

BASIC: Boosting Visual Alignment with Intrinsic Refined Embeddings in Multimodal Large Language Models

Jianting Tang^{1,2}Yubo Wang^{1,2}Haoyu Cao^{1,2}Linli Xu^{1,2*}

¹University of Science and Technology of China, ²State Key Laboratory of Cognitive Intelligence
{jiantingtang, wyb123, caohaoyu}@mail.ustc.edu.cn, linlixu@ustc.edu.cn

Abstract

*Mainstream Multimodal Large Language Models (MLLMs) achieve visual understanding by using a vision projector to bridge well-pretrained vision encoders and large language models (LLMs). The inherent gap between visual and textual modalities makes the embeddings from the vision projector critical for visual comprehension. However, current alignment approaches treat visual embeddings as contextual cues and merely apply auto-regressive supervision to textual outputs, neglecting the necessity of introducing equivalent direct visual supervision, which hinders the potential finer alignment of visual embeddings. In this paper, based on our analysis of the refinement process of visual embeddings in the LLM’s shallow layers, we propose **BASIC**, a method that utilizes refined visual embeddings within the LLM as supervision to directly guide the projector in generating initial visual embeddings. Specifically, the guidance is conducted from two perspectives: (i) optimizing embedding directions by reducing angles between initial and supervisory embeddings in semantic space; (ii) improving semantic matching by minimizing disparities between the logit distributions of both visual embeddings. Without additional supervisory models or artificial annotations, BASIC significantly improves the performance of MLLMs across a wide range of benchmarks, demonstrating the effectiveness of our introduced direct visual supervision.*

1. Introduction

Multimodal Large Language Models (MLLMs) [12, 35, 57, 58, 63] have demonstrated impressive performance on tasks requiring strong visual perception and logical reasoning, marking a solid step towards general artificial intelligence (AGI). To efficiently construct high-performance MLLMs, a simple vision projector [10, 35, 36] is typically used to bridge well-pretrained powerful vision encoders [30, 41, 48,

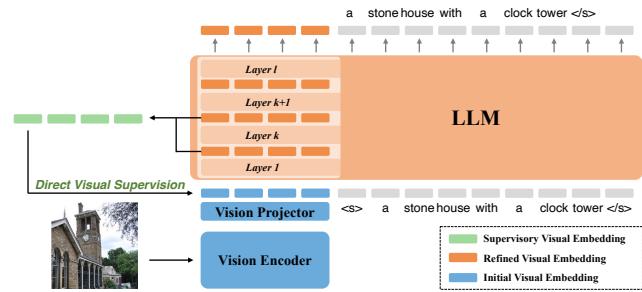


Figure 1. Overview of the proposed BASIC framework. Conventional MLLM training treats visual embeddings derived from the vision projector as contextual cues, only applying auto-regressive supervision to text tokens. Beyond that, BASIC leverages the refined visual embeddings from the LLM’s shallow layers to provide direct visual supervision to the initial visual embeddings.

68] and large language models (LLMs) [9, 13, 18, 61]. To fully leverage both models, it is crucial to effectively align the visual and textual modalities through the vision projector [32, 33, 43].

Current leading MLLMs, such as LLaVA [35], InternVL [12] and Qwen2-VL [57], employ a multistage training paradigm to achieve modality alignment. Generally, the early stages establish basic multimodal understanding using large-scale image-caption pairs, aligning the vision projector’s output with the LLM’s input embedding space. The subsequent stages refine this alignment by tuning on high-quality visual instruction-response data, developing the model’s capacity for specific visual tasks. However, due to the continuity of visual embeddings and the discreteness of text tokens, the current training approaches treat visual embeddings purely as contextual cues and apply auto-regressive supervision to text tokens, which implies lack of equivalent direct supervision for visual embeddings. The asymmetric supervision adapted from the training paradigm targeted for LLMs leads to two key problems. First, it fails to fully utilize the rich information present in visual data [60]. Second, it limits the model’s ability to achieve fine-grained alignment between visual and linguistic representations [45].

*Corresponding author.

Recently, alternative works such as Chameleon [54], SEED-LLaMA [20] and LaVIT [28] employ pre-trained image tokenizers to obtain discrete visual tokens, achieving unified auto-regressive modeling. However, while these approaches treat both modalities equally, the discretization process introduces significant information loss. Additionally, Emu1 [52] and Emu2 [53] use the ℓ_2 regression loss to encourage each continuous visual representation output by the LLM to directly fit the input value at the next position. Despite their impressive image generation ability benefiting from this unique design, they still lag behind mainstream MLLMs in visual comprehension. To enhance visual comprehension by introducing direct visual supervision, it is essential to ensure the quality of the constructed supervisory signals and the rationality of the optimization objective during the training process.

In this paper, we first analyze the visual perception process in the well-established MLLMs. Based on our findings, we propose BASIC, a novel approach that leverages refined visual embeddings within LLMs as supervisory signals to boost modality alignment in the input space, as illustrated in Figure 1.

Our analysis begins by examining how visual embeddings relate to textual concepts. For each initial visual embedding derived from the vision projector, we calculate its cosine similarity with all text token embeddings in the LLM’s vocabulary and visualize the most matching token. As shown in Figure 2, for certain visual embeddings, the most matching text tokens directly reflect attributes of corresponding image patches, such as color, shape, and object class. This partially reveals the internal visual perception mechanism of MLLMs: LLMs interpret the textual concepts within visual embeddings to understand images. However, there are still plenty of initial visual embeddings associated with irregular text tokens. Tracking these embeddings through the LLM’s layers, we observe that, in shallow layers, initially misaligned embeddings often gradually align with more meaningful text tokens, which should be attributed to the LLM’s strong semantic modeling capabilities by considering the visual context. Despite the gradual refinement of visual embeddings within the LLM, the visual embeddings seen by the *questions* in the early stages are inaccurate, which can lead to confusion in image understanding and impair the final *answers*. Therefore, it is crucial that initial visual embeddings exhibit high quality from the outset.

