



# Knowledge Transfer from Interaction Learning

Yilin Gao<sup>1</sup> Kangyi Chen<sup>1</sup> Zhongxing Peng<sup>1</sup> Hengjie Lu<sup>1</sup> Shugong Xu<sup>2,\*</sup>  
<sup>1</sup>Shanghai University <sup>2</sup>Xi'an Jiaotong-Liverpool University

<sup>1</sup>{gaoyilin, 2582712917, pzx, luhengjie}@shu.edu.cn    <sup>2</sup>shugong.xu@xjtlu.edu.cn

## Abstract

Current visual foundation models (VFM) face a fundamental limitation in transferring knowledge from vision language models (VLMs): while VLMs excel at modeling cross-modal interactions through unified representation spaces, existing VFMs predominantly adopt result-oriented paradigms that neglect the underlying interaction processes. This representational discrepancy hinders effective knowledge transfer and limits generalization across diverse vision tasks. We propose **Learning from Interactions (LFI)**, a cognitive-inspired framework that addresses this gap by explicitly modeling visual understanding as an interactive process. Our key insight is that capturing the dynamic interaction patterns encoded in pre-trained VLMs — beyond their final representations — enables more faithful and efficient knowledge transfer to VFMs. The approach centers on two technical innovations: (1) *Interaction Queries*, which maintain persistent relational structures across network layers, and (2) *interaction-based supervision*, derived from the cross-modal attention mechanisms of VLMs. Comprehensive experiments demonstrate consistent improvements across multiple benchmarks: achieving  $\sim 3.3\%$  and  $+1.6$  mAP/ $+2.4$  AP<sup>mask</sup> absolute gains on TinyImageNet classification and COCO detection/segmentation respectively, with minimal parameter overhead and faster convergence ( $7\times$  speedup). The framework particularly excels in cross-domain settings, delivering  $\sim 2.4\%$  and  $\sim 9.3\%$  zero-shot improvements on PACS and VLCS. Human evaluations further confirm its cognitive alignment, outperforming result-oriented methods by  $2.7\times$  in semantic consistency metrics.

## 1. Introduction

*“To teach someone how to fish is better than to just give him a fish.”*

– *Huainanzi*

\*Corresponding author

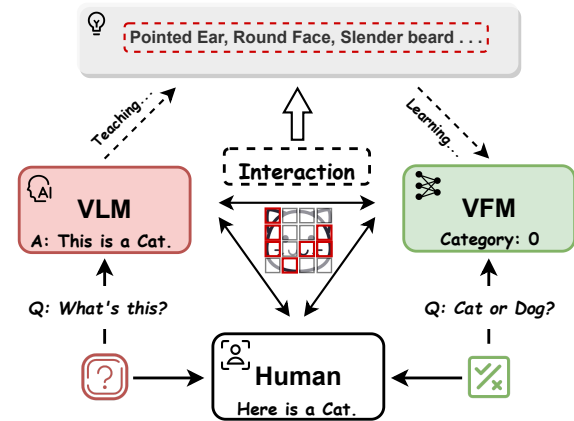


Figure 1. Humans and intelligent agents, including Vision-Language Models (VLMs) and Visual Foundation Models (VFM), comprehend the world through analogous cognitive processes, irrespective of the final form of representation. We denote this cognitive process as *interaction*. By refining these representation-agnostic interactions, we facilitate cross-modal and cross-task knowledge transfer.

Recent advancements in Multi-modal Large Language Models (MLLMs) [5, 6] have marked a transformative era in Artificial Intelligence (AI), enabling unprecedented capabilities in processing and integrating diverse data modalities such as text, images, sound, and video. These models have demonstrated remarkable proficiency in understanding and synthesizing human-generated content, thereby absorbing a vast spectrum of human knowledge. However, despite these achievements, a critical challenge remains: facilitating **effective knowledge transfer** between heterogeneous models while ensuring **cognitive alignment**. Current methodologies often prioritize superficial output imitation, neglecting the fundamental processes that underlie knowledge acquisition.

Amidst this rapid progress, a fundamental question arises: Do these models truly acquire *\*knowledge\**? And is the cognitive process of *\*knowledge\** consistent across different models? While MLLMs exhibit formidable capa-

bilities, the mechanisms underlying their knowledge acquisition remain underexplored. Knowledge, being an abstract concept, is inherently difficult to define. Its representation varies across models and even among humans, manifesting through language, writing, painting, and other forms. Consequently, measuring knowledge solely based on its representational form is inherently problematic.

Although knowledge itself may not be directly measurable, the cognitive processes that facilitate its understanding can be represented. For instance, as illustrated in Figure 1, both humans and models recognize a cat through a set of definitive interactions (e.g., pointed ears, round face, long whiskers). This perspective is supported by extensive research on model interpretability [23, 38]. We posit that these definitive interaction sets constitute an alternative representation of knowledge, and importantly, that this cognitive process is quantifiable. Following [13], we term these sets **interaction**. The essence of interactions lies in their shareability and complementarity [1–3], implying that the cognitive processes of different agents regarding the same knowledge are analogous and can mutually enhance one another. This mirrors the pedagogical process where an experienced teacher translates knowledge into concepts and effectively imparts them to students.

Building on this understanding, we argue that although Vision Language Models (VLMs) and Visual Foundation Models (VFMs) differ in their representational approaches, VLMs can still impart their cognitive processes of knowledge to VFMs through an interaction-based perspective. This approach diverges from traditional result- or representation-oriented learning methods that rely on explicit labels. Moreover, given the extensive multi-modal knowledge (e.g., vision and language) absorbed during VLM training, VLMs can function as seasoned educators, refining the cognitive processes of VFMs. This aligns with the adage, “teach someone to fish.” To operationalize these ideas, we propose a novel knowledge transfer methodology termed **Learning from Interaction**. By emulating the cognitive processes of VLMs for various tasks, this methodology facilitates knowledge transfer from VLMs to VFMs. The primary goal is to leverage the diverse knowledge sources of VLMs to enhance VFMs, enabling them to comprehend the natural world more holistically.

