HAWAII: Hierarchical Visual Knowledge Transfer for Efficient Vision-Language Models

Yimu Wang, Mozhgan Nasr Azadani, Sean Sedwards, Krzysztof Czarnecki
University of Waterloo, Canada
{yimu.wang,mnasraza,sean.sedwards,k2czarne}@uwaterloo.ca

Abstract

Improving the visual understanding ability of vision-language models (VLMs) is crucial for enhancing their performance across various tasks. While using multiple pretrained visual experts has shown great promise, it often incurs significant computational costs during training and inference. To address this challenge, we propose HAWAII, a novel framework that distills knowledge from multiple visual experts into a single vision encoder, enabling it to inherit the complementary strengths of several experts with minimal computational overhead. To mitigate conflicts among different teachers and switch between different teacher-specific knowledge, instead of using a fixed set of adapters for multiple teachers, we propose to use teacher-specific Low-Rank Adaptation (LoRA) adapters with a corresponding router. Each adapter is aligned with a specific teacher, avoiding noisy guidance during distillation. To enable efficient knowledge distillation, we propose fine-grained and coarse-grained distillation. At the fine-grained level, token importance scores are employed to emphasize the most informative tokens from each teacher adaptively. At the coarse-grained level, we summarize the knowledge from multiple teachers and transfer it to the student using a set of general-knowledge LoRA adapters with a router. Extensive experiments on various vision-language tasks demonstrate the superiority of HAWAII compared to popular open-source VLMs. The code is available at https://github.com/yimuwangcs/wise-hawaii.

1 Introduction

Vision-language models (VLMs) [1, 2] enable machines to perform complex reasoning over multimodal inputs by combining the powerful language reasoning capabilities of pretrained large language models (LLMs) [3, 4, 5] with the rich perceptual understanding offered by vision foundation models [6, 7, 8]. These two components are connected through alignment modules, such as Q-Formers [9] or MLP projections [10], which map visual tokens into a representation space compatible with LLMs. At the heart of this pipeline, the vision encoder plays a central role, as its ability to extract semantically rich visual features directly impacts the generation and reasoning capabilities of the VLM.

Recent studies have shown that incorporating multiple vision experts improves performance by a large margin [11, 12, 13, 14, 15]. Nevertheless, these gains in effectiveness often come at the cost of efficiency [16, 17, 18, 19, 20]: multi-expert setups require computing visual tokens from all vision experts during both training and inference, making them expensive and less practical for deployment, especially in latency-sensitive or resource-constrained settings [21, 22, 23]. As a result, there is growing interest in approaches that can retain the benefits of multiple vision experts while avoiding their substantial inference-time costs.

Knowledge distillation (KD) [24], as a general framework for transferring knowledge from a larger model (teacher) to a smaller model (student), has been widely used in various domains [25, 26, 27, 28]. As a pioneer study of KD in VLMs, MoVE-KD [29] distills knowledge from multiple visual experts

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

into a single vision encoder using a *fixed set of Low-Rank Adaptation (LoRA) adapters [30]* for all teachers, enhancing visual understanding while only adding a small set of trainable parameters. However, learning from multiple teachers is challenging [31, 32], as the training data, model architecture, and training objectives of each teacher could be different. It can lead to noisy and redundant knowledge transfer, which can hinder the learning process with suboptimal performance [33].

To this end, we propose a novel **h**ierarchical visu**a**l kno**w**ledge tr**a**nsfer method for eff**ici**ent VLMs, namely HAWAII. It is designed to distill knowledge from multiple visual experts, *i.e.*, SAM [6], ConvNext [34], EVA [8], and Pix2Struct [35], into a single vision encoder, specifically, CLIP's vision encoder, enabling it to inherit the complementary strengths of these experts with minimal computational overhead. HAWAII consists of a novel *mixture-of-LoRA-adapter* (MOLA) module and a *hierarchical knowledge distillation* (HKD) mechanism that enables the student encoder to distill knowledge at coarse-grained and fine-grained levels.

Fine-grained distillation. As each teacher's knowledge is different, due to the heterogeneity of training data, architecture, and optimization methods, in MOLA, teacher-specific LoRA adapters are employed to avoid conflicts between teachers' knowledge. Each adapter is aligned with its teacher separately, allowing the student encoder to learn from diverse teachers while mitigating noisy distillation. Moreover, to emphasize the informative tokens generated by each teacher, at the fine-grained level, HKD utilizes a new token importance scoring method, which assigns weights to tokens according to the similarity to the text instructions and visual features.

Coarse-grained distillation. To obtain the collective consensus among visual teachers, HKD summarizes the knowledge from multiple teachers using a projector. Then, MOLA incorporates a set of general-knowledge LoRA adapters and a router to align the student with the collective consensus for a global alignment.

In summary, the main contributions of this work are:

- We propose HAWAII, a novel framework that distills knowledge from multiple pretrained visual experts into a single vision encoder, improving the visual understanding ability of VLMs without incurring substantial computational overhead.
- The proposed MOLA module consists of teacher-specific LoRA adapters and general-knowledge LoRA adapters that enable the student encoder to learn from diverse teachers separately (fine-grained) and globally (coarse-grained), avoiding noisy and redundant knowledge transfer.
- HKD distills knowledge from multiple teachers at coarse-grained and fine-grained levels. At the fine-grained level, HKD utilizes teacher-specific LoRA adapters and token importance scoring to select and learn from the most informative tokens from each teacher, as indicated by the visual and text tokens. At the coarse-grained level, HKD summarizes the knowledge from multiple teachers and transfers it to the student encoder globally using general-knowledge LoRA adapters.
- Extensive experiments on various vision-language tasks [36, 37, 38, 39, 40, 41, 42, 43, 44, 45] show that HAWAII achieves better performance on all the benchmark datasets compared to the baseline model (LLaVA-1.5 [46]). In particular, the performance on VizWiz, SQA, and MMBench is improved by 7.8%, 5.5%, and 4.0%, respectively.

2 HAWAII

In this section, we introduce the HAWAII framework, which learns from multiple powerful visual teachers for a better visual perception ability. HAWAII inherits the complementary strengths of several experts without incurring substantial computational overhead. First, we introduce the architecture of HAWAII. Second, we present the details of MOLA, which consists of a set of teacher-specific and general-knowledge LoRA adapters in Section 2.2. Last, we provide the details of our hierarchical knowledge distillation method, which contains the coarse- and fine-grained distillation in Section 2.3.