To this end, we construct direct visual supervision by weighted summation of refined visual embeddings from the shallow layers. Firstly, we optimize the directional alignment between initial and supervisory visual embeddings by minimizing their angular distances in semantic space. Secondly, to ensure the consistency of semantic distribution, we compute logit distributions across the entire vocabulary

for both initial and supervisory visual embeddings, and then minimize their KL divergence. Our approach ensures the visual embeddings maintain consistent high quality in the LLM’s shallow layers, enabling the *questions* to acquire accurate image information at an early stage.

Notably, our method exhibits the following advantages. Firstly, it does not rely on additional supervisory models or artificial annotations, thus saving the resource overhead. Furthermore, it is generally applicable to a wide range of MLLMs adopting the vision encoder-vision projector-LLM architecture. Comprehensive experiments demonstrate the effectiveness of our introduced direct visual supervision.

In summary, our contributions are as follows:

- We systematically analyze the association between the visual embeddings within different LLM layers and the text token embeddings, which provides valuable insights into the internal visual perception mechanism of MLLMs.
- We propose BASIC, an effective direct visual supervision method that leverages the LLM’s internal refined visual embeddings to guide initial visual embeddings in the input space from two perspectives: the directional alignment and semantic distribution.
- BASIC notably improves the performance of a series of MLLMs across a wide range of benchmarks, demonstrating its applicability and robustness.

2. Related Work

2.1. Multimodal Large Language Models

At present, most prevailing MLLMs, such as LLaVA [35], InternVL [12] and Qwen2-VL [57], adopt a vision encoder-vision projector-LLM architecture. The vision projector [2, 26, 32, 36] is responsible for mapping the image features encoded by the vision encoders into the LLM’s input embedding space. This simple approach broadens the LLM’s comprehension to images. However, the approach of independently training and then grafting inherently leads to the difficulty of modality alignment. Recently, another technical route [20, 28, 54] employs an image tokenizer [4, 46, 56, 64] to obtain discrete visual tokens and conducts unified auto-regressive modeling to create native MLLMs. However, it requires extensive training to converge, and the discretization process leads to substantial visual information loss. Therefore, this paper follows the first technical route, and utilize refined visual embeddings within the LLM to construct additional direct visual supervision to boost modality alignment.

2.2. Mechanistic Interpretability

Understanding the internal visual perception mechanism of MLLMs is crucial for constructing effective direct visual supervision. Mechanistic interpretability [7, 15, 42] aims to uncover the internal mechanisms that drive the input-

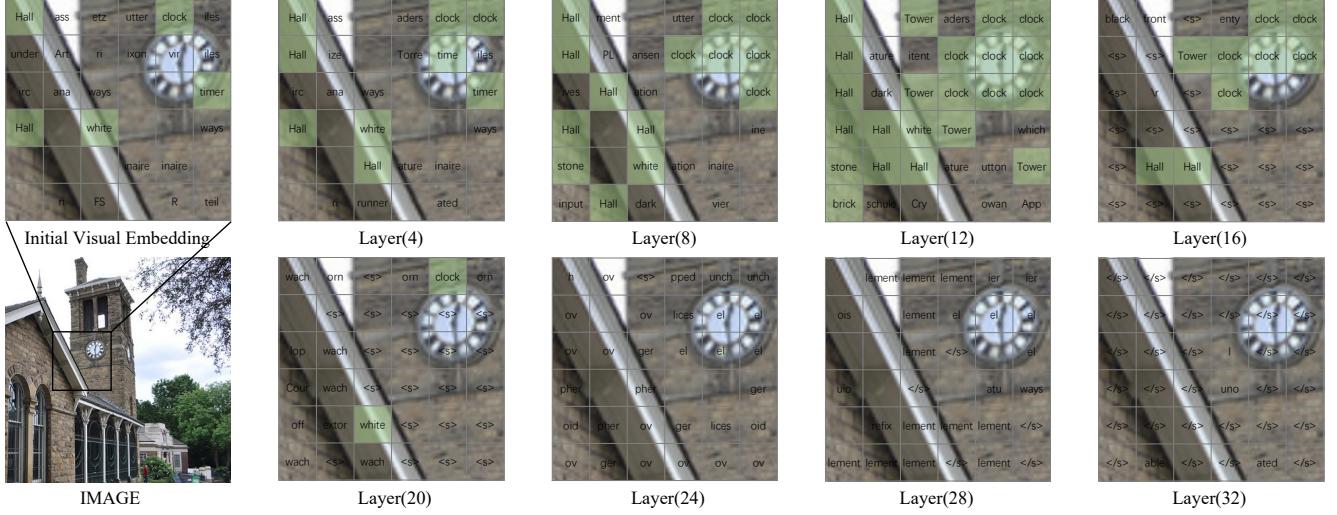


Figure 2. Visualization of the closest matching text token for each visual embedding of the LLM. Green patches indicate semantically meaningful matches. The initial visual embeddings are derived from the vision projector and have not yet entered the LLM. Layer(\cdot) indicates the matching results of the visual embeddings from the corresponding LLM layer. There are two notable patterns: (i) in the shallow layers of the LLM, visual embeddings that initially correspond to irregular tokens gradually align with more meaningful tokens; (ii) in the deep layers of the LLM, visual embeddings tend to correspond with the special end token $</s>$.

output transformation. Sparse autoencoder(SAE) based methods [8, 16, 19, 67] utilize the representation reconstruction and sparsification to facilitate discovering semantic features in sparse representations. Logit lens based methods [6, 34, 59, 66] use the language model head to project hidden states to interpret the prediction process. Currently, most interpretability studies mainly focus on LLMs, which only involves the text modality. The systematic analysis of visual perception process within MLLMs remains a rather unexplored field.

