Self-Improving Teacher Cultivates Better Student: Distillation Calibration for Multimodal Large Language Models

Xinwei Li
Southeast University
Nanjing, China
seulixinwei@seu.edu.cn

Li Lin*
Southeast University
Nanjing, China
linli321@seu.edu.cn

Shuai Wang
Southeast University
Nanjing, China
shuaiwang@seu.edu.cn

Chen Qian
Tsinghua University
Beijing, China
qianc62@tsinghua.edu.cn

ABSTRACT

Multimodal content generation, which leverages visual information to enhance the comprehension of cross-modal understanding, plays a critical role in Multimodal Information Retrieval. With the development of large language models (LLMs), recent research has adopted visual instruction tuning to inject the knowledge of LLMs into downstream multimodal tasks. The high complexity and great demand for resources urge researchers to study efficient distillation solutions to transfer the knowledge from pre-trained multimodal models (teachers) to more compact student models. However, the instruction tuning for knowledge distillation in multimodal LLMs is resource-intensive and capability-restricted. The comprehension of students is highly reliant on the teacher models. To address this issue, we propose a novel Multimodal Distillation Calibration framework (MmDC). The main idea is to generate high-quality training instances that challenge student models to comprehend and prompt the teacher to calibrate the knowledge transferred to students, ultimately cultivating a better student model in downstream tasks. This framework comprises two stages: (1) multimodal alignment and (2) knowledge distillation calibration. In the first stage, parameter-efficient fine-tuning is used to enhance feature alignment between different modalities. In the second stage, we develop a calibration strategy to assess the student model's capability and generate high-quality instances to calibrate knowledge distillation from teacher to student. The experiments on diverse datasets show that our framework efficiently improves the student model's capabilities. Our 7B-size student model, after three iterations of distillation calibration, outperforms the current state-of-the-art LLaVA-13B model on the ScienceQA and LLaVA Test datasets and also exceeds other strong baselines in a zero-shot setting.

*Corresponding author

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657692

CCS CONCEPTS

• **Information systems** → **Question answering**; **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Computer vision**.

KEYWORDS

multimodal reasoning; knowledge distillation; large language models

ACM Reference Format:
Xinwei Li, Li Lin, Shuai Wang, and Chen Qian. 2024. Self-Improving Teacher Cultivates Better Student: Distillation Calibration for Multimodal Large Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657692

1 INTRODUCTION

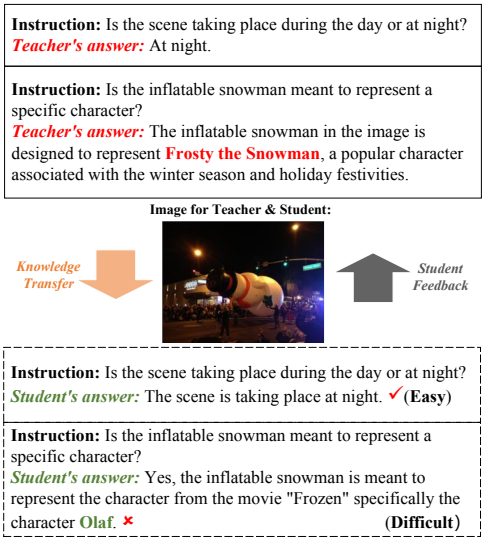

Figure 1: Comparison of our approach with previous methods.

Advancements in large language models (LLMs) significantly influence the domain of information retrieval (IR), encompassing a

wide range of practical applications in recommender systems [2], relevant feedback generation [26], and fact-checking [49]. To further enhance LLMs' capabilities, researchers focus on the development of open-source multimodal LLMs, the multimodal capabilities substantially extend the development of IR, as evidenced by innovations in multimodal recommendation [44] and information retrieval systems [8, 14]. The prevailing approach for constructing large multimodal models, known as instruction-tuning, involves distilling multimodal knowledge by aligning the responses of the more compact LLMs (student models) with those of the larger size LLMs (teacher models) in response to a set of instructions, such as BLIP-2 [21], LLaVA [23], and MiniGPT-4 [52].

The instructions are typically generated through GPT-4 [29] or derived from a manually constructed dataset. Constructing multimodal instructions using these methods is costly or requires significant effort [42]. Additionally, the instruction tuning-based knowledge distillation method is unidirectional, which causes the student models' capability restricted by the pre-trained teacher model, as illustrated by the orange arrow in Figure 1. The instructions formulated from the teacher model's answers allow the student model to learn how to discern the time of day in an image. Yet, the student model usually fails when tackling more challenging questions, such as "identifying the specific cartoon figure corresponding to the snowman in the image." Current model distillation methods overlook the assessment of the student's capability in the learning process, indicated by the gray arrow in Figure 1, which provides feedback about whether the student has learned the knowledge from the teacher model in specific instances. By incorporating this feedback, the teacher model can offer tailored training that concentrates on these difficult examples, thus improving the student model's performance. Recent studies also show evidence that the accuracy of student models has a strong correlation with knowledge calibration during distilling [35, 46]. Thus, to efficiently distill the knowledge from the large-scale teacher model to the compact student model for multimodal reasoning, it not only requires learning the feature mapping between different modalities but also calibrating knowledge distillation to provide high-quality instructions according to the student's feedback.

Aiming to address the two main challenges above, we propose a **D**istillation **C**alibration framework for **M**ulti**m**odal large language models (**MmDC**). We develop an assessment module to measure how well the student model learns from the teacher and prompt the teacher model to facilitate self-improvement in knowledge transferring by feeding the high-quality instructions of difficult instances to both teacher and student models. There are two stages in our framework as shown in Figure 2. In Stage 1, we conduct the **Multimodal alignment**, where parameter-efficient fine-tuning is used by updating a small set of parameters to enhance the feature alignment between image and textual modalities. In Stage 2, we design the three-phase **Multimodal distillation calibration** to prompt the teacher model to facilitate its self-improving and enhance the capability of the student model. Specifically, they are: 1) *Multimodal instruction tuning*, which only keeps the visual encoder frozen and continues to update the weights of the alignment layer using generated instructions, thereby enabling the teacher to better cultivate the student model in response to the challenging instructions when student does not perform well; 2) *Multimodal-assessment*, which identifies easy and difficult multimodal instructions by evaluating the performance of both student and teacher models; and 3) *Multimodal-augmentation*, which generates high-quality instructions and combine them with original images to build a new multimodal instruction dataset to train the student model. Essentially, the multimodal distillation calibration stage establishes a cycle of efficient training based on not only the knowledge transfer from teacher to student but also the student's feedback, which effectively enhances the multimodal capabilities of the student model.