The contributions of this work are threefold:

1. We empirically validate the consistency of the cognitive process of knowledge at the interaction level between VLMs and VFMs.
2. We introduce **Interaction Learning**, a novel framework that enhances the task cognition capabilities of VFMs through additional interactive supervision from VLMs, enabling cross-task and cross-modal knowledge transfer.
3. We demonstrate the effectiveness of our approach on downstream visual foundation tasks. Through Interac-

tion Learning, the model exhibits significant improvements in accuracy, convergence speed, and generalization. Human evaluations further confirm that our approach aligns more closely with human cognitive processes for tasks.

## 2. Related Works

### 2.1. Foundation Visual Models

The transformer architecture [43] has revolutionized computer vision, with numerous adaptations tailored to diverse tasks. The Vision Transformer (ViT) [17] pioneered the use of transformers in vision by segmenting images into patches as inputs. The Swin Transformer [32] enhanced efficiency through shifted windows, while the DETection TRansformer (DETR) [10] extended transformers to object detection, eliminating the need for Non-Maximum Suppression (NMS) and anchor designs. Subsequent works [24, 31, 37, 48, 54] addressed slow convergence by refining object queries and Hungarian matching algorithms. These advancements underscore the versatility of transformers, unifying model architectures across tasks.

### 2.2. Vision Language Models

Vision Language Models (VLMs) [15, 16, 25, 27, 29, 30, 44, 46, 50, 53] bridge visual and textual modalities, integrating diverse perceptual and representational approaches. BLIP-2 [25] introduced the Q-Former to align these modalities, while LLaVA and MiniGPT-4 [29, 30, 53] employed linear projection layers for alignment. Mini-Gemini [27] enhanced VLMs with an additional visual encoder for high-resolution refinement. Despite their strengths, VLMs’ large scale limits their practical application. DeCo [46] addressed this by compressing visual tokens at the patch level, and SparseVLM [50] reused self-attention matrices to prune insignificant vision tokens. These works inspire us to transfer VLMs’ rich cognitive understanding to Visual Foundation Models (VFMs), enhancing their comprehension of the world.

### 2.3. Model Interpretability

Model interpretability is crucial for understanding the inner workings of Deep Neural Networks (DNNs), which are often regarded as *black-box* systems. Recent research has increasingly focused on this area. [22, 26] demonstrated that DNNs can learn symbolic interactions, distilling transferable knowledge from raw data. Similarly, transformer-based vision models exhibit interactions between patches, as analyzed by [4, 7, 11–13, 21, 34]. Multi-modal systems have also been explored [33, 40, 42, 45], with LVLM-Interpret [42] designing interactive tools to uncover the mechanisms of large vision-language models. These efforts

collectively aim to reveal the cognitive processes, or *interactions*, between inputs and outputs.

## 2.4. Knowledge Transfer

Knowledge distillation, introduced by [20], is a widely used method for knowledge transfer, enabling student networks to learn feature distributions or labels from teacher networks. Early works like [47] used attention maps for transfer, while [14] leveraged multi-level teacher information. In open-vocabulary object detection, [8, 19, 35] distilled VLM knowledge to align region-level and image-level embeddings. Recent works [9, 41] extended distillation to transfer knowledge between large-scale and small-scale VLMs. However, these methods are limited to single-modality distillation and cannot achieve cross-modal or cross-task transfer.

Building on the success of Vision-Language Model pre-training, transfer learning has become a dominant approach due to its scalability and efficiency. Techniques like prompt tuning [51, 52] and visual adaptation [18, 49] adapt pre-trained VLMs for downstream tasks. [39] demonstrated that pre-trained LLMs can handle purely visual tasks without language reliance, while LOAT [28] combined LLMs with historical experiences for commonsense reasoning. Inspired by these works, we propose leveraging the extensive knowledge embedded in VLMs and transferring it to VFMs to enhance their capabilities in downstream tasks.

## 3. Preliminary

### 3.1. Vision Transformer

The Vision Transformer (ViT) [17] processes an image as a sequence of non-overlapping patches. For an image divided into  $N$  patches  $\{p_1, \dots, p_N\}$ , each patch is linearly projected into a token:

$$\mathbf{x}_i = \text{Embedding}(p_i) + \mathbf{e}_i,$$

where  $\mathbf{e}_i$  denotes positional embeddings. These tokens are processed through stacked Transformer layers, where the multi-head self-attention mechanism computes dependencies between tokens:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V.$$

This operation can be factorized into two components by identity transformation:

$$\text{Attention}(Q, K, V) = \underbrace{\sigma(Q, K)}_{\text{Structure}} \odot \underbrace{\phi\left(\frac{QK^\top}{\sqrt{d_k}}\right)}_{\text{Strength}} \cdot V, \quad (1)$$

where: -  $\sigma(Q, K) \in \{0, 1\}^{N \times N}$  is a binary matrix indicating token relationships (1: related, 0: unrelated), -  $\phi(\cdot)$  computes normalized attention weights, -  $\odot$  denotes element-wise multiplication.

### 3.2. Definition of Interaction

To establish a unified framework for cross-modal knowledge transfer, we first formalize the concept of **interaction** – the fundamental cognitive process underlying both human and machine perception. Consider the human process of recognizing a cat: it emerges from synergistic relationships between features like pointed ears (A), round face (B), and long whiskers (C). These features do not act in isolation; their combinatorial relationships (e.g., “A AND B AND C”) collectively define the concept. Analogously, in deep neural networks, outputs are determined by structured interactions between input elements.