2.3. Self-Distillation

Self-distillation [40, 70, 71] is a unique instance of knowledge distillation [14, 21, 44, 72] where the teacher and student models share the same architecture. The network leverages its internally learned knowledge to guide its own training. Some works [50, 62] employ models updated at earlier steps as teachers for the current step, facilitating knowledge transfer across the temporal dimension. Some studies [11, 24, 47, 69] divide the model into different parts and use deeper blocks as teachers for shallower blocks, achieving knowledge transfer within the spatial dimension. From this perspective, utilizing the refined visual embeddings within LLM’s shallow layers to guide the vision projector in generating better-aligned initial visual embeddings can be regarded as a form of self-distillation.

3. Visual Perception Process Analysis

3.1. Preliminary

MLLMs typically comprise three components: a vision encoder $F_v(\cdot)$, a vision projector $F_p(\cdot)$, and an LLM $F_t(\cdot)$.

Specifically, the vision encoder $F_v(\cdot)$ extracts visual features from raw images. The vision projector $F_p(\cdot)$ transforms these image features into initial visual embeddings $V \in \mathbb{R}^{m \times d}$, where m denotes the number of image patches and d is the embedding dimension. The textual inputs are tokenized into token ids, which are then used to retrieve the corresponding token embeddings from the LLM’s embedding layer $E \in \mathbb{R}^{N \times d}$, where N represents the vocabulary size. The resulting textual embeddings are denoted as $T \in \mathbb{R}^{n \times d}$, where n denotes the number of text tokens. The LLM $F_t(\cdot)$ is responsible for processing the initial visual embeddings V and textual embeddings T to generate the final output. This can be formalized as:

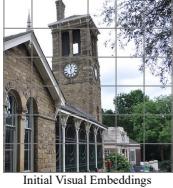
$$V = F_p(F_v(\text{image})), \quad T = E(\text{text}); \quad \text{output} = F_t(V, T). \quad (1)$$

In the forward propagation process, the LLM gradually refines the initial visual embeddings V , and the resulting hidden states are referred as refined visual embeddings $\tilde{V} \in \mathbb{R}^{l \times m \times d}$, where l represents the number of LLM layers.

3.2. Similarity-Based Analysis

To construct effective direct supervision for each visual embedding, we first analyze the visual perception process within MLLMs. For each initial visual embedding v_i derived from the vision projector, the most semantically matching text token embedding e_i can be obtained as:

$$e_i = \arg \max_j \frac{v_i \cdot e_j}{\|v_i\|_2 \|e_j\|_2} \quad (2)$$



Initial Visual Embeddings

The image features a large brick clock tower with a clock on each of its sides. The clock tower is situated in front of a building, possibly a church, and is surrounded by trees. The clocks on the tower are visible from different angles, making it a prominent and eye-catching structure. There are two people in the scene, one standing closer to the left side of the image and the other near the right side. They appear to be observing the clock tower or the building it is in front of.

....
....	Tower	tree
....	clock
....	Hall	tree
....	Hall	stone
....
....

Replaced Token Embeddings

The image depicts a large building with a clock tower prominently visible on its side. The clock is situated near the top of the tower, making it easily visible to those passing by. The building appears to be a mix of stone and concrete, giving it a sturdy and impressive appearance. The clock tower stands out as a focal point of the building, drawing attention to its presence.

Figure 3. Visual embeddings from the vision projector are replaced with their closest matching text token embeddings in the LLM’s vocabulary. The model is then prompted to generate descriptions. The generated descriptions (bottom) demonstrate strong semantic consistency with the raw image content.

where $0 \leq j < N$, $v_i \in V$ and $e_j \in E$. As illustrated in Figure 2 (upper left), for certain initial visual embeddings, the most similar tokens directly reflect specific attributes of the corresponding image patches, such as *clock* expressing the class of object, and *white* reflecting the color of object. Accordingly, we replace initial visual embeddings $[v_1, v_2, \dots, v_m]$ with the most matched token embeddings $[e_1, e_2, \dots, e_m]$, and prompt the LLM to generate image descriptions based on these replaced token embeddings. As shown in Figure 3, the descriptions are highly consistent with the content of the raw image. This indicates that LLMs might interpret the textual concepts within visual embeddings to understand images, and it also inspires us that direct visual supervision should guide the initial visual embeddings to establish more accurate text token associations.

Furthermore, to fully reveal the processing path of visual information within the LLM, we apply the same operations on all refined visual embeddings from different layers of the LLM. As illustrated in Figure 2, there are significant pattern differences between the LLM’s shallow and deep layers. In the LLM’s shallow layers, plenty of visual embeddings that initially correspond to irregular tokens will gradually align with more meaningful tokens. In the LLM’s deep layers, visual embeddings tend to correspond with the special end token $\langle /s \rangle$. Existing interpretability analyses targeted for LLMs [34, 59, 66] have shown that, when processing text, the shallow layers mainly focus on constructing better textual semantic representations by considering the context, while the deep layers concentrate on predicting the next token. In terms of visual embeddings, this enhanced semantic representations in the LLM’s shallow layers offer an intuitive presentation, namely a better relevance of associated text tokens. Meanwhile, due to the lack of token-level labels for visual inputs, the LLM’s deep layers tend to predict $\langle /s \rangle$ for each visual embedding to terminate the output.

Despite the refinement of visual embeddings in the shallow layers, the *questions* perceive plenty of inaccurate visual embeddings in the early stages, which would cause significant confusion in image understanding. Therefore, we leverage the refined visual embeddings from the LLM’s shallow layers to guide the initial visual embeddings, boosting vision-language alignment from the early stage.