To evaluate the effectiveness of our framework, we employ it to distill the knowledge from LLaVA-13B to a 7B-size student model (**DiLLaVA-7B**). Our dataset was initialized using llava-80K (which contains only images and corresponding instructions without answers). We conduct three iterations of distillation calibration, which generate 504K multimodal instructions. The experimental results show that our distillation calibration framework consistently improves the capabilities of the student model, and **DiLLaVA-7B** has superior performance surpassing multimodal large language model as LLaVA [23]. Our main contributions can be summarized as follows:

- Our work is the first attempt to adopt the idea of data distillation calibration to open-source multimodal large language models.
- Our proposed framework demonstrates impressive efficiency and effectiveness. Without any human annotations, our model outperforms the current SOTA model on multimodal reasoning and outperforms the baselines with larger parameter sizes in the zero-shot setting.
- Our distillation calibration method demonstrates the possibility of the student model outperforming the teacher model in downstream tasks, and it can be easily adapted to fit a variety of other open-source multimodal LLMs.

2 RELATED WORK

2.1 Multimodal Instruction Tuning

Instruction-tuning is an efficient method to train LLM by utilizing datasets with diverse NLP tasks, which has been successfully applied to well-known LLMs like InstructGPT [30] and FLAN-T5 [6]. It significantly improves the performance and generalization capabilities of the tuned models. Building on this success, instruction-tuning has been extended to the visual domain recently. MiniGPT4 [52] utilizes ChatGPT to enhance the detailed description of image captions and generate high-quality instruction data. LLaVA [23] generates multimodal instruction data by prompting plain text GPT-4 [29] along with bounding boxes of objects and image captions. LLaMA-Adapter [11, 50] aligns text and image features using the COCO dataset and leverages only text data for instruction tuning. mPLUG-owl [48] pre-trains the model with over 1000M image-text pairs and constructs a 400M hybrid dataset comprising plain text and multimodal instruction fine-tuning data. InstructBLIP [7] converts 13 visual language tasks into multimodal instruction tuning data format for instruction tuning. Such work usually constructs instruction datasets through closed-source large models or manual efforts, which is costly and labor-intensive. Therefore, it is crucial to prioritize the quality of instructions over their quantity as this

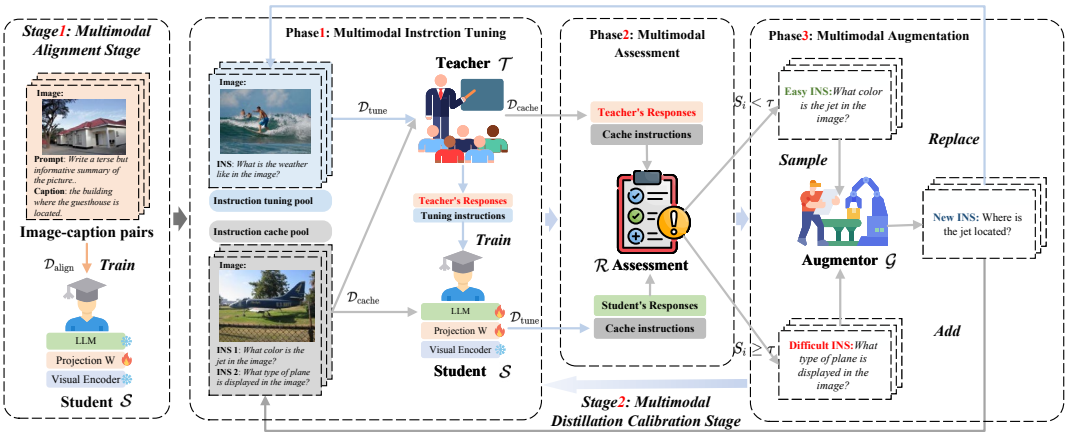

Figure 2: The overview of our multimodal distillation calibration framework. The first stage enhance fearure alignment between different modalities. The second stage generates high-quality instances to calibrate knowledge distillation from teacher to student, which consists of three phases in an iteration: 1) Multimodal instruction tuning; 2) Multimodal assessment; 3) Multi-modal augmentation.

can enhance the capability of the multimodal model and reduce the cost of constructing instruction data.

2.2 Knowledge distillation

Knowledge distillation (KD) aims to transfer knowledge from a teacher model to a student model. Currently, knowledge distillation can be classified into two categories: black-box distillation and white-box distillation [53]. In black-box KD, the student model only has access to the teacher's predictions, while white-box KD allows the student model to utilize the weights of the teacher model [53]. Typically, the prevailing distillation method for large language models is black-box distillation, which can be divided into three subcategories: In context learning (ICL) distillation [9, 41], Chain-of-Thought (CoT) distillation [34, 40, 43], and Instruction Following (IF) distillation [3, 17, 31]. ICL distillation transfers contextual few-shot learning and language modeling abilities from the teacher model to the student model [16]. In contrast, CoT distillation takes a different approach, MT-COT [22] aims to enhance the reasoning performance of the student model by utilizing the CoT generated by the teacher model. Step-by-Step distillation [13] employs chain-of-thought arguments generated by LLM as additional guidance for training student models within a multi-task framework. IF distillation [53] uses instruction data containing a variety of NLP tasks to train models. LaMini-LM [45] model utilizes ChatGPT as its teacher model and generates new instruction by prompting ChatGPT, resulting in a comprehensive dataset comprising 2.58 million instructions, covering a diverse array of topics. Although these methods successfully distill the knowledge of teacher models into the student models, they still strictly follow the unidirectional knowledge transfer, neglecting the potential benefits of learning from students' feedback to enhance teaching efficacy and facilitate the calibration of knowledge.

3 METHODOLOGY

The purpose of our work is to prompt the teacher model to facilitate self-improving by continuously challenging the questions that students struggle to solve, then calibrate the knowledge acquired by the student model, denoted as $\mathcal{S}$. Since the training of the student model in our framework varied in different stages, we first introduce the architecture and training methods of the student model. Then we illustrate the multimodal distillation calibration framework, which involves two stages: the multimodal alignment stage and the multimodal distillation calibration stage.

3.1 Architecture and training methods of student model

We design a unified multimodal model architecture to accept both textual and image inputs. Figure 3 illustrates the architecture of our student model $\mathcal{S}$. For the textual input, the student model is initialized using Vicuna-7B-1.1 [5], which was fine-tuned using supervised data based on LLaMA-7B [36]. We name it **DiLLaVA-7B**. For the image input, we utilize the pre-trained CLIP visual encoder ViT-L/14 [32] to extract the visual features.