Recent work [26] rigorously proves that the output of any DNN  $v : \mathbb{R}^n \rightarrow \mathbb{R}$  can be decomposed into logical interactions:

$$v(p) = \sum_{S \in \Omega_{\text{and}}} \underbrace{I_{\text{and}}(S|p)}_{\text{Conjunctive Interaction}} + \sum_{S \in \Omega_{\text{or}}} \underbrace{I_{\text{or}}(S|p)}_{\text{Disjunctive Interaction}},$$

where  $I_{\text{and}}$  encodes **joint activation** of variable subsets  $S$  (e.g., “A AND B”), while  $I_{\text{or}}$  represents **alternative activation** (e.g., “A OR B”). The sets  $\Omega_{\text{and}}$  and  $\Omega_{\text{or}}$  enumerate all valid AND/OR relationships.

### Bridging Logical Interactions to Transformer Attention

In Transformer-based models, these logical interactions are explicitly manifested through attention mechanisms. To align the decomposition with architectural primitives, we re-express interactions at the token-pair level. For each attention head  $h$ , the logical subsets  $S \in \Omega_{\text{and/or}}$  can be mapped to pairwise interactions between query-key pairs  $(q_i, k_j)$ . Specifically: - **Conjunctive Interaction** ( $I_{\text{and}}$ ): Implemented as the *product* of binarized logic gates  $\sigma(q_i, k_j)$  and contextualized signals  $\phi(\cdot)W_o v_j$ . - **Disjunctive Interaction** ( $I_{\text{or}}$ ): Corresponds to the *sum* of independent pairwise activations.

This yields the unified formulation:

$$v(p) = \sum_{h=1}^H \sum_{i,j} \underbrace{\sigma(q_i, k_j)}_{\text{Logic Gate (AND/OR)}} \cdot \underbrace{\phi\left(\frac{q_i^\top k_j}{\sqrt{d_k}}\right)W_o v_j}_{\text{Contextualized Signal (Content \& Strength)}}. \quad (2)$$

Here,  $\sigma(\cdot)$  acts as a binary selector (1 for active interactions, 0 otherwise), while  $\phi(\cdot)$  modulates interaction strength. Notably, setting  $\phi(\cdot) = 1$  simplifies the model to pure symbolic reasoning (e.g., “A AND B”), enhancing interpretability without loss of generality.

Critically, while  $V$  varies across tasks (e.g., classification vs. detection), the interaction structure  $C_{\text{struct}}$  and strength  $C_{\text{strength}}$  remain spatially consistent due to the shared visual grounding in natural images. This spatial coherence allows interactions to serve as universal carriers for cross-modal knowledge transfer.

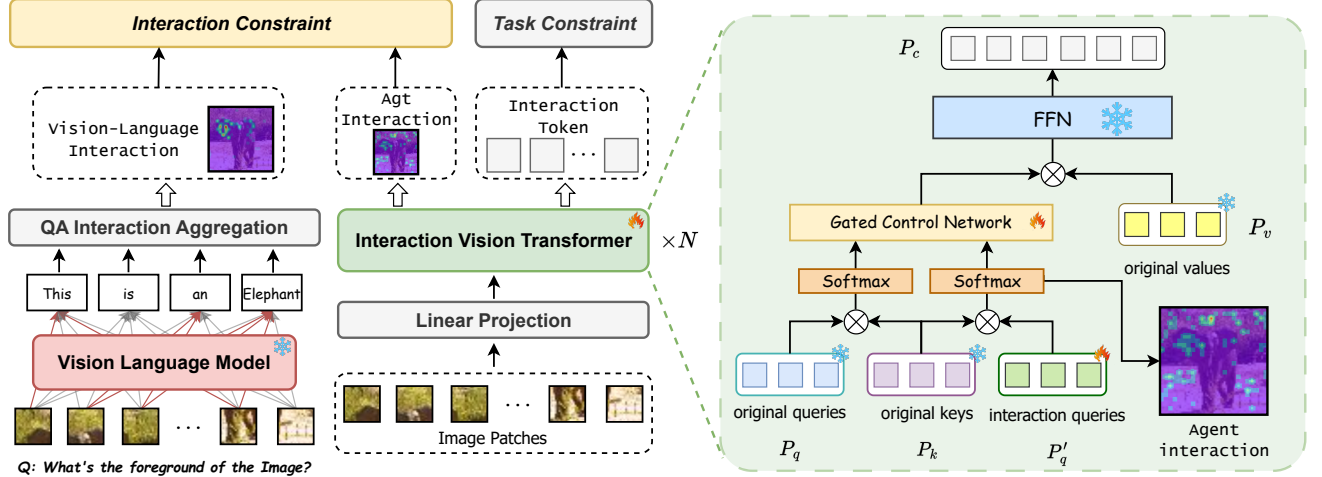


Figure 2. **Pipeline of Learning from Interaction.** The process consists of several steps: (1) Extracting Vision-Language Interactions from pre-trained Vision Language Models (VLMs) using a Visual Question Answering (VQA) framework; (2) Generating additional interaction queries during query formulation in the Vision Transformer, supervised by the extracted Vision-Language Interactions; (3) Integrating both sets of interactions via a Gated Control Network; (4) Producing the final Interaction Tokens, which serve as the basis for task predictions.

## 4. Interaction Learning

Building on the theoretical foundation of structured interactions (Section 3.2), we present the Interaction Vision Transformer (I-ViT) that operationalizes cross-modal knowledge transfer through explicit interaction modeling. The overall pipeline is shown in Figure 2.