2.1 Architecture

The overall architecture of HAWAII is presented in the upper part of Figure 1. It follows the general design (vision expert-projector-LLM) of existing MLLMs [2, 10, 47]. The vision expert is trained

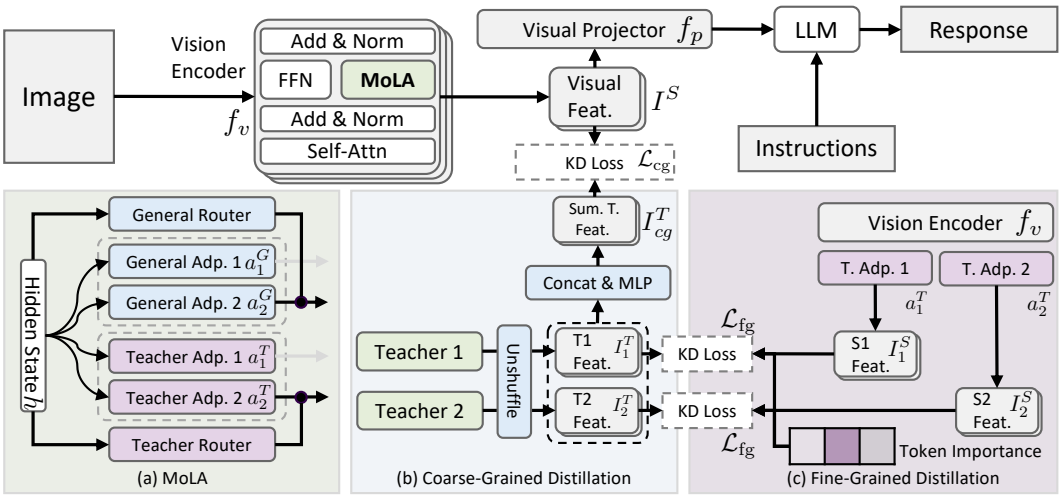

Figure 1: The overall architecture of HAWAII. We use two teachers for simplicity. (a) MOLA (Section 2.2) consists of teacher-specific LoRA adapters (Teacher Adp.) and general-knowledge LoRA adapters (General Adp.) with two routers controlling the activation of adapters. (b) Coarse-grained distillation (Section 2.3.1) first summarizes the knowledge from multiple teachers and then transfers it to the student encoder globally. "T1 Feat.", "T2 Feat,", and "Sum. T. Feat." represents the visual features I_*^T generated by different teachers and the summarized teacher features I_{cg}^T. (c) In the fine-grained distillation (Section 2.3.2), teacher-specific LoRA adapters (T. Adp.) and token importance scoring (Figure 2) are employed to select and learn from the most informative tokens.

to distill knowledge from multiple pretrained vision experts and produces visual tokens used for visual comprehension. The projector maps the visual tokens to the LLM input space, and the LLM generates the instruction-following response.

The *vision encoder* $f_v(\cdot)$ takes the input image and generates a set of visual tokens $I^S \in \mathcal{R}^{m \times D}$, where m is the number of visual tokens and D is the dimension of each token. To boost the performance, instead of using multiple vision encoders [19, 47, 48], which would be computationally expensive, only one student vision encoder is employed. And, it is trained to distill knowledge from multiple pretrained vision experts [6, 7]. We introduce the mixture-of-LoRA-adapter (MOLA, Section 2.2) module that enables the student encoder to learn from diverse teachers in a fine-grained (Section 2.3.2) and coarse-grained (Section 2.3.1) manner.

A *visual projector* $f_p(\cdot)$ is applied to project the generated visual token I^S to the LLM input space.

An *LLM* $f_{\text{LLM}}(\cdot)$ then takes the mapped visual tokens $f_p(I^S)$ and the textual instruction tokens T as input to generate the instruction-following response $Y = \{y_i\}_{i \in [L]}$ as

$$p(Y|f_p(I^S), T) = \prod_{i=1}^{L} p(y_i|f_p(I^S), T, y_{<i}), \tag{1}$$

where L is the length of the response and $y_{<i}$ is the previous tokens of y_i.

2.2 Mixture of LoRA Adapters

Directly fine-tuning the student encoder is challenging, as it often leads to overfitting on the limited fine-tuning data and catastrophic forgetting [29, 49]. To avoid this, we propose a mixture-of-LoRA-adapter (MOLA) module consisting of teacher-specific LoRA adapters and general-knowledge LoRA adapters [30] to enable the student encoder to learn from diverse teachers without forgetting. MOLA is illustrated in Figure 1 (a).

Teacher-specific LoRA adapters. Learning from multiple teachers is challenging [31, 32, 33], as each teacher [6, 7, 8] might have different training data, model architecture, and training objectives. Directly transferring diverse teachers' knowledge to the student could lead to noisy distillation and performance drop. To avoid this, we introduce a set of teacher-specific LoRA adapters $\{a_i^T\}_{i=1}^{N_t}$,

where N_t is the number of teachers. Each adapter is designed to align with *one teacher* only, which avoids the conflicts between multiple teachers (see Section 2.3.2). Those adapters are applied to each feedforward layer of the student encoder $f_v(\cdot)$.

General-knowledge LoRA adapters. For learning the collective consensus from teachers and the training data, we introduce a set of general-knowledge LoRA adapters $\{a_i^G\}_{i=1}^{N_g}$ that are applied to each feedforward layer of the student encoder $f_v(\cdot)$, where N_g is the number of general-knowledge LoRA adapters. The details of this general (global) knowledge transfer are provided in Section 2.3.1.

We adopt the general (sparse) design of mixture-of-experts (MoE) [50, 51] to select the LoRA adapters based on the hidden inputs of each layer. Specifically, we employ two sparse routers, *i.e.*, $f_r^T(\cdot)$ and $f_r^G(\cdot)$, to select the teacher-specific LoRA adapters and general-knowledge LoRA adapters, respectively. Formally, for each feedforward layer of the student encoder, the MoE output $F^*(\cdot)$ is computed as

$$\begin{aligned} &F^*(h) = F(h) + a_i^T(h) + a_j^G(h), \\ &\text{with } i = \textsc{argmax}(f_r^T(h)) \text{ and } j = \textsc{argmax}(f_r^G(h)), \end{aligned} \tag{2}$$

where h is the hidden input of the current layer and $F(\cdot)$ is the current layer. We denote the visual tokens generated by the student encoder with MoLA as I^S.

2.3 Hierarchical Knowledge Distillation

To integrate diverse teachers' knowledge into a single student encoder, we propose a hierarchical knowledge Distillation (HKD) mechanism that transfers knowledge at two levels of granularity, *i.e.*, coarse-grained and fine-grained levels. Specifically, for coarse-grained distillation (Section 2.3.1), we summarize the knowledge from multiple teachers (collective consensus) and transfer it to the student encoder globally. For fine-grained distillation (Section 2.3.2), teacher-specific LoRA adapters are employed to align with each teacher separately for a precise noise transfer. Moreover, to attend to the most informative tokens during knowledge transfer, we introduce a token importance scoring method (Figure 2) based on the similarity among teachers' visual tokens and the input instructions.

2.3.1 Coarse-Grained Distillation (CGKD)

To globally distill the knowledge from multiple teachers to the student encoder, we propose a coarse-grained distillation (CGKD) mechanism that first summarizes the knowledge from multiple teachers and then transfers it to the student encoder.

