapping knowledge neurons excited from

	Neurons	Modality	$S_{sup.}$ ↓	$S_{en.}$ ↑	S_L ↑
(i)	$\mathcal{N}_L$	Lan.	-0.234	0.075	0.309
	$\mathcal{N}_V$	Vis.	-0.121	0.069	0.190
(ii)	$\mathcal{N}_L \cup \mathcal{N}_V$	Lan.	-0.249	0.125	0.374
	$\mathcal{N}_L \cup \mathcal{N}_V$	Vis.	-0.107	0.041	0.148
(iii)	$\mathcal{N}_L \cup \mathcal{N}_F$	Lan.	-0.251	0.129	**0.380**
	$\mathcal{N}_V \cup \mathcal{N}_E$	Vis.	-0.145	0.072	**0.217**
(iv)	$\mathcal{N}_L \cup \mathcal{N}_B$	Lan.	-0.032	0.011	0.043
	$\mathcal{N}_V \cup \mathcal{N}_A$	Vis.	-0.017	0.005	0.022
(v)	$\mathcal{N}_V - \mathcal{N}_L$	Lan.	0.002	-0.004	-0.006
	$\mathcal{N}_L - \mathcal{N}_V$	Vis.	0.016	0.001	-0.015
(vi)	$\mathcal{N}_M - \mathcal{N}_L$	Lan.	-0.104	0.029	0.133
	$\mathcal{N}_M - \mathcal{N}_V$	Vis.	-0.049	0.014	0.063

Table 4: The performance of different sets of knowledge neuron combinations in predicting answer correct samples on **MM-LAMA** for knowledge localization.

the multimodal space still positively impact single-modality, indicating that multimodal settings integrate knowledge useful for single-modality.

In summary, complementarity neurons do exist for single-modal knowledge probe tasks and are highly likely to exist in regions E and F. This discovery can help more accurately locate a certain piece of multimodal knowledge for downstream tasks such as local knowledge editing.

5 Related Work

5.1 Large Vision Language Models

The vision and language models (VLMs) aim to leverage a single model to process both image and text information concurrently. Early works, such as VL-BERT (Su et al., 2020), UNITER (Chen et al., 2020), OFA (Wang et al., 2022), etc., followed the paradigm of pre-trained language models (Devlin et al., 2019), incorporating a multi-modal pre-training objective, pre-training on image-text datasets, and subsequently evaluating on downstream multimodal tasks (Antol et al., 2015; Agrawal et al., 2019; Plummer et al., 2015). The advent of large language models (LLMs) (Hoffmann et al., 2022; Brown et al., 2020; Zhang et al., 2022) has significantly transformed natural language processing. Recent vision and language models, such as BLIP2 (Li et al., 2023), LLaVA (Liu et al., 2023), Instruct BLIP (Dai et al., 2023), minigpt4 (Zhu et al., 2023), among others, seek to integrate the extensive knowledge present in large language models and train multi-modal interfaces for these pre-trained models. This integration enables language models to comprehend multimodal data, defining them as large vision and language models (LVLMs). These models have demonstrated significant success in various multimodal knowledge-intensive tasks (Schwenk et al., 2022; Marino et al., 2019; Wang et al., 2018; Agrawal et al., 2019). Our work builds upon the advancements made by these models.

5.2 Knowledge Localization

Knowledge localization methods can be broadly categorized into two groups. The first category comprises gradient-based methods (Ancona et al., 2019). Some researchers (Dai et al., 2022; Chen et al., 2023) argue that the integral gradient (Sundararajan et al., 2017) of a parameter can effectively indicate its importance. In the context of knowledge probes (Petroni et al., 2019), these parameters are considered Knowledge Neurons. The second category involves causal-based methods (Li et al., 2022; Meng et al., 2022b; Zhao et al., 2023), which aim to provide explanations for the reasons behind specific outputs. For instance, Zhao et al. (Zhao et al., 2023) attempts to define the output as a combination of certain concepts. However, these methods are often tailored to a single modality and may struggle when dealing with multi-modal models and data.

6 Conclusion and Future Work

In this paper, we introduce the **MM-LAMA** dataset and a novel gradient-based method **MIG**, contributing to addressing the challenges in understanding LVLMs. Furthermore, our exploration of LVLMs has provided valuable insights into the intricate knowledge storage mechanisms within these models. The identification of Synesthesia Neurons, demonstrating consistent activation across diverse knowledge settings, and Complementarity Neurons, enhancing cross-modal knowledge probing, marks a significant step towards unraveling the complexity of LVLMs.

7 Limitations

However, it's important to note that our method has some limitations. For instance, it may not fully consider other parts of LVLMs besides the language model. Additionally, further investigation is needed to explore how the identified types of neurons and

other phenomena observed can be practically applied, such as in knowledge editing applications.

References

Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956. IEEE.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus H. Gross. 2019. Gradient-based attribution methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pages 169–191. Springer.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. *CoRR*, abs/2308.13198.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Song Han, Huizi Mao, and William J. Dally. 2016. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In

4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. 2017. Pruning filters for efficient convnets. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022. How pre-trained language models capture factual knowledge? A causal-inspired analysis. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1720–1732. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *CoRR*, abs/2304.08485.

Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019. MMKG: multi-modal knowledge graphs. In *The Semantic Web - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings*, volume 11503 of *Lecture Notes in Computer Science*, pages 459–474. Springer.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in GPT. In *NeurIPS*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022b. Locating and editing factual associations in GPT. In *NeurIPS*.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan

Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. 2020. Richpedia: A large-scale, comprehensive multi-modal knowledge graph. *Big Data Res.*, 22:100159.

Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2018. FVQA: fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2413–2427.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. 2021. Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. *CoRR*, abs/2109.12761.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Ruochen Zhao, Shafiq R. Joty, Yongjie Wang, and Tan Wang. 2023. Explaining language models' predictions with high-impact concepts. *CoRR*, abs/2305.02160.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2022. Mmchat: Multi-modal chat dataset on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 5778–5786. European Language Resources Association.

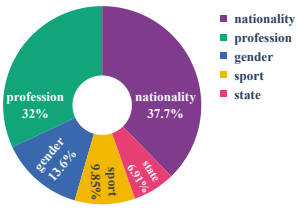
Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

A MM-LAMA Dataset

This Section will introduce how to construct the MM-LAMA dataset for studying multimodal knowledge probe and localization tasks. We filter some triples from the FB15k-237-IMG dataset (Liu et al., 2019; Bordes et al., 2013), a subset of Freebase (Bollacker et al., 2008), which automatically assigns images to entities from the Internet, as our dataset MM-LAMA. We filter triples that meet the following characteristics:

- The tail entity is unambiguous, that is, there will not be multiple values from the tail entity that have a relation with the head entity.
- Head entities can be represented by images or texts.
- The subject indicated by the image is clear, that is, the subject contained in the image is single and occupies most of the image area.
- The text representation of the tail entity is simple, that is, the text corresponding to the tail entity should be less than 3 words.

Triples with these characteristics can be beneficial for the analysis of multimodal knowledge neurons. For each relation involved, we manually design two different prompt templates, as shown in Table 5. The specific statistical information of MM-LAMA is shown in Figure 5. The left side shows the proportion of different relations contained in the dataset, and the right side shows some basic statistical information about the dataset.



	MM-LAMA
#samples	1564
#relations	5
#images	4957
#head entity	1254
#tail entity	149

Figure 5: **Left:** The proportion of samples with different relations in MM-LAMA. **Right:** Statistics of MM-LAMA dataset.

B Multimodal Knowledge Neurons Location Analysis

As depicted in Figure 6, we visualized the distribution of activated neurons in different layers under different settings. It is evident that the neurons activated in the encoder are more concentrated, while the decoder shows a more even distribution. Further analysis of an individual sample, as shown in Figure 7, also confirms this conclusion. The centralization of knowledge in the encoder may be leveraged for model pruning (Li et al., 2017; Han et al., 2016).

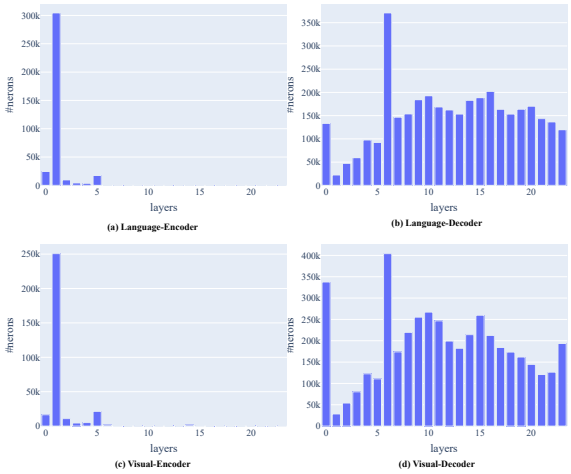


Figure 6: The distribution of all knowledge neurons of **MM-LAMA** in different layers under different modal settings. (a) Knowledge neurons in encoder under language setting. (b) Knowledge neurons in decoder under language setting. (c) Knowledge neurons in encoder under visual setting. (d) Knowledge neurons in decoder under visual setting.

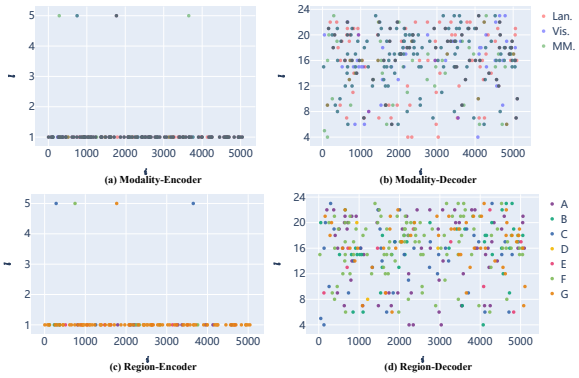


Figure 7: The distribution of knowledge neurons ($\lambda_m = 0.2$) activated by one sample in **MM-LAMA** on the Encoder and Decoder, where i represents the neuron index in l-th layer of the transformer. (a) Knowledge neurons in the encoder under different settings. (b) Knowledge neurons in the decoder under different settings. (C) Different regions of neuron sets in the encoder. (d) Different regions of neuron sets in the decoder.

Relation	Prompt Funtion	Prompt Template
nationality	p_1	What country does h^t come from?
	p_2	What country does the person in the picture come from?
state	p_1	Which state does h^t belong to?
	p_2	Which state does the place shown in the image belong to?
sport	p_1	What sport does the team h^t play?
	p_2	What sport does the team shown in the image play?
profession	p_1	What is the career of h^t?
	p_2	What is the career of the person in the image?
gender	p_1	What do you think the gender of h^t is? (Male or Female)
	p_2	What do you think the gender of the person in the image is? (Male or Female)

Table 5: Different prompt templates are manually designed, among which p_1 is for language setting and multimodal setting, p_2 is for visual setting, and h^t is the text of the head entity.