

# Align-KD: Distilling Cross-Modal Alignment Knowledge for Mobile Vision-Language Large Model Enhancement

Qianhan Feng<sup>1,2\*</sup>, Wenshuo Li<sup>2</sup>, Tong Lin<sup>1†</sup>, Xinghao Chen<sup>2†</sup>

<sup>1</sup>State Key Laboratory of General Artificial Intelligence,  
 School of Intelligence Science and Technology, Peking University, China

<sup>2</sup>Huawei Noah's Ark Lab, China

fengqianhan@stu.pku.edu.cn, lintong@pku.edu.cn, {liwenshuo, xinghao.chen}@huawei.com

## Abstract

Vision-Language Models (VLMs) bring powerful understanding and reasoning capabilities to multimodal tasks. Meanwhile, the great need for capable artificial intelligence on mobile devices also arises, such as the AI assistant software. Some efforts try to migrate VLMs to edge devices to expand their application scope. Simplifying the model structure is a common method, but as the model shrinks, the trade-off between performance and size becomes more and more difficult. Knowledge distillation (KD) can help models improve comprehensive capabilities without increasing size or data volume. However, most of the existing large model distillation techniques only consider applications on single-modal LLMs, or only use teachers to create new data environments for students. None of these methods takes into account the distillation of the most important cross-modal alignment knowledge in VLMs. We propose a method called Align-KD to guide the student model to learn the cross-modal matching that occurs at the shallow layer. The teacher also helps student learn the projection of vision token into text embedding space based on the focus of text. Under the guidance of Align-KD, the 1.7B MobileVLM V2 model can learn rich knowledge from the 7B teacher model with light design of training loss, and achieve an average score improvement of 2.0 across 6 benchmarks under two training subsets respectively.

## 1. Introduction

Vision Language Model (VLM) is an important Multimodal technology, which build a bridge between vision and text data, and facilitate many real world tasks and applications [14, 26]. Based on the success of Large Language Models (LLMs) [43, 44, 52], efforts have been done to integrate vi-

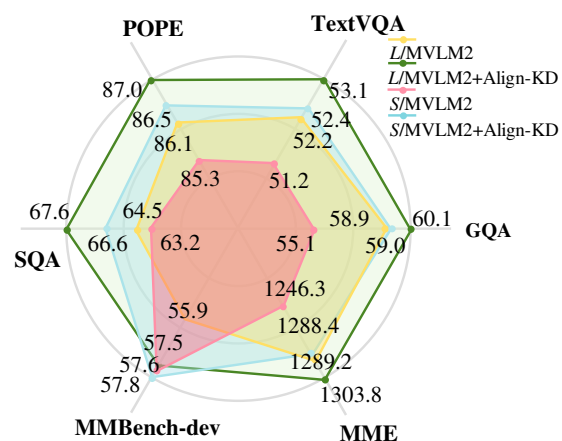


Figure 1. Radar plot of MobileVLM V2 1.7B model' performances with Align-KD policy under different settings. *L* and *S* refer to different *Long* and *Short* subdatasets for training, and MVLM2 refers to MobileVLM V2 model.

sion modal features with LLMs to extend models' capability and their application potential and build up new Vision-Language Models (VLMs) [12, 22, 25, 28, 56]. However, new issues rise up: as the input features become more complex, the structures of VLMs also become deeper and heavier, since they have to digest information from different modalities and face even more various scenes [2, 4]. The growing size and complexity of VLMs makes them difficult to be accessed outside the server or high-speed Internet, which limit the development of these cutting-edge artificial intelligence under different scenarios, especially their deployment in off-line devices like mobile phones and robots, or some confidential application devices.

Growing attentions have been focused on compressing VLMs while maintaining their remarkable capability as better as possible. MobileVLM family models [8, 10] are the first works to scale down VLMs to be able to run on

\*Work done during the internship at Noah's Ark Lab.

†Corresponding authors.

mobile devices. Both MobileVLM V1 and MobileVLM V2 model contain special designed lightweight downsample vision projector to embed tokens into text embedding dimensions. With novel training strategy, MobileVLM V2 1.7B model outperforms a bunch of VLMs at 7B scale. Even so, continue to scale down the model would encounter more tough situations, where more severe performance drop would occur. In this case, new strategies are in need to aid the development of mobile VLMs.

In contrast to cutting down the scale of mobile VLMs with the risk of significant performance drop, we are motivated to boost the model without enlarging the scale of data amount. Knowledge distillation (KD) method is widely used to increase the capability of neural networks, aims to instruct smaller student model with a larger and stronger teacher model, thus learning teacher’s behaviors or hidden representations [19]. Previous knowledge distillation method for large model are mainly designed for single NLP modality [18, 20, 24]. In the field of vision-language multi-modal models, most works are done before the large model era, focusing on aligning the vision proposals at the front side [13], which are no longer used in VLM technologies. Other works apply MSE loss on every transformer layers between models, but is not suitable for VLMs with significantly more layers with even different structures.