To obtain the collective consensus, *i.e.*, summarized teacher feature, each teacher's visual features are first unshuffled [2, 47, 52] to have the same length [2] as the student's visual features $I^S \in \mathcal{R}^{m \times D}$. Then, those visual tokens are channel-wise concatenated and the summarized feature I_{cg}^T is obtained by applying a two-layer MLP $f_{cg}(\cdot)$ as

$$I_{cg}^T = f_{cg}\left(\textsc{Concat}\left(I_1^T, I_2^T, \ldots, I_{N_t}^T\right)\right) \in \mathcal{R}^{m \times D}, \tag{3}$$

where I_i^T is the unshuffled visual tokens from the i-th teacher, and $\textsc{Concat}(\cdot)$ is the channel-wisely concatenation operation.

Next, we apply the coarse-grained distillation loss $\mathcal{L}_{\text{cg}}$ to transfer the collective consensus by minimizing the mean square error loss (MSE) between the summarized features I_{cg}^T and the student encoder output I^S as

$$\mathcal{L}_{\text{cg}} = \text{MSE}(I^S, I_{cg}^T). \tag{4}$$

2.3.2 Fine-Grained Distillation (FGKD)

Using LoRA adapters [30] to transfer knowledge from one teacher to a student has proven to be successful. However, transferring knowledge from multiple teachers to a single student is challenging, especially when using a fixed set of LoRA adapters [29] for all the teachers. The reason is that the noisy and redundant teachers' knowledge can hinder the learning process and lead to suboptimal performance [31, 32, 33], due to the conflicts among teachers, which arises from the heterogeneity of training data, architectures, and the training algorithms.

To address this challenge, we propose the fine-grained distillation (FGKD) that exploits teacher-specific LoRA adapters and token importance scoring. Each teacher-specific LoRA adapter is designed to align with one teacher only, allowing the student to learn from each teacher separately. Token importance scoring is used to select and attend to the most informative tokens from each teacher during knowledge transfer, reducing the noise and redundancy.

Teacher-specific LoRA adapters. We expect each teacher-specific LoRA adapter to learn the knowledge from one teacher *only*, such that the knowledge transfer is more effective and less noisy. We denote the output of the student encoder with only the i-th teacher-specific LoRA adapter a_i^T being activated for each layer as I_i^S. Specifically, at each feedforward layer of the student encoder, we apply the LoRA adapter $a_i^T(\cdot)$ as $F(h) + a_i^T(h)$, where $F(\cdot)$ is the current layer and h is the input to the layer. In that case, I_i^S only needs to align with the i-th teacher's visual feature I_i^T, making the knowledge transfer procedure smooth and precise.

Token importance scoring. The key to knowledge distillation is to transfer the most important information [24]. As previous studies show that not all tokens are equally informative [18, 19, 20, 29], to identify the most informative tokens, we introduce a new similarity-based importance score that considers teachers' visual tokens and the input instructions T, allowing us to prioritize tokens that are more relevant to the task context. Specifically, for the i-th teacher, we compute the token importance score $s_i \in \mathcal{R}^{1\times m}$ as

$$s_i = \text{MEAN}\left(\text{SOFTMAX}\left(\frac{\text{CONCAT}\left(\hat{I}_i^T, \hat{T}\right)(\hat{I}_i^T)^\top}{\sqrt{D}}\right)\right), \tag{5}$$

where $\hat{I}_i^T \in \mathcal{R}^{m\times D}$ and $\hat{T}$ are the visual tokens and input instructions projected by a learnable two-layer MLP to have the same dimension of the student features. D is the dimension of the visual tokens.

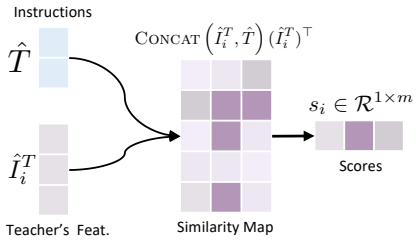

Figure 2: The calculation of token importance score s_i. To focus on the most informative tokens, we consider the similarity among the teacher's features and the input instructions T.

Now, with the token importance scores $\{s_i\}_{i\in[m]}$, the fine-grained distillation loss $\mathcal{L}_{\text{fg}}$ is calculated as

$$\mathcal{L}_{\text{fg}} = \frac{1}{N_t}\sum_{i=1}^{N_t} s_i \cdot \text{MSE}(I_i^S, \hat{I}_i^T)\,. \tag{6}$$

2.4 Training Objectives

The overall training objective of HAWAII is to minimize the loss consisting of the text generation loss $\mathcal{L}_{\text{gen}}$ [4, 53], the coarse-grained distillation loss $\mathcal{L}_{\text{cg}}$ (Equation (4)), the fine-grained distillation loss $\mathcal{L}_{\text{fg}}$ (Equation (6)), and the MoE balance loss $\mathcal{L}_{\text{mb}}$ [51, 54, 55]. This is given as

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_1(\mathcal{L}_{\text{fg}} + \mathcal{L}_{\text{cg}}) + \lambda_2\mathcal{L}_{\text{mb}}, \tag{7}$$

where λ_1 and λ_2 are the hyper-parameters to balance the losses. We set $\lambda_1 = 0.5$ and $\lambda_2 = 0.05$ for all our experiments.

3 Experiments

3.1 Experimental setup

Implementation details. We use Vicuna-v1.5-7B [3] as the LLM and use CLIP [1] for the vision encoder, with the teachers of CLIP (as CLIP is updated) being ConvNeXt [34], Pix2Struct [35], SAM [6], and EVA-02 [8]. The base version of HAWAII uses CLIP, ConvNeXt, and EVA-02 as the vision teachers, while for HAWAII †, we further add Pix2Struct as the teacher. To understand how different teachers contribute to the performance, we also conduct experiments with CLIP, ConvNeXt, EVA-02, and SAM as the teachers, denoted as HAWAII ‡. The visual projector is a 2-layer MLP with the GELU activation function [56]. For MOLA, we use three (or four) teacher-specific LoRA adapters and three general-knowledge LoRA adapters for each FFN layer of the student encoder.

