Master's Thesis

INVESTIGATING MULTIMODAL INTERACTION IN VISION LARGE LANGUAGE MODELS

2310041 WEI Houjing

Supervisor INOUE Naoya

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

February 2025

Abstract

Vision **L**arge **L**anguage **M**odels (VLLMs) extend **L**arge **L**anguage **M**odels (LLMs) by equipping them with the ability to perceive and process both textual and visual data, enabling impressive capabilities such as drafting stories based on images and building a website based on handcrafted images. In recent years, the development of VLLMs has advanced rapidly, yielding a substantial body of remarkable research. For instance, MiniGPT-4 [1] introduces a simple linear mapping to align visual information from a pre-trained vision encoder with a frozen large language model, concretely, this linear mapping transfers the output vectors (image features) into suitable inputs for language model, effectively functioning as a **modality connector**. By fine-tuning this linear layer using a high-quality instruction-following dataset, their work demonstrates the effectiveness of leveraging a modality connector to elicit multimodal capabilities from frozen models. Meanwhile, the LLaVA series [2–4] follows a similar architecture, utilizing a linear projection or a multilayer perceptron as the modality connector. Through fine-tuned on visual instruction-tuning data, LLaVA models achieve remarkable multimodal conversational capabilities, setting new state-of-the-art benchmarks in multimodal reasoning tasks. In addition, other works [5, 6] such as InstructBLIP [5] utilize a **Q**uery trans**F**ormer (Q-Former) as the cross-modal interface, wherein the query-based mechanism enables a more selective extraction of visual features tailored to language instruction. When combined with vision-language instruction tuning, the model attains impressive zero-shot performance across a variety of vision-language tasks.

As mentioned before, a common architecture of these VLLMs consists of a frozen visual encoder, a pre-trained **L**arge **L**anguage **M**odel (LLM), and a learnable cross-modal projector for mapping representations from different modalities. Such a connector-based framework is the basis of most modern VLLMs [1, 3, 5, 7],and has demonstrated significant efficiency and remarkable performance in many vision-language task scenarios. In detail, during the inference, VLLMs are fed with both visual and textual inputs, and (1) the frozen image encoder first encodes the image into a set of visual representations, then (2) the visual representations are transferred by a cross-modal projector aiming an alignment with the distribution of typical text token representation, and (3) the projected visual representations (tokens) are then concatenated with instruction tokens (if any), and fed into the pre-trained LLM for a causal language modeling operation.

Given the remarkable progress of VLLMs across various vision-language tasks, another line of work has emerged, focusing on investigating the inner work of the VLLMs. A pioneering study [8] approached this problem by identifying multimodal neurons within the Transformer's MLP layers and mapping them to semantically related text. Their experiments empirically showed that image tokens, which have been projected into LM embedding space, do not effectively encode interpretable semantics. Similarly, another study [9] found that language models inherently capture domain-specific visual attributes, while fine-tuning the cross-modal projector does not enhance this capability. Compared to the above work interpreting visual feature encoding in a human-readable format, more recent research adopts a mechanistic interpretation approach to examine the internal processes of VLLMs. In [10], the authors demonstrate that VLLMs encode factual associations within early multi-layer perceptron (MLP) layers and subsequently transfer this information to the final position token through intermediate Multi-Head Self Attention(MHSA) modules. Similarly, paper [11] finds that object-specific information is primarily localized within visual tokens corresponding to object patches and is later propagated into text tokens to facilitate language token predictions. Additional studies [12, 13] contribute to this research domain by investigating the internal workings of VQA in LLaVA, employing methodologies such as log-probability analysis, parameter projections into the unembedding space, and the examination of multimodal information flow across LM layers. These studies have significantly advanced our understanding of the internal mechanisms of VLLMs.

In VLLMs, the LLM is fed with a concatenation of visual token representations and textual token embeddings to perform the causal language modeling operation, indicating that the internal mechanisms of the language model, particularly the attention module, are required to leverage information from the visual modality to refine representations. However, existing research has primarily focused on interpreting the projected image tokens or examining how multimodal information flows throughout the LLM, leaving interaction between the vision token and text token (multimodal interaction) unexplored. Moreover, considering that the image tokens are obtained from encoders pre-trained exclusively on visual data, how they are progressively processed within the LLM representation space is a crucial indicator for revealing the aforementioned multimodal interaction.

Thus, this thesis highlights the interaction between the image token and text token. Especially, we focus on investigating how image representations evolve along Transformer-based autoregressive text decoders in modern VLLMs. To this end, we first map the projected visual representations into textual tokens by LM heads in LLMs (*logit lens*) to examine how

encoded representations from the visual encoder, after being projected, are progressively transformed into language semantics along the layers of the LLMs. Our experiments reveal two key findings:

1. Although visual representations are not explicitly trained for next-token prediction, LLMs can decode the visual representations into related text tokens. Additionally, in mid-to-late layers, the hidden states of visual tokens become more semantically aligned with the textual modality compared to the early layers.
2. The correctness of the LM decoding of the visual token's hidden states appears largely independent of the instruction tokens.

Next, we employ the cosine similarity between the hidden states of visual tokens and textual tokens to characterize the magnitude of the multimodal interaction, using the aligning dynamics of visual token representations toward text token embeddings as an indicator. Specifically, our experimental results, conducted on four models across two datasets, reveal the following findings:

1. Despite differences in designs of cross-modal projectors and the size of LM components, these models exhibit consistent trends in their inter-modal similarity curves, suggesting a general aligning dynamics of visual token representations towards textual token embeddings.
2. Similarity curves exhibit a bimodal pattern and increase rapidly in mid-to-late layers, suggesting three-stage multimodal interaction dynamics.
3. Regardless of varying types of textual prompt tokens, the layer-wise changes in inter-modal similarity values remain consistent, suggesting that the inter-modal interaction dynamics are largely independent of the specific prompt used.

Moreover, from another perspective, we conducted a layer-wise attention visualization analysis. Our analysis reveals that: (1) Attention scores from instruction tokens (as attention queries) to visual tokens strengthen starting from the middle layers. (2) Certain visual tokens at specific positions receive significantly higher attention than others from textual tokens. Such observation motivated our investigation into the relationship between the number of visual tokens and language modeling loss, aiming to provide empirical insights for balancing the effectiveness (lower forward loss) and efficiency (fewer image tokens) during model inference. In detail, by modifying the forward function during model inference, we investigate the impact of varying the number of visual tokens on the model's forward computation loss. Extensive experiments reveal that

1. Once the quantity of image tokens surpasses a certain threshold, loss

reduction goes slowly or even stops, indicating that subsequent image tokens may have limited contribution and could be redundant.
2. The contribution of visual tokens to loss reduction is not uniform, with certain tokens at specific positions playing a significantly greater role.

This finding provides valuable insights for optimizing and accelerating inference in VLLMs.

Keywords: Interpretability, Vision Large Language Models, Inference Dynamics.

Acknowledgment

First, I would like to express my heartfelt gratitude to my supervisor, Prof. INOUE Naoya. During my two years at JAIST, he has been always patient, providing me with guidance and support with his extensive professional knowledge. As a student who switched to this field from a completely different discipline, I faced many academic and technical challenges, such as unfamiliar concepts and methodologies, and on many occasions, I found myself unsure of how to proceed. In those moments, discussions—even casual conversations—with Prof. INOUE Naoya always offered valuable insights, helping me regain confidence and see things. As I continuously refine my understanding of these concepts and methodologies, I have experienced great growth. I feel so honored and thankful to complete my master's research under his supervision.

At the same time, I would also like to thank Prof. Shogo Okada and Prof. NGUYEN, Minh Le. I was quite nervous during my mid-term presentation, but they supported me with kindness and provided constructive feedback to help me advance my research. I am also grateful to my peers in the laboratory. They are friendly, smart, and deeply passionate about their academic pursuits. Talking with them has always been a great experience, as they offer fresh ideas and help me think things through from different perspectives. Their kindness and patience have motivated me, allowing me to face challenges and difficulties in research with a more positive mindset.

Lastly, I would like to acknowledge everyone who has supported me in any way—whether it was offering kind words during stressful times, exchanging ideas, or simply expressing faith in my ability to succeed. I also must thank my family, whose unwavering support has been my greatest source of strength. No matter how lost or uncertain I felt, my parents were always there to uplift me, providing reassurance and the motivation to keep moving forward.

List of Figures

Contents

Chapter 1

Introduction

1.1 Background

Vision **L**anguage (VL) research lies at the intersection of computer vision and natural language processing, aiming to endow computers with the dual capabilities of visual perception and textual understanding. This field has long been regarded as a crucial step toward achieving general artificial intelligence. However, because images and text differ significantly in their feature representations, early approaches were restricted to task-specific model designs or training objectives, severely limiting research progress [14–16]. Subsequently, the Transformer architecture, originally introduced to address sequence-to-sequence tasks in the NLP realm, demonstrated a remarkable ability to improve model generalization capability [17–19]. Inspired by these successes, researchers extended the Transformer architecture to image recognition, achieving promising initial results. Building on this foundation, a new paradigm of **V**ision-**L**anguage **P**retraining (VLP) emerged, driving remarkable progress in vision-language joint learning.

In recent years, substantial advancements have been made in the realm of **L**arge **L**anguage **M**odels (LLMs). By extending the impressive emergent capabilities of LLMs to multimodal scenarios, researchers have developed a variety of Vision Large Language Models (VLLMs), demonstrating amazing emergent abilities in recent studies [1–3, 5–7, 20, 21]. Compared to traditional VLP approaches, which involve pretraining vision-language representations, the aforementioned VLLMs typically consist of three key components: a pre-trained visual encoder for extracting visual representation from the image, a pre-trained large language model for producing text output, and a cross-modal projector (or connector) accounts for mapping visual features to LMs' embedding space. These models are called connector-based VLLMs, requiring training only a lightweight projection layer—often implemented as a simple linear mapping, a multilayer perceptron, or a cross-modal extractor—or, in some cases, fine-tuning the LM decoder.

During inference, a typical pipeline of implementing VLLMs is to (1)

utilize a frozen image encoder to extract visual features from the input image, then (2) leverage a cross-modal connector to project these visual features into LMs' embedding space, and (3) the projected visual token representations are concatenated with text token embedding and fed into the LM decoder to perform causal language modeling. Despite this simplified training process, VLLMs perform tasks beyond the reach of earlier VLP approaches, such as generating detailed recipes from food images or creating advertising promotions for products showcased in images. Furthermore, since these models do not conduct extensive end-to-end training to fuse visual and linguistic representations at a massive scale, they significantly reduce computational and memory overhead while maintaining impressive performance across a range of vision-language downstream tasks.