Among these works, we notice that the alignment of vision and text inputs, the most important aspect of VLMs, is not considered in distillation. Poorly aligned cross-modal features could lead to difficulties in comprehending or reasoning. In this paper, we propose a lightweight knowledge distillation method, namely Align-KD, to let 1.7B student model learn the alignment knowledge from much stronger teachers. Firstly, we conduct several experiments on well-trained VLMs, and find that the first and last Transformer layer brings the largest shift on the features, similar as the trends in LLMs [42]. This contributes helps us to develop our belief that the alignment of modalities mainly happens at the shallow layers, where the input embeddings are projected to high dimension space for comprehending and reasoning. Then, given that the natural cross-modal querying mechanism of Attention block, we let student mimic teacher’s text-query-vision attention distribution at the first layer. What’s more, considering that the importance of vision tokens in the queue are different according to different text prompts, we inject teacher’s informative vision embeddings unbalancedly into student’s vision projector’s output. Finally, we follow latest LLM research [18] to calculate reversed Kullback-Leibler divergence (R-KLD) between outputs to aid more general mean-seeking learning.

We apply Align-KD policy to distill MobileVLM V2 1.7B, the state-of-the-art open-source VLM for mobile devices, from MobileVLM V2 7B teacher. We formulate two different subdatasets with increasing limitation of prompt

maximum length, thus testing the effectiveness of Align-KD and VLMs under resource-limited scenarios. Following the multi-step training strategy of MobileVLM V2, Align-KD helps MobileVLM V2 1.7B model obtains universal promotion across 6 different benchmarks. The results show that Align-KD has great potential to help mobile VLMs to get enhancement with limited computation resource.

The main contributions of our work are the followings:

- We propose a knowledge distillation method Align-KD for mobile VLMs, which is the first work to distill the key cross-modal alignment knowledge.
- Align-KD helps cutting-edge MobileVLM V2 1.7B model obtains stable enhancement across different settings and benchmarks, largely facilitates the application of VLMs on edge devices.
- Align-KD doesn’t rely on specific design of VLMs, and only requires light training designs, which gives it great potential to expand to various resource limited scenes.

## 2. Related Works

### 2.1. Large Models and Their Boosting

In recent years, Large Language Models (LLMs) like GPT-3 [3], OPT [53] and LLaMA [43] significantly break through the borderline of deep learning and its applications. ChatGPT [36] set off a new wave and inspired follow-up work such as Vicuna [7]. Besides, some works try to introduce multimodal knowledge into the large model [2, 4, 6, 35, 49, 50, 56]. LLaVA [28] feed visual tokens into LLM and build up a comprehensive reasoning between visual and text contents, and many other works [12, 22, 25, 48] also approach to balance between vision and text understanding. However, the growing size of LLMs leads to a high demand on computing resources, which limits their applications. TinyLLaMA [52] and MobileLLaMA [9] scale down the architectures and maintain relatively good performance. Meanwhile in vision-language model field, MobileVLM family [8, 10] is the first open source work to facilitate the Vision-Language Model on mobile devices. Except for the development of training strategy and special architectures for large model, model compression techniques including quantization and pruning [15, 16] also thrive and provide solutions to relief burden of the computation resources. Equipped with these methods, LLMs are able to inference faster and lighter with little drop in accuracy.

### 2.2. Large Language Model Distillation

While former techniques are trying to do the subtraction, knowledge distillation (KD) [19] techniques are trying to do adding. In KD, weaker student model tries to learn from a stronger teacher model from different aspects, like the output or hidden representations. MiniLLM [18] studies the Kullback-Leibler divergence (KLD) loss on the output dis-

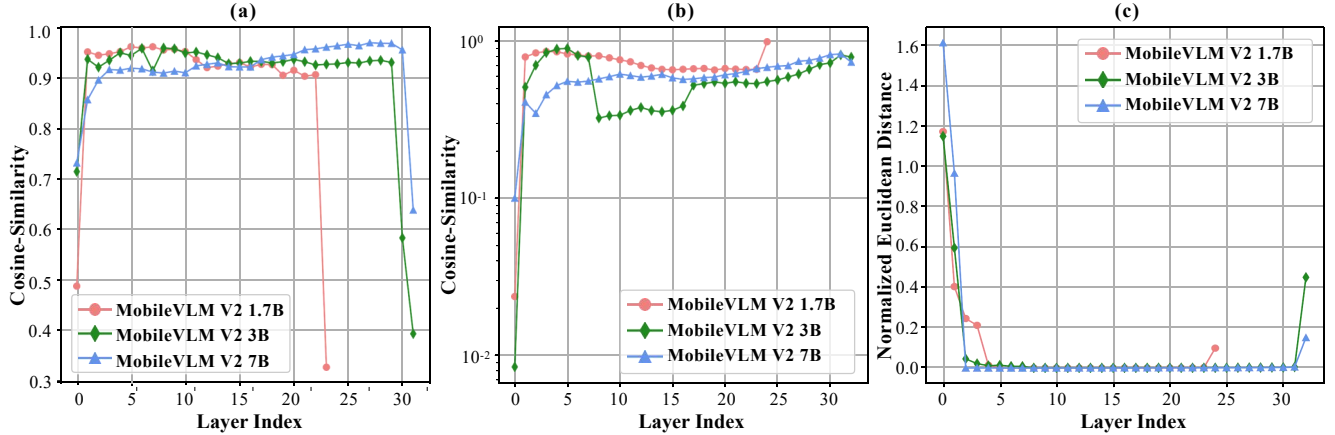


Figure 2. Exploration of feature changing trend at different layers of MobileVLM V2 model families. (a) Cosine similarity of features from every two adjacent layers. (b) Cosine similarity of features from original vision and text embedding positions within each same layer. The data is presented in an order of magnitude to highlight the trend of change. (c) Normalized Euclidean distance of features from original vision and text embedding positions within each same layer. The experiments are conducted on ShareGPT4V-PT dataset, and all the calculations are averaged after conducted on each tokens.