The training data for our student model can be divided into two categories: single-turn dialogue instruction data and multi-turn dialogue instruction data. We unify the formats of training data to facilitate model training in different stages without the need to modify the model architecture. Specifically, the unified training instance X containing a pair of multimodal data can be denoted as:

$$\begin{aligned} X_i &= \{(V_i, C_i)\}_{i\in[1,N]} \\ C_i &= \left\{(Q_i^t, A_i^t)\right\}_{t\in[1,H]} \end{aligned} \tag{1}$$

where N is the total number of training instances in the dataset, V_i and C_i represent the image and text respectively. For each C_i, it may contain multiple turns of instruction Q_i^t and answer A_i^t, where the length H of C_i is the count of (Q, A) pairs. We organize the input U_i^t at the t-th turn as a unified format sentence C_i^t:

$$U_i^t = \begin{cases} [token(V_i), Q_1] \text{ or } [Q_1, token(V_i)], & t = 1 \\ [C_i^{t-1} || Q_t], & t > 1 \end{cases} \tag{2}$$

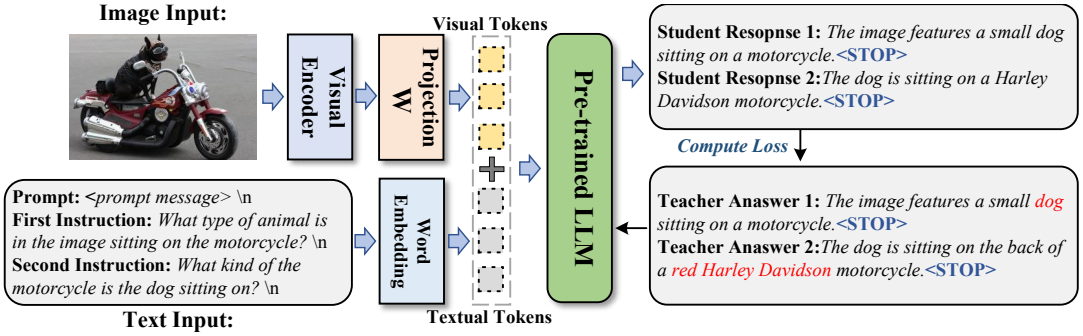

Figure 3: The architecture of DiLLaVA-7B and multi-turn dialogue instruction data training method. For the textual input, we choose Vicuna-7B-1.1 to isinitialize our pre-trained LLM, and utilize pre-trained CLIP visual encoder ViT-L/14 to extract the visual features. we demonstrate a specific example of training two turns of instruction data. DiLLaVA-7B is trained to both predict the answers and determine where to stop.

the initial input is a random order of the tokenized image data $token(V_i)$ and the first instruction Q_i^1. When the turns of instruction tuning grow, the instructions Q_i^t are concatenated to the end of text vectors continuously. Figure 3 illustrates a specific example of training using two turns of instruction data. The student model $\mathcal{S}$ is trained to both predict the answers and determine where to stop, so we add the <STOP> token to indicate the end of an instruction or answer as shown in Figure 3. Consequently, the loss of our model is computed using the response from $\mathcal{S}$, and the representation of <STOP> token.

We compute the probability of generating target answers A_i^t by:

$$p\left(A_i^t \mid V_i, Q_i^{<t}\right) = \prod_{m=1}^{M} p_\theta\left(x_m \mid U_i^t, A_{i,<m}^t\right) \tag{3}$$

Here θ represents the trainable parameter of the student model. U_i^t refers to the initial image and text, as well as the generated instructions in all previous turns $< t$. $A_{i,<m}^t$ represents the predicted tokens of the before m steps of auto-regressive decoder in the student model.

3.2 Multimodal alignment stage

The multimodal alignment stage aims to train the feature alignment layer (projection matrix $\mathbf{W}$) on a combined dataset $\mathcal{D}_{align}$, where we filter the samples from Conceptual Caption 3M [4], SBU [37], and LAION [33]. Specifically, we employ Spacy [1] to extract noun phrases from each caption in the combined dataset and calculate the frequency of each phrase. Noun phrases with frequencies less than 3 are excluded since they typically represent rare concepts and attributes that are already covered by other pairs. Pairs containing these excluded noun phrases are sequentially added to the candidate pool, starting with the noun phrase with the lowest remaining frequency. If a noun phrase occurs more than 100 times, we randomly select a subset of 100 pairs that contain that noun phrase. By applying this filtering method, the combined dataset yields approximately 885K image-text pairs.

The filtered dataset is then used to train the projection matrix $\mathbf{W}$ to initialize $\mathcal{S}$. We adopt the parameter-efficient fine-tuning solution by keeping the weights of the visual encoder and pre-trained LLM frozen and updating the projection layer $\mathbf{W}$. This stage can be interpreted as the process of training a visual tokenizer for the student model $\mathcal{S}$.

3.3 Multimodal distillation calibration stage

The multimodal distillation calibration stage consists of three phases: 1) Multimodal instruction tuning phase, which aligns students' responses with teachers' responses; 2) Multimodal Assessment, which identifies difficult instructions; and 3) Multimodal-Augmentation, which generates instructions to increase the challenges faced by student models. To achieve the cycle of distillation calibration, we introduce four roles, including Teacher ($\mathcal{T}$), Student ($\mathcal{S}$), Assessment Agent ($\mathcal{R}$), and Augmentor ($\mathcal{G}$). To maintain all raw inputs of training data as well as the generated instructions, we design two data pools: i) Instruction Tuning Pool and Instruction Cache Pool for $\mathcal{T}$ and $\mathcal{S}$ respectively. Figure 2 illustrates the establishment of four roles and two data pools in our framework. Inspired by the self-distillation solutions [10], where student and teacher models possess identical architectures. We make the models of different roles share identical architectures. Specifically, we initialize the Teacher $\mathcal{T}$, Assessment $\mathcal{R}$, and Augmentor $\mathcal{G}$ using the same multimodal open-source large model, i.e., LLaVA-13B [23]. Our data pools are built on LLaVA-80K [23], which consists of 80,000 multi-turn dialogue data. This dataset can also be represented as Formula (1). To initialize our multimodal instruction tuning pool, we first unfold LLaVA-80K to single-turn dialogue data, and then we remove the answers in the dialogue. This results in a single-turn multimodal instruction dataset denoted as $\mathcal{D}_{sin}$:

$$D_{sin} = \{(V_i, Q_i)\}_{i \in [1, |\mathcal{D}_{sin}|]} \tag{4}$$

where $|\mathcal{D}_{sin}| = 220,000$.