Meanwhile, various benchmark evaluations have been developed to specifically assess the visual and language understanding capabilities of VLLMs. These benchmarks cover a comprehensive evaluation of different aspects of the model, including verb understanding, spatial cognition, visual compositional reasoning, and overall performance evaluation, spanning multiple task levels. While these evaluations highlight the potential of VLLMs, they also reveal critical issues such as inconsistent adherence to instructions, hallucination generation, and difficulty in performing complex reasoning, which significantly hinder the further development and safe deployment of these models.

Against this backdrop, research on the interpretability of VLLMs becomes increasingly urgent, leading to a growing focus on understanding their internal representations and improving model transparency. A pioneering study [8] localized multimodal neurons in MLP layers, translating them into related text tokens. The authors empirically showed that visual feature outputs of the cross-modal projector do not effectively encode interpretable linguistic semantics. A following work [9] found that it is the LM decoder, not the cross-modal projector, that encodes domain-specific visual attributes. In comparison with the above work, where feature encoding is interpreted in a human language form, more recent research explores the inner workings of VLLMs in a broader scope via a mechanistic interpretation lens. In [10], the authors show that VLLMs store factual associations in earlier Multilayer Perceptron (MLP) layers and transfer them to the final token position via middle-layer Multi-Head Self Attention (MHSA) blocks. Meanwhile, Paper [11] discovers that object-specific information is concentrated in visual tokens spatially corresponding to object patches, then integrated into text tokens for language token predictions. Other works [12, 13] enrich this research area by examining the decision-making mechanism of Vision Question Answering (VQA) in LLaVA through approaches such as log-probability

changes, parameter projection to the unembedding space, and multimodal information flow along the layers of LM decoer.

These investigations have greatly improved our understanding of the internal mechanics of VLLMs.

1.2 Research Objective

Among existing studies on interpreting the internal behavior of VLLMs, our research is particularly aligned with investigations into the evolution of visual token representations across LM layers [11] and the flow of cross-modal information integration [13].

However, prior studies have primarily focused on interpreting projected visual features or examining how crossmodal information flows within the LLM component, leaving the systematic analysis of multimodal interaction largely unexplored. As introduced in § 1.1, the connector-based VLLMs allow the LLM component to implicitly refine representations from both modalities to complete causal language modeling computation, leading to multimodal interaction that naturally happens within the LLM. Furthermore, since the visual features are extracted from frozen vision encoders pre-trained solely on visual data, understanding how these visual token representations are sequentially transformed throughout the LLM layers is a fundamental aspect to understanding the multimodal interaction within the LLMs.

Therefore, this thesis aims to analyze the interaction between image tokens and text tokens. Specifically, we systematically and quantitatively investigate the evolution of image token representations, referred to as the dynamics of visual representation, across the different layers in modern VLLMs. To this end, we mainly utilize two approaches described as follows:

1. Use *logit lens* to investigate whether visual token representations are transformed into text tokens.
2. Apply cosine similarity to investigate how and to what extent visual token representation interacts with textual token embeddings.

The first approach focuses on interpreting the changing process of visual token representation using human language, while the second method utilizes similarity metrics to quantify the degree of such an evolving procedure by focusing on the interaction between image token representations and text token embeddings.

The experimental results of *logit lens* show that while visual representations are not explicitly trained for next-token prediction, LLMs can decode them into related text tokens. Additionally, the hidden states of visual tokens

become more semantically aligned with the textual semantics in mid-to-late layers compared to early layers. Finally, the correctness of the LM decoding of the hidden states appears to be almost independent of the instruction tokens, since the correctness remains the same across different instructions, exhibiting negligible variance.

Cosine similarity is a widely used metric for assessing the semantic similarity between high-dimensional vectors in the representation space, reflecting the degree of contextualization encoded by the language model. Building on this, analyzing how the similarity between visual token representations and text token embeddings evolves layer by layer within the LM provides an indirect perspective on the aligning dynamics of visual token representations toward text token embeddings. Our experiments reveal several key findings. Despite variations in the design of cross-modal projectors and the size of LM components, all models exhibit consistent trends in their inter-modal similarity curves, indicating a general alignment dynamic where visual token representations progressively converge toward textual token embeddings. Moreover, similarity values increase as the layers deepen, suggesting that cross-modal alignment strengthens throughout the model's depth. Additionally, this layer-wise trend in cross-modal similarity remains consistent across different textual input sequences, implying that cross-modal alignment is largely independent of the specific prompt used.

In addition, a norm-based attention analysis is conducted to visualize the alignment dynamics along LM decoder layers. Our qualitative analysis reveals that, first, attention scores from instruction tokens (serving as attention queries) to visual tokens become increasingly prominent starting from the middle layers. Second, specific visual tokens at certain positions receive substantially higher attention from textual tokens compared to others. This result leads us to look into the connection between the number of visual tokens and language modeling loss to offer empirical guidance for achieving a balance between effectiveness and efficiency (by shortening the input length) during model inference.

Chapter 2

Related Work

2.1 Transformer Architecture

A standard Transformer architecture lies in an encoder-decoder framework, where $\mathbf{L} \in \{0, ..., L\}$ stacked blocks (layers) are designed for both encoder and decoder sides. Consider the input and output sequences are represented as $\mathbf{X} = [x_1, ..., x_N] \in \mathbb{R}^{N \times d}$ and $\mathbf{Y} = [y_1, ..., y_T] \in \mathbb{R}^{T \times d}$, respectively. The input is first tokenized, then projected to get the embeddings by applying an embedding matrix $\mathbf{E} \in \mathbb{R}^{d \times e}$, followed by adding positional embedding, resulting in the input representations $\mathbf{H^0} = \mathbf{E^0} + \mathbf{P^0} \in \mathbb{R}^{N \times d}$, where $\mathbf{E^0}$, $\mathbf{P^0}$ are token embeddings and its positional embeddings when $l = 0$. Each encoder block comprises two sub-layers, called **M**ulti-**H**ead **S**elf-**A**ttention (MHSA) and **F**eed-**F**orward **N**etwork (FFN), each sub-layer followed by a residual connection and a layer normalization operation.

MHSA takes as input the sequence of representations $\mathbf{H^0} \in \mathbb{R}^{N \times d}$, and multiplies them by four matrices $\mathbf{W}_Q^l, \mathbf{W}_K^l, \mathbf{W}_V^l, \mathbf{W}_O^l \in \mathbb{R}^{d \times d}$ in each layer l (we henceforth exclude the layer superscript for conciseness). This produces queries, keys, and values for the subsequent attention computation: $\mathbf{Q} = \mathbf{H}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{H}\mathbf{W}_K$, $\mathbf{V} = \mathbf{H}\mathbf{W}_V$. These queries, keys, and values are split along the columns to form H attention heads with $\frac{d}{H}$ dimensions, denoted as $\mathbf{Q}^h \in \mathbb{R}^{N \times \frac{d}{H}}$, $\mathbf{K}^h \in \mathbb{R}^{N \times \frac{d}{H}}$, $\mathbf{V}^h \in \mathbb{R}^{N \times \frac{d}{H}}$, respectively. The attention maps are computed as:

$$\mathbf{A}^h = \text{softmax}\left(\frac{\mathbf{Q}^h \mathbf{K}^{h^{\mathrm{T}}}}{\sqrt{d/H}} + \mathbf{M}\right) \in \mathbb{R}^{N \times N} \tag{2.1}$$

where $\mathbf{M} \in \mathbb{R}^{N \times N}$ refers to the attention mask. Next, we multiply each attention map with its corresponding values to get weight-attended intermediate representations $\mathbf{Z}^h = \mathbf{A}^h \mathbf{V}^h$. We repeat this operation on H heads and concatenate their outcomes along columns, followed by multiplying with $\mathbf{W}_O$. Afterward, the residual connection and layer normalization are applied

to form the output of the attention module in layer l as $\tilde{\mathbf{H}}^l \in \mathbb{R}^{N\times d}$:

$$\tilde{\mathbf{H}}^l = \text{LayerNorm}\left(\mathbf{H}^{l-1} + \text{Concat}\left[\mathbf{A}^1\mathbf{V}^1, \ldots, \mathbf{A}^h\mathbf{V}^h, \ldots, \mathbf{A}^H\mathbf{V}^H\right]\mathbf{W}_O^l\right) \tag{2.2}$$

FFN is generally a two-layer linear transformation with an activation function between them, parameterized by two learnable matrices: $\mathbf{W}_{in}^l \in \mathbb{R}^{d\times d_{\text{ffn}}}$, $\mathbf{W}_{out}^l \in \mathbb{R}^{d_{\text{ffn}}\times d}$ to form as follows:

$$\text{FFN}^l(\tilde{\mathbf{H}}^l) = \text{ReLU}(\tilde{\mathbf{H}}^l\mathbf{W}_{in}^l)\mathbf{W}_{out}^l \in \mathbb{R}^{N\times d} \tag{2.3}$$

Same as the MHSA module, the output of the FFN module also applies a residual connection and layer normalization, resulting in the hidden states matrix $\mathbf{H}^l \in \mathbb{R}^{N\times d}$:

$$\mathbf{H}^l = \tilde{\mathbf{H}}^l + \text{LayerNorm}\left(\mathbf{FFN}^l(\tilde{\mathbf{H}}^l)\right) \in \mathbb{R}^{N\times d} \tag{2.4}$$

Each decoder layer possesses a design analogous to the encoder block but with an additional cross-attention sublayer between the MHSA and FFN sub-layers.

Cross-attention primarily models the correspondence between the source and target sequence. Let the self-attention and cross-attention in a decoder block be defined as:[1]

$$\mathbf{S}_{\text{self}} = \text{LayerNorm}\left(\mathbf{Attn}_{\text{self}} + \mathbf{S}\right) \tag{2.5}$$

$$\mathbf{S}_{\text{cross}} = \text{LayerNorm}\left(\mathbf{Attn}_{\text{cross}}\left(\mathbf{H}_{enc}, \mathbf{S}_{dec}\right) + \mathbf{S}_{\text{self}}\right) \tag{2.6}$$

where $\mathbf{S}$ denotes the input representation of the self-attention sub-layer, $\mathbf{S}_{\text{self}}$ and $\mathbf{S}_{\text{cross}}$ are the outputs of the self-attention and cross-attention sub-layer, and $\mathbf{H}_{enc}$ is the output of the encoder.