tillation, and suggests that reverse KLD teaches LLM student better on mean-seeking than forward KLD. Other researchers pay attention to the hidden features, training a task-aware filter to distill knowledge from teacher to student at different middle layers [24]. Teachers can also be used to create a more suitable data environment for student model, especially in LLM background where data are diverse and sometimes polluted. Hsieh et al. [20] devise a step-by-step distillation strategy to use teacher model’s inference ability to provide training data for student, thus injecting both label noise and text inference into the data. Meta delves deeply into black-box systems [51], distilling high-quality outputs generated by System 2 techniques, such as Chain-of-Thought, Rephrase and Respond, etc., back into the standard large language model generation.

### 2.3. Distillation for Vision-Language Model

Most distillation methods for Vision-Language Model (VLM) are designed before Large Model Era. Considering that traditional VLMs rely on vision proposals, Fang et al. [13] propose to align the input proposals between teacher and student, and enable following transformer blocks to align their attention distributions. To compress VLM, Wang et al. [47] combine pruning with distillation, conducting easy output logits imitation and distillation on attention and hidden states. Although these works provide a thinking of VLM distillation, they are restrained within the field outside Vision-Language Model (VLM), which usually consists of more transformer layers and more complex alignment between vision and language modalities. In VLM-KD [54], researchers use VLM like LLaVA-NeXT [30] to generate text prompts and using contrastive learning to promote long-tail recognition ability of vision mod-

els. LLaVA-MoD [39] minimizes the Kullback-Leibler divergence between output distributions and utilizes Direct Preference Optimization (DPO) to enhance the ability of the s-MLLM to discern high-quality from low-quality samples. But LLaVA-MoD relies on integrating the Sparse Mixture of Experts (MoE) architecture into language model, and also neglects the gap of alignment knowledge in distillation, which leaves a huge space for further explorations.

## 3. Align-KD

In Vision-Language Models (VLMs), vision and text embeddings comprise the input of the large model. However, it is obvious that the embedding mechanisms are different for the two modalities, which means the embeddings have to go through cross-modal alignment in the feature space. The cross-modal alignment ability is crucial for VLMs, but previous works mainly focus on single modal LLMs distillation and neglect the importance of teaching student about the alignment knowledge. Here we first explore the cross-modal alignment in VLMs, and then propose our Align-KD method step by step based on MobileVLM family.

### 3.1. Where Does the Alignment Happen?

Most VLMs like MobileVLM [8, 10] design special vision projectors to project the vision embeddings, but this operation mainly align the dimension of embedded tokens. The alignment of vision and text embeddings into the same high dimension space is almost a black box system.

Sun et al. [42] try to figure out the internal working mechanism of Transformer layers in LLMs. The researchers perform both skip and switch operations on every Transformer layer in LLaMA 2 7B, 13B and 70B [44] mod-



attention  $A_{t-t}$  and text-query-vision attention  $A_{t-v}$ .

Previous knowledge distillation methods usually teach student model to learn from the whole attention values of teacher’s different layers. On the contrary, we propose to let student model mimic the first layer’s text-query-vision attention ( $A_{1,t-v}$ ) only. Given a knowledgeable teacher model  $T$ , we pick out its attention matrix  $A_1^T$  from its first layer, then split out the text-query-vision part to form  $A_{1,t-v}^T$  and do the same to get student  $S$ ’s first layer’s text-query-vision attention  $A_{1,t-v}^S$ . Then, we apply an  $1 \times 1$  convolution projector  $P_{attn}$  to align the dimension of these two, followed by a simple mean square error (MSE) loss to get the first layer  $A_{1,t-v}$  only KD loss:

$$\mathcal{L}_{A_{1,t-v}} = MSE(P_{attn}(A_{1,t-v}^T), A_{1,t-v}^S). \quad (1)$$

This first layer  $A_{1,t-v}$  only KD has several advantages. (1) First and foremost, thanks to the cross-attention mechanism, the text-query-vision attention matrix of the first layer naturally implies how the input tokens of two modalities perceive each other, as well as the alignment projection scheme that tokens are going to take to project into more aligned feature space. On the contrary, it is difficult to extract useful knowledge from  $A_{1,t-t}$  when then text tokenizer of VLMs are usually fixed, while distilling  $A_{1,v-v}$  only lead to more vision-only enhancement, which has already been reinforced by front vision projector. And both of them lack the key cross-modal knowledge. (2) Secondly, first layer  $A_{1,t-v}$  only KD significantly reduces computing workload, making the whole method efficient. Some works distill attention matrixes at multiple layers, and some of them even device special downstream tasks to help the student learning, which can cause excessive computational pressure. Even within the first attention matrix, first layer  $A_{1,t-v}$  only KD also saves up to 50% calculation compared with full distillation. This lightweight design enable potential chances to conduct training with limited computation resources. (3) What is more, considering that the design of different VLMs could be different in both the model depth and the block details, first layer  $A_{1,t-v}$  only KD exhibits greater flexibility and can be easily migrated to different models regardless of their particular structural design or the meaning of different depth transformer layers.