4. Methodology

In the general training process of MLLMs, the auto-regressive supervision is performed on all input text tokens, and the training loss can be expressed as:

$$\mathcal{L}_{lm} = - \sum_{i=1}^n \log p(t_i | V, t_{<i}) \quad (3)$$

4.1. Supervisory Visual Embedding

This section details the construction of supervisory visual embeddings. In order to comprehensively utilize refined visual embeddings from the LLM’s shallow layers to provide robust direct visual supervision, we assign weight w_i for each LLM layer and obtain the final supervisory visual embedding \hat{V} :

$$\hat{V} = \sum_{i=1}^k w_i \tilde{V}_i; \quad w_i = \frac{i^2}{\sum_{i=1}^k i^2} \quad (4)$$

where $\tilde{V}_i \in \mathbb{R}^{m \times d}$, $\hat{V} \in \mathbb{R}^{m \times d}$, $\sum w_i = 1$ and we use the refined visual embeddings from layer $1 \sim k$. Considering the refining effect of LLM on visual embeddings, w_i is set to increase quadratically with the layer number i .

In each data sample, image patches contribute unequally to the model’s text outputs, with more important image patches having a greater impact on the outputs. Therefore, we apply stronger supervisions to the initial visual embeddings from more important patches. Specifically, we use the attention scores from the text part across the LLM’s layers to measure the importance of i th image patch:

$$a_i = \frac{1}{kn} \sum_{h=1}^k \sum_{j=1}^n a_{h,j,i}; \quad a_i = \frac{a_i}{\sum_{i=1}^m a_i} \quad (5)$$

where $a_{h,j,i}$ denotes the attention score from the j th text token to i th image patch in the LLM’s h th layer. Finally, a_i denotes the mean attention score on the i th image patch and serves as the degree of supervision.

4.2. Directional Alignment Supervision

In the embedding space of LLMs, semantically similar embeddings could be measured by the cosine similarity[34, 59]. The cosine similarity essentially characterizes the angular relationship between embeddings. Therefore, we

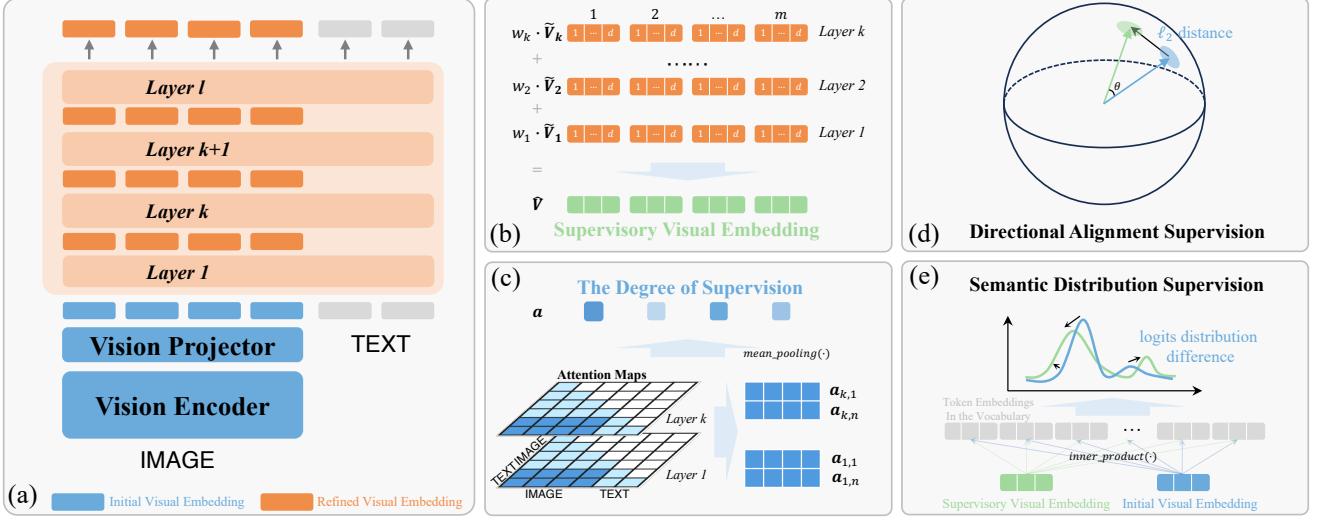


Figure 4. The construction of supervisory visual embeddings and two optimization objectives. (a) **Embedding Refinement**: The LLM processes the initial visual embeddings derived from the vision projector, generating refined visual embeddings at each layer. (b) **Supervisory Signal Construction**: Refined visual embeddings \tilde{V}_i from layer 1 to k are weighted averaged by w_i to serve as the supervisory visual embedding. (c) **Attention-based Supervision Degree**: Attention scores from text tokens to visual embeddings are mean-pooled to serve as the degree of supervision on each visual embedding. (d) **Directional Alignment Supervision**: Initial and supervisory embeddings are aligned by narrowing the angle θ on the unit hypersphere. (e) **Semantic Distribution Supervision**: Logit distributions are computed by projecting both initial and supervisory embeddings against the LLM’s vocabulary, then aligned by the minimizing KL divergence .

guide the initial visual embeddings V to align with the direction of supervisory visual embeddings \hat{V} to improve the semantic of V . Specifically, we first normalize the embeddings to eliminate magnitude effects and then reduce the angle by minimizing their ℓ_2 distance on the unit hypersphere:

$$\mathcal{L}_{das} = \sum_{i=1}^m a_i \left\| \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2} - \frac{\hat{\mathbf{v}}_i}{\|\hat{\mathbf{v}}_i\|_2} \right\|_2^2 \quad (6)$$

where $\mathbf{v}_i \in \mathbb{R}^d$, $\hat{\mathbf{v}}_i \in \mathbb{R}^d$, and a_i is used to control the degree of supervision.