3.3.1 Multimodal instruction tuning phase. In this phase, we prompt the teacher model $\mathcal{T}$ to generate the corresponding answer $A_i = \mathcal{T}(V_i, Q_i)$ for each multimodal instruction in the instruction tuning pool $\mathcal{D}_{tune}$, which is initialized with $\mathcal{D}_{sin}$. It is further refreshed by replacing a current instruction with a newly generated one by Augmentor $\mathcal{G}$. Then we convert all single-turn dialogues (Q_i^t, A_i^t) corresponding to the image V_i into multi-turn dialogue form C_i as

[1]https://github.com/explosion/spaCy

defined in Equation 1. We use the training method in subsection 3.1 to instruction tuning our student model $\mathcal{S}$.

3.3.2 Multimodal assessment phase. The multimodal assessment phase begins with the instruction cache pool, denoted as $\mathcal{D}_{cache}$. It is also initialized with $\mathcal{D}_{sin}$, which continuously merges the generated instructions from Augmentor $\mathcal{G}$. Based on the data in the cache pool, we prompt the assessment agent $\mathcal{R}$ to evaluate the difficulty degree of multimodal instructions. To accomplish this, we input each multimodal instruction from the cache pool into both the $\mathcal{T}$ and $\mathcal{S}$, then we prompt each of them to generate an answer. Subsequently, we prompt the assessment to score the answers provided by the $\mathcal{T}$ and $\mathcal{S}$ as follows:

$$\begin{aligned} R_i^S &= \mathcal{R}\left(\mathcal{S}\left(V_i, Q_i\right) \mid \left(V_i, Q_i, \mathcal{T}\left(V_i, Q_i\right)\right)\right) \\ R_i^T &= \mathcal{R}\left(\mathcal{T}\left(V_i, Q_i\right) \mid \left(V_i, Q_i, \mathcal{S}\left(V_i, Q_i\right)\right)\right) \end{aligned} \tag{5}$$

where each $(V_i, Q_i) \in \mathcal{D}_{cache}$.

To mitigate any positional bias of the LLM referee [39], we repeat the process twice by exchanging the positions of the teacher's response and the student's response. The final score R_i^S and R_i^T are then calculated as the average of the two runs. Once the scores are obtained, we calculate the degree of difficulty as follows:

$$S_i = \frac{abs\left(R_i^S - R_i^T\right) + 1}{\max\left(R_i^S, R_i^T\right)} \tag{6}$$

This equation first calculates the absolute value of the difference between the scores of student and teacher, then normalizes it to the range of 0 to 1, so that the impact of the score difference on different levels is consistent. Finally, we obtain a score that can reflect the difficulty of the problem. The higher the score, the more difficult the instruction. We add 1 to smooth the numerator to avoid the situation when the final degree equals 0. For example, given two different instructions Q_m and Q_n, the teacher and the student achieve the difficulty score as R_m^S=1, R_m^T=1, and R_n^S=9, R_n^T=9. We expect the assessment agent to give the conclusion that Q_n is much more difficult than Q_m because both the student and teacher struggle to give the correct answer, which is consistent with common sense. But the final degrees will both get 0 without the additive smoothing, failing to distinguish the differences. Finally, we set the threshold $\tau = 0.33$ to classify instructions into two categories: difficult instructions ($S_i \geq \tau$), and easy instructions ($S_i < \tau$).

3.3.3 Multimodal augmentation phase. After assessing the difficulty degree of the multimodal instructions, the objective of the augmentation phase is to generate new instructions that differ in content but similar in difficulty degree to the original instructions. This process is conducted by prompting the augmentor $\mathcal{G}$ to generate new instructions based on each difficult instruction and its corresponding image. To alleviate catastrophic forgetting of the model and enhance the diversity of the instructions tuning pool, we sample the set of identified easy instructions to update the tuning pool and cache pool so that the number of difficult and easy instructions for training is balanced.

To ensure instruction diversity, each newly generated instruction is considered valid only if its ROUGE-L score with all other instructions of the corresponding image is below 0.7 following the previous work [17]. Finally, as described in Figure 2, the original instructions in the tuning pool are replaced with the new instructions, while simultaneously adding the newly generated instructions into the cache pool to enhance the student's capability when using the updated dataset to distill knowledge from the teacher.

4 EXPERIMENTS

4.1 Experimental Settings

In our experiment, we comprehensively evaluated the multimodal student model $\mathcal{S}$ after three iterations of distillation. We consider two types of tasks: fine-tuning and zero-shot reasoning. These tasks involve three distinct datasets, which encompass various capabilities such as complex reasoning, scene understanding, and scientific question answering. We selected these datasets as they allow us to closely align with the challenges presented by numerous other multimodal datasets, e.g. OK-VQA [27], within the constraints of our paper's length, and they provide most up-to-date test cases for our model.

4.1.1 Datasets. **ScienceQA** [25] is a large-scale multimodal dataset utilized for scientific question-answering, including three subjects, 26 themes, 127 categories, and 379 skills. ScienceQA is composed of plain text and text-image examples, containing 12,726, 4,241, and 4,241 examples for training, validation, and testing, respectively.

SEED-Bench [19] includes 19K multiple-choice questions, covering 12 evaluation dimensions across image and video modalities. We chose the visual modality (SEED-IMG) to evaluate our model under a zero-shot setting, which includes 9 dimensions and 14K multiple-choice questions.

LLaVA Test Set [23] comprises 90 multimodal questions, covering three categories: conversation, complex reasoning, and detail description. Primarily, the LLaVA Test Set evaluates the performance of the model in multimodal conversations.

4.1.2 Baselines. For ScienceQA dataset, we select powerful VQA models: MM-CoT Base &Large [51], as well as the SOTA multimodal LLM LLaVA-13B [23] (also our teacher model) and the Open AI GPT model (LLaMA-Adapter [50], GPT3.5 [28], GPT-4 [29]) as our baseline models. For the text-only baselines, we use the image caption to prompt the model.

For SEED-IMG dataset and LLaVA Test Set, we choose the mainstream 7B multimodal LLMs, including Otter [20], OpenFlamingo [1], MultiModal-GPT [12], mPLUG-Owl [48], LLaMA-Adapter V2 [11], InstructBLIP [7], GVT [38], VisualGLM [15], MiniGPT-4 [52], Ziya-Visual [24] and our teacher model LLaVA-13B [23]as our baseline models.

4.1.3 Implementation Details. The three roles in our MmDC framework: Teacher, Assessment, and Augmentor, use the prompt templates as shown in Table 6. The student model uses the standard prompt templates specified by each downstream task.

Our multimodal distillation calibration framework's second stage involved three iterations to strike a balance between computational efficiency and enhanced multimodal reasoning performance. Each iteration begins with the multimodal instruction tuning phase, where the student model is instruction-tuned to enhance its multimodal reasoning capabilities. Subsequently, the multimodal assessment phase evaluates the instructions in the instruction cache pool

Table 1: Performance comparison (accuracy %) of DiLLaVA-7B on scienceQA with baselines. (NAT = natural science, SOC= social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6/7-12 = grades 1-6/7-12).