### 3.3. Vision Enhancement Based on Text’s Focusing

The vision projector that generates the input vision tokens also takes up the responsibility to do rudimentary cross-modal alignment. For example, the LDP in MobileVLM model downsample the embeddings while extracting both detail and semantic features, and project the embeddings into less tokens with same dimension as text embeddings. This rough alignment lacks the perception of text modal information, only receiving backward gradients from downstream Transformer layers’ cross-attention to get in touch

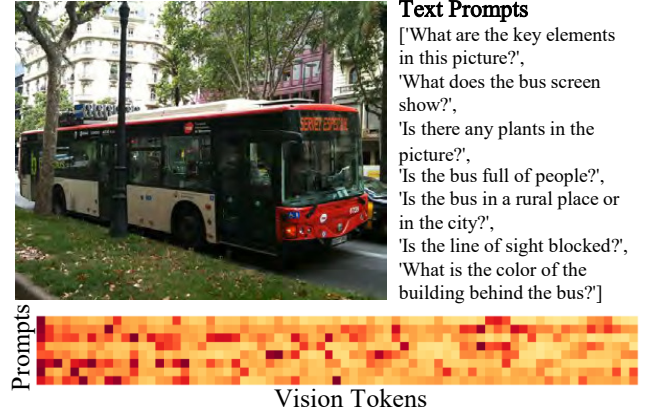


Figure 4. Different text prompts cause different attention on vision tokens of the picture. The vision tokens with high attention activation also distribute sparsely.

with another modality.

To mitigate this shortage, we propose to leverage the cross-modal attention to help the projector get more exposure to cross-modal information. Although the vision tokens are already refined by the projector, the text’s attention on different vision tokens is still rather sparse because of the directional indication of text prompts. The  $A_{1,t-v}$  not only contains information about cross-modal aligning projection, but also reveals which vision tokens the text prompts pay most attention to. Different prompts may focus on different embeddings, but some of the tokens are significantly left behind. Instead of inhibiting the learning of temporarily unpopular vision tokens, we propose to enhance most popular tokens at current based on teacher model’s  $A_{1,t-v}$ , thus preventing hurting the learning of others. Having the text-query-vision attention from the first layer of teacher model, we add-up the attention value  $A_{1,t-v}^T$  along the text dimension to get the attention score  $Score_n$  of vision token  $n$ ,

$$Score_n = \sum_{m=1}^M A_{1,t-v,(n,m)}^T, m \in M, \quad (2)$$

$M$  is the number of text tokens. Then, the indexes of tokens whose attention score  $Score_n$  if the top- $K$  are sorted out, namely  $Idx_K$ . Finally, we conduct knowledge injection from teacher to student on vision tokens listed in  $Idx_K$ :

$$\mathcal{L}_{V-focus} = MSE(P_V(Emb_{Idx_K}^T), Emb_{Idx_K}^S), \quad (3)$$

where  $Emb_{Idx_K}^T$  and  $Emb_{Idx_K}^S$  are teacher’s and student’s vision token embeddings within the range of  $Idx_K$ , and  $P_V$  is an  $1 \times 1$  convolution projector.

Except for the knowledge injection on current popular tokens, the rest should not be overlooked. Low-ranked attention from the current text prompt does not mean the stable low popularity under other scenes. The focus on vision

tokens may dramatically shift, e.g., when from ‘*What’s the name of book’s author on the cover?*’ to ‘*What is the color of the desk under the book?*’. What is more, the teacher vision token naturally contains stronger knowledge because of the instruction of stronger LLM and higher dimension in most cases. In this case, we add the general knowledge distillation on all vision tokens from teacher to student, i.e.,

$$\mathcal{L}_{V-all} = MSE(P_V(Emb_T), Emb_S), \quad (4)$$

which prevents harmful suppression of other tokens. And we combine two losses using a weight  $\lambda$  to form the vision token enhancement loss based on text’s focusing as

$$\mathcal{L}_V = \mathcal{L}_{V-all} + \lambda \mathcal{L}_{V-focus}. \quad (5)$$

### 3.4. Overall Knowledge Distillation Strategy

Except for alignment distillation in the front of VLM, we follow MiniLLM to add reverse Kullback-Leibler divergence (RKLD) loss, which uses the predicted distribution of student model as target, between the outputs of student and teacher. Since the text prompts and input images are varied, RKLD loss is more suitable than forward Kullback-Leibler divergence (FKLD) for VLM to learn about the mean-seeking instead of mode-seeking, preventing from overfitting on specific scene. Extract the output prediction distribution  $p_T$ ,  $p_S$  from teacher and student, the reverse Kullback-Leibler divergence loss can be formulated as

$$\mathcal{L}_{RKLD} = p_S \log \frac{p_S}{p_T}. \quad (6)$$

The alignment loss and RKLD loss work together with the original supervised loss  $\mathcal{L}_{Sup}$ , and our overall Align-KD loss on student model can be formulated as:

$$\mathcal{L} = \mathcal{L}_{Sup} + \mathcal{L}_{A_{1,t-v}} + \mathcal{L}_V + \mathcal{L}_{RKLD}. \quad (7)$$