Autoregressive Large Language Models are trained to predict a probability distribution of the next token based on the preceding tokens. The autoregression is achieved by picking the next token from the model's predicted distribution (e.g., greedy, beam search, etc), appending it to the sequence, and feeding that extended sequence back into the model for the next forward pass. The process repeats until reaching an end-of-sequence condition or a certain length. The training process usually involves maximizing the log-likelihood objective: $\text{L} = -\sum_{t=1}^{T}\log(p(y_t|x, y_{<t}))$

[1]We don't dive into details about the decoder blocks in this work, as autoregressive large language models use an encoder-only architecture. Listing them here is merely for the sake of completeness for the introduction of Transformer architecture.

2.2 Vision Language Models

Vision Language Pretraining (VLP) Models. Building on the success of pretraining techniques in the NLP domain, numerous studies have leveraged large-scale image-text pair datasets for pertaining Vision-and-Language models, leading to the development of a wide range of Vision Language Pretraining (VLP) models.

A typical architecture of VLP models consists of three main components: 1) Vision and Language Encoding, which converts the raw input data into modality-specific representations; 2) Vision-Language Modeling, where multimodal representations are learned during pre-training; 3) Vision-Language (multimodal) Representation, which serves as the model output or is optionally fed into a decoder to conduct text generation. A common pipeline in most VLP models begins with tokenizing textual input and converting the resulting tokens into embeddings while the image is processed into visual features using a vision encoder. Both text input and visual input follow a BERT-like format, where each representation is a summation of three types of learnable embeddings, *i.e.*, token embedding, position embedding, and segment embedding for text modality and visual feature, spatial position embedding, and segment embedding for vision modality. Then these two modality representations are fed into the vision-language modeling module to produce vision-language (multimodal) representations. Based on how vision-language representations are modeled, we classify existing VLP models into three categories:

Early-fusion VLP models handle the concatenation of text embeddings and image features directly in a unified modality fusion module, usually consisting of a stack of several Transformer blocks. The cross-modal representations are learned by pre-training on several objectives, e.g., image-text matching, cross-modal masked language modeling, and cross-modal masked region prediction. For instance, VL-BERT [42] extends the Transformer backbone to take regional visual features and linguistic embeddings as input. To produce visual-linguistic representations, the model is pre-trained on massive-scale image-text datasets with two carefully designed pre-training tasks: cross-modal masked language modeling and cross-modal masked regional feature classification. VisualBERT [41] employs a stack of Transformer layers to leverage self-attention to align textual tokens with corresponding image regions. They propose two visually-grounded language model objectives for pre-training on image caption data.

Dual-encoder VLP modles commonly adopt two separate encoders for learning joint representations of images and text via large-scale contrastive

learning. One outstanding work is CLIP [43], which is trained on a vast dataset of image-text pairs crawled from websites using a contrastive objective that drives semantically related image-text pairs closer in embedding space while pushing unrelated pairs apart. Their work demonstrates a strong zero-shot transfer ability on image classification tasks. Similarly, ALIGN [44] adopts a dual-encoder architecture for aligning visual and language representations in a shared latent embedding space. The image and text encoders were trained through a contrastive objective, effectively attracting matched image-text pairs while repelling unmatched ones. Their work suggests that scaling corpus size can compensate for noise and yield powerful, general-purpose embeddings across various vision and language tasks.

Cross-attention VLP models adopt a cross-attention mechanism to model the interaction between vision and language, which usually contains two unidirectional cross-attention sub-layers: one from language to vision and another from vision to language. The cross-attention module is responsible for exchanging information and aligning the semantics between the two modalities. ViLBERT [45] treats image features and text token embeddings as two parallel streams that are fed into two cross-modal modules for cross-modal interaction. LXMERT [46] utilizes a combination of three encoders: an object relationship encoder, a language encoder, and a cross-modality encoder. By pre-training with five proxy tasks, they aim to achieve thorough learning of interaction between visual and textual representations.

Vision Large Language Models (VLLMs). Recently, GPT-4V has exhibited surprising multimodal abilities that are rarely observed in previous VLP models. This inspired many researchers to develop multimodal systems that can benefit from the emerging capabilities of large language models, leading to various works for VLLMs. In contrast to previous VLP models, VLLMs have two key features: first, they are equipped with a large-scale language model with billions of parameters; second, they follow a novel training paradigm based on multimodal instruction tuning.

A typical VLLM architecture comprises three components, *i.e.*, a pre-trained vision encoder, a pre-trained LLM, and a cross-modal connector serving as an interface to connect different modalities. In general, ViT [47] or CLIP image encoder and its variants are widely used as the vision encoder to extract image patch representations. Most VLLMs utilize open-sourced LLMs, e.g., LLaMA series [24] and Vicuna family [48] as text decoders. A learnable connector between the pre-trained visual encoder and LLM accounts for projecting visual representation into the LLM space for bridging the modality gap.

There are several prominent representative VLLMs. InstructBLIP [5] proposes an instruction-aware Q-Former module that leverages a set of learnable query tokens to extract instruct-aware image features. Around 26 publicly available VL datasets are transformed into the instruction-following format for conducting multimodal instruction tuning. MiniGPT-4 [1] employs a single linear projection layer to align the visual features with the Vicuna [48]. Only training this linear projection layer in two stages enables MiniGPT-4 to exhibit impressive capabilities absent in previous VLP models. Another line of work [2] adopts one/two linear MLP to project visual tokens for feature dimension alignment.

2.3 VLLMs Interpretation

As introduced in §2.2, we have witnessed impressive success and efficiency in mapping image features to language model soft prompts, denoted as visual tokens, as cross-modal connectors for instruction tuning of VLLMs. Meanwhile, these remarkable advances also motivate researchers to explore the underlying workings behind such cross-modal mapping and, furthermore, to interpret how these VLLMs work.

Interpreting Vision Large Language Models The cross-modal projector in VLLMs takes input as the visual features from an off-the-shelf vision encoder and maps them to have the same dimension with language model embeddings. In work [21], the authors trained a linear projector $\mathbf{P}$ to project the visual representations of a pre-trained image encoder into the input space of a generative language model using an objective of image captioning. During training, they use different image encoders that accept different levels of linguistic supervision; they demonstrate the effectiveness of linear projection on image captioning as well as VQA tasks, suggesting that a language model structurally represents visual concepts in a manner resembling that acquired by a vision encoder.

However, another line of study [8, 9] demonstrated that post-projection presentations do not encode interpretable semantics for language models and contain fewer task-relevant visual attributes, showing that the update of cross-modal projector weights does not lead to the correspondence between image tokens and discrete language tokens. Specifically, Work [8] identifies multimodal neurons inside Transformer MLP parameters (*i.e.*, row vectors of weight matrix $\mathbf{W}_{out}$), proposing a method of attributing neuron effects from image patches for localizing them and empirically demonstrating the effectiveness of translating them into semantically related text. Another

consistent work [9] indicates that LMs account for modeling domain-specific visual attributes while fine-tuning the cross-modal projector does not enhance such capability.

Compared with the studies mentioned above, where feature encoding is interpreted in a human language form, more recent research explores the inner workings of VLLMs in a broader scope via a mechanistic interpretation lens. In paper [10], the authors explore the mechanism of multi-modal knowledge storage and transfer in a factual VQA task setting, revealing that VLLMs retrieve factual associations from much earlier MLP layers and transfer them to the last token position of the input prompt via MHSA blocks in the middle layers. Subsequently, [11] reveal that object-specific information is concentrated in visual tokens spatially corresponding to object patches, and representations at visual token positions are iteratively refined to align with interpretable textual concepts. Other papers, such as investigating the mechanism of VQA in LLaVA models via methods like calculating the log probability increase or projecting model parameters into the unembedding space of the language model [12]; examining the information flow between vision and language across LM decoder layers when solving the QVA task [13], also enrich the research progress in this area.

Chapter 3

Investigating Verbalization across Transformer Layers

In contrast with previous work focused on investigating how information flow propagates in VLLMs [13], we highlight our research target as systematically investigating how visual representations are transformed and shaped across different layers of the LM decoder during model inference. This chapter comprises three parts, demonstrating to what extent the hidden representation of visual tokens can be converted into linguistic concepts represented in the language vocabulary. We first describe the *logit lens* technique in §3.1, followed by the experimental setup in §3.2. Then, we present a series of findings and observations in §3.3.

Notations. For notation consistency with §2.1, the concatenation of **image tokens** and **instruction tokens** is denoted as $\mathbf{H} = [\mathbf{H}_{vis},\ \mathbf{H}_{inst}]$. We use $\mathbf{H}^0$ and $\mathbf{H}^L$ to represent the outputs from the input embedding layer and final output layer, respectively. The intermediate representation of the concatenation is denoted as $\mathbf{H}^l$, where $l \in \{1, \ldots, L-1\}$.

3.1 Logit Lens

logit lens [22] is commonly used to intuitively examine the language semantics encoded in a model's intermediate representations. This technique treats each Transformer block in a decoder-only language model accounting for implementing an incremental update to a probable next-token prediction. Such an update is achieved by multiplying hidden states at any layer with the unembedding matrix, producing the unnormalized logits of each token, which are then turned into a probability distribution via a softmax function for next-token prediction. This yields a sequence of most likely next tokens, called a prediction trajectory, which shows a tendency within the model to converge to the final output.

Specifically, consider an arbitrary hidden state of the visual token at the layer l of LLM in a VLLM as $h_{vis}^{(l)}$, the multimodal version of *logit lens* can be formulated as:

$$\text{LogitLens}\left(h_{\text{vis}}^{(l)}\right) = \text{LayerNorm}\left[h_{\text{vis}}^{(l)}\right] W_U, \tag{3.1}$$

where W_U is the unembedding matrix of the LLM.

3.2 Experimental Setting

We provide detailed information on VLLMs studied and evaluation datasets used in our experiments.

Models. InstructBLIP [5], which is initialized from the pre-trained BLIP-2 model [23], pioneers the exploration of vision language instruction tuning. By using their collected instruction-response multimodal dataset to train an instruction-aware **Q**uery **T**ransformer (Q-Former), they enable the model to extract flexible and informative visual features from the output of a frozen image encoder according to the given instructions. Extensive experiments demonstrate the strong performance of InstructBLIP models in both zero-shot and fine-tuning settings on a wide range of vision-language tasks. InstructBLIP is implemented with the same image encoder but different pre-trained LLLMs, including instruction-tuned encoder-decoder LLMs—FlanT5XL (3B), FlanT5-XXL (11B), and decoder-only Transformer instruction-tuning from the LLaMA family [24]-Vicuna—7B and Vicuna-13B.