Method	Subject			Context Modality			Grade		Average
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Human	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [28]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [28]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
GPT-4 [29]	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
LLaMA-Adapter [50]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [51]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [51]	**95.91**	82.00	**90.82**	**95.26**	88.80	**92.89**	92.44	90.31	91.68
LLaVA [23]	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
DiLLaVA-7B	91.83	**95.95**	88.91	90.91	**89.94**	91.08	**92.47**	**90.97**	**91.94**

to obtain their degrees of difficulty. Finally, during the multimodal augmentation phase, we create new instructions based on the degree of difficulty. Our instruction tuning pool sequentially increased by 84K, 90K, and 110K instruction-tuning data respectively, which resulted in our model being trained sequentially for three iterations. The instruction-tuning pool contains a total of 504K multimodal instruction data (initialized with 220K LLaVA single-turn multimodal instructions, by removing the answers), and the pre-training data set contains a total of 885K multimodal dialogue data $\mathcal{D}_{align}$.

We adopt AdamW as the optimizer, with the batch size and warmup ratio set to 16 and 0.03, respectively. For the first and second stages, we set the learning rate to 2e-3 and 2e-5. In our multimodal knowledge transfer framework, the temperature of the Teacher, Assessment, and Augmentor all are 0.5. For the ScienceQA dataset, we trained on the training set for 6 epochs and set the learning rate to 2e-5, keeping the remaining hyperparameters unchanged. And in the testing phase, for ScienceQA, SEED-IMG, and LLaVA Test Set, the temperatures are 0.5, 0.1, and 0.7 respectively. All our experiments were conducted on 6 V100 (32G) GPUs, we use DeepSpeed [47]and Xformer [18]to optimize GPU memory usage. The first stage takes around 37h, and the second stage takes around 340h.

4.2 Experimental Results

4.2.1 ScienceQA. We evaluate the performance of our framework on ScienceQA by comparing **DiLLaVA-7B** with the strong baseline methods and the current SOTA models. The results are shown in Table 1, where we find that the current LLMs, such as GPT3.5 (COT) [28], GPT4 [29], still underperform compared to humans in few-shot or zero-shot settings, indicating that ScienceQA still presents a significant challenge for these models. In contrast, existing supervised methods yield better results.

Notably, MM-CoT Large [51] achieves previous state-of-the-art results, with an average accuracy of 91.68%. LLaVA-13B [23], serves as our teacher model and adopts a model architecture similar to ours, which is more closely aligned with our work. The results suggest that LLaVA [23]remains competitive compared to MM-CoT Large [51], particularly in the SOC category. Our model achieves better feature alignment by pre-training on a richer dataset (885K compared to LLaVA's 556K). Importantly, based on our multimodal distillation calibration framework, **DiLLaVA-7B** is trained on a larger and higher-quality instruction dataset without any human annotation, allowing it to surpass LLaVA's performance in nearly all categories with fewer parameters (7B compared to LLaVA's 13B). Furthermore, **DiLLaVA-7B** outperforms the current SOTA method, MM-CoT Large [51], in the SOC, IMG, G1-6, G7-12 categories, as well as in the final average accuracy. This makes it the first model of 7B size to surpass MM-CoT Large [51].

By prompting the teacher model, our method enhances the teacher model's capability for self-improvement by continuously assigning it more and more challenging questions. These include calibrating the data and generating higher-quality instructional data, thereby transferring increasingly precise knowledge to the student model. The efficacy of our multimodal distillation calibration framework is corroborated by the results. Furthermore, our experimental findings support the notion that, *within a self-distillation paradigm, the student model has the potential to outperform the teacher model when high-quality data augmentation techniques are employed[35].*

4.2.2 SEED-IMG. We evaluate the multimodal reasoning performance of **DiLLaVA-7B** in the zero-shot setting on the SEED-IMG dataset [19]. We select the mainstream 7B-size models and our teacher model LLaVA-13B [23]as our baseline models. The results demonstrate that our proposed model, **DiLLaVA-7B**, exhibits competitive performance, achieving the highest accuracy in visual reasoning and text recognition. It surpasses the current SOTA model, InstructBLIP [7], by 16.62% and 3.52%, respectively, as well as the teacher model LLaVA-13B [23] by 3.03% and 8.82%, respectively. This underlines the superior visual reasoning and text recognition capabilities of **DiLLaVA-7B**, which is attributed to our multimodal knowledge distillation framework. Through continual iterative distillation, **DiLLaVA-7B** is trained with more instruction data, enhancing the model's understanding of different scenarios and text instructions. However, its performance in instance location and spatial relations is slightly inferior to InstructBLIP Vicuna [7] and LLaVA [23]. This may be due to LLaVA's larger parameter size, which is advantageous for fine-grained spatial position recognition, and InstructBLIP Vicuna's larger multimodal instruction dataset (16M), which enables the model to learn more visual knowledge.

Table 2: Performance comparison (accuracy %) of DiLLaVA-7B on the SEED-IMG with 7B multimodal LLMs and 13B LLaVA. (SUG = Scene Understanding, IIY = Instance Identity, IAS = Instance Attributes, ILN = Instance Location, ICG = Instance Counting, SRS = Spatial Relations, IIN = Instance Interaction, VRG = Visual Reasoning, TRN = Text Recognition).

Model	Language Model	SUG	IIY	IAS	ILN	ICG	SRS	IIN	VRG	TRN	Average
Otter [20]	LLaMA-7B	44.90	38.56	32.24	30.88	26.28	31.81	31.96	51.36	31.76	35.16
OpenFlamingo [1]	LLaMA-7B	43.86	38.12	31.28	30.06	27.30	30.59	29.90	50.15	20.00	34.51
MultiModal-GPT [12]	LLaMA-7B	43.64	37.85	31.45	30.78	27.34	30.14	29.90	51.36	18.82	34.54
mPLUG-Owl [48]	LLaMA-7B	49.68	45.33	32.52	36.71	27.26	32.72	44.33	54.68	18.82	37.88
LLaMA-Adapter V2 [11]	LLaMA-7B	45.22	38.50	29.30	33.03	29.67	35.46	39.18	51.96	24.71	35.19
InstructBLIP Vicuna [7]	Vicuna-7B	60.20	**58.93**	**65.63**	**43.56**	**57.05**	**40.33**	**52.58**	47.73	43.53	**58.76**
GVT [38]	Vicuna-7B	41.74	35.50	31.79	29.45	36.17	31.96	31.96	51.06	27.06	35.49
LLaVA [23]	Vicuna-13B	**63.43**	49.10	49.04	43.04	30.93	38.35	45.36	61.32	38.82	48.43
DiLLaVA-7B	Vicuna-7B	63.10	51.50	53.80	42.23	34.36	38.20	51.54	**64.35**	**47.05**	50.90