### 4.1. Interaction Vision Transformer (I-ViT)

The I-ViT enhances standard ViT through dual interaction pathways that preserve both visual and linguistic reasoning patterns. Given input tokens  $\mathcal{X}$ , it generates:

- **Original Queries** ( $P_q$ ): Maintain task-specific interaction structures learned from visual data
- **Interaction Queries** ( $P'_q$ ): Encode cross-modal interaction patterns distilled from VLMs

**Interaction Strength Modeling** Both query types interact with shared keys  $P_k$  to compute interaction strength matrices:

$$C_{\text{VFM}} = \phi \left( \frac{P_q P_k^\top}{\sqrt{d_k}} \right) \quad (\text{Visual Foundation Model}) \quad (3)$$

$$C_{\text{AGT}} = \phi \left( \frac{P'_q P_k^\top}{\sqrt{d_k}} \right) \quad (\text{VLM Guidance}) \quad (4)$$

where  $\phi(\cdot)$  denotes softmax normalization. Here,  $C_{\text{VFM}}$  captures visual-centric interaction strengths, while  $C_{\text{AGT}}$  incorporates linguistic-aware interaction intensities from VLMs.

**Structural Interaction Fusion** The Gated Control Network (GCN) dynamically fuses interaction structures across modalities:

$$[g_1, g_2] = \text{sigmoid}(\text{FFN}([C_{\text{AGT}}; C_{\text{VFM}}])), \quad (5)$$

$$C_F = g_1 \odot C_{\text{AGT}} + g_2 \odot C_{\text{VFM}}, \quad (6)$$

where  $g_1, g_2 \in (0, 1)$  are continuous gating weights. This implements soft structure selection:

$$\text{Fused Structure} = \underbrace{g_1 \cdot \sigma_{\text{VLM}}}_{\text{Linguistic-Aware Interaction Importance}} + \underbrace{g_2 \cdot \sigma_{\text{VFM}}}_{\text{Vision-Aware Interaction Importance}}, \quad (7)$$

preserving the AND/OR logic through  $\sigma_{\text{VLM}}, \sigma_{\text{VFM}} \in \{0, 1\}$  (from Eq. (2)) while allowing adaptive importance weighting. Here: -  $\sigma_{\text{VLM}}/\sigma_{\text{VFM}}$ : Binary interaction structures (0/1) from VLMs/VFMs -  $g_1/g_2$ : Continuous importance scores for cross-modal alignment

This formulation separates *what to interact* (binary  $\sigma$ ) from *how importantly to interact* (continuous  $g$ ), where the former enforces logical rules and the latter enables context-aware fusion.

**Content-Aware Interaction Generation** The fused interaction matrix modulates value transformations:

$$P_c = \text{FFN}(C_F \odot P_v), \quad (8)$$

where  $P_v$  carries task-specific content information. This implements:

$$\text{Interaction Tokens} = \underbrace{C_F}_{\text{Structure+Strength}} \odot \underbrace{P_v}_{\text{Content}}, \quad (9)$$



strictly adhering to the interaction formulation  $C_{\text{struct}} \odot C_{\text{strength}} \cdot V$  from Section 3.1.

## 4.2. Vision-Language Interaction Extraction

**Unified Interaction Extraction Protocol** We establish a task-agnostic framework for extracting vision-language interactions from pre-trained VLMs, applicable to classification, detection, and segmentation. The pipeline comprises three stages: 1. **Task-Conditioned Prompting**: Inject task semantics through language-vision prompts 2. **Cross-Modal Activation**: Compute token-level interactions between visual and linguistic modalities 3. **Cognitive Alignment Filtering**: Refine interactions to match human perceptual patterns

For all tasks, the VLM processes concatenated inputs:

$$Q_{\text{input}} = [\mathcal{X}; \mathcal{Q}] \in \mathbb{R}^{(N+L) \times D}, \quad (10)$$

where  $Q_{\text{input}}$  represents the concatenated sequence of image tokens  $\mathcal{X}$  includes visual prompts (e.g., bounding boxes for detection) as part of the input image and question token  $\mathcal{Q}$ .

**Task-Specific Formulations** The unified interaction extraction framework adapts to different tasks through specialized components in the response template:

- **Classification**: *Language Prompt*: “Classify foreground vs. background” *Response Structure*:  $\{\text{foreground: [class]}, \text{background: [...]}\}$

$$C_{\text{VLM}} = \frac{\max(0, C_{\text{fore}} - C_{\text{back}})}{\|\max(0, C_{\text{fore}} - C_{\text{back}})\|_1} \quad (11)$$

where  $C_{\text{fore}}$  and  $C_{\text{back}}$  denote interaction strengths for foreground class and background clusters respectively.

- **Dense Prediction (Detection/Segmentation)**: *Language Prompt*: “Locate all  $\{\text{tgt\_obj}\}$  instances” *Response Structure*:  $\{\text{objects: [list]}, \text{background: [...]}\}$

$$C_{\text{VLM}} = \frac{\max\left(0, \sum_{k=1}^K C_{\text{obj}}^{(k)} - C_{\text{back}}\right)}{\left\|\max\left(0, \sum_{k=1}^K C_{\text{obj}}^{(k)} - C_{\text{back}}\right)\right\|_1} \quad (12)$$

where  $C_{\text{obj}}^{(k)}$  represents interaction strength for the  $k$ -th target object instance, aggregated across all  $K$  instances.

**Visual Prompt Integration** For detection/segmentation, spatial constraints **Bounding Boxes** (drawn on input images) are *pre-encoded* in the input image through.