LLaVA [2] is developed by integrating the open-set visual encoder of CLIP [19] with the language decoder Vicuna [25], followed by end-to-end fine-tuning on their generated instructional vision-language data. The model is trained in a two-stage instruction-tuning way, where in the first stage only the linear projection layer is trained while both the projection layer and the text decoder are fine-tuned. The model is optimized by maximizing the likelihood of next-token prediction probability. **LLaVA-**1.5 [3] is an enhanced version with simple modifications, incorporating a bigger vision encoder CLIP-ViT-L-336px and an MLP projection, further improving the performance on various VL tasks.

Datasets. We leverage two common vision-language datasets in this experiment. **1) MS COCO Caption** dataset [26] annotates images from **M**icrosoft **C**ommon **O**bjects in **CO**ntext (COCO) [27], where each image is

accompanied by five human-generated captions using Amazon's Mechanical Turk (AMT), resulting in 413K captions for 82K images in training, 202K captions for 40K images in validation, and 379K captions for 40K images in testing. In our experiment, we use the preprocessed Karpathy's split[1] [28] derived from original MS COCO captions, which is predominantly created for benchmarking the image captioning task, comprising 82K/5K /5K for the train/validation/test sets. **2) Winoground** dataset [29] is a carefully handcrafted probing dataset for evaluating the ability of vision and language models to conduct visio-linguistic compositional reasoning. It comprises 400 items, each including two pairs of images and their corresponding captions. While MS COCO caption dataset features include images containing multiple objects in their natural context, Winoground presents difficulties as effective matching necessitates the model to discern nuanced distinctions between the image and the caption, with both captions containing a completely identical set of words in a different order.

Implementation Details. We use pre-trained InstructBLIP equipped with ViT-g/14 [30] as the image encoder and Vicuna-13B [25] as the text decoder. We randomly picked 400 images from the COCO Captions and Winoground datasets. For InstructBLIP, we extract the hidden representations of 32 visual tokens at each layer of the LM decoder, then decode them into language words using *logit lens*. We repeat the same visual hidden representation procedure for pre-trained LLaVA, which applies the pre-trained CLIP visual encoder ViT-L/14 as a vision encoder and Vicuna-13B [25] as the text decoder.

We use three different types of prompts for the image captioning task: 1) normal instruction formed as a statement or a question; 2) noisy prompt formed by randomly sampling from LM vocabulary. 3) empty prompt without any text instruction.

We apply precise lexical overlap as our evaluation criteria and define a visual hidden state as being decoded correctly if its decoded word matches the ground-truth caption. Therefore, precision indicates how many correctly decoded words overlap with all decoded words, while recall refers to the proportion of correctly decoded words to ground-truth caption words. Before calculating the precision and recall, we preprocess both decoded words and caption words, such as lowercase initials, filtering out stop words and punctuation.

[1] https://cs.stanford.edu/people/karpathy/deepimagesent/

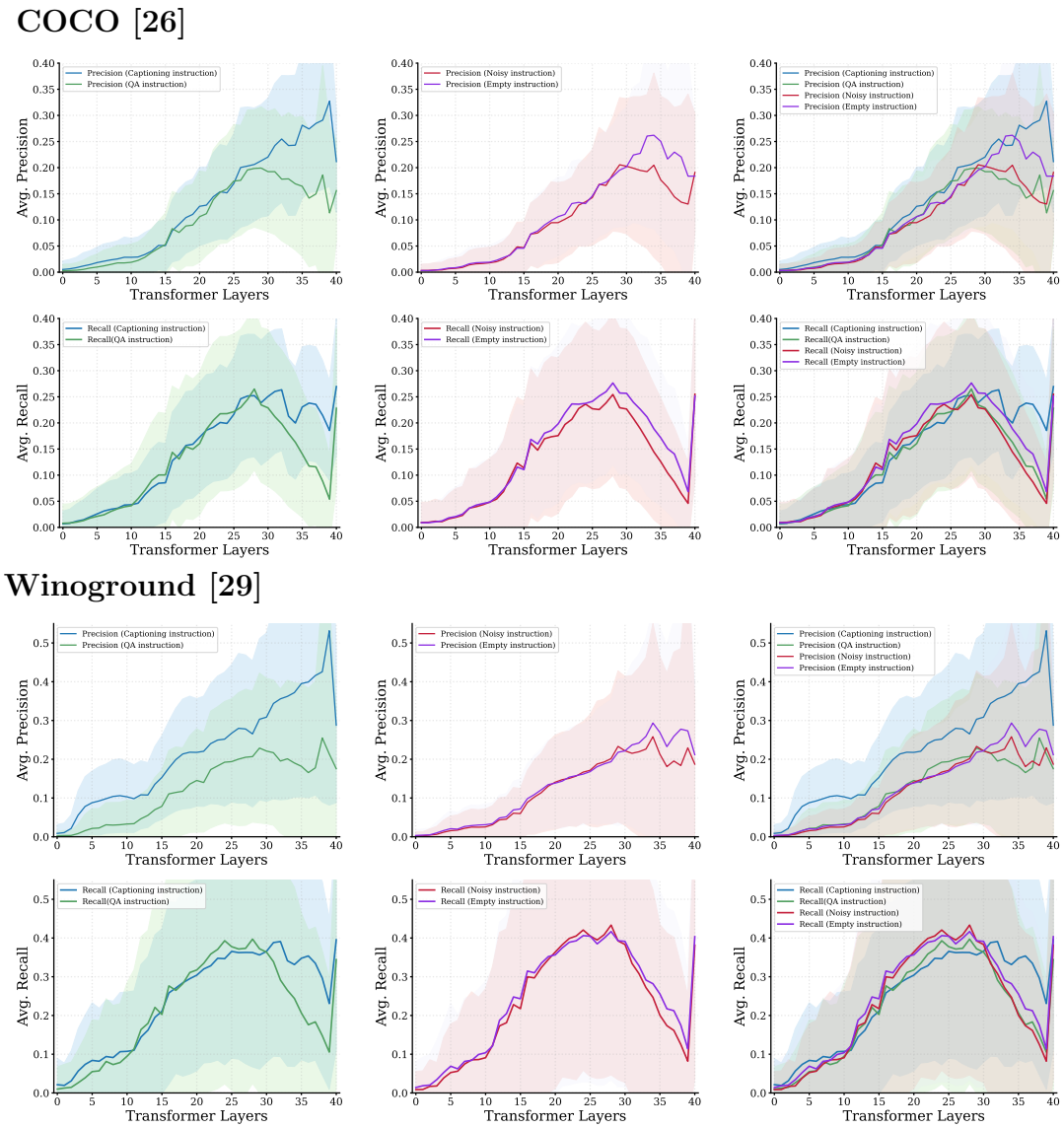

Figure 3.1: Precision and recall of decoded visual tokens along LM layers on COCO and Winoground for InstructBLIP (Vicuna-13B).

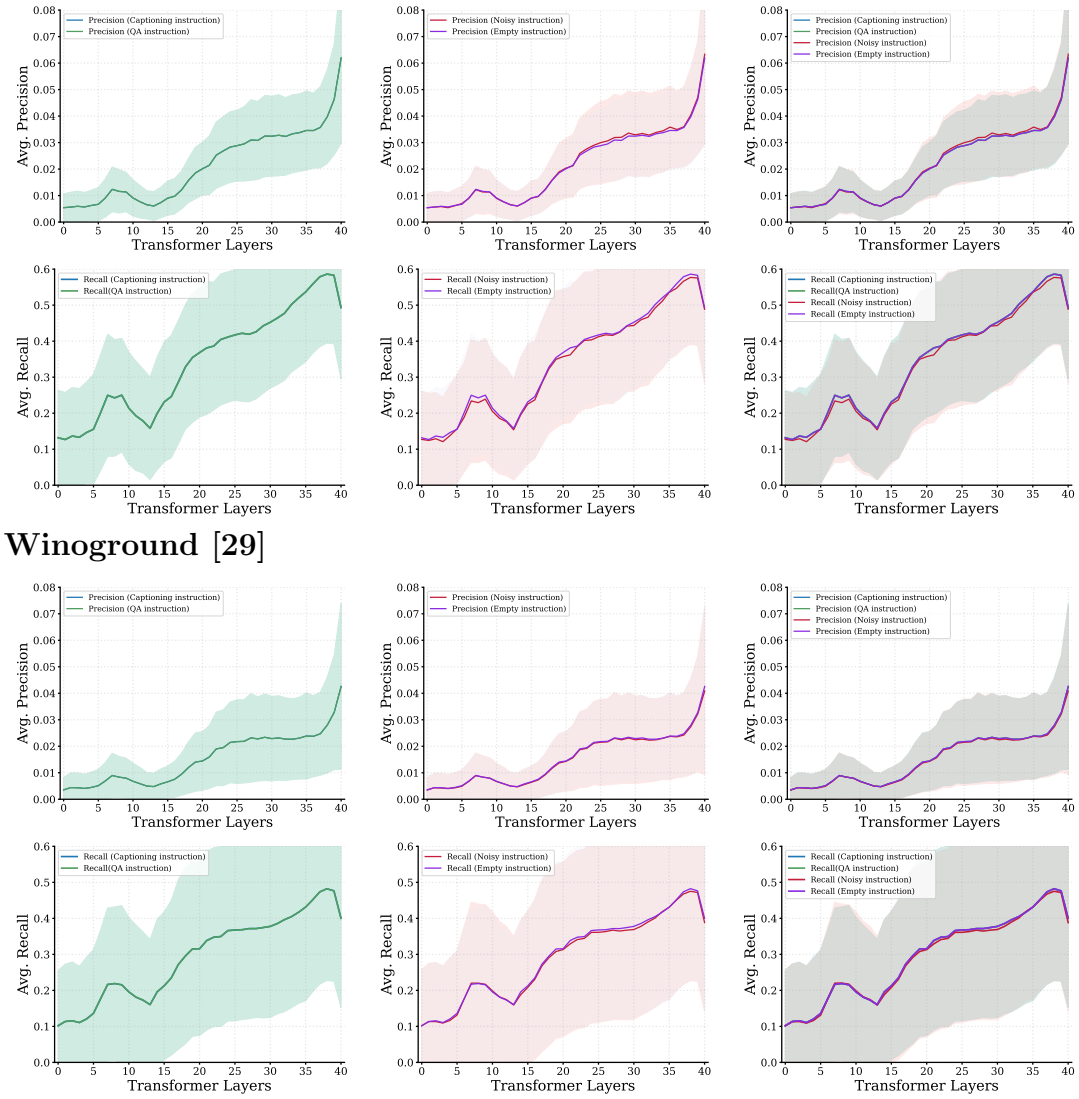

Figure 3.2: Precision and recall of decoded visual tokens along LM layers on COCO and Winoground for LLaVA-1.5 (Vicuna-13B)

3.3 Results

Fig. 3.1 and Fig. 3.2 present average precision and recall scores using different types of prompts on the COCO Captions and Winoground datasets for InstructBLIP and LLaVA-1.5, respectively.