Methods	VQA^{Text}	VizWiz	GQA	SQA	POPE	MME	MMBench	MMMU	AI2D	SeedBench^{I}
BLIP-2 [9] (Vicunna-13B)	42.5	-	-	61.0	85.3	1293.8	-	-	-	-
IDEFICS-9B [57] (LLaMA-7B)	25.9	35.5	38.4	-	-	-	48.2	-	-	-
Qwen-VL [58] (Qwen-7B)	63.8	35.2	59.3	67.1	-	-	38.2	-	-	56.3
mPLUG-Owl2 [59] (LLaMA-7B)	54.3	54.5	56.1	68.7	-	1450.2	64.5	-	-	57.8
Vicunna-1.5-7B										
InstructBLIP [60]	50.1	34.5	49.2	60.5	-	-	36.0	30.6	-	53.4
Video-LLaVA [61]	51.8	48.1	60.3	66.4	84.4	-	60.9	-	-	-
MoVE-KD [29]	58.3	52.3	63.2	69.4	86.9	1524.5	66.3	-	-	-
LLaVA-1.5 [46] (Baseline)	58.2	50.0	62.0	66.8	85.9	1510.7	64.3	34.7	55.5	66.1
HAWAII	58.7	53.9	62.8	70.5	87.3	**1540.2**	66.9	36.6	**56.2**	67.5
Δ	↑ 0.4	↑ 1.6	↓ 0.4	↑ 1.1	↑ 0.4	↑ 15.7	↑ 0.6	↑ 1.9	↑ 0.7	↑ 1.4
HAWAII† (HAWAII + Pix2Struct)	58.6	54.2	**63.6**	69.5	86.8	1533.6	**67.1**	36.0	**56.2**	67.2
HAWAII‡ (HAWAII + SAM)	**59.2**	**54.3**	63.3	**70.8**	**87.5**	1528.6	66.7	**36.9**	55.8	**67.9**

Table 1: Performance comparisons of HAWAII and the baseline VLMs using Vicunna-1.5-7B (if not specified). HAWAII utilize CLIP, ConvNeXt, and EVA-02 as the teachers (the same setting with MoVE-KD), HAWAII† further adds Pix2Struct as the teacher, and HAWAII‡ uses CLIP, ConvNeXt, EVA-02, and SAM as the teachers. The best results are in **bold** and the second best results are underlined.

Each adapter is a LoRA block [30] with rank of 32. The routers are sparse and 2-layer MLPs with the GELU activation function. Each router selects only the LoRA adapter with the highest probability. Models are run on eight NVIDIA A6000 GPUs with 48GB of memory.

Training stages. We follow the standard paradigm of LLaVA-1.5 [46]. The training of HAWAII consists of two stages, *i.e.*, pretraining and fine-tuning. The pretraining stage is to align the vision encoder with the LLM. During this stage, only the vision projector, LoRA adapters, and the routers are trained. The supervised fine-tuning stage is to align the vision encoder with the LLM and the instruction-following response. In this stage, the whole model is trained.

Training datasets. HAWAII uses the same training data as LLaVA-v1.5 [46]. Specifically, in the pretraining stage, we use 558K image-text pairs, while in the supervised fine-tuning stage, we use 665K instruction-following image-text data to boost the performance.

Benchmarks and baselines. We evaluate HAWAII on several image understanding tasks [36, 37, 38, 39, 40, 41, 42, 43, 44, 45]. Details are deferred to the Appendix. We compare HAWAII with several baseline methods, including general VLMs [9, 57, 58, 59, 60, 61] and a VLM with knowledge distillation [29].

3.2 Main Results

The results are shown in Table 1. Compared to the baseline method (LLaVA-1.5), HAWAII achieves significant improvements on most benchmarks, demonstrating its effectiveness. Results also demonstrate that compared to the existing knowledge distillation method [29] that uses the same teachers as HAWAII, HAWAII achieves better performance on most benchmarks, demonstrating the effectiveness of the proposed MOLA module and HKD mechanism.

3.3 Ablation Studies

In this part, we conduct ablation studies to analyze the effectiveness of the proposed components in HAWAII.

Ablation on FGKD, CGKD, and MOLA. The results are shown in Table 2. When all components are included, HAWAII achieves the best performance on most tasks (highlighted row), with an average of 63.7% across all tasks. The baseline model (LLaVA-1.5) with only FGKD (w/o token scoring) and teacher-specific LoRA adapters achieves 63.2% on average. Further adding the token importance scoring mechanism improves the performance to 63.5%. However, we also observe that the performance on GQA is slightly decreased, which might be due to the fact that GQA requires more general knowledge rather than specific knowledge from vision teachers. Adding CGKD and general-knowledge LoRA adapters further improves the performance to 63.7% on average.

Number of visual teachers. To understand how different teachers provide complementary knowledge for visual understanding, we conduct experiments with different teachers, as shown in Table 1. The

Methods	VQA^{Text}	VizWiz	GQA	SQA	POPE	MME	MMBench	MMMU	AI2D	SeedBench^{I}	Avg.
LLaVA-1.5	58.2	50.0	62.0	66.8	85.9	1510.7	64.3	34.7	55.5	66.1	61.9
+ FGKD (w/ot token scoring)	59.0	52.5	**63.1**	70.1	86.6	1532.1	66.8	**36.7**	54.6	66.3	63.2
+ token scoring	**59.1**	52.5	62.8	70.2	**87.4**	**1541.7**	**67.3**	35.9	56.1	67.0	63.5
+ CGKD	58.7	**53.9**	62.8	**70.5**	87.3	1540.2	66.9	36.6	**56.2**	**67.5**	**63.7**
w. DoRA	58.4	53.2	61.8	69.3	**87.7**	**1558.5**	66.9	35.2	55.5	**67.8**	63.4

Table 2: Ablation study on various vision-language tasks of HAWAII. We normalize the results of MME to compute the average results. FGKD and CGKD denote fine-grained distillation with teacher-specific LoRA adapters and coarse-grained distillation with general-knowledge LoRA adapters. W. DoRA represents the variant trained with DoRA for comparison.

#	VQA^{Text}	GQA	SQA	POPE	MME	MMMU	AI2D	SeedBench^{I}
1	58.7	62.6	70.1	84.5	1516.2	37.0	55.5	67.4
3	58.7	62.8	70.5	87.3	1540.2	36.6	56.2	67.5
5	58.6	62.8	70.4	85.2	1530.2	36.4	55.0	66.9

Table 3: Performance of HAWAII with different numbers of general-knowledge adapters.

	VQA^{Text}	GQA	SQA	POPE	MME	MMMU	AI2D	SeedBench^{I}
LLaVA-1.5-13B	61.3	63.3	71.6	85.9	1531.3	35.5	59.3	68.2
MoVE-KD-13B	59.7	64.2	73.2	85.7	1568.1	-	-	-
HAWAII-13B	**61.7**	**64.7**	**75.0**	**86.6**	**1568.7**	**35.7**	**60.0**	**68.5**

Table 4: Performance comparison using the Vicunna-1.5-13B.