## 4. Experiments

### 4.1. Basic Experiment Settings

MobileVLM family [8, 10] is the latest and cutting-edge Vision-Language Model for mobile devices. The MobileVLM V2 1.7B model shows remarkable performance with lightweight design. Considering the limitation of further compression of the model, we propose to apply Align-KD on MobileVLM V2 1.7B to help obtain better performance on edge devices. We choose the well trained MobileVLM V2 7B model as the teacher to conduct the knowledge distillation. Align-KD is used as an extra strategy working together with the common student training.

We use 8 NVIDIA V100 GPUs to conduct our distillation training. We follow MobileVLM V2’s work to divide the training stage into pre-training and multi-task finetuning. Because of the limited memory space, we use gradient

accumulation to achieve a global batch size of 256 for the pre-training stage and 128 for the multi-task finetuning stage. For both stage, we use the ZeRO2 strategy of DeepSpeed [1], and run all experiments under half-precision floating-point. We follow MobileVLM V2 to freeze the vision encoder and tokenizer while training, only train rest of the whole network. The maximum learning rate for the projector and other components are set to  $1e^{-3}$  and  $2e^{-5}$  respectively during pretraining and  $4e^{-5}$  during multi-task finetuning, using a cosine schedule. The convolutional projector for dimension alignment are randomly initialized and trained together. The weight  $\lambda$  in vision token loss  $\mathcal{L}_V$  is set to 0.1 to adapt to unstable changing of the tokens been focused. For top- $K$  selection in text-focus-based vision token knowledgeable distillation, we select the top 16 tokens with highest attention score. We follow the original MobileVLM V2 work on the rest of settings.

### 4.2. Data and Dataset Reforming

Align-KD follows MobileVLM V2 to train on various datasets. During the pretraining stage, ShareGPT4V-PT [4] is used to give the student a brief knowledge of vision and text. It is a caption dataset and comprises 1.2 million image-text pairs. In the multi-task training stage, more data from different tasks like conversation and VQA are provided: COCO[5], SBU[37], Visual Dialog[11], ShareGPT4V[4], SQA[34], IConQA[33], TextVQA[40], VSR[27], VIGC[46]. Note that SBU is a re-collected dataset and is updated from time to time, therefore some of the original data might have been removed. In this case, we washed the data list in the dataset, which is different from the original MobileVLM training. To evaluate the effectiveness of our Align-KD method, we test the performance on different benchmarks, including GQA[21], SQA[34], TextVQA[40], MME[17], MMBench[32] and POPE[23].

While mobile VLMs are designed for mobile devices deployment with limited resources, the training of VLMs also faces challenges when the computational resources is not that adequate. The computation workload can be extremely high when the input text prompts are too long, which sometimes causes ‘Out of Memory’ error during training. We formulate two different subdatasets based on the original data listed above, each contains data with different maximum prompt lengths: *Short* with maximum lengths of 512 embedded tokens, and *Long* of 2048. Considering that the training data comprises many Visual Question Answering (VQA) tasks, we drop the overlong data instead of truncating them, making two subsets also vary in the data amount. This setting is different from the original MobileVLM work, but can help examine the effectiveness of Align-KD strategy under different resource-limited scenarios, and also helps extend the application scenarios of VLMs. The subset details are shown in Table 2.

Method	LLM	#Samples	MME <sup>P</sup>	GQA	VQA <sup>T</sup>	POPE	MMB <sup>dev</sup>	SQA <sup>I</sup>	Avg.
MiniGPT-4	Vicuna-7B	5.0M	581.7	32.2	-	-	23.0	-	-
LLaVA-1.5[29]	Vicuna-7B	1.2M	1510.7	62.0	58.2	85.9	64.3	66.8	68.8
ShareGPT4V	Vicuna-7B	1.9M	1567.4	63.3	60.4	85.7	68.8	68.4	70.8
MoE-LLaVA-1.6B×4	StableLM-1.6B	2.2M	1300.8	60.4	47.8	84.3	59.4	62.6	63.3
MoE-LLaVA-2.7B×4	Phi-2.7B	2.2M	1396.4	61.1	50.2	85.0	65.5	68.7	66.7
MobileVLM 1.7B	MobileLLaMA 1.4B	3.6M	1196.2	56.1	41.5	84.5	53.2	57.3	58.7
MobileVLM 3B	MobileLLaMA 2.7B	3.6M	1288.9	59.0	47.5	84.9	59.6	61.2	62.8
MobileVLM V2 3B	MobileLLaMA 2.7B	3.6M	1440.5	61.1	57.5	84.7	63.2	66.7	68.1
MobileVLM V2 7B	Vicuna-7B	3.6M	1559.0	62.6	62.3	86.6	69.2	74.7	72.2
Subset / #Samples	Student Model	Align-KD	MME <sup>P</sup>	GQA	VQA <sup>T</sup>	POPE	MMB <sup>dev</sup>	SQA <sup>I</sup>	Avg.
<i>Short</i> / 3.6M	MobileVLM V2 1.7B	-	1246.3	55.1	51.2	85.3	57.6	63.2	62.4
	MobileVLM V2 1.7B	✓	<b>1288.4</b>	<b>58.9</b>	<b>52.4</b>	<b>86.5</b>	<b>57.8</b>	<b>66.6</b>	<b>64.4</b>
<i>Long</i> / 3.6M	MobileVLM V2 1.7B	-	1289.2	59.0	52.2	86.1	55.9	64.5	63.7
	MobileVLM V2 1.7B	✓	<b>1303.8</b>	<b>60.1</b>	<b>53.1</b>	<b>87.0</b>	<b>57.5</b>	<b>67.7</b>	<b>65.1</b>