In general, all subplots present a continuously rising tendency in mid-to-late layers, indicating the intermediate representations of visual tokens are progressively morphed into linguistic forms that match correct ground-truth captions. In the lower layers (near embedding space), both precision and recall are nearly negligible, suggesting that raw image tokens tend to produce irrelevant word distributions. In the mid-to-late layers (starting from around layer 10), both lines continuously climb, reflecting an ongoing process of refinement where the visual token representations become more semantically coupled with the textual domain. Around the deepest layers (after layer 30), we observe a slight variability between precision and recall, indicating a possible reduction in correctly decoded words. Such a trend is observed in both InstructBLIP and LLaVA-1.5, demonstrating that VLLMs can correctly assign probabilities to vision tokens, even though they are not trained to make the next word prediction.

Next, let us take a closer look at how different prompts affect the correctness of decoded words. For InstructBLIP, comparing a meaningful prompt to a noisy or empty one, we observe that in mid-to-late layers (from layer 13 to layer 30), the empty prompt curve performs best among the three prompt types, meaning the intermediate hidden states of vision tokens are converted into text tokens more successfully. However, in later layers (around layer 30 and onwards), curves of empty and noisy prompts go downward and fall significantly below the line using normal instruction prompts, indicating that few words are being decoded correctly. This observation is intriguing and, to some extent, counterintuitive. We conjecture that in those mid-range layers (around layers 10 to 30), having no prompt at all sometimes allows the model to "free-associate" from the image tokens without being constrained (or misled) by a partially relevant or noisy instruction. However, in the last few layers, such "free-associate" seems to be diminishing, leading to a gap in precision and recall scores. On the other hand, we observe completely different behavior in the LLaVA-1.5 model, where the evaluation performance remains unchanged across various prompts and datasets. As shown in Fig. 3.2, across both datasets, regardless of the type of prompt given, the precision and recall at all Transformer layers exhibit a consistent trend, nearly overlapping. This phenomenon aligns with the aforementioned hypothesis that there is a "free association" between image tokens and

instruction tokens.

In summary, this chapter describes our extensive exploration of interpreting the intermediate representation of visual tokens directly using human language, demonstrating the refinement process of visual tokens toward language's next token embedding space. Additionally, it empirically revealed that a decoder trained merely on text data can nevertheless process image tokens into meaningful language words, revealing a capacity for cross-modal integration within a text-only backbone.

Chapter 4

Investigating Aligning Dynamics across Transformer Layers

Since the LLM component in modern VLLMs processes a concatenation of visual token representations and textual token embeddings to perform causal language modeling, its internal mechanisms—particularly the attention module—require leveraging information from the visual modality to refine the intermediate representations. Therefore, instead of solely analyzing the evolution of visual representations in isolation, we focus on examining the layer-wise evolution of similarity between visual token representations and text token embeddings within LLMs, i.e., inter-modal similarity, aiming to provide an indirect yet informative perspective on understanding the dynamics of visual representations towards textual embeddings.

4.1 Measuring the Interaction via Cosine Similarity

Cosine similarity is a commonly used metric for measuring the semantic similarity between high-dimensional vectors within a representation space. It serves as an indicator of how well contextual information is encoded by the language model, offering insights into the extent to which different tokens interact within the model computation process. Drawn inspiration from prior work in contextual representation analysis [31], we utilize cosine similarity as a measure of contextuality to capture how intermediate representations (hidden states) originally trained on different modality data affect one another's representation in the language model. Specifically, by quantifying the contextualization between image token representation and text token embeddings, we demonstrate the magnitude and progression of alignment dynamics of visual representations towards text token embeddings across Transformer blocks in VLLMs, aiming to shed light on the deeper processes that enable text-only decoders to handle visual information effectively.

4.1.1 Experimental Setting

Let $v_i^{(l)}$ and $w_j^{(l)}$ denote the hidden state vectors of tokens i and j, respectively. The average cosine similarity for the hidden states at each layer l in LMs is thus defined as follows:

$$s^{(l)} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \cos\left(v_i^{(l)}, w_j^{(l)}\right), \tag{4.1}$$

where m and n indicate the number of tokens in two sets. Inter-modal similarity is computed by choosing $v_i^{(l)}$ from vision tokens and $w_j^{(l)}$ from text tokens. Higher similarity suggests that the two sets of vectors occupy closely related subspaces in the representation space, indicating that they may encode similar features.

Models and Datasets. We implement experiments for exploring and evaluating the dynamics of visual token representations on four variants of two VLLMs, i.e., InstructBLIP(Vicuna-7B), InstructBLIP(Vicuna-13B), LLaVA-1.5(Vicuna-7B), and LLaVA-1.5(Vicuna-13B). The details of these models and datasets are introduced in §3.1.

Implementation Details. We compute the cosine similarity between the hidden states of image tokens and text tokens at each layer of the Transformer-based text decoder, averaging the resulting similarity scores to derive an alignment metric. This metric serves as an indicator of the degree of alignment dynamics, reflecting the extent to which visual token representations align with textual token embeddings.

In our implementation for LLaVA-1.5 models, after extracting hidden representations of 576 image tokens and corresponding instruction hidden states from a specific layer of the model, we calculate the cosine similarity for each pair of image and text token hidden states. We then average these similarity scores to obtain an inter-modal similarity measure for that layer. This experiment is conducted on four VLLMs, each evaluated using a set of randomly sampled 600 images for the image captioning task. We employ two types of prompts: 1) a normal instruction prompt, constructed as either a statement or a question for the image caption task; 2) a noisy prompt, generated by randomly sampling tokens from the language model's vocabulary. The prompts used in our experiments are included in the Appendix A.

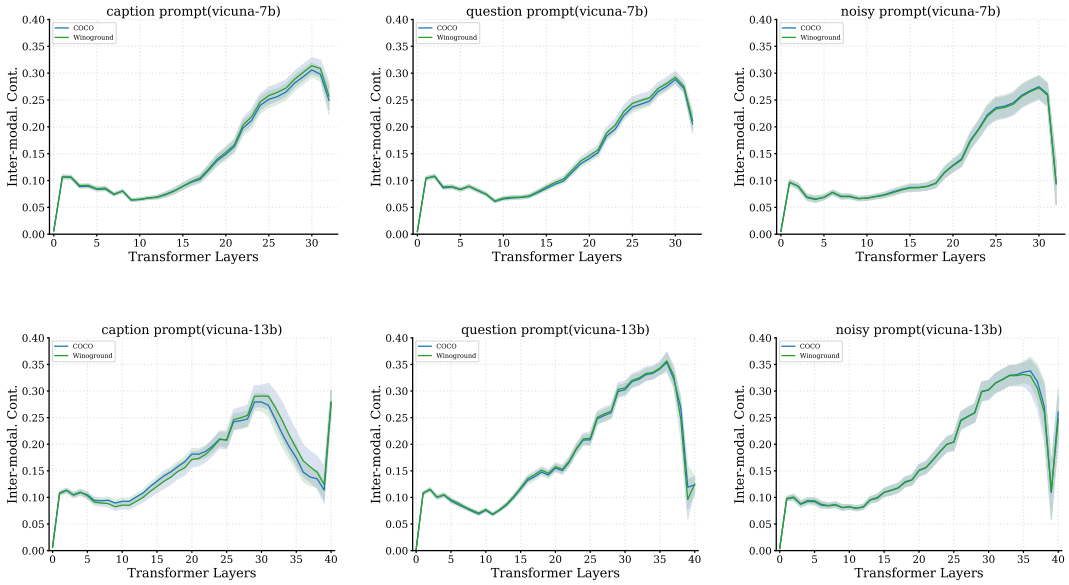

Figure 4.1: Alignment dynamics of visual token representation in InstructBLIP with two different LM decoder sizes

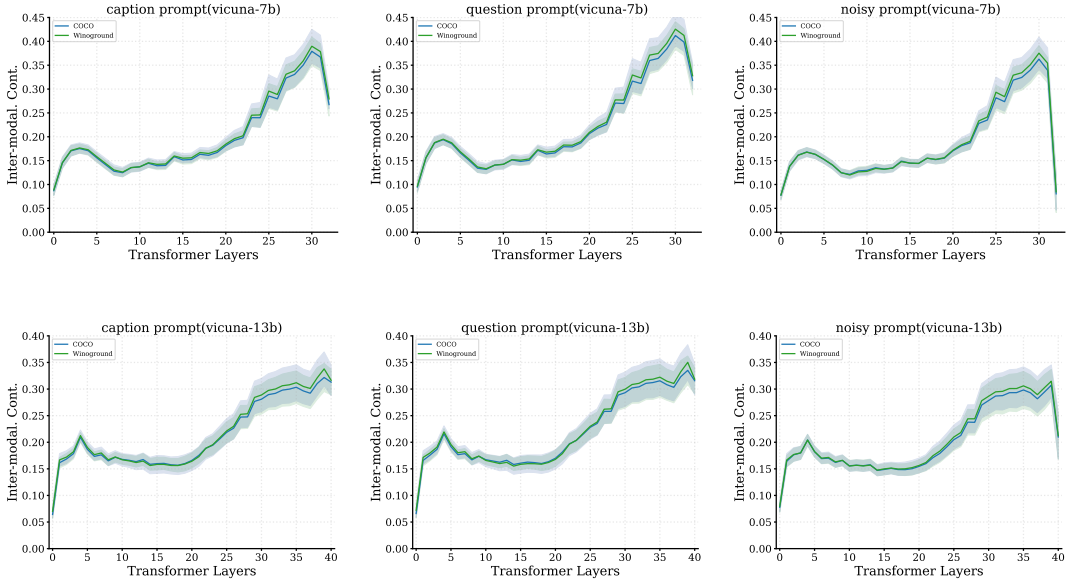

Figure 4.2: Alignment dynamics of visual token representation in LLaVA-1.5 with two different LM decoder sizes