Table 3: Comparison of the results (Score rated by GPT-4) of DiLLaVA-7B on the LLaVA Test Set with other powerful baselines. Question classes: Con: conversation category. CR: complex reasoning category. DD: detail description category

Model	Con	CR	DD	AVG
VisualGLM [15]	65.8	80.6	64.5	70.3
MiniGPT-4 [52]	65.3	75.6	66.3	69.1
mPLUG-owl [48]	69.0	84.1	59.0	70.8
Ziya-Visual [24]	82.3	90.2	71.2	81.3
InstructBLIP [7]	82.2	90.2	68.4	80.7
LLaVA [23]	83.1	**96.5**	75.3	85.1
DiLLaVA-7B	**86.4**	93.0	**77.5**	**85.7**

Our model ranks second in average accuracy among all current 7B-size models, next to the InstructBLIP [7]. The superiority of InstructBLIP [7]is mainly due to its tuning data, which includes 16M multimodal samples (**30** times more than ours), covering a wide range of multimodal tasks, including OCR and visual reasoning QA data. Our work does not require the construction of a large instruction dataset through manual labor or other closed-source large models, hence our main contribution is orthogonal to that of InstructBLIP [7]. Our multimodal distillation framework is applicable to other open-source large models, allowing the performance to be continually improved at a minimal cost.

Finally, we found that multimodal LLMs still perform poorly on fine-grained visual reasoning tasks, such as Instance Counting, Spatial Relations, Instance Interaction, and Text Recognition. This suggests that fine-grained visual question-answering tasks still pose a significant challenge to multimodal large models. However, our DiLLaVA-7B shows improvement in fine-grained tasks compared to the teacher model LLaVA-13B [23], with increases in performance of 3.43%, 6.18%, and 8.23% respectively. This indicates that our method can help enhance the performance of models on fine-grained visual tasks.

4.2.3 LLaVA Test Set. As displayed in Table 3, the results of our proposed model, **DiLLaVA-7B**, are compared with other leading baseline models on the LLaVA Test Set across three question categories: conversation (Con), complex reasoning (CR), and detail description (DD). The scores are rated by GPT-4[29].

Table 4: Ablation study of the threshold τ.

Threshold τ	Science QA	SEED-IMG	LLaVA Test Set
0 (**w/o** easy. Inst.)	91.08	48.21	85.10
0.33 (**Ours**)	**91.94**	**50.90**	**85.70**
0.67	91.32	49.37	84.80
1 (**w/o** diff. Inst.)	91.06	48.79	85.30

In the conversation category, **DiLLaVA-7B** outperforms all other models with a score of 86.4. For the complex reasoning category, the highest score of 96.5 is achieved by LLaVA[23], with **DiLLaVA-7B** scoring slightly lower at 93.0. When it comes to the detail description category, **DiLLaVA-7B** again leads with a score of 77.5. This may be attributed to the fact that **DiLLaVA-7B**, in the iterative distillation process, is trained with more complex instructions (including detailed description tasks) as well as simpler instructions (conversation tasks). As a result, it is able to provide more detailed descriptions of images and generate richer dialogue content. Taking the average scores into account, **DiLLaVA-7B** demonstrates superior overall performance with a score of 85.7, slightly surpassing LLaVA's average score of 85.1 and other baseline models.

In summary, our proposed model, **DiLLaVA-7B**, exhibits robust performance across all categories, especially in conversation and detail description categories, indicating its effectiveness and versatility in handling different types of questions in the context of natural language processing.

4.2.4 Ablation Results. **The parameter τ that differentiates between difficult and easy instructions.** As shown in Table 4, We conducted a systematic investigation of τ ranging from 0.0 to 1.0, observing its impact on average performance across three datasets. τ = 0 implies that all newly generated instructions are considered difficult, thus excluding any simple instructions. Conversely, τ =1 implies that all newly generated instructions are considered simple, thereby excluding any difficult instructions. The experimental results demonstrate that a lack of diversity in difficult and simple instructions can lead to decreased model performance. Remarkably, our model exhibits optimal performance when τ = 0.33. This suggests that our parameter settings effectively discriminate between difficult and easy instructions.

Table 5: Ablation study of multimodal allignment stage.

Method	Science QA	SEED-IMG	LLaVA Test Set
w/o Allignment	86.43	42.31	78.88

Multimodal alignment stage. We skip the first stage and utilize all the generated instructions datasets to train the base model, maintaining consistent other parameters. As shown in Table 5, it is evident that the model's accuracy on the three datasets has significantly declined (-5.51%, -8.59%, -6.82%). This underscores the efficacy and efficiency of our alignment layer. Through the fine-tuning of a limited set of parameters, student model acquires extensive multimodal alignment knowledge during the initial stage, thereby establishing a solid foundation for enhancing multimodal reasoning capabilities in the subsequent stage.

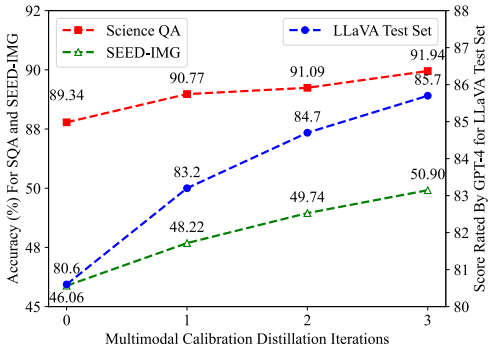

Figure 4: Performance of DiLLaVA-7B on ScienceQA, SEED-IMG, and LLaVA Test Set through the distillation iterations.

4.2.5 Effect of number of iterations. Figure 4 illustrates the performance of **DiLLaVA-7B** on the ScienceQA, SEED-IMG, and LLaVA test sets over three distillation calibration iterations. The results indicate a consistent enhancement of the student model's performance as the number of iterations escalates, with the most substantial improvement occurring in the initial iteration. This result underscores the efficacy of our multimodal two-stage distillation calibration framework.

5 CASE STUDY

5.1 Qualitative Comparison

To better understand the multimodal capabilities of our **DiLLaVA-7B** model, we selected representative examples from three different datasets and compared the responses of the **DiLLaVA-7B** model with those of the previous SOTA models within each dataset, we find that DiLLaVA-7B is superior in *fine-grained understanding*, *complex content perception* and *detailed description*. The comparison results are shown in the Figure 5.