Table 1. Test of Align-KD strategy’s effectiveness on MobileVLM V2 1.7B model. *Long* and *Short* refer to two subsets with different maximum prompt lengths limitations. MME<sup>P</sup> refers to MME Perception, MMB<sup>dev</sup> refers to MMBench-dev, SQA<sup>I</sup> refers to SQA-IMG. The score of MME<sup>P</sup> is divided by 20 when calculating the average performance.

Datasets	<i>Long</i> Samples	<i>Short</i> Samples
<b>Pretraining</b>		
ShareGPT4V-PT	1.25M	1.24M
<b>Multi-task Finetuning</b>		
COCO	592K	589k
SBU	837K*	822k*
Visual Dialog	123K	115K
ShareGPT4V	665K	655K
SQA	13K	5K
IConQA	107K	107K
TextVQA	35K	33K
VSR	13K	13K
VIGC	37K	35K
<b>Total</b>	<b>3.67M</b>	<b>3.61M</b>

Table 2. Details of datasets used in different stages. \*SBU dataset is re-collected, some datas are removed from the original links.

### 4.3. Effectiveness of Align-KD

After formulating our *Long* and *Short* subdatasets, we use them to train MobileVLM V2 1.7B model with proposed Align-KD strategy. Note that we use fully-trained and open-sourced MobileVLM V2 7B model provided by MobileVLM work as our knowledge distillation teacher. The results across 6 different benchmarks and two subsets are shown in Table 1. Trained with *Long* set, MobileVLM V2 1.7B model achieves an average score of 63.7. When trained with Align-KD policy, the student model witnesses a universal promotion across all benchmarks, achieving an average score of 65.1. To be more specific, Align-KD helps improve the MobileVLM V2 1.7B model to obtain an im-

provement of 3.2 on SQA benchmark, as well as 1.6 on MMBench. Under resource-limited scenarios, Align-KD gives MobileVLM V2 1.7B model an even better average improvement of 2 across all benchmarks, from 62.4 to 64.4. On GQA, Align-KD gives a promotion of 3.8. And it also brings notable promotion of 1.2 on POPE, which is a rather challenging hallucination testing benchmark.

The comprehensive results are visualized in a radar plot in Figure 1. Our Align-KD brings stable and good bonus under different vision-language tasks. What’s more, when the student suffers from the performance drop brought by the absence of long prompts, Align-KD successfully injects the knowledge into the model and makes student model after knowledge distillation achieve performances comparable with the model trained with long texts.

### 4.4. Ablation Study

We take a step forward to conduct ablation studies to testify the effectiveness of each method in Align-KD. We run all ablation experiments on *Short* subdataset for fairness, and the results are shown in Table 3. The reverse Kullback-Leibler divergence (RKLD) loss helps student learn to mimic the outputs of stronger teacher model, but brings somehow biased improvement. On the contrary, applying first layer  $A_{1,t-v}$  only loss provides notable and more balanced promotion to an average of 63.6. The combination of distillation on focused and all vision tokens further increase the performance by 0.8 on average.

During cross-modal alignment learning, we apply the knowledge distillation only on the text-query-vision part of the first layer’s attention. We further testify the rationality of this design by comparing with other methods, and the re-

Method	MME <sup>P</sup>	GQA	VQA <sup>T</sup>	POPE	MMB <sup>dev</sup>	SQA <sup>I</sup>	Avg.
MVLM2 1.7B	1246.3	55.1	51.2	85.3	57.6	63.2	62.4
+ $\mathcal{L}_{RKLD}$	1223.2	55.9	52.5	85.5	58.5	64.7	63.0
+ $\mathcal{L}_{A_{1,t-v}}$	1263.6	57.8	53.1	85.7	57.2	64.6	63.6
+ $\mathcal{L}_{V-all}$	1288.2	57.7	52.3	86.2	57.8	64.3	63.8
+ $\mathcal{L}_{V-focus}$	1288.4	58.9	52.4	86.5	57.8	66.6	64.4

Table 3. The ablation results of testing each component in Align-KD. MVLM2 refers to MobileVLM V2 model.

sults are shown in Table 4. Replacing first layer  $A_{1,t-v}$  only KD strategy with distillation of vision-query-vision attention triggers extreme drop in performances. We suggest that this is because the self-attention of vision tokens is not helpful for initial cross-modal alignment, and could lead to very unstable fluctuation in the attention embedding. However, changing into text-query-text attention is remarkably better than vision-query-vision training, with a slight drop from 64.4 to 63.2. We believe this phenomenon can be attributed to the fixed text tokenizer while training, and this lead to indirect stable learning of projection of cross-modal attention. What’s more, the learning on full attention performs better than learning only  $A_{1,v-v}$  and  $A_{1,t-t}$ , which further improves our conjectures above: the cross-modal  $A_{1,t-v}$  is the most important factor while learning  $A_{1,t-t}$  helps mitigate the negative influence brought by  $A_{1,v-v}$  distilling.