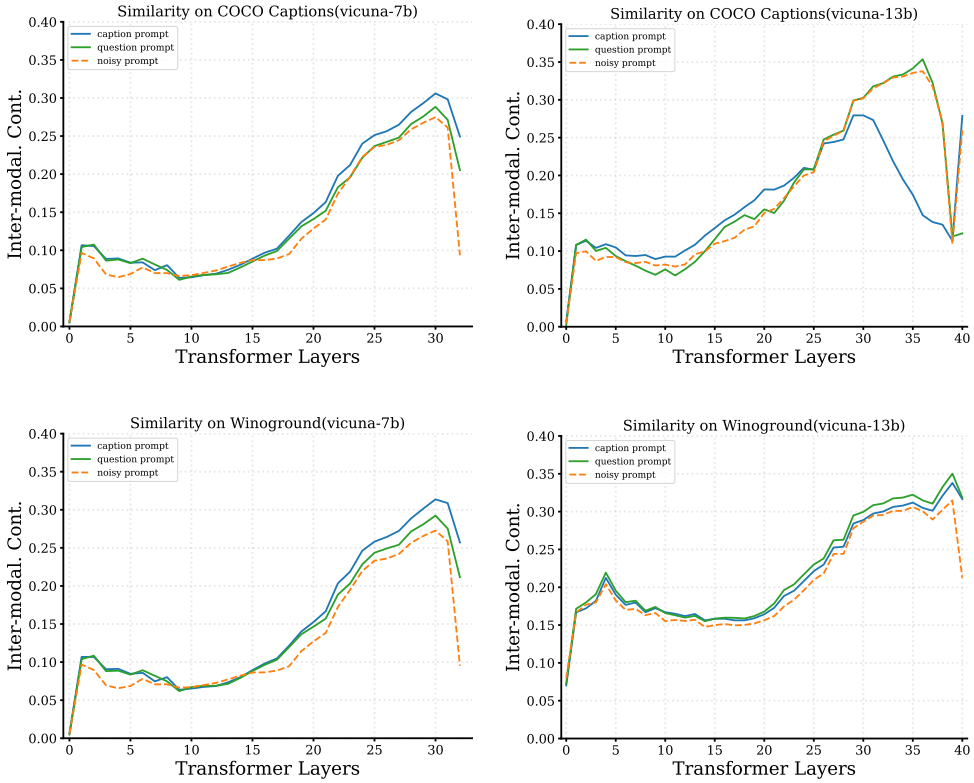

Figure 4.3: Alignment dynamics of visual token representation in InstructBLIP models under different prompts

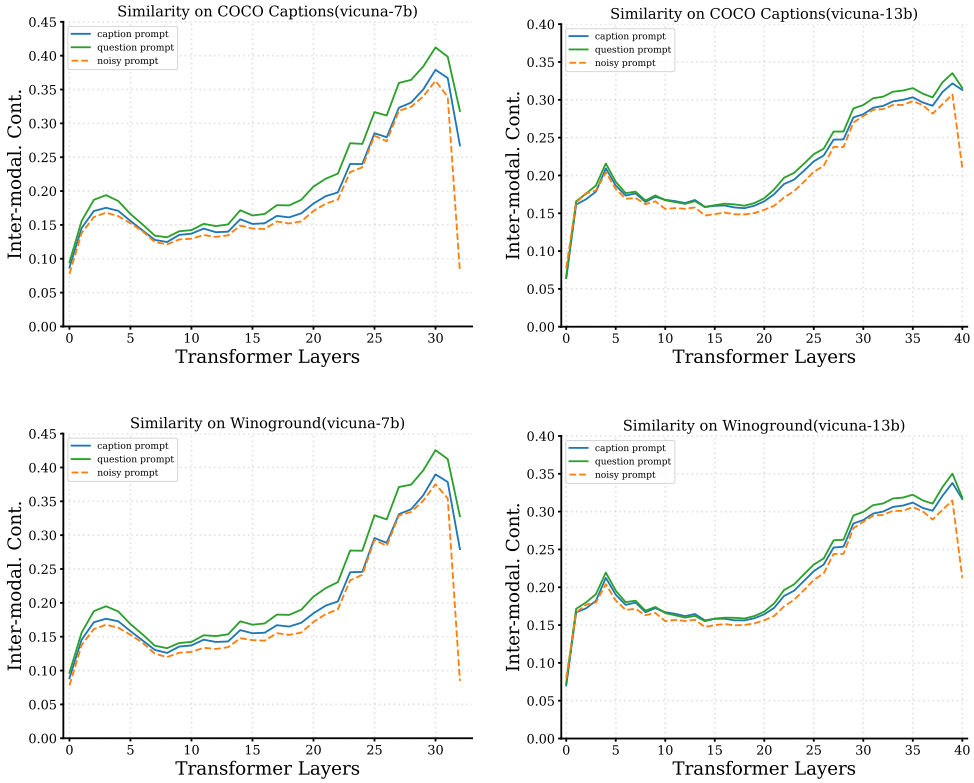

Figure 4.4: Alignment dynamics of visual token representation in LLaVA-1.5 models under different prompts

4.1.2 Results

The results of our similarity experiments are shown in Fig. 4.1 for InstructBLIP(Vicuna-7B), InstructBLIP(Vicuna-13B) and Fig. 4.2 for LLaVA-1.5(Vicuna-7B), LLaVA-1.5(Vicuna-13B).

Combining the results from InstructBLIP and LLaVA-1.5 models, we observe that the patterns in multi-modal similarity remain consistent across all settings, regardless of the different sizes of LLM components and datasets.

In general, all models exhibit an upward trend as expected; meanwhile, three distinct intervals are observed, demonstrating the alignment dynamics of visual token representation in language model representation space. In early layers (below layer 5), a small initial peak is observed, indicating an early-stage alignment between the two modalities. In mid-to-late layers (layer 10 to layer35), we observe a continuous rise in inter-modal similarity, suggesting a progressive interaction between image token representations and instruction token hidden states. In the last deep layers (after layer 35), it presents the global decline in inter-modal similarity, implying the model shifts its focus away from multimodal interaction.

Additionally, from Fig. 4.3 and Fig. 4.4, we find that the aforementioned trends remain consistent across different types of prompts. Moreover, even when using prompts that are considered meaningless, the overall trend persists. Although we compare three distinct types of prompts, i.e., caption, question, and noisy, the corresponding curves maintain the same general shape, with minor differences in peak magnitudes and slopes. These findings point to a notable prompt robustness within the evaluated VLLMs. In other words, the inter-modal interaction captured by cosine similarity remains relatively stable across varying linguistic inputs, including prompts that lack coherent semantics. Thus, the internal alignment dynamics of visual representation towards text token embeddings appear to be governed by the model's intrinsic mechanism rather than being invoked by input prompts.

In conclusion, our findings highlight the following insights:

1. Despite architectural differences in cross-modal projectors and LLMs' sizes, the inter-modal similarity curves follow a consistent three-stage trend, suggesting a universal alignment process where visual token representations gradually converge toward textual embeddings.
2. The increasing similarity values in deeper layers suggest that cross-modal alignment strengthens as information propagates through the LM decoder.
3. The layer-wise evolution of cross-modal similarity remains stable across various prompts, indicating that alignment is primarily governed by internal model mechanisms rather than input phrasing.

4.2 Visualization via Norm-based Attention

Norm-based Attention. The MHSA module is widely regarded as a pivotal mechanism for contextualizing intermediate representations in language models. Extensive research has been conducted to explore how this mechanism enables language models to acquire various linguistic capabilities. This mechanism computes a global update for input tokens by aggregating relevant information from a sequence of input vectors at the previous layer. This process involves two primary steps: first, attention weights are assigned to each input token; second, the input vectors are aggregated through a weighted summation based on these attention weights.

In the context of VLLMs, the input sequence to the language decoder is composed of a concatenation of image tokens and text tokens. This allows the model to utilize information from both modalities to generate attention-weighted outputs for subsequent computational steps. To better understand how multimodal information interacts within the text decoder, we propose employing attention analysis as an investigative tool. Given the challenges associated with the faithfulness of attention scores as an explanation [32–34], we adopt the norm-based attention approach proposed by [35]. This method leverages the norm of multi-head attention's output transformation to scale the attention score, enabling a more faithful investigation of the linguistic capabilities of the Transformer. By incorporating the magnitudes of transformed vectors, this norm-based attention analysis provides a more reliable interpretation of the contribution of the input vector to the final output.

Experiments and Results. To analyze how attention allocation between the two modalities changes across language model (LM) decoder layers in Vision Large Language Models (VLLMs), we conducted a detailed investigation. Specifically, we randomly selected 100 images each from the COCO and Winoground datasets and extracted norm-based attention results from two VLLMs: InstructBLIP (Vicuna-13B) and LLaVA-1.5 (Vicuna-13B). The attention heatmaps were then generated for visualization.