In the first sample, the task is to analyze four different objects in the image (rock, tin foil, binder, ceramic mug) and infer their common characteristics. While the multimodal COT large model analyzed the four objects, it failed to correctly deduce their shared characteristics. In contrast, the **DiLLaVA-7B** correctly inferred that "An opaque object does not let light through. All four objects are opaque." In the second example, the task involves perceiving complex motion states, and then determining the color of a sportsman's gloves. The InstructBLIP model incorrectly identified the color of one sportsman's wristband (white) as the answer. However, the **DiLLaVA-7B** correctly identified and discerned the color of the gloves (black) in the image, demonstrating superior comprehension of complex images. The third example requires a detailed description of the image content. We observed that while the LLaVA model described some of the main elements in the image (elephant, sandy area), its answer is hallucinatory, including a non-existent person as a key element in its description. On the other hand, the **DiLLaVA-7B** avoided this issue, accurately described the main elements (elephant, dirt area), and also identified the "tire" element that was missing in the answer.

5.2 Error Analysis

To further understand the behavior of the **DiLLaVA-7B** and facilitate future studies, we present some errors that both the **DiLLaVA-7B** and previous SOTA models tend to make. These errors include challenges in *counting numbers*, *event reasoning*, and *complex image description*, as illustrated in Figure 6. In the first example, although neither multimodal COT large nor **DiLLaVA-7B** identified the average velocity information of the particles, **DiLLaVA-7B** provided a more detailed inference path. In the second example, **DiLLaVA-7B** accurately identified the gingerbread house in the image, but due to the "celebrations and entertainment" information associated with this element, **DiLLaVA-7B** mistakenly inferred that the scene in the image took place in a living room. In the third example, facing a complex image description task, while **DiLLaVA-7B** lacked descriptive details about the state and behavior of the people in the image, it correctly described the main elements in the image, including "cars, motorcycle, truck, large orange statue."

6 CONCLUSION

This paper proposes a novel framework for multimodal large model knowledge distillation, addressing the challenge of the expensive and labor-intensive construction of multimodal instructions and the capability restriction caused by the unidirectional learning of instruction tuning-based knowledge distillation. Our method prompts the teacher model to facilitate its self-improving based on the student's feedback, which in turn constructs a multimodal distillation calibration cycle, effectively enhancing the student model's capabilities. Our work is the first to apply distillation calibration to open-source multimodal LLMs. With the ability of high-quality data augmentation without any human annotations, our model outperforms the current state-of-the-art models on the reasoning task and even larger parameter size models in the zeroshot setting. This framework can be adapted to fit a variety of other open-source multimodal LLMs, paving the way for further advancements in this field.

7 ACKNOWLEDGMENTS

This work was supported in part by Science and Technology Innovation 2030-Major Project under Grant 2021ZD0114202, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20230815.

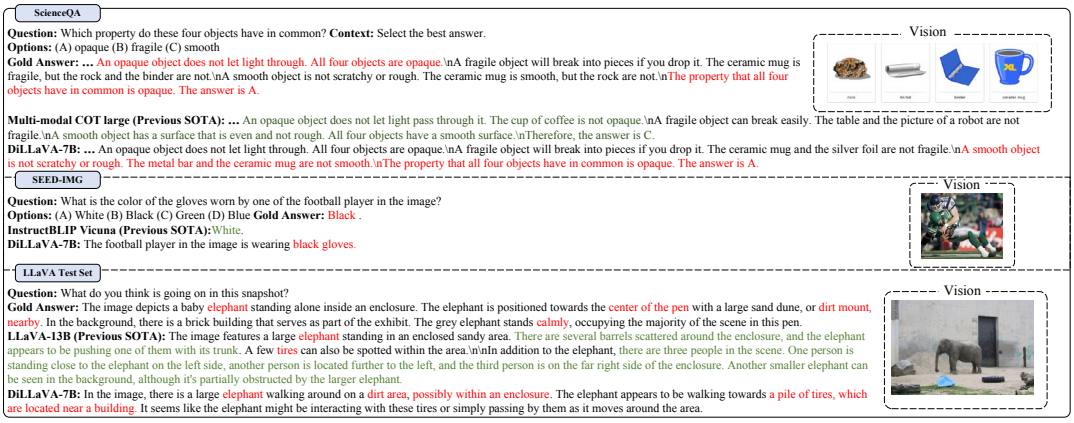

Figure 5: Qualitative comparisons of multimodal instructions. We compare the performance of DiLLaVA-7B with previous SOTA models on selected samples across three different datasets, with the correct responses marked in red and incorrect marked in green.

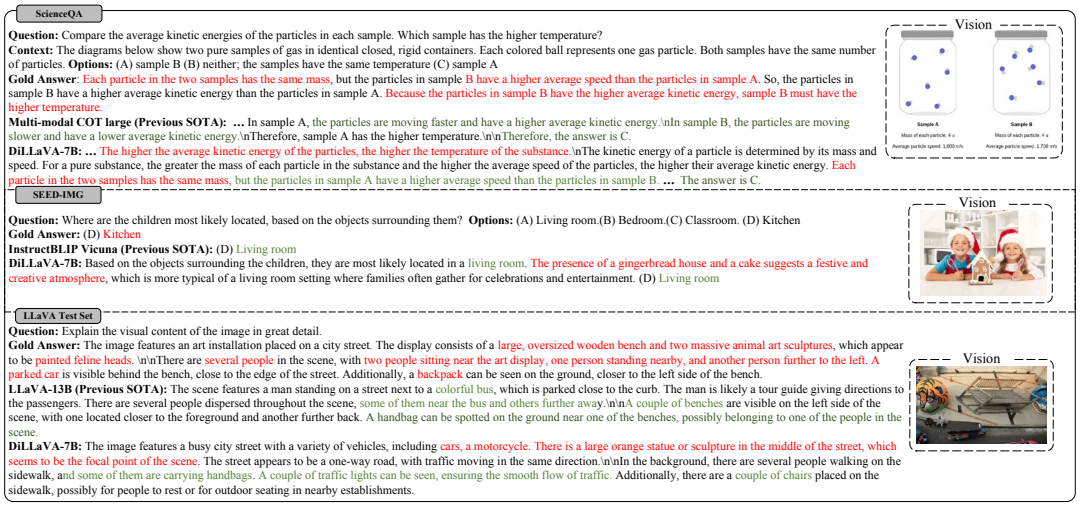

Figure 6: Error analysis of multimodal instructions. We compare the performance of DiLLaVA-7B with previous SOTA models on selected samples across three different datasets, with the correct responses marked in red and incorrect marked in green.