	VQA^{Text}	GQA	SQA	POPE	MME	MMMU	AI2D
LLaVA-Next-7B	64.9	64.2	70.1	86.5	1519.0	35.8	64.9
MOVE-KD (LLaVA-Next-7B)	63.7	64.5	70.7	86.7	1537.2	-	-
HAWAII (LLaVA-Next-7B)	**65.5**	**65.2**	**72.0**	**87.8**	**1551.3**	**37.4**	**65.6**

Table 5: Perofmance of HAWAII on LLaVA-Next-7B.

basic version of HAWAII uses CLIP, ConvNeXt, and EVA-02 as the teachers. Further adding Pix2Struct as the teacher improves the performance on VizWiz, GQA, and MMBench, compared to HAWAII. However, maybe due to the redundancy of knowledge, the performances on VQA^{Text}, SQA, and SeedBench^{I} are slightly decreased. We further test the performance of HAWAII with CLIP, ConvNeXt, EVA-02, and SAM as the teachers, denoted as HAWAII ‡ in Table 1. Results show that HAWAII ‡ improves performance on VQA^{Text}, VizWiz, GQA, SQA, POPE, MMMU, and SeedBench^{I}, compared to HAWAII, as SAM might bring strong fine-grained descriptive visual understanding ability to the model. However, we also observe that the performance on MME decreases with adding more teachers, which might be due to the fact that MME requires more general common sense knowledge for reasoning rather than specific knowledge from vision teachers.

Number of general-knowledge adapters. The number of teacher-specific LoRA adapters is dependent on the number of visual teachers, whereas the number of general-knowledge LoRA adapters is a hyperparameter. To understand the optimal number of general-knowledge adapters, we present an ablation in Table 3. The results show that increasing the number of adapters to three improves performance on most benchmarks, while five adapters can lead to slight degradation, indicating that excessive redundancy may introduce overfitting.

Generalizing to larger base models. To test the efficiency of our proposed method, we conducted experiments with Vicunna-1.5-13B. The results in Table 4 show that HAWAII achieves significant improvements. Specifically, HAWAII improves the performance on SQA from 71.6 to 75.0. However, we also notice that with larger base models, the performance on POPE decreases as compared to that with a 7B model.

The impact of the base method. To understand how our proposed knowledge distillation generalizes across different base models, we conducted experiments with LLaVA-Next-7B [62]. The results in

Figure 3: HAWAII is able to perform vision-language understanding tasks, such as emotion understanding, OCR, spatial reasoning, attribute reasoning, and relation reasoning. The examples are from the following benchmarks: VQA$^{\text{Text}}$ [37], MMBench [42], and SeedBench [45].

Table 5 show that HAWAII achieves significant improvements on most benchmarks, compared to the baselines.

The impact of different LoRA methods. We use LoRA in our design because of its generalizability. To understand how different LoRA adapters impact the performance, we conducted experiments with DoRA [63] replacing LoRA. The results are shown in Table 2. DoRA, which is more advanced than LoRA, is less generalizable than LoRA, as evidenced by the performance degradation on some benchmarks.

3.4 Qualitative Results

Visualization of inference examples. We perform qualitative evaluation to highlight the diverse reasoning capabilities of our model across a range of challenging visual understanding tasks [37, 42, 45]. As illustrated in Figure 3, HAWAII demonstrates strong attribute reasoning, accurately identifying fine-grained visual characteristics such as color, texture, and shape. For tasks involving OCR and mathematical content, the model effectively reads and interprets text in images. Beyond factual perception, HAWAII is capable of higher-level understanding, such as inferring image emotion and reasoning about contextual relationships and spatial arrangements. For instance, it can assess emotional tone from facial expressions and body language, and discern nature-related dependencies. These examples showcase the model's comprehensive visual-language understanding, grounded in both low-level perception and abstract reasoning.

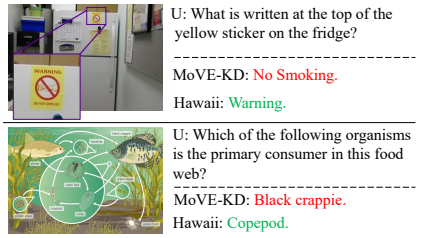

Figure 4: Comparison between HAWAII and MoVE-KD [29] on OCR and visual-semantic reasoning capabilities.

Moreover, a comparison with MoVE-KD [29] (Figure 4) highlights HAWAII 's stronger visual-semantic reasoning, as it accurately interprets ecological relationships in complex diagrams and effectively minimizes text hallucinations in OCR tasks.

Visualization of token importance scores by different teachers and the instructions. Different teachers and instructions typically attend to different regions of the image, providing diverse visual cues that are important for the model to develop a comprehensive understanding. To understand how the token importance scores distribute, we visualize similarity scores between different teachers and the instructions in Figure 5. As shown, the teachers and the instruction exhibit distinct preferences. Text instructions usually focus on the center objects in the images, which are usually indicated by the

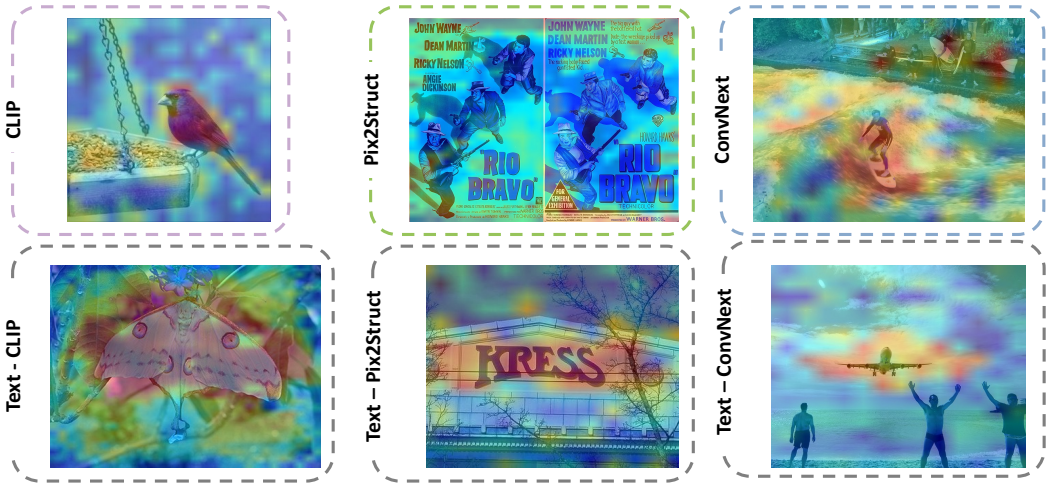

Figure 5: Visualization of the similarity score used in calculating importance score (Section 2.3.2) using HAWAII†.

questions. For visual teachers, CLIP usually attends to the center objects, while ConvNext tends to care more about the common objects in an image (for example, people in the top right image). In contrast to CLIP and ConvNext, Pix2Struct focuses on the small signs and texts in the image, which is useful for OCR-related tasks.

4 Related work

Multi-expert knowledge. In the context of vision-language learning, multi-expert knowledge typically refers to the use of multiple pretrained visual models, each specialized in a particular domain or task, to provide richer and more diverse visual understanding. One common strategy for incorporating such knowledge is through auxiliary supervision or multitask learning [64, 65], where expert models trained on tasks such as segmentation, object detection, or depth estimation provide additional learning signals during training. These experts are typically integrated via auxiliary losses or parallel task-specific heads [66], allowing the model to benefit from complementary visual perspectives. While this approach has shown effectiveness, it often requires task-specific annotations and careful balancing of multiple objectives, which can complicate training and limit scalability.