In particular, we focused on plotting the attention assignments from the final position token of the input prompt to all preceding tokens across the LM decoder layers. For qualitative analysis, we highlight the norm-based attention heatmaps for three images (id_200, id_237, id_323), as shown in Fig. 4.5. These heatmaps illustrate how the last text token distributes its norm-based attention over preceding tokens at different layers, revealing two rough trends where 1) progression of attention against the Transformer

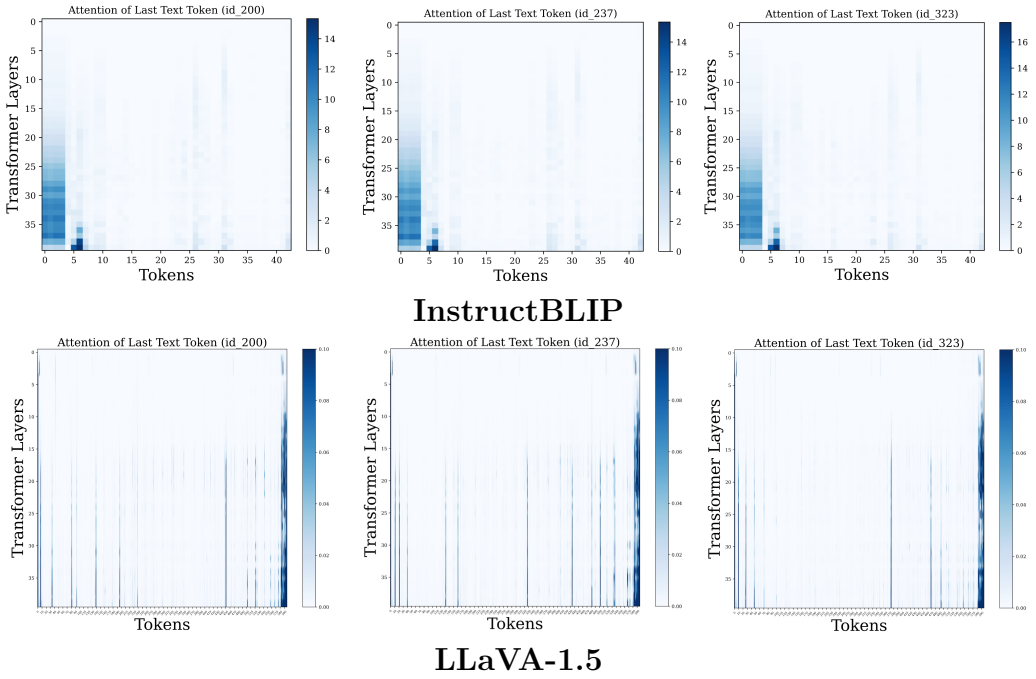

Figure 4.5: Qualitative analysis of norm-based attention results on InstructBLIP and LLaVA-1.5

blocks and 2) uneven distribution of attention assignment among individual tokens. This observation holds for both models despite their architectural differences. In the early layers, attention to image tokens tends to be diffuse and relatively weak, suggesting that the model has not yet fully integrated the visual information. However, we observe that attention allocation is accumulated as the model proceeds to the middle and deeper layers. Meanwhile, more focused attention is assigned to several specific tokens. The above two observed patterns visualize that crucial cross-modal interaction, i.e., attention from the last text token to image tokens, is likely to intensify in those mid-to-late layers. Moreover, the model tends to focus on particular visual patches while suppressing those deemed less relevant for the final textual prediction.

4.3 Application

The visualization of attention allocation during the model's forward pass demonstrates that the model assigns varying levels of attention to different image tokens, indicating that individual image tokens may play distinct roles

in the forward computation process. To further investigate this phenomenon, we conducted an ablation study here.

We employ a log-likelihood evaluation metric to quantify the impact of varying numbers of image tokens on the model loss during forward pass computation. This criterion computes the log-likelihood of a given sentence, measuring how effectively the model predicts the reference text. Specifically, during inference, we feed the ground-truth caption as part of the input into the text decoder of VLLMs. At each time step t, the model sees the true token x_i, and the forward pass calculates a vector of logits over the vocabulary for next-token prediction. The loss is then obtained by calculating the cross-entropy between the model's distribution of next-token prediction and that of ground-truth tokens. The resulting loss tells how well the model performs at predicting each token given the perfect preceding tokens. A small loss means the model assigns a high probability to the actual ground-truth token at each step; a large loss means the model's predictions deviate from the ground truth.

Experimental Details. We test two VLLMs: **1)** InstructBLIP (Vicuna-13B) and **2)** LLaVA-1.5 (Vicuna-13B), using different numbers of image tokens across multiple prompts that differ only in phrasing. For each forward pass computation, we provide the VLLMs with a combination of an image, an instruction, and a ground-truth caption. Prior to concatenating the image tokens and text tokens for input to the LM decoder, we truncate the image tokens, limiting the amount of visual information available to the model. Specifically, we directly truncate the sequence of image tokens and increase the number of image tokens by one for each forward pass setting. This leads to 32 settings for InstructBLIP and 576 settings for LLaVA-1.5, respectively. For each forward pass setting, we randomly select 1200 images from the COCO Captions dataset and calculate the average loss. Such a setting is repeated using five paraphrased prompts for both models.

Results. Fig. 4.6 illustrates how the forward loss changes as the number of image tokens increases, comparing InstructBLIP (above) and LLaVA-1.5 (bottom). We define the threshold as the mean loss across all forward pass settings spanning five prompts, representing the average performance across various ablation study configurations. Our extensive experiments reveal the following key findings:

1. Both models exhibit a nearly identical overall tendency: once the quantity of image tokens surpasses a certain threshold (6.68 for InstructBLIP and 2.86 for LLaVA-1.5), loss reduction goes slowly or even

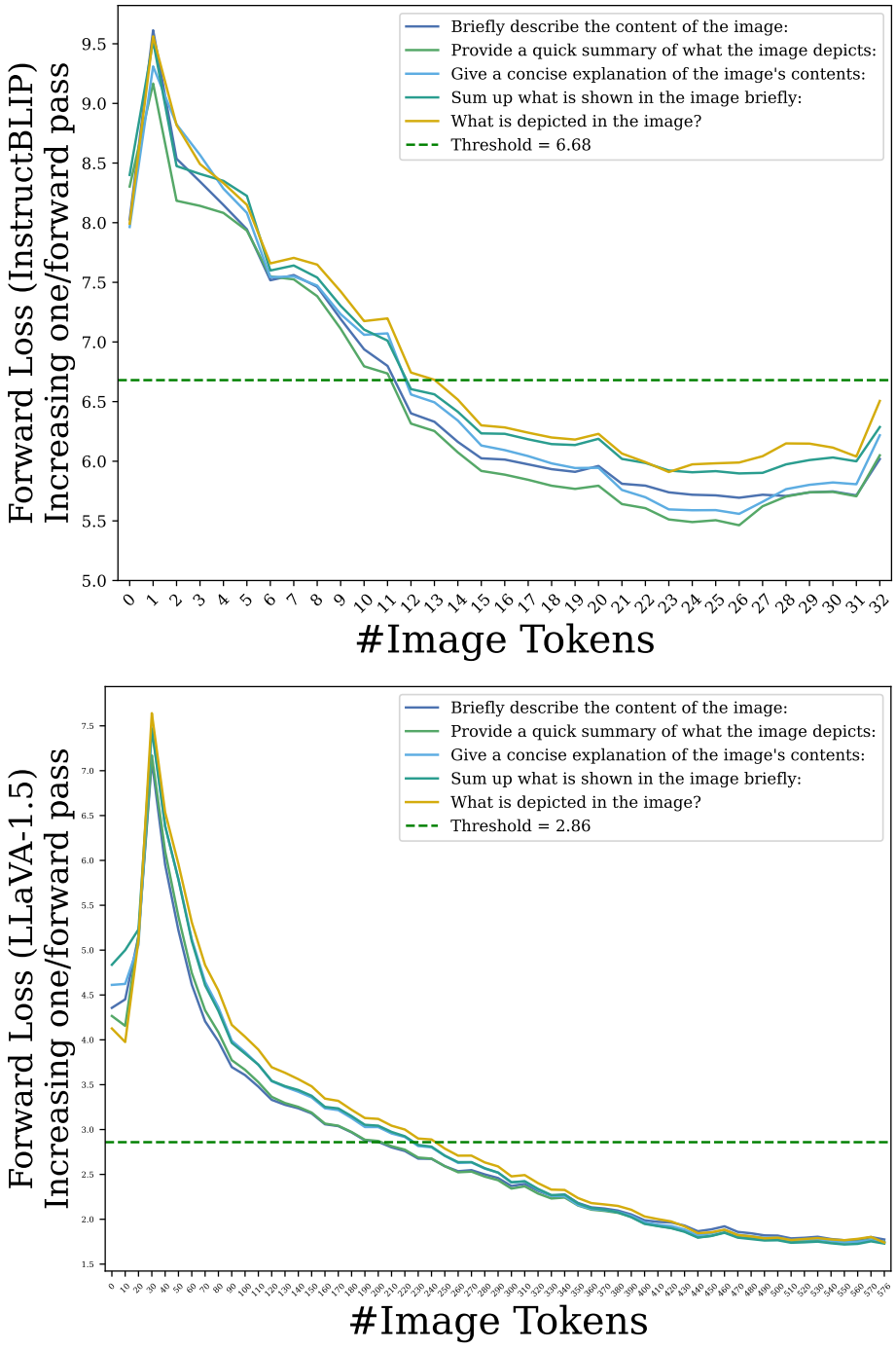

Figure 4.6: Impact on forward loss with varying numbers of image tokens in LLMs on InstructBLIP and LLaVA-1.5

stops, suggesting that image tokens in subsequent positions may carry minimal useful information.

2. Across different paraphrased prompts, the curves follow a similar downward trend, suggesting that VLLMs are robust to minor textual variations during inference.

Based on the above findings, we empirically reveal that during forward computation, not all visual tokens contribute equally to loss reduction. Notably, using only 40% of the visual tokens achieves 70% of the total loss reduction.

Additional Experiments. From Fig. 4.6, we observe a sharp spike in loss at very low image token counts (e.g., only the first position token seen for InstrctBLIP), suggesting image tokens in specific positions might play a more significant role in affecting loss computation compared to other positions. To examine the above hypothesis, we then design a controlled experiment where masks are applied to those non-trivial image tokens to observe the changes in forward pass loss. Specifically, we predefine masking intervals for several early image tokens and apply them during the model's forward pass. Same as before, we run forward computation using 1200 images for each masking interval setting.

Results. The results shown in Fig. 4.7 demonstrate the effect of masking image tokens at different positions on the model's forward computation loss. In detail, masking image tokens at early positions results in a greater loss reduction than masking tokens at other positions, suggesting that these tokens are non-trivial but have a negative impact.