4.3. Semantic Distribution Supervision

To analyze the association between visual and token embeddings, we compute the inner product of the visual embedding with the entire LLM vocabulary, yielding an interpretable logits vector. Each dimension of this logits vector reflects the semantic association between the visual embedding and the corresponding token, with higher values indicating tighter associations. The logits vector comprehensively characterizes the global semantic distribution of each visual embedding on the whole vocabulary. Therefore, we utilize the supervisory visual embeddings to guide the initial visual embeddings to learn better textual associations. Specifically, we first compute the logits vectors P between the supervisory visual embedding \hat{V} and token embeddings E , as well as the logits vectors Q between the initial visual embeddings V and token embeddings E . We then minimize

the KL divergence between both logits vectors to match the semantic distribution:

$$P = \text{softmax}(\hat{V}E^\top), \quad Q = \text{softmax}(VE^\top) \quad (7)$$

$$\mathcal{L}_{sds} = \sum_{i=1}^m a_i \text{KL}(\mathbf{p}_i \parallel \mathbf{q}_i)$$

where $P, Q \in \mathbb{R}^{m \times N}$, $\mathbf{p}_i, \mathbf{q}_i \in \mathbb{R}^N$, m represents the number of image patches and N represents the vocabulary size. Overall, the total training loss used in the pre-training stage and instruction-tuning stage is:

$$\mathcal{L} = \mathcal{L}_{lm} + \lambda_1 \mathcal{L}_{das} + \lambda_2 \mathcal{L}_{sds} \quad (8)$$

where λ_1 and λ_2 are used to balance different losses.

5. Experiments

5.1. Experimental Setup

Model Settings. In main experiments, we adopt CLIP-ViT-L/14-336px [48] as the vision encoder, a two-layer MLP with GeLU activation function as the vision projector and Vicuna-v1.5 [13] as the LLM.

Data and Training Details. In the pre-training stage, only the vision projector is trainable. The dataset used is LLaVA-1.5-558k [35], composed of image and caption pairs. The training epoch is 1, with a batch size of 256. We utilize the weighted summation of refined visual embeddings from the LLM’s shallow layers (specifically, layers 1 ~ 16/32 of the Vicuna-v1.5-7B and layers 1 ~ 20/40

Method	LLM	Res.	VQA ^{v2}	GQA	SQA ^I	VQA ^T	MMB ^{EN}	MMB ^{CN}	MM-Vet	VizWiz
Models using 7B LLM										
Fuyu [5]	Fuyu-8B	-	74.2	-	-	-	10.7	-	21.4	35.9
LaVIT-v2 [28]	LLaMA2-7B	224	68.2	48.0	-	-	-	-	-	41.0
IDEFICS [31]	LLaMA-7B	224	50.9	38.4	-	25.9	48.2	25.2	-	35.5
InstructBLIP [17]	Vicuna-7B	224	-	49.2	60.5	50.1	36.0	23.7	26.2	34.5
Qwen-VL-Chat [3]	Qwen-7B	224	78.2	57.5	68.2	61.5	60.6	56.7	-	38.9
VW-LMM [45]	Vicuna-7B	336	78.9	62.7	68.1	57.6	65.9	59.8	31.3	48.3
LLaVA-1.5 [35]	Vicuna-7B	336	78.5	62.0	66.8	58.2	64.3	58.3	31.1	50.0
BASIC	Vicuna-7B	336	79.2_{±0.7}	63.5_{±1.5}	70.6_{±3.8}	58.0_{±0.2}	68.8_{±4.5}	62.1_{±3.8}	33.8_{±2.7}	52.5_{±2.5}
Models using 13B LLM										
Emu-I [52]	LLaMA-13B	224	62.0	46.0	-	-	-	-	36.3	38.3
BLIP-2 [32]	Vicuna-13B	224	65.0	41	61	42.5	-	-	22.4	19.6
InstructBLIP [17]	Vicuna-13B	224	-	49.5	63.1	50.7	-	-	25.6	33.4
LLaVA-1.5 [35]	Vicuna-13B	336	80.0	63.3	71.6	61.3	67.7	63.6	36.1	53.6
BASIC	Vicuna-13B	336	80.6_{±0.6}	64.6_{±1.3}	73.1_{±1.5}	61.0_{±0.3}	69.6_{±1.9}	64.9_{±1.3}	37.2_{±1.1}	55.8_{±2.2}

Table 1. Comparison with leading representative MLLMs on 8 popular benchmarks. Res. represents the image resolution of vision encoder. The **best results** are highlighted. Fuyu [5] discards the vision encoder and merely relies on the LLM to model both images and text. LaVIT-v2 [28] discretizes the image using a pre-trained image tokenizer. Emu1 [52] drives each continuous visual representation outputted by the LLM to fit the input value at the next position.

of the Vicuna-v1.5-13B) as the supervisory visual embedding, with a quadratically increasing weighting coefficient w_i . The loss coefficients λ_1 and λ_2 are set to 1 and 0.01 respectively. The learning rate is 1e-3.

In the instruction-tuning stage, both the vision projector and LLM are trainable. The multi-modal instruction dataset adopted is LLaVA-1.5-mix-665k [35], comprising visual instruction-response pairs from various sources including VQAv2 [22], ShareGPT [49], RefCOCO[29, 39] and others. We finetune BASIC for 1 epoch with a batch size of 128. The learning rate is 2e-5. All other settings remain the same as in the previous stage.

Evaluations. To thoroughly assess the effectiveness of our method, we conduct evaluations across a wide range of benchmarks. This includes four popular general VQA benchmarks: VQA-v2 [22], GQA [25], ScienceQA-IMG [38] and TextVQA [51]. In addition, we adopt four benchmarks specifically targeting MLLMs and involving more comprehensive ability assessments: MMBench [37], MMBench-CN [37], MM-Vet [65] and VisWiz [23].