In Figure 2, we infer that the first and last layer act as the most outstanding characters. Align-KD mainly focuses on the first layer, since it is exposed the most to the inputs that need to be aligned, while the last layer is mainly in charge of projecting into output space. However, we also testify how the student would behave with the knowledge of output attention. As it is shown in the last row of Table 4, after adding the distillation of the last attention to Align-KD, student model witnesses a drop in performance. A possible explanation is that the deep feature is already well mixed and extracted, and there is some overlap with functionality of the RKLD on outputs. Besides, the distillation on extra attention would lead to significant growth in the calculation, which is a burden for resource-limited scenarios, and the first layer  $A_{1,t-v}$  only KD in Align-KD demonstrates huge advantage in both effectiveness and efficiency.

#### 4.5. Discussions and Limitations

Align-KD strategy brings benefit to MobileVLM V2 model under both long and short prompt limitations. Align-KD is relatively light designed, which enable possible expansion to resource-limited scenarios. We provide the working expense comparisons in Table 5, including the total training time and maximum memory occupied during training. The experiments are conducted on 8 Nvidia V100 GPUs. It costs around 296 GPU hours and 676 GPU hours to train Align-KD MVLM2 1.7B on *Short* and *Long* subdatasets.

Method	MME <sup>P</sup>	GQA	VQA <sup>T</sup>	POPE	MMB <sup>dev</sup>	SQA <sup>I</sup>	Avg.
$A_{1,t-v}$	1288.4	58.9	52.4	86.5	57.8	66.6	64.4
$A_{1,v-v}$	1036.5	52.0	27.3	81.1	13.8	36.1	43.7
$A_{1,t-t}$	1228.2	61.0	49.7	86.2	56.4	64.3	63.2
$A_{1,all}$	1265.0	59.8	52.7	86.3	54.4	64.6	63.5
+ $A_{last,all}$	1281.1	58.7	53.8	86.1	57.7	64.1	64.1

Table 4. The ablation of attention distillation strategy. ' $A_{1,t-v}$ ' is our first layer  $A_{1,t-v}$  only KD, and ' $A_{1,v-v}$ ', ' $A_{1,t-t}$ ' are in the same way. ' $A_{1,all}$ ' refers to applying attention distillation on full attention of the first layer. '+ $A_{last,all}$ ' refers to adding an extra KD loss on full attention of the last layer.

Settings	Total Training Time	Max Memory Occupied
<i>Short</i> MVLM2 1.7B	176 GPU hours	-
<i>Long</i> MVLM2 1.7B	228 GPU hours	-
<i>Short</i> w/ Align-KD	296 GPU hours	22.3 GB/device*
<i>Long</i> w/ Align-KD	676 GPU hours	30.7 GB/device†

Table 5. The comparison of training time and memory workload. 'w/ Align-KD' refers to training with Align-KD policy. \*Set batch size to 4 per iteration. †Set batch size to 1 per iteration.

For some examples that would cause instantaneous overload due to the gradient accumulation, we save the output representations from teacher and let them join training later when we can remove the teacher model from the device. When dealing with *Short*, Align-KD can achieve training with maximum memory workload of 22.3 GB per GPU, which is acceptable for commercial devices like RTX 3090. As for *Long* subset, the workload of 30.7 GB per GPU is also acceptable for devices like NVIDIA V100. The memory load of vanilla MobileVLM V2 1.7B is not presented here since it can be flexibly tuned. Despite the extra workload brought, Align-KD still enables universal improvements across different vision-language tasks.

## 5. Conclusion

We propose a knowledge distillation method for MobileVLM V2 model, namely Align-KD in this paper. Based on the conjecture that the alignment mainly happens at the front layer of LLM in VLMs, Align-KD proposes to conduct knowledge distillation only on the text-query-vision part of the first attention. The vision tokens are also unbalancedly enhanced according to the text tokens' focusing. Using MobileVLM V2 7B model as teacher, Align-KD enables universal improvements across benchmarks under both regular training setting and resource-limited setting.

## Acknowledgment

This work was supported by Sichuan Science and Technology Program (No.2024YFHZ0026).