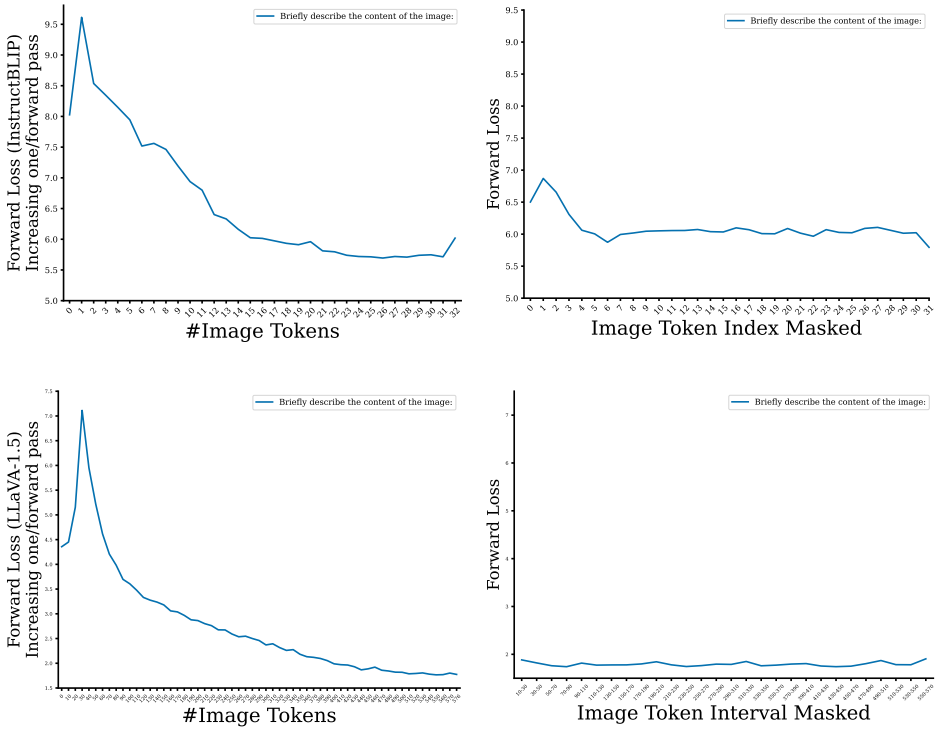

Figure 4.7: **Above**: Loss when masking one image token per forward pass on InstructBLIP. **Below**: Loss when masking non-trivial image token intervals per forward pass on LLaVA-1.5.

Chapter 5

Conclusion

This thesis investigates the interaction between the image token and text token, especially focusing on the evolution of image representations in modern VLLMs. Specifically, we systematically and quantitatively investigate how image representations evolve across Transfermer-based autoregressive LLMs in modern VLLMs.

Chapter 3 describes our exploration of interpreting the intermediate representation of visual tokens directly using human language, empirically revealing that a decoder trained merely on text data can nevertheless process image tokens into meaningful language words. Chapter 4 investigate the alignment dynamics of visual token representation towards text token embeddings along the layer of LLM decoders. Our extensive experiments reveal that a consistent three-stage trend in the alignment dynamics of visual representations holds universally, regardless of architectural differences in cross-modal projectors and LLM sizes. In addition, our findings on the invariance of inter-modal interaction trends across different types of prompts underscore the strong prompt robustness of VLLMs. Based on observations from attention analysis, Section 4.3 examines the relationship between model forward computation loss and the number of image tokens, aiming to provide valuable insight for balancing the effectiveness (lower loss) and efficiency (fewer image tokens) to enable inference acceleration. Our empirical analysis reveals that not all visual tokens contribute equally to loss reduction during forward computation.

Future Work. We empirically identified the existence of a consistent three-stage trend of multi-modal alignment during model inference, regardless of distinct designs of projectors, LM decoder size, and linguistic input. Building on this, we could hypothesize that such three-stage inference dynamics of VLLMs are inner-intrinsic rather than input-evoked. We leave this for future work.

Appendix A

Prompts

A.1 Normal Prompts

```
{
    "0": "Briefly describe the content of the image:",
    "1": "Provide a quick summary of what the image depicts:",
    "2": "Give a concise explanation of the image content:",
    "3": "Sum up what is shown in the image briefly:",
    "4": "What is depicted in the image?",
    "5": "USER: <image>\nWhat is the content of the image?
       ASSISTANT:",
    "6": "USER: <image>\nBriefly describe the content of the
       image. ASSISTANT:",
    "7": "USER: <image>\nProvide a quick summary of what the
       image depicts. ASSISTANT:",
    "8": "USER: <image>\nGive a concise explanation of the image
        content. ASSISTANT:",
    "9": "USER: <image>\nSum up what is shown in the image
       briefly. ASSISTANT:"
}
```

A.2 Noisy Prompts


```
[
    {
        "noisy_prompt": "env \u0432\u043d\u0435 \u010casyidense\
           u0434\u0438\u0438\u04d9 aqu Angnihkins CV pendant"
    },
    {
        "noisy_prompt": "ethe d\u00e9cco demolaqu allerdings\u
           044d\u0439ipo SET \u0434\u0435\u044f\u0442\u0435\u043
           b\u044ccile cin\u00e9ma Edinburgh"
    },
    {
        "noisy_prompt": "minipage ports\u0441\u043e\u043d\u0430
           infl open tantoconsole \u0434\u0435\u044f sua einer\u
           2500\u2500 street its"
    },
    {
        "noisy_prompt": "\u674eSQL refreshirmingham Seine\u00e1
           ntittel '_ tabstha — pode Package"
    },
    {
        "noisy_prompt": "Mel bleUnsadj]), pdf windows\u2153
           Einzelusch Bl n\u00fa efect"
    },
    {
        "noisy_prompt": "rad\u226b nations Bron notreares
           Finalieved converter south Ninica relation"
    },
    {
        "noisy_prompt": "ellschaft Peg ricdomain Et usesthrow
           demonstrated daughters Karriere \u0430\u043d\u0442\u
           0438 Ach region"
    },
    {
        "noisy_prompt": "IdcommandsmeisterschaftSocket \u00e4r
           GesellschaftITYetryIds gradleotic reli ownership"
    },
    {
        "noisy_prompt": "revision js \u041e sign keyboard Prime
           galaxies../\u041f\u0435\u0442\u0435\u0440 nat Prz ver
            removing"
    }
```

References

[1] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=1tZbq88f27

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

[3] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2024. [Online]. Available: https://arxiv.org/abs/2310.03744

[4] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[5] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: towards general-purpose vision-language models with instruction tuning," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2024.

[6] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, and F. Huang, "mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 13 040–13 051.

[7] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qian, J. Zhang, F. Huang, and J. Zhou, "mplug-owl: Modularization empowers large language models with multimodality," 2024. [Online]. Available: https://arxiv.org/abs/2304.14178

[8] S. Schwettmann, N. Chowdhury, S. Klein, D. Bau, and A. Torralba, "Multimodal neurons in pretrained text-only transformers," 2023. [Online]. Available: https://arxiv.org/abs/2308.01544

[9] G. Verma, M. Choi, K. Sharma, J. Watson-Daniels, S. Oh, and S. Kumar, "Cross-modal projection in multimodal llms doesn't really project visual attributes to textual space," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2024, pp. 657–664.

[10] S. Basu, M. Grayson, C. Morrison, B. Nushi, S. Feizi, and D. Massiceti, "Understanding information storage and transfer in multi-modal large language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.04236

[11] C. Neo, L. Ong, P. Torr, M. Geva, D. Krueger, and F. Barez, "Towards interpreting visual information processing in vision-language models," *arXiv preprint arXiv:2410.07149*, 2024.

[12] Z. Yu and S. Ananiadou, "Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering," 2025. [Online]. Available: https://arxiv.org/abs/2411.10950

[13] Z. Zhang, S. Yadav, F. Han, and E. Shutova, "Cross-modal information flow in multimodal large language models," 2024. [Online]. Available: https://arxiv.org/abs/2411.18620

[14] K. Xu, "Show, attend and tell: Neural image caption generation with visual attention," *arXiv preprint arXiv:1502.03044*, 2015.

[15] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," 2018. [Online]. Available: https://arxiv.org/abs/1707.07998

[16] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," 2018. [Online]. Available: https://arxiv.org/abs/1803.09845

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[18] A. Radford, "Improving language understanding by generative pre-training," 2018.

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

[20] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," 2023. [Online]. Available: https://arxiv.org/abs/2310.09478

[21] J. Merullo, L. Castricato, C. Eickhoff, and E. Pavlick, "Linearly mapping from image to text space," *arXiv preprint arXiv:2209.15162*, 2022.

[22] nostalgebraist, "logit lens on non-gpt2 models," https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2021.

[23] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023. [Online]. Available: https://arxiv.org/abs/2301.12597

[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[25] Vicuna. (2023) Vicuna: Fastchat. Accessed on March 6, 2023. [Online]. Available: https://github.com/lm-sys/FastChat

[26] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," 2015. [Online]. Available: https://arxiv.org/abs/1504.00325

[27] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[28] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[29] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross, "Winoground: Probing vision and language models for visio-linguistic compositionality," 2022. [Online]. Available: https://arxiv.org/abs/2204.03162

[30] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 358–19 369.

[31] K. Ethayarajh, "How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings," 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, pp. 55–65.

[32] K. Clark, "What does bert look at? an analysis of bert's attention," *arXiv preprint arXiv:1906.04341*, 2019.

[33] S. Serrano and N. A. Smith, "Is attention interpretable?" *arXiv preprint arXiv:1906.03731*, 2019.

[34] S. Jain and B. C. Wallace, "Attention is not explanation," 2019. [Online]. Available: https://arxiv.org/abs/1902.10186

[35] G. Kobayashi, T. Kuribayashi, S. Yokoi, and K. Inui, "Attention is not only a weight: Analyzing transformers with vector norms," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 7057–7075. [Online]. Available: https://aclanthology.org/2020.emnlp-main.574

[36] Z. Wu and M. Palmer, "Verb semantics and lexical selection," *arXiv preprint cmp-lg/9406033*, 1994.

[37] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[38] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "What does BERT with vision look at?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 5265–5275. [Online]. Available: https://aclanthology.org/2020.acl-main.469

[39] Y. Guan, J. Leng, C. Li, Q. Chen, and M. Guo, "How far does BERT look at: Distance-based clustering and analysis of BERT's attention," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 3853–3860. [Online]. Available: https://aclanthology.org/2020.coling-main.342

[40] H. Pan, Y. Cao, X. Wang, X. Yang, and M. Wang, "Finding and editing multi-modal neurons in pre-trained transformers," 2024. [Online]. Available: https://arxiv.org/abs/2311.07470

[41] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," 2019. [Online]. Available: https://arxiv.org/abs/1908.03557

[42] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," 2020. [Online]. Available: https://arxiv.org/abs/1908.08530

[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[44] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," 2021. [Online]. Available: https://arxiv.org/abs/2102.05918

[45] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019. [Online]. Available: https://arxiv.org/abs/1908.02265

[46] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," 2019. [Online]. Available: https://arxiv.org/abs/1908.07490

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[48] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

Publications

[1] Wei, H., Cho, H., Shi, Y. and Inoue, N., 2024. Phase Diagram of Vision Large Language Models Inference: A Perspective from Interaction across Image and Instruction. arXiv preprint arXiv:2411.00646.

[2] Houjing Wei, Hakaze Cho, Y Shi and Naoya Inoue. A Study on Multimodal Interaction in Vision Large Language Models. To appear in The Association for Natural Language Processing. 2025.