Table 6: Prompt template of Teacher, Assessment and Augmentor for generating responses.

Teacher	**System Content**	You are a helpful assistant. You are able to understand the visual content that the user provides, and answer the uesr questions based on the image content.
	User Content	[Question] \n {*User Question*}
Assessment	**System Content**	We would like to request your feedback on the performance of two AI assistants in response to the user question displayed following. The user asks the question on observing an image.
	User Content	[Question] \n {*User Question*} \n [Assistant 1] \n {*Assistant 1 Answer*} \n [End of Assistant 1] \n [Assistant 2] \n {*Assistant 2 Answer*} \n [End of Assistant 2] Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment. Please output in the following format: "The scores for Assistant 1 and 2 are [Assistant 1 score] and [Assistant 2 score], respectively."
Augmentor	**System Content**	We would like you to act as a Question Creator, and you are seeing a single image. Your goal is to draw inspiration from the given question.
	User Content	[Question] \n {*User Question*} Please Design a Created Question between you and a person asking about this image. Please output a created question according to the following requirements: (1)The difficulty level of the Created Question must be similar to the Given Question. (2)The Created Question must be based on the content in the image. (3)The output must contain only one single sentence and do not output the answer.

REFERENCES

[1] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. arXiv:2308.01390 [cs.CV]

[2] Keqin Bao, Jizhi Zhang, Yang Zhang, Wang Wenjie, Fuli Feng, and Xiangnan He. 2023. Large Language Models for Recommendation: Progresses and Future Directions. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 306–309.

[3] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning To Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18392–18402.

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3558–3568.

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling Instruction-Finetuned Language Models. arXiv:2210.11416 [cs.LG]

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500 [cs.CV]

[8] Yashar Deldjoo, Johanne R Trippas, and Hamed Zamani. 2021. Towards multimodal conversational information seeking. In *Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval*. 1577–1587.

[9] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. arXiv:2301.00234 [cs.CL]

[10] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International Conference on Machine Learning*. PMLR, 1607–1616.

[11] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. arXiv:2304.15010 [cs.CV]

[12] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. MultiModal-GPT: A Vision and Language Model for Dialogue with Humans. arXiv:2305.04790 [cs.CV]

[13] Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. 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 2023*. Association for Computational Linguistics, Toronto, Canada, 8003–8017. https://doi.org/10.18653/v1/2023.findings-acl.507

[14] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. 2019. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 635–644.

[15] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language Is Not All You Need: Aligning Perception with Language Models. arXiv:2302.14045 [cs.CL]

[16] Yukun Huang, Yanda Chen, Zhou Yu, and Kathleen McKeown. 2022. In-context Learning Distillation: Transferring Few-shot Learning Ability of Pre-trained Language Models. arXiv:2212.10670 [cs.CL]

[17] Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. Lion: Adversarial Distillation of Proprietary Large Language Models. arXiv:2305.12870 [cs.CL]

[18] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, and Daniel Haziza. 2022. xFormers: A modular and hackable Transformer modelling library. https://github.com/facebookresearch/xformers.

[19] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. arXiv:2307.16125 [cs.CL]

[20] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023. Otter: A Multi-Modal Model with In-Context Instruction Tuning. arXiv:2305.03726 [cs.CV]

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs.CV]

[22] Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from Large Language Models Make Small Reasoners Better. arXiv:2210.06726 [cs.CL]

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. arXiv:2304.08485 [cs.CV]

[24] Junyu Lu, Dixiang Zhang, Xiaojun Wu, Xinyu Gao, Ruyi Gan, Jiaxing Zhang, Yan Song, and Pingjian Zhang. 2023. Ziya-Visual: Bilingual Large Vision-Language Model via Multi-Task Instruction Tuning. arXiv:2310.08166 [cs.CL]

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

[26] Iain Mackie, Shubham Chatterjee, and Jeffrey Dalton. 2023. Generative Relevance Feedback with Large Language Models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Taipei, Taiwan,) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2026–2031. https://doi.org/10.1145/3539618.3591992

[27] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 3195–3204.

[28] OpenAI. 2023. ChatGPT. https://chat.openai.com.

[29] OpenAI. 2023. GPT-4 Technical Report.

[30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[31] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems(NIPS)* 35 (2022), 27730–27744.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[33] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. arXiv:2111.02114 [cs.CV]

[34] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language Models are Multilingual Chain-of-Thought Reasoners. arXiv:2210.03057 [cs.CL]

[35] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. 2021. Does Knowledge Distillation Really Work?. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 6906–6919. https://proceedings.neurips.cc/paper_files/paper/2021/file/376c6b9ff3bedbbea56751a84fffc10c-Paper.pdf

[36] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[37] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. 2016. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 816–832.

[38] Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. 2023. What Makes for Good Visual Tokenizers for Large Language Models? arXiv:2305.12223 [cs.CV]

[39] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large Language Models are not Fair Evaluators. arXiv:2305.17926 [cs.CL]

[40] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171 [cs.CL]

[41] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023. Large Language Models Are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. arXiv:2301.11916 [cs.CL]

[42] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. arXiv:2212.10560 [cs.CL]

[43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing*

Systems(NIPS) 35 (2022), 24824–24837.

[44] Chuhan Wu, Fangzhao Wu, Tao Qi, Chao Zhang, Yongfeng Huang, and Tong Xu. 2022. Mm-rec: Visiolinguistic model empowered multimodal news recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2560–2564.

[45] Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions. arXiv:2304.14402 [cs.CL]

[46] Lehan Yang and Jincen Song. 2021. Rethinking the Knowledge Distillation From the Perspective of Model Calibration. arXiv:2111.01684 [cs.CV]

[47] Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, Zhongzhu Zhou, Michael Wyatt, Molly Smith, Lev Kurilenko, Heyang Qin, Masahiro Tanaka, Shuai Che, Shuaiwen Leon Song, and Yuxiong He. 2023. DeepSpeed-Chat: Easy, Fast and Affordable RLHF Training of ChatGPT-like Models at All Scales. arXiv:2308.01320 [cs.LG]

[48] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. arXiv:2304.14178 [cs.CL]

[49] Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large Language Models Are Versatile Decomposers: Decomposing Evidence and Questions for Table-Based Reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Taipei, Taiwan,) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 174–184. https://doi.org/10.1145/3539618.3591708

[50] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. arXiv:2303.16199 [cs.CV]

[51] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal Chain-of-Thought Reasoning in Language Models. arXiv:2302.00923 [cs.CL]

[52] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592 [cs.CV]

[53] Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2023. A Survey on Model Compression for Large Language Models. arXiv:2308.07633 [cs.CL]