Another strategy for incorporating multi-expert knowledge into vision components of VLMs involves using multiple visual encoders to extract diverse representations, which are then fused to form a unified visual understanding. These methods [13, 47, 48, 67] typically focus on efficiently integrating visual tokens generated by a mixture of pretrained visual experts. By drawing on the complementary strengths of these encoders, such approaches aim to enhance the model's visual perception capabilities. However, they often introduce substantial computational overhead due to the large number of visual tokens, particularly in approaches that concatenate token sequences [11, 13, 68]. In contrast, HAWAII adopts a different strategy by using multiple vision encoders as teachers to distill their knowledge into a single student encoder, enabling it to inherit their complementary strengths while maintaining efficiency.

Knowledge distillation. Knowledge distillation (KD) [24] is a process where a smaller, more efficient model called the student learns from the output logits or feature representations of a larger, pretrained model known as the teacher. In the context of vision-language learning, KD has been explored in several directions. Some approaches [69, 70] focus on distilling large vision-language models into smaller ones. This line of work aims to compress the knowledge of powerful multimodal models into more compact and efficient versions that can still perform effectively on vision-language tasks. In contrast to our work, these methods prioritize reducing the overall model size. Instead, our approach focuses on enhancing the visual capabilities of the vision encoder within a VLM by distilling knowledge from multiple expert teachers without necessarily reducing the VLM itself. Another common use of KD is to train efficient vision foundation models [7] by distilling smaller vision

backbones from larger teacher(s) in a standalone setting, separate from the VLM training pipeline. For example, InternViT-300M [71] is distilled from InternViT-6B using feature distillation with a cosine similarity loss applied between the hidden states of the final transformer layers. Similarly, RADIO [72] trains a vision model from scratch by merging multiple backbone models into a unified architecture through multi-teacher distillation. It employs feature-level distillation using cosine distance loss, with equal weighting applied to the outputs of each teacher. While effective, these approaches are highly computationally intensive and require massive datasets and substantial compute resources. In contrast to these standalone approaches, our work focuses on optimizing the student vision encoder within the training loop of a vision-language model, allowing it to benefit directly from multimodal supervision and alignment during training.

The work closest to ours is MoVE-KD [29], which distills knowledge from multiple visual experts into a single vision encoder using a weighted distillation loss with a fixed set of LoRA adapters [30]. The weights are shared between different teachers based on the attention weights from CLIP [1], which introduces a bias toward CLIP. In contrast, HAWAII introduce teacher-specific LoRA adapters which are aligned with each teacher separately, allowing the student encoder to learn from diverse teachers while avoiding noisy distillation. Moreover, the token importance scoring in HAWAII is based on each teacher's visual features and the input instructions, which helps to select the most informative tokens from each teacher without introducing bias toward any specific teacher.

5 Conclusion, Limitations, and Societal Impacts

Conclusion. We introduced HAWAII, a novel framework that distills knowledge from multiple pretrained visual experts into a single vision encoder. HAWAII consists of a novel mixture-of-LoRA-adapter (MOLA) module and a new hierarchical knowledge distillation (HKD) mechanism. MOLA consists of teacher-specific LoRA adapters and general-knowledge LoRA adapters that enable the student encoder to learn from diverse teachers while learning general knowledge from the training data. HKD distills knowledge from multiple teachers at coarse-grained and fine-grained levels. The coarse-grained distillation summarizes the knowledge from multiple teachers and transfers it to the student encoder globally. The fine-grained distillation utilizes teacher-specific LoRA adapters and token importance scoring to select the most informative tokens from each teacher for distillation. Extensive experiments on various vision-language tasks demonstrate the superiority of HAWAII over existing methods with minimal computational overhead.

Limitations. Due to the limitation of computational resources, we only used five pretrained vision experts in our experiments. Also, we only evaluated HAWAII using the Vicuna-v1.5-7B [3] as the LLM. In the future, it would be interesting to explore the performance of HAWAII with more pretrained vision experts and different LLMs. We only distill knowledge from the visual experts to the vision encoder, while the knowledge distillation from a bigger LLM to a smaller LLM is not considered. We believe that further improvements can be achieved by distilling knowledge from a bigger LLM to a smaller LLM.

Societal impacts and safeguards. The proposed HAWAII framework is designed to enhance the performance of VLMs. Thus, it inherits the same societal impacts as existing VLMs. The use of HAWAII and VLMs in general may raise concerns related to bias, misinformation, and privacy. However, we have taken steps to mitigate these risks by carefully curating the training data and implementing safeguards to ensure responsible use.

Acknowledgement

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC)-CSE Research Community project entitled "An End-to-End Approach to Safe and Secure AI Systems" and NSERC's Postdoctoral Fellowship. Researchers funded through the NSERC-CSE Research Communities Grants do not represent the Communications Security Establishment Canada or the Government of Canada. Any research, opinions, or positions they produce as part of this initiative do not represent the official views of the Government of Canada.

References

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

[2] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks, January 2024. arXiv:2312.14238 [cs] version: 3.

[3] Wei-Lin Chiang, Zhuohan Li, Ziqing 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.

[4] 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. LLaMA: Open and Efficient Foundation Language Models, February 2023. arXiv:2302.13971 [cs].

[5] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything, April 2023. arXiv:2304.02643 [cs].

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

[8] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369, Vancouver, BC, Canada, June 2023. IEEE.

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

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. November 2023.

[11] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

[12] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[13] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision*, pages 113–132. Springer, 2024.

[14] Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models. *arXiv preprint arXiv:2407.12709*, 2024.

[15] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025.

[16] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15139, pages 19–35, Cham, 2024. Springer Nature Switzerland. Series Title: Lecture Notes in Computer Science.

[17] Bo Tong, Bokai Lai, Yiyi Zhou, Gen Luo, Yunhang Shen, Ke Li, Xiaoshuai Sun, and Rongrong Ji. FlashSloth: Lightning Multimodal Large Language Models via Embedded Visual Compression, December 2024. arXiv:2412.04317 [cs].

[18] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models, May 2024. arXiv:2403.15388 [cs].

[19] Yimu Wang, Mozhgan Nasr Azadani, Sean Sedwards, and Krzysztof Czarnecki. Leo-mini: An efficient multimodal large language model using conditional token reduction and mixture of multi-modal experts. *arXiv preprint arXiv:2504.04653*, 2025.

[20] Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token, January 2025. arXiv:2501.03895 [cs].

[21] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, Elahe Arani, and Oleg Sinavski. LingoQA: Visual Question Answering for Autonomous Driving. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 252–269, Cham, 2024. Springer Nature Switzerland. TLDR: This work introduces LingoQA, a novel dataset and benchmark for visual question answering in autonomous driving and proposes a truthfulness classifier, called Lingo-Judge, that achieves a 0.95 Spearman correlation coefficient to human evaluations, surpassing existing techniques like METEOR, BLEU, CIDEr, and GPT-4.