## References

- [1] Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, et al. Deepspeed-inference: enabling efficient inference of transformer models at unprecedented scale. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15, 2022. 6
- [2] 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. 1, 2
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS 2020, Virtual*, 2020. 2
- [4] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*, abs/2311.12793, 2023. 1, 2, 6
- [5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 6
- [6] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Januspro: Unified multimodal understanding and generation with data and model scaling. *CoRR*, abs/2501.17811, 2025. 2
- [7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023. 2
- [8] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, and Chunhua Shen. Mobilevlm : A fast, strong and open vision language assistant for mobile devices. *CoRR*, abs/2312.16886, 2023. 1, 2, 3, 6
- [9] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023. 2
- [10] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. Mobilevlm V2: faster and stronger baseline for vision language model. *CoRR*, abs/2402.03766, 2024. 1, 2, 3, 6
- [11] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 6
- [12] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *CoRR*, abs/2401.16420, 2024. 1, 2
- [13] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *IEEE/CVF International Conference on Computer Vision, ICCV, Montreal, QC, Canada*, pages 1408–1418, 2021. 2, 3
- [14] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Injecting semantic concepts into end-to-end image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, New Orleans, LA, USA*, pages 17988–17998, 2022. 1
- [15] Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning, ICML, Honolulu, Hawaii, USA*, pages 10323–10337, 2023. 2
- [16] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: accurate post-training quantization for generative pre-trained transformers. *CoRR*, abs/2210.17323, 2022. 2
- [17] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiaowu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023. 6
- [18] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR, Vienna, Austria*, 2024. 2
- [19] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 2
- [20] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL, Toronto, Canada*, pages 8003–8017, 2023. 2, 3
- [21] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Long Beach, CA, USA*, pages 6700–6709, 2019. 6

- [22] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *CoRR*, abs/2301.13823, 2023. 1, 2
- [23] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP, Singapore*, pages 292–305, 2023. 6
- [24] Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning, ICML, Honolulu, Hawaii, USA*, pages 20852–20867, 2023. 2, 3
- [25] Bin Lin, Zhenyu Tang, Yang Ye, Jiayi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *CoRR*, abs/2401.15947, 2024. 1, 2
- [26] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. Interbert: Vision-and-language interaction for multi-modal pretraining. *CoRR*, abs/2003.13198, 2020. 1
- [27] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Trans. Assoc. Comput. Linguistics*, 11:635–651, 2023. 6
- [28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, NeurIPS, New Orleans, LA, USA*, 2023. 1, 2
- [29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Seattle, WA, USA*, pages 26286–26296, 2024. 7
- [30] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge. *arXiv preprint arXiv:2403.05530*, 2024. 3
- [31] Langming Liu, Liu Cai, Chi Zhang, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Yifu Lv, Wenqi Fan, Yiqi Wang, Ming He, Zitao Liu, and Qing Li. Linrec: Linear attention mechanism for long-term sequential recommender systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, Taipei, Taiwan*, pages 289–299, 2023. 4
- [32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV - 18th European Conference, Milan, Italy, Proceedings, Part VI*, pages 216–233, 2024. 6
- [33] Pan Lu, Liang Qiu, Jiaqi Chen, Tanglin Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks, virtual*, 2021. 6
- [34] 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. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems, NeurIPS, New Orleans, LA, USA*, 2022. 6
- [35] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *CoRR*, abs/2411.07975, 2024. 2
- [36] OpenAI. Chatgpt. Online, 2023. 2
- [37] Vicente Ordóñez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems, Granada, Spain*, pages 1143–1151, 2011. 6
- [38] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, and Chengqi Zhang. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI, Stockholm, Sweden*, pages 4345–4352, 2018. 4
- [39] 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
- [40] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Long Beach, CA, USA*, pages 8317–8326, 2019. 6
- [41] Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. Adaptive attention span in transformers. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL, Florence, Italy, Volume 1: Long Papers*, pages 331–335, 2019. 4
- [42] Qi Sun, Marc Pickett, Aakash Kumar Nain, and Llion Jones. Transformer layers as painters. *CoRR*, abs/2407.09298, 2024. 2, 3
- [43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. 1, 2
- [44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutie Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn,

- Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. [1](#), [3](#)
- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA*, pages 5998–6008, 2017. [4](#)
- [46] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, and Conghui He. VIGC: visual instruction generation and correction. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI, Vancouver, Canada*, pages 5309–5317, 2024. [6](#)
- [47] Tiannan Wang, Wangchunshu Zhou, Yan Zeng, and Xinsong Zhang. Efficientvlm: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning. In *Findings of the Association for Computational Linguistics: ACL, Toronto, Canada*, pages 13899–13913, 2023. [3](#)
- [48] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *CoRR*, abs/2312.06109, 2023. [2](#)
- [49] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *CoRR*, abs/2410.13848, 2024. [2](#)
- [50] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *CoRR*, abs/2412.10302, 2024. [2](#)
- [51] Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1. *CoRR*, abs/2407.06023, 2024. [3](#)
- [52] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tynllama: An open-source small language model. *CoRR*, abs/2401.02385, 2024. [1](#), [2](#)
- [53] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022. [2](#)
- [54] Zaiwei Zhang, Gregory P. Meyer, Zhichao Lu, Ashish Shrivastava, Avinash Ravichandran, and Eric M. Wolff. VLM-KD: knowledge distillation from VLM for long-tail visual recognition. *CoRR*, abs/2408.16930, 2024. [3](#)
- [55] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, NeurIPS, New Orleans, LA, USA*, 2023. [4](#)
- [56] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations, ICLR, Vienna, Austria*. OpenReview.net, 2024. [1](#), [2](#)