This directly shape the VLM’s cross-attention patterns. As shown in Figure 9 (in *Supplementary Material*), bounding boxes induce stronger activations on target regions during  $C_{\text{obj}}$  computation and act as *attentional anchors* that guide the VLM to: - Attend to object parts within prompted regions (high  $C_{\text{obj}}$  values) - Suppress irrelevant background (near-zero  $C_{\text{obj}}$ ).

**Cognitive Filtering Mechanism** The operator  $\frac{\max(0, \cdot)}{\|\cdot\|_1}$  implements two biological principles: 1. **Non-Negative Competition**: Neural activations are constrained to non-negative values to align with the stability conditions of **softmax** in Equation (4), preventing instability. 2. **Contrast Enhancement**: Amplify foreground-background differentiation through selective signal amplification.

## 4.3. Optimization Objective

The learning objective combines task performance and cognitive alignment through a dual-loss framework:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{align}}, \quad (13)$$

$$\mathcal{L}_{\text{align}} = D_{\text{KL}}(C_{\text{AGT}} \parallel C_{\text{VLM}}), \quad (14)$$

where  $\mathcal{L}_{\text{task}}$  is task-specific, and  $\mathcal{L}_{\text{align}}$  measures the KL-divergence between VLM-guided interactions ( $C_{\text{AGT}}$ ) and visual foundation model’s intrinsic patterns ( $C_{\text{VFM}}$ ). The Gated Control Network dynamically adjusts the contribution of these interactions, eliminating the need for manual hyperparameter tuning.

## 5. Experiments

### 5.1. Experimental Setup

#### 5.1.1. Datasets and Metrics.

Evaluations cover **TinyImageNet** (classification), **COCO/VOC** (detection/segmentation), and **PACS/VLCS** (zero-shot domain generalization). Metrics include *Top-1* accuracy (classification), mAP (detection),  $AP_{\text{mask}}$  (segmentation), and cross-domain variance (zero-shot).

#### 5.1.2. Implementation.

We adopt LLaVa-1.5 [30] as VLM and ViT [17] as classification model. For detection/segmentation, DETR[10] is used for COCO benchmark while Conditional-DETR[36] is used for few-shot evaluations. Training uses  $4 \times \text{A6000}$  GPUs with 50 epochs (25 for segmentation), **freezing original transformer layers** and only updating the GCN, interaction queries, and task heads. Other hyperparameters follow ViT/DETR defaults.

### 5.2. Experimental Results

#### 5.2.1. Comparison with baseline methods.

Given the limited research on knowledge transfer from VLMs to VFMs, we compared Vision Transformers (ViT) and Interaction Vision Transformers (I-ViT) on fundamental visual tasks. As shown in Table 1, I-ViT significantly outperforms ViT with a reasonable increase in parameters. We also hypothesized that Interaction Learning, which focuses on teaching cognitive processes rather than results, would lead to faster convergence and better performance

Table 1. Comparison of I-ViT and ViT on TinyImageNet.

Model	Top-1 Acc. (%)	#params	FLOPs
ViT-S/16 [17]	79.94	22M	15.5B
I-ViT-S/16	<b>81.52 (+1.6)</b>	24M	18.1B
ViT-B/16	80.24	86M	55.5B
I-ViT-B/16	<b>82.14 (+1.9)</b>	93M	62.7B
ViT-L/16	87.83	304M	191.2B
I-ViT-L/16	<b>91.08 (+3.3)</b>	329M	213.9B

with fewer training samples. Figure 3 demonstrates that I-ViT converges approximately 7x faster than result-oriented learning.

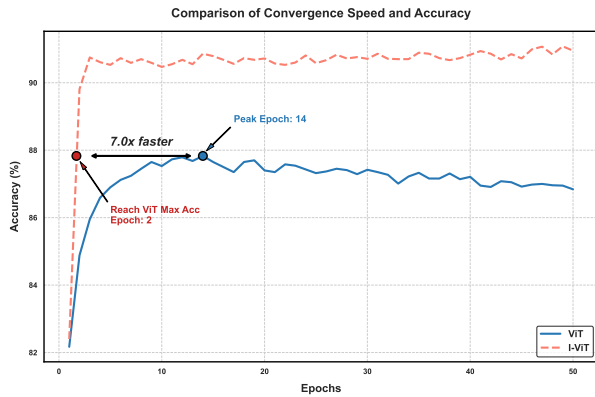


Figure 3. **Comparison of Convergence Speed.** I-ViT-L/16 exhibits faster convergence and achieves higher accuracy compared to the traditional Vision Transformer (ViT-L/16), further highlighting the efficiency of learning from interactions.

We further evaluated our method on reduced-scale datasets (1%, 10%, and 50% of the training data). Table 2 shows that Interaction Learning consistently achieves substantial performance gains, even with limited data. Training on 100% of the data from scratch also confirms the indispensability of  $C_{VFM}$ .

Table 2. Comparison of I-ViT-L/16 and ViT-L/16 at Different Training Data Ratios on TinyImagenet.

Model	Training Data Ratio (%)			
	100%	50%	10%	1%
ViT	32.67	27.04	12.64	2.86
I-ViT	<b>34.12</b>	<b>29.04</b>	<b>15.60</b>	<b>3.07</b>

### 5.2.2. Ablations on Interaction Vision Transformer.

The I-ViT introduces additional queries and networks, increasing parameters and computational overhead. To

demonstrate that performance improvements stem from effective learning rather than increased parameters, we conducted ablation studies, as detailed in Table 3.

Table 3. Ablations on Interaction Queries (IQ), Interaction Constraint (IC), and Gated Control (GC).