[22] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M. Rehg, and Chao Zheng. MAPLM: A Real-World Large-Scale Vision-Language Benchmark for Map and Traffic Scene Understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21819–21830, Seattle, WA, USA, June 2024. IEEE. TLDR: A new vision-language benchmark that can be used to finetune traffic and HD map domain-specific foundation models, and annotate and leverage large-scale, broad-coverage traffic and map data extracted from huge HD map annotations is proposed.

[23] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2Tell: A Multimodal Driving Dataset for Joint Importance Ranking and Reasoning. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7498–7507, Waikoloa, HI, USA, January 2024. IEEE. TLDR: A novel dataset, Rank2Tell 1, a multi-modal ego-centric dataset for Ranking the importance level and Telling the reason for the importance is introduced, which provides dense annotations of various semantic, spatial, temporal, and relational attributes of various important objects in complex traffic scenarios.

[24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[25] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R. Qi, Xinchen Yan, Scott Ettinger, and Dragomir Anguelov. Unsupervised 3D Perception with 2D Vision-Language Distillation for Autonomous Driving, September 2023. arXiv:2309.14491 [cs].

[26] Yi Xie, Yihong Lin, Wenjie Cai, Xuemiao Xu, Huaidong Zhang, Yong Du, and Shengfeng He. D3still: Decoupled Differential Distillation for Asymmetric Image Retrieval. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17181–17190, Seattle, WA, USA, June 2024. IEEE.

[27] Yi-Ting Hsiao, Siavash Khodadadeh, Kevin Duarte, Wei-An Lin, Hui Qu, Mingi Kwon, and Ratheesh Kalarot. Plug-and-Play Diffusion Distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13743–13752, Seattle, WA, USA, June 2024. IEEE.

[28] Jing Xu, Jiazheng Li, and Jingzhao Zhang. Scalable Model Merging with Progressive Layer-wise Distillation, February 2025. arXiv:2502.12706 [cs].

[29] Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning MA, and Shanghang Zhang. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. *arXiv preprint arXiv:2501.01709*, 2025.

[30] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021. arXiv:2106.09685 [cs].

[31] Sihui Luo, Xinchao Wang, Gongfan Fang, Yao Hu, Dapeng Tao, and Mingli Song. Knowledge Amalgamation from Heterogeneous Networks by Common Feature Learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 3087–3093, Macao, China, August 2019. International Joint Conferences on Artificial Intelligence Organization.

[32] Chengchao Shen, Mengqi Xue, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Customizing Student Networks From Heterogeneous Teachers via Adaptive Knowledge Amalgamation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3503–3512, Seoul, Korea (South), October 2019. IEEE.

[33] Chengchao Shen, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Amalgamating Knowledge towards Comprehensive Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3068–3075, July 2019.

[34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[35] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding. In *Proceedings of the 40th International Conference on Machine Learning*, pages 18893–18912. PMLR, July 2023. ISSN: 2640-3498.

[36] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, Salt Lake City, UT, USA, June 2018. IEEE.

[37] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8309–8318, 2019.

[38] Drew A. Hudson and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, Long Beach, CA, USA, June 2019. IEEE.

[39] Pan Lu, Swaroop Mishra, Tony 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 *The 36th conference on neural information processing systems (NeurIPS)*, 2022.

[40] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In Houda Bouamor, Juan Pino, and

Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, December 2023. Association for Computational Linguistics.

[41] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models, March 2024. arXiv:2306.13394 [cs].

[42] 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 2024*, pages 216–233. 2024.

[43] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, Seattle, WA, USA, June 2024. IEEE.

[44] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram is Worth a Dozen Images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 235–251, Cham, 2016. Springer International Publishing.

[45] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking Multimodal Large Language Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308, 2024.

[46] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. arXiv, October 2023. arXiv:2310.03744 [cs].

[47] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *The Thirteenth International Conference on Learning Representations*, 2025.

[48] Mozhgan Nasr Azadani, James Riddell, Sean Sedwards, and Krzysztof Czarnecki. Leo: Boosting mixture of vision encoders for multimodal large language models. *arXiv preprint arXiv:2501.06986*, 2025.

[49] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning, July 2022. arXiv:2203.12119 [cs].

[50] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA Experts. October 2023.

[51] Damai Dai, Chengqi Deng, Chenggang Zhao, R.x. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.k. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[52] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, June 2016. ISSN: 1063-6919.

[53] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti 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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. arXiv:2307.09288 [cs].

[54] Tianlin Liu, Mathieu Blondel, Carlos Riquelme Ruiz, and Joan Puigcerver. Routers in Vision Mixture of Experts: An Empirical Study. *Transactions on Machine Learning Research*, February 2024.

[55] Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. LoRAMoE: Alleviating World Knowledge Forgetting in Large Language Models via MoE-Style Plugin. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[56] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs), June 2023. arXiv:1606.08415 [cs].

[57] Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. November 2023.

[58] 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, October 2023. arXiv:2308.12966 TLDR: The Qwen-VL series is introduced, a set of large-scale vision-language models designed to perceive and understand both text and images that outperforms existing Large Vision Language Models (LVLMs).

[59] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. pages 13040–13051, 2024.

[60] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. November 2023.

[61] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[62] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

[63] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-Decomposed Low-Rank Adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 32100–32121. PMLR, July 2024. ISSN: 2640-3498.

[64] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. Prismer: A vision-language model with multi-task experts. *Trans. Mach. Learn. Res.*, pages 2835–8856, 2024.

[65] Yuanchen Wu, Junlong Du, Ke Yan, Shouhong Ding, and Xiaoqiang Li. Tove: Efficient vision-language learning via knowledge transfer from vision experts. *arXiv preprint arXiv:2504.00691*, 2025.

[66] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022.

[67] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

[68] Xiaoran Fan, Tao Ji, Changhao Jiang, Shuo Li, Senjie Jin, Sirui Song, Junke Wang, Boyang Hong, Lu Chen, Guodong Zheng, et al. Mousi: Poly-visual-expert vision-language models. *arXiv preprint arXiv:2401.17221*, 2024.

[69] Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, and Xiang Bai. Llava-kd: A framework of distilling multimodal large language models. *arXiv preprint arXiv:2410.16236*, 2024.

[70] Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Lei Zhang, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, et al. Llava-mod: Making llava tiny via moe knowledge distillation. *arXiv preprint arXiv:2408.15881*, 2024.

[71] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):1–17, 2024.

[72] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12490–12500, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. We focus on distilling visual perception knowledge from multiple pretrained visual experts into a single vision encoder, enabling it to inherit the complementary strengths of these experts while maintaining efficiency.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper has no theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided all the information needed to reproduce the main experimental results of the paper. The code and models will be released upon acceptance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data is publicly available. The code and models will be released upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify all the training and test details necessary to understand the results in Section 3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to the limited computational resources, we do not report error bars in our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information on the computer resources needed to reproduce the experiments in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that the research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed both potential positive societal impacts and negative societal impacts of the work in Section 5.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We have discussed it in Section 5.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly cited the assets used in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code and models will be released upon acceptance with sufficient documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Experiments

A.1 Benchmark Datasets

We evaluate our method on the following benchmark datasets: MME [41], MMBench [42], Seed-Bench [45], GQA [38], SQA [39], MMMU [43], POPE [40], AI2D [44], VizWiz [36], and TextVQA [37].