5.2. Main Results

As shown in Table 1, BASIC achieves superior results on various benchmarks. Notably, BASIC utilizes the same model settings and training data as LLaVA-1.5 [35], the obviously improved performance highlighting the effectiveness of our method in promoting modality alignment and this basic factor benefiting a wide range of benchmarks. Additionally, we notice that BASIC’s performance on VQA^T [51] slightly declines. This benchmark primarily assesses the model’s textual content recognizing ability, with questions generally like “*What is the year on the calendar?*” and “*What’s the letter next to the z on the ma-*

chine?”. As the supervisory signal from the LLM’s shallow layers focuses more on “semantic concepts”, it might blur the abstract textual information within initial visual embeddings and affect scenarios with tiny text. To understand the reason behind the performance differences between BASIC and LLaVA [35], we visualize the closest matching text token for each initial visual embedding, respectively. As illustrated in Figure 5, the initial visual embeddings generated by BASIC have improved semantic and associate with more meaningful tokens. For an quantitative evaluation, we randomly select 30 images from the MS COCO dataset and have 2 master students count the number of “meaningful” initial visual embeddings in LLaVA and BASIC, respectively. The results show that BASIC outperforms in 100% images, increasing the average ratio of meaningful embeddings from 74/576 to 217/576. The improved initial visual embeddings ensure the consistent high quality of visual embeddings in the LLM’s shallow layers, allowing the *questions* perceive accurate visual embeddings from the outset and reducing the confusion in image understanding.

Fuyu [5] completely discards the well-trained vision encoder and directly uses a simple linear layer to transform image patches into input embeddings, entirely relying on the LLM to model both images and text. LaVIT-v2 [28] employs an image tokenizer to encode images into discrete visual tokens, thus enabling unified auto-regressive modeling. These works represent initial efforts to develop native MLLMs, but they still significantly lag behind those adopting a vision encoder-vision projector-LLM architecture in image comprehension. VW-LLM [45] introduces an additional VM-head on the basis of LLaVA to provide auto-regressive supervision for the continuous image fea-

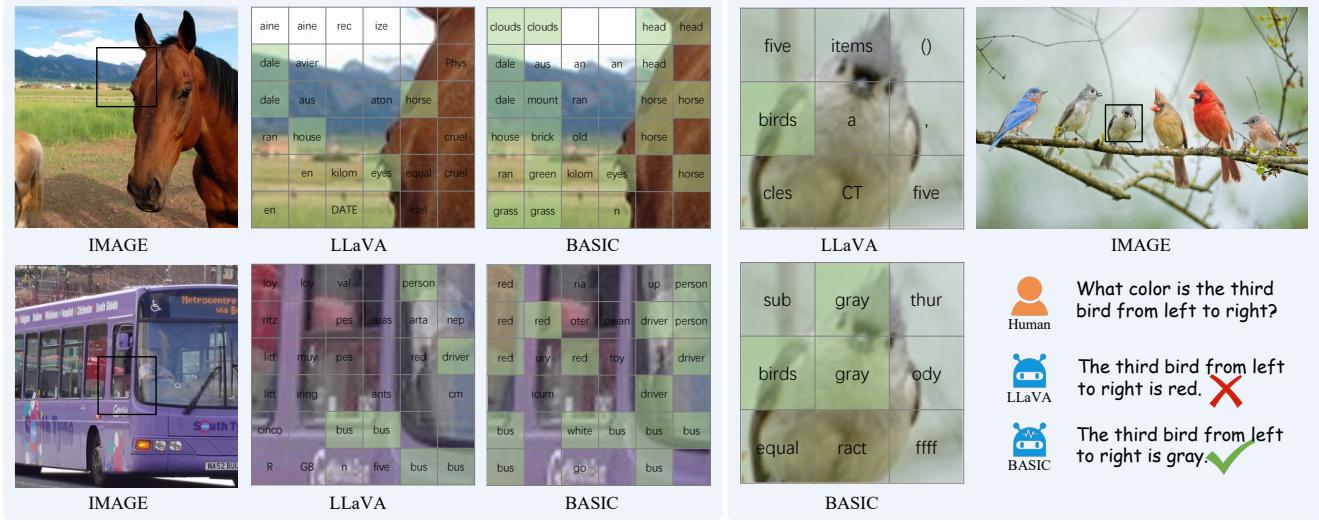


Figure 5. Comparisons between LLaVA [35] and BASIC in comprehending images. (Left) The respective closest matching text token for each initial visual embedding. (Right) More reasonable textual associations help BASIC solve visual comprehension tasks.

Method	\mathcal{L}_{das}	\mathcal{L}_{sds}	VQA ^{v2}	GQA	SQA ^I	VQA ^T	MMB ^{EN}	MMB ^{CN}	MM-Vet	VizWiz
BASIC-7B	✗	✗	78.5	62.0	66.8	58.2	64.3	58.3	31.1	50.0
	✓	✗	78.9	63.0	68.5	57.7	68.6	61.4	33.1	51.4
	✗	✓	79.1	63.3	68.1	57.6	68.0	60.6	32.5	51.2
	✓	✓	79.2	63.5	70.6	58.0	68.8	62.1	33.8	52.5

Table 2. Contributions of each supervisory loss to the performance of MLLM. Results demonstrate the effectiveness of both losses.

tures outputted by the LLM. However, the requirement for specialized training for the VM-head finally resulting in a complex four-stage training pipeline.

5.3. Ablation Studies

Analysis of Supervisory Losses. To assess the contributions of two proposed optimization objectives respectively, we compare the performance of models obtained under three different training settings: using only \mathcal{L}_{das} , only \mathcal{L}_{sds} , and a combination of both. The experiments are conducted on BASIC-7B, composed of CLIP-ViT-L/14-336px [48] and Vicuna-7B [13]. The training process is consistent with the main experiment, first pre-training on the LLaVA 1.5-558k dataset, and then instruction-tuning on the LLaVA 1.5-mix-665k dataset. As shown in Table 2, when \mathcal{L}_{das} and \mathcal{L}_{sds} are not used, BASIC represents the original LLaVA [35]. The additional introduction of either \mathcal{L}_{das} or \mathcal{L}_{sds} can effectively enhance the model’s performance across various benchmarks, demonstrating the effectiveness of guiding initial visual embeddings using supervisory visual embeddings from two distinct perspectives: the directional alignment and the semantic distribution. Additionally, the simultaneous utilization of both losses can further enhance the model’s performance.