Model	Interaction Query	Interaction Constraint	Gated Control	Acc.
ViT-L/16				87.8
	✓		✓	89.6
I-ViT-L/16		✓		80.7
	✓	✓		90.3
	✓	✓	✓	<b>91.1</b>

Specifically, the *second* row of Table 3 shows the effect of removing the supervision of  $C_{VLM}$  in I-ViT and introducing an extra set of learnable queries to expand the network’s scale. It is observed that this additional set of learnable queries indeed enhances performance. However, the lack of additional supervision and the parallel operation of Interaction Queries and Original Queries mean that the network does not learn distinct and complementary information during backpropagation.

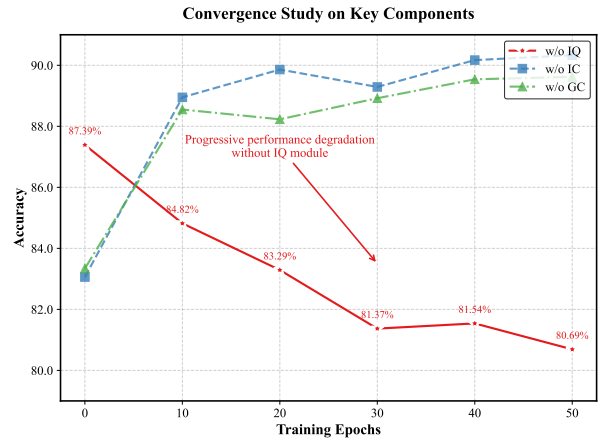


Figure 4. **Component Ablation Analysis** reveals distinct failure patterns. While removing INTERACTION CONSTRAINT (w/o IC, dashed blue) or GATED CONTROL (w/o GC, dash-dot green) causes temporary accuracy drops, both variants recover to 90.3% and 89.6% through parameter adaptation. The IQ-ablated model (w/o IQ, solid red) shows deceptive initial competence comparable to standard ViT, benefiting from shared visual feature learning objectives. Progressive divergence between  $C_{VFM}$ ’s task-specific representations and  $C_{VLM}$ ’s semantic structures leads to **irreversible degradation** (6% relative drop to 80.7%), demonstrating IQ’s indispensable role in cross-modal interaction reconciliation.

Furthermore, we explored directly supervising the Original Queries with Vision-Language Interactions, as shown in the *third* row. The results indicate not only a failure to

improve performance but also a degradation, further supporting our hypothesis that interactions derived by different models for the same task, due to different training sources, are complementary rather than directly interchangeable.

As shown in Fig. 4, removing INTERACTION CONSTRAINT (IC) or GATED CONTROL (GC) causes only transient accuracy fluctuations (90.3% and 89.6% final accuracy respectively), confirming their non-essential role in convergence. The critical exception is IQ ablation – despite initial ViT-level performance (87.4%), it suffers catastrophic decline to 80.7% accuracy as modality conflicts between  $C_{VFM}$  and  $C_{VLM}$  escalate. This proves IQ’s unique necessity in sustaining stable multimodal integration.

The *fourth* row validates the necessity of the Gated Control Network (GCN). By adaptively adjusting the weights of the two interactions, the model can synthesize more reliable interactions, thereby enhancing its cognitive capabilities.

### 5.2.3. Visualization on Interaction

In this section, we compare  $C_{VFM}$ ,  $C_{VLM}$ , and  $C_{AGT}$ , as shown in Figure 5.

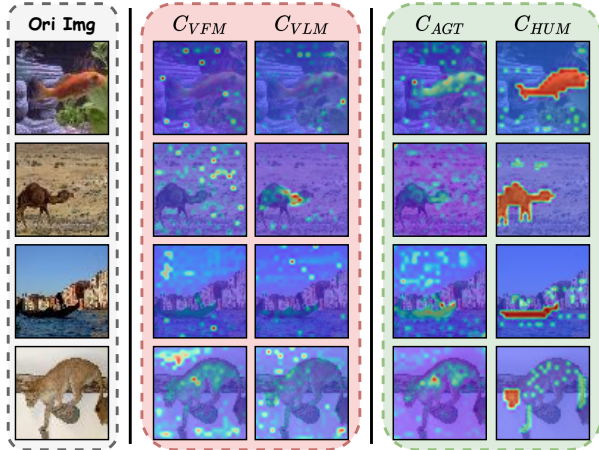


Figure 5. **Interaction Visualization Analysis.**  $C_{VFM}$  predominantly emphasizes background information, whereas  $C_{VLM}$  concentrates more on the instances themselves.  $C_{AGT}$  synthesizes the cognitive processes and capabilities of both  $C_{VFM}$  and  $C_{VLM}$ , enabling object recognition through a balance of minimal background context and a focus on the instances, thereby aligning more closely with human cognitive patterns.

It is observed that  $C_{VFM}$  tends to focus more on background information compared to  $C_{AGT}$  and  $C_{VLM}$ . This bias is likely due to backgrounds occupying a larger area in the training images, leading the model to learn through background cues with minimal instance information. Conversely,  $C_{VLM}$ , which benefits from extensive visual-linguistic training data, exhibits a nuanced understanding of interactions, focusing more on instances rather than environmental context.

Despite being trained on the same data as  $C_{VFM}$ ,  $C_{AGT}$  integrates the cognitive processes and capabilities of both  $C_{VFM}$  and  $C_{VLM}$ , recognizing objects through a combination of minimal background information and a greater focus on instances. We also compared  $C_{AGT}$  with the human cognitive process  $C_{HUM}$ , and as expected,  $C_{AGT}$  aligns more closely with human cognition.