MME [41]. The MME benchmark is designed to rigorously evaluate a model's perceptual and cognitive abilities through 14 subtasks. It employs carefully constructed instruction-answer pairs and concise instructions to minimize data leakage and ensure fair evaluation. This setup provides a robust measure of a model's performance across various tasks.

MMBench [42]. MMBench offers a hierarchical evaluation framework, categorizing model capabilities into three levels. The first level (L-1) focuses on perception and reasoning. The second level (L-2) expands this to six sub-abilities, while the third level (L-3) further refines these into 20 specific dimensions. This structured approach allows for a nuanced and comprehensive assessment of a model's multifaceted abilities.

Seed-Bench [45]. SEED-Bench consists of 19K multiple-choice questions with accurate human annotations, covering 12 evaluation dimensions including both the spatial and temporal understanding.

GQA [38]. GQA is structured around three core components: scene graphs, questions, and images. It includes not only the images themselves but also detailed spatial features and object-level attributes. The questions are crafted to assess a model's ability to comprehend visual scenes and perform reasoning tasks based on the image content.

ScienceQA [39]. ScienceQA spans a wide array of domains, including natural, language, and social sciences. Questions are hierarchically categorized into 26 topics, 127 categories, and 379 skills, providing a diverse and comprehensive testbed for evaluating multimodal understanding, multi-step reasoning, and interpretability.

MMMU [43]. MMMU includes 11.5K meticulously collected multimodal questions from college exams, quizzes, and textbooks, covering six core disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering. These questions span 30 subjects and 183 subfields, comprising 30 highly heterogeneous image types, such as charts, diagrams, maps, tables, music sheets, and chemical structures.

POPE [40]. POPE is tailored to assess object hallucination in models. It presents a series of binary questions about the presence of objects in images, using accuracy, recall, precision, and F1 score as metrics. This approach offers a precise evaluation of hallucination levels under different sampling strategies.

AI2D [44]. AI2D is a dataset of over 5000 grade school science diagrams with over 150000 rich annotations, their ground truth syntactic parses, and more than 15000 corresponding multiple choice questions.

VizWiz [36]. VizWiz consists of over 31,000 visual questions originating from blind people who each took a picture using a mobile phone and recorded a spoken question about it, together with 10 crowdsourced answers per visual question.

TextVQA [37]. TextVQA emphasizes the integration of textual information within images. It evaluates a model's proficiency in reading and reasoning about text embedded in visual content, requiring both visual and textual comprehension to answer questions accurately.

A.2 Comparison with MLLMs with Multiple Vision Encoders

To better understand how HAWAII compares with the existing MLLMs with multiple vision encoders [13, 14, 47, 68], we present the comparison in Table 6. Results show that HAWAII achieves competitive or significant improvements on most benchmarks, demonstrating the effectiveness of HAWAII. However, we also notice performance degradation on some benchmarks, such as POPE, GQA, and SeedBench.

	VizWiz	GQA	SQA	POPE	MME	AI2D	MMMU	SeedBench
Eagle-X5 [47]	**54.4**	**64.9**	69.8	**88.8**	1528	-	36.3	**73.9**
MoME [14] (CLIP + DINO + Pix2Struct)	-	59.7	-	-	-	-	-	-
MouSi [68] (LayoutLMv3+DINOv2+CLIP)	-	63.6	69.0	86.5	-	-	-	67.5
Brave [13]	54.2	52.7	-	87.6	-	-	-	-
LLaVA-1.5 (CLIP)	50.0	62.0	66.8	85.9	1510.7	55.5	34.7	66.1
MoVE-KD	52.3	63.2	69.4	86.9	1524.5	-	-	-
HAWAII	53.9	62.8	70.5	87.3	**1540.2**	**56.2**	36.6	67.5
HAWAII† (HAWAII + Pix2Struct)	54.2	63.6	69.5	86.8	1533.6	**56.2**	36.0	67.2
HAWAII‡ (HAWAII + SAM)	54.3	63.3	**70.8**	87.5	1528.6	55.8	**36.9**	67.9
Δ	↓ 0.1	↓ 1.3	↑ 1.0	↓ 1.3	↑ 12.2	-	↑ 0.6	↓ 6.0

Table 6: Comparison with MLLMs with multiple vision encoders.

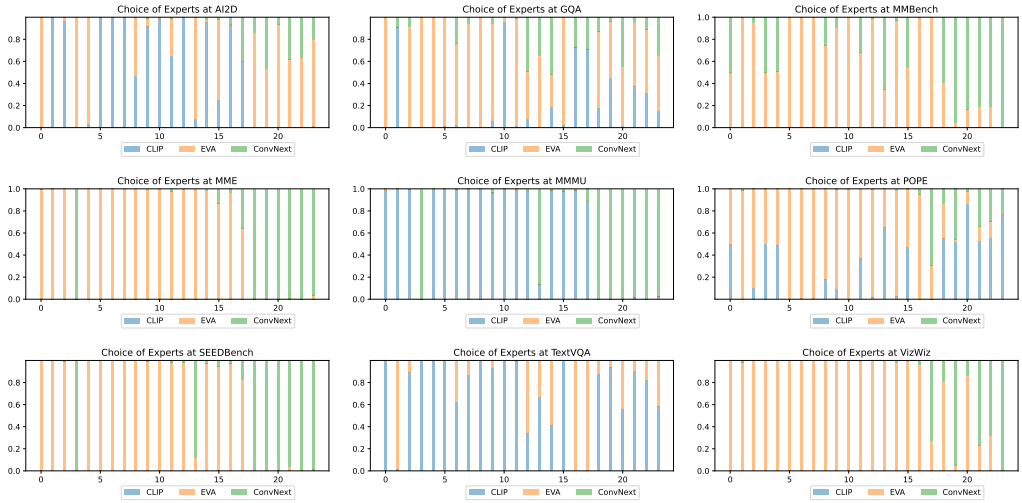

Figure 6: Visualization of the routing choice using HAWAII-v1.0. Best viewed in color.

A.3 Ablation Study

Routing between specific teachers' knowledge. To further understand how HAWAII switches between different teachers' knowledge, we visualize the routing results in Figure 6. It is obvious that HAWAII selects different expert's knowledge across different benchmark datasets and different layers. A notable observation is for most of the cases, HAWAII does not choose CLIP for understanding visual contents. We observe that for MME, VizWiz, and SEEDBench, the model has similar selection preference, while for MMMU, model mainly choose CLIP and ConvNext.