Analysis of Supervisory Visual Embeddings. We utilize refined visual embeddings in the LLM’s shallow layers to establish the direct visual supervision. We analyze three key decision considerations in constructing the supervisory visual embeddings: (i) the selection of refined visual embeddings; (ii) the integration method of these refined visual embeddings; (iii) the degree of supervision on each initial visual embedding. The experiments are conducted using BASIC-7B, with the involved LLM Vicuna-7B [13] consisting of 32 layers. We utilize 10% of LLaVA-1.5-mix-665k [35] in the instruction-tuning stage.

(i) *The Selection of Refined Visual Embeddings.* We explore constructing the supervisory visual embeddings using refined visual embeddings from multiple layers (specifically from the 1 ~ 4th, 1 ~ 8th, ..., and 1 ~ 32th layers). As illustrated in Figure 6, when more refined visual embeddings from the shallow layers (generally 1 ~ 16th layers) are adopted, the model performance can gradually improve. However, introducing embeddings from deep layers (generally 16 ~ 32th layers) degrades the performance. This implies the pattern difference across the LLM layers. Generally, using refined visual embeddings from the LLM’s lower half can construct an appropriate supervision.

(ii) *The Integration of Refined Visual Embeddings.* In

Method	VE	LLM	PT+IT	VQA ^{v2}	GQA	SQA ^I	VQA ^T	MMB ^{EN}	MMB ^{CN}	MM-Vet	VizWiz
LLaVA BASIC	CLIP-L [48]	Gemma-2B [55]	558K+665K	72.5 73.1	56.0 56.8	61.3 62.5	43.7 43.7	54.0 55.8	49.5 51.2	23.9 24.6	38.7 40.4
LLaVA BASIC	CLIP-L [48]	Phi3-3.8B [1]	558K+665K	77.4 77.6	60.8 61.5	73.0 74.6	54.6 55.2	68.7 70.2	59.9 61.1	35.4 35.8	37.1 39.2
LLaVA BASIC	CLIP-L [48]	Mistral-7B [27]	558K+665K	79.1 80.3	62.5 64.1	72.4 74.5	56.6 57.2	70.0 72.1	63.6 65.1	36.3 36.8	47.6 48.5
LLaVA BASIC	SigLIP-SO [68]	Vicuna-7B [13]	558K+665K	80.8 81.2	63.2 64.1	70.6 71.4	62.0 62.0	68.0 69.6	58.6 61.3	32.9 34.2	51.1 52.3
LLaVA BASIC	SigLIP-SO [68]	Vicuna-13B [13]	558K+665K	81.8 82.4	64.3 65.5	73.8 75.1	64.4 64.4	69.5 70.8	65.8 66.7	37.6 38.2	54.2 55.3

Table 3. Comparisons between LLaVA [35] and BASIC when adopting various combinations of vision encoders and LLMs.

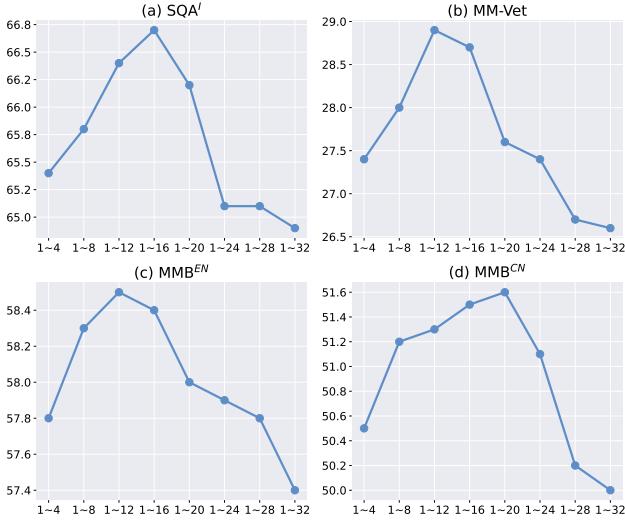


Figure 6. The results of utilizing refined visual embeddings from different layers to construct the supervisory visual embedding. The horizontal axis indicates the specific source layers.

this paper, w_i is used to integrate refined visual embeddings from different layers. With utilizing embeddings from the LLM’s lower half ($1 \sim l/2$ layers), we explore three settings of the layer weights w_i : first, w_i decreases quadratically, $w_i^{\text{dec}} = \frac{(l/2-i+1)^2}{\sum_{i=1}^{l/2} i^2}$; second, w_i remains constant, $w_i^{\text{const}} = \frac{1}{l/2}$; third, w_i increases quadratically, $w_i^{\text{inc}} = \frac{i^2}{\sum_{i=1}^{l/2} i^2}$. As shown in Table 4, w_i^{inc} maintains better results across various benchmarks. This is largely due to the gradual refinement process within the LLM’s shallow layers.