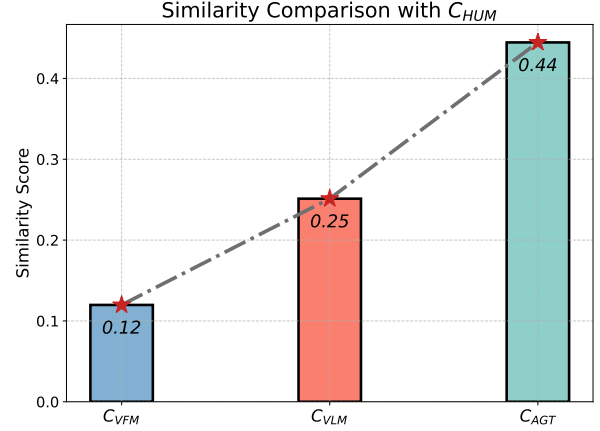


Figure 6. **Comparison with Human Cognition ( $C_{HUM}$ ).**  $C_{VLM}$  is closer to human cognition than  $C_{VFM}$ , but  $C_{AGT}$  achieves the highest score, effectively simulating human cognitive processes.

Furthermore, we quantitatively compared the proximity of  $C_{VFM}$ ,  $C_{VLM}$ , and  $C_{AGT}$  to human cognitive process using *Cosine Similarity*, as depicted in Figure 6. Consistent with the qualitative analysis,  $C_{VLM}$  is closer to the human cognitive approach than the result-oriented  $C_{VFM}$ . However,  $C_{AGT}$  achieved a significantly higher score, effectively simulating the human cognitive process. Additional experimental settings and details are provided in Appendix 9.

### 5.2.4. Experiments on Cross-Domain Datasets

To validate the generalizability of Interaction Learning, we conducted zero-shot experiments on the cross-domain dataset **PACS**. As shown in Table 4, I-ViT outperforms ViT across nearly all domains, reinforcing the efficiency of Interaction Learning.

Table 4. **Comparison of ViT[17] and I-ViT on the PACS Dataset.** I-ViT demonstrates excellent domain generalization capabilities, validating that learning from interactions is a more efficient and robust learning approach.

Model	Art Painting	Cartoon	Photo	Sketch
ViT-L/16	45.28	29.89	49.86	<b>15.49</b>
I-ViT-L/16	<b>51.89</b>	<b>31.90</b>	<b>51.22</b>	15.18

On the **Art Painting** subset, which includes human-

created subcategories, I-ViT achieved a  $\sim 6.6\%$  improvement. This subset represents human interpretations of natural objects, with an abstraction level between **Photo** and **Sketch**. However, I-ViT experienced slight performance degradation in the **Sketch** category, likely due to the simplistic content of sketches, which makes consistent interactions challenging.

Table 5. **Comparison of VLM, ViT and I-ViT on the VLCS Dataset.** I-ViT’s behavioral trends align with LLaVa, demonstrating effective knowledge transfer through interaction mechanisms.

Model	PASC.	CALT.	LABE.	SUN
LLaVa-1.5	56.40	62.89	<b>52.58</b>	32.71
ViT-L/16	60.42	58.59	30.49	31.86
I-ViT-L/16	<b>63.98</b>	<b>68.75</b>	49.35	<b>36.61</b>

In VLCS (Table 5), we further compared the accuracy of LLaVa. It can be found that while I-ViT surpasses ViT, and its behavioral trends align with LLaVa, proving the effective transfer between interactions.

### 5.3. Extension on Dense Prediction

#### 5.3.1. Main Results on COCO

Our experiments on COCO 2017 reveal consistent performance gains across both detection and segmentation tasks. As shown in Table 6, the integration of vision-language interactions ( $C_{VLM}$ ) elevates detection mAP from 42.0 to 43.6 (+1.6 relative improvement). Notably, the segmentation task benefits more substantially in mask-level accuracy (+2.4  $AP_{mask}$ ), suggesting that cross-modal cues particularly enhance boundary-sensitive predictions.

Table 6. COCO 2017 Benchmark Results (val set)

Task	$C_{VLM}$	mAP	$AP_{mask}$	#param	FLOPs
Det	✓	<b>43.6</b>	-	41.9M	55.65B
	✗	42.0	-	41.5M	55.62B
Seg	✓	-	<b>33.5</b>	43.3M	162.6B
	✗	-	31.1	42.9M	162.4B

This cross-task success comes with minimal computational overhead - the added parameters less than  $\sim 1\%$  of base models (41.9M vs 41.5M for detection). The FLOPs analysis confirms our design’s efficiency, showing less than  $\sim 0.6\%$  increase. These findings collectively validate that vision-language interactions provide generalized benefits beyond specific task formulations.

#### 5.3.2. Data-Efficient Learning on VOC

The VOC experiments in Table 7 demonstrate our method’s effectiveness in data-scarce scenarios (Condition-DETR[36]; Train from scratch).

Table 7. Few-Shot Detection Analysis on VOC

Model	mAP	$AP_S$	$AP_M$	$AP_L$
w/ $C_{VLM}$	<b>40.9</b>	4.9	<b>24.2</b>	<b>53.5</b>
w/o $C_{VLM}$	38.6	<b>5.0</b>	22.4	51.2

With only 10% of COCO’s training data, the  $C_{VLM}$  integration achieves 40.9 mAP, outperforming the baseline by 2.3 points. This significant gap (+6.0% relative) highlights the method’s ability to compensate for limited annotations through linguistic knowledge transfer. The  $C_{VLM}$  integration primarily enhances medium and large object detection ( $AP_M$ : +1.8,  $AP_L$ : +2.3), as the query attention mechanism naturally prioritizes salient regions where these objects dominate. Small object performance remains comparable ( $AP_S$ : -0.1), indicating preserved spatial sensitivity in the visual foundation model. This phenomenon aligns with the inherent bias of vision-language interactions toward capturing semantically prominent patterns in images.