(iii) *The Degree of supervision.* In the main experiment, we implement fine control of the supervision degree for distinct image patches, which is based on the attention scores from the text to the image. The underlying motivation is that image patches vary in importance and should be treated differently. We compare it with the scenario where the degree of supervision is equal, namely a_i are all the same. As shown in Table 4, $\text{Sup.}^{\text{auto}}$ achieves better results compared with $\text{Sup.}^{\text{equal}}$, demonstrating the effectiveness of taking the

Benchmark	w_i^{dec}	w_i^{const}	w_i^{enc}	$\text{Sup.}^{\text{equal}}$	$\text{Sup.}^{\text{auto}}$
VQA ^{v2}	72.9	73.2	73.4	73.0	73.4
GQA	55.1	55.6	55.9	55.6	55.9
SQA ^I	65.6	66.0	66.7	66.1	66.7
MMB ^{EN}	57.4	57.9	58.4	57.6	58.4
MMB ^{CN}	51.0	51.5	51.5	51.0	51.5
MM-Vet	27.8	28.3	28.7	27.6	28.7

Table 4. Results under different settings of layers weights and supervision degrees. w_i^{dec} , w_i^{const} and w_i^{enc} denote w_i decreasing quadratically, remaining constant, and increasing quadratically respectively. $\text{Sup.}^{\text{equal}}$ and $\text{Sup.}^{\text{auto}}$ denote supervising each image patch equally and based on the importance respectively.

importance of image patches into account.

Analysis of Model Settings. To evaluate the applicability and robustness of our method, we conduct experiments on a series of MLLMs. The adopted LLMs include Gemma-2B [55], Phi3-3.8B [1], Mistral-7B [27], Vicuna-7B [13], and Vicuna-13B [13]. The vision encoders include CLIP-L [48] and SigLIP-SO [68]. As shown in Table 3, under the combinations of various sizes of LLMs and various resolutions of vision encoders, the introduction of direct visual supervision from the LLM’s shallow layers can robustly enhance the model’s performance.

6. Conclusion

Modality alignment is a basic issue for MLLMs. The prevailing aligning approach, which solely applies autoregressive supervision on the texts, has long neglected the introduction of direct supervision for the visual embeddings. In this work, we conduct a detailed analysis of the visual perception process of MLLMs, revealing the refinement of visual embeddings in the LLM’s shallow layers. Based on this, we utilize the refined visual embeddings from the LLM’s shallow layers to improve the initial visual embeddings from the perspective of directional alignment and semantic distribution. Our method effectively enhances the model’s performance, providing valuable insights into the construction of effective direct visual supervision.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No.62276245).

References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 8
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 6
- [4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2
- [5] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Soman, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 6
- [6] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. 3
- [7] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety—a review. *arXiv preprint arXiv:2404.14082*, 2024. 2
- [8] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemantics: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2, 2023. 3
- [9] Zheng Cai, Maosong Cao, Haojong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 1
- [10] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827, 2024. 1
- [11] Wei-Chi Chen and Wei-Ta Chu. Sssd: Self-supervised self distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2770–2777, 2023. 3
- [12] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 1, 2
- [13] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1, 5, 7, 8
- [14] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019. 3
- [15] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36: 16318–16352, 2023. 2
- [16] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. 3
- [17] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 6
- [18] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1
- [19] Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer ElShowk, Nicholas Joseph, Nova DasSarma, Ben Mann, Danny Hernandez, Amanda Askell, Kamal Ndousse, Andy Jones, et al. Softmax linear units. *Transformer Circuits Thread*, 2022. <https://transformer-circuits.pub/2022/solu/index.html>. 3
- [20] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023. 2
- [21] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. 3
- [22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 6
- [23] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 6
- [24] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self at-

tention distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1013–1021, 2019. 3

[25] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 6

[26] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2

[27] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lampe, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 8

[28] Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chengru Song, Dai Meng, Di Zhang, Wenwu Ou, Kun Gai, and Yadong Mu. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. In *ICLR*. OpenReview.net, 2024. 2, 6

[29] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 6

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[31] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2, 6

[33] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 1

[34] Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023. 3, 4

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 2, 5, 6, 7, 8

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2

[37] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 6

[38] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 6

[39] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 6

[40] Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361, 2020. 3

[41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[42] Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2856–2861, 2023. 2

[43] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridging vision and language spaces with assignment prediction. *arXiv preprint arXiv:2404.09632*, 2024. 1

[44] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019. 3

[45] Tianshuo Peng, Zuchao Li, Lefei Zhang, Hai Zhao, Ping Wang, and Bo Du. Multi-modal auto-regressive modeling via visual words. *CoRR*, abs/2403.07720, 2024. 1, 6

[46] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2

[47] Mary Phuong and Christoph H Lampert. Distillation-based training for multi-exit architectures. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1355–1364, 2019. 3

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 5, 7, 8

[49] ShareGPT. <https://sharegpt.com/>, 2023. 6

[50] Yiqing Shen, Liwu Xu, Yuzhe Yang, Yaqian Li, and Yandong Guo. Self-distillation from the last mini-batch for consistency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11943–11952, 2022. 3

[51] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 6

[52] Quan Sun, Qiyi Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 6

[53] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiyi Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 2

[54] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2

[55] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 8

[56] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[57] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2

[58] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1

[59] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*, 2024. 3, 4

[60] Yifan Xu, Xiaoshan Yang, Yaguang Song, and Changsheng Xu. Libra: Building decoupled vision system on large language models. In *ICML*. OpenReview.net, 2024. 1

[61] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1

[62] Chenglin Yang, Lingxi Xie, Chi Su, and Alan L Yuille. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2859–2868, 2019. 3

[63] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 1

[64] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Verhari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 2

[65] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 6

[66] Zeping Yu and Sophia Ananiadou. How do large language models learn in-context? query and key matrices of in-context heads are two towers for metric learning. *arXiv preprint arXiv:2402.02872*, 2024. 3, 4

[67] Zeyu Yun, Yubei Chen, Bruno A Olshausen, and Yann LeCun. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. *arXiv preprint arXiv:2103.15949*, 2021. 3

[68] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 1, 8

[69] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3713–3722, 2019. 3

[70] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2021. 3

[71] Zhilu Zhang and Mert Sabuncu. Self-distillation as instance-specific label smoothing. *Advances in Neural Information Processing Systems*, 33:2184–2195, 2020. 3

[72] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022. 3