## 6. Conclusions and Future Work

Multi-modal language models (MLLMs) have achieved remarkable progress in various visual tasks, largely due to the availability of large-scale image-text datasets. However, the inherent discrepancy between input and output modalities hinders the effective transfer of their natural scene understanding capabilities to downstream Visual Foundation Models (VFMs). To address this challenge, we propose **Learning from Interactions**, a novel paradigm that emphasizes the alignment of cognitive processes in knowledge transfer between Vision Language Models (VLMs) and Visual Foundation Models (VFMs). By introducing additional interaction-based supervision from VLMs to VFMs, we enable cross-modal and cross-task knowledge transfer. Our experiments demonstrate that interaction-based learning outperforms traditional result-oriented approaches in terms of accuracy, convergence speed, and generalization, while also aligning more closely with human cognitive processes.

**Future Work:** Building upon the current framework, we focus on knowledge transfer from VLMs to VFMs. In future, we aim to explore bi-directional knowledge transfer, including from VFMs to VLMs and VFMs to VFMs, to uncover a more fundamental knowledge system and the underlying knowledge modeling process within models. Additionally, we plan to extend this approach to other foundational tasks, such as video understanding, to fully leverage the world understanding capabilities of MLLMs. This will further bridge the gap between human-like cognitive processes and machine learning models, paving the way for more robust and interpretable AI systems.



## 7. Acknowledgment

This work was supported in part by the 6G Science and Technology Innovation and Future Industry Cultivation Special Project of Shanghai Municipal Science and Technology Commission under Grant 24DP1501001, in part by the National High Quality Program under Grant TC220H07D.

## References

- [1] The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*. 2
- [2] Frozen transformers in language models are effective visual encoder layers. *arXiv preprint arXiv:2310.12973*.
- [3] Can the inference logic of large language models be disentangled into symbolic concepts? *arXiv preprint arXiv:2304.01083*. 2
- [4] S Abnar and W Zuidema. Quantifying attention flow in transformers. *arxiv* 2020. *arXiv preprint arXiv:2005.00928*, 2022. 2
- [5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [6] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 1, 2023. 1
- [7] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 2
- [8] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022. 3
- [9] Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, and Xiang Bai. Llava-kd: A framework of distilling multimodal large language models. *arXiv preprint arXiv:2410.16236*, 2024. 3
- [10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 5
- [11] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 2
- [12] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.
- [13] Haozhe Chen, Junfeng Yang, Carl Vondrick, and Chengzhi Mao. Interpreting and controlling vision foundation models via text explanations. *arXiv preprint arXiv:2310.10591*, 2023. 2
- [14] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5008–5017, 2021. 3
- [15] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2
- [16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2
- [17] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 5, 6, 7
- [18] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 3
- [19] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3
- [20] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3
- [21] Jinfan Hu, Jinjin Gu, Shiyao Yu, Fanghua Yu, Zheyuan Li, Zhiyuan You, Chaochao Lu, and Chao Dong. Interpreting low-level vision models with causal effect maps. *arXiv preprint arXiv:2407.19789*, 2024. 2
- [22] Seonggyeom Kim and Dong-Kyu Chae. What does a model really look at?: Extracting model-oriented concepts for explaining deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [23] Brenden M Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023. 2
- [24] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627, 2022. 2
- [25] 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. 2

- [26] Mingjie Li and Quanshi Zhang. Does a neural network really encode symbolic concepts? In *International conference on machine learning*, pages 20452–20469. PMLR, 2023. 2, 3
- [27] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 2
- [28] Mengying Lin, Yaran Chen, Dongbin Zhao, and Zhao-ran Wang. Advancing object goal navigation through llm-enhanced object affinities transfer. *arXiv preprint arXiv:2403.09971*, 2024. 3
- [29] 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. 2
- [30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 5
- [31] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 2
- [32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2
- [33] Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 455–467, 2022. 2
- [34] Jie Ma, Yalong Bai, Bineng Zhong, Wei Zhang, Ting Yao, and Tao Mei. Visualizing and understanding patch interactions in vision transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2
- [35] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2022. 3
- [36] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3651–3660, 2021. 5, 8
- [37] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660, 2021. 2
- [38] Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, and Keyu Chen. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint arXiv:2409.02387*, 2024. 2
- [39] Ziqi Pang, Ziyang Xie, Yunze Man, and Yu-Xiong Wang. Frozen transformers in language models are effective visual encoder layers. *arXiv preprint arXiv:2310.12973*, 2023. 3
- [40] Ali Rasekh, Sepehr Kazemi Ranjbar, Milad Heidari, and Wolfgang Nejdl. Ecor: Explainable clip for object recognition. *arXiv preprint arXiv:2404.12839*, 2024. 2
- [41] Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, Siming Fu, et al. Llava-mod: Making llava tiny via moe knowledge distillation. *arXiv preprint arXiv:2408.15881*, 2024. 3
- [42] Gabriela Ben Melech Stan, Raanan Yehezkel Rohekar, Yaniv Gurwicz, Matthew Lyle Olson, Anahita Bhiwandiwalla, Estelle Aflalo, Chenfei Wu, Nan Duan, Shao-Yen Tseng, and Vasudev Lal. Lvlm-intrepret: An interpretability tool for large vision-language models. *arXiv preprint arXiv:2404.03118*, 2024. 2
- [43] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2
- [44] 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. 2
- [45] Xin Xiao, Bohong Wu, Jiacong Wang, Chunyuan Li, Xun Zhou, and Haoyuan Guo. Seeing the image: Prioritizing visual correlation by contrastive alignment. *arXiv preprint arXiv:2405.17871*, 2024. 2
- [46] Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024. 2
- [47] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 3
- [48] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2
- [49] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 3
- [50] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024. 2
- [51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 3
- [52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

- [53] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [2](#)
- [54] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. [2](#)