Towards One-to-Many Visual Question Answering

Huishan Ji[1,2], Qingyi Si[1,2], Zheng Lin[1,2]*, Yanan Cao[1], Weiping Wang[1]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{jihuishan, siqingyi, linzheng, caoyanan, wangweiping} @iie.ac.cn

Abstract

Most existing Visual Question Answering (VQA) systems are constrained to support domain-specific questions, i.e., to train different models separately for different VQA tasks, thus generalizing poorly to others. For example, models trained on the reasoning-focused dataset GQA struggle to effectively handle samples from the knowledge-emphasizing dataset OKVQA. Meanwhile, in real-world scenarios, it is user-unfriendly to restrict the domain of questions. Therefore, this paper proposes a necessary task: One-to-Many Visual Question Answering, of which the ultimate goal is to enable a single model to answer as many different domains of questions as possible by the effective integration of available VQA resources. To this end, we first investigate into ten common VQA datasets, and break the task of VQA down into the integration of three key abilities. Then, considering assorted questions rely on different VQA abilities, this paper proposes a novel dynamic Mixture of LoRAs (MoL) strategy. MoL mixes three individually trained LoRA adapters (corresponding to each VQA ability) dynamically for different samples demanding various VQA abilities. The proposed MoL strategy is verified to be highly effective by experiments, establishing SOTAs on four datasets. In addition, MoL generalizes well to three extra zero-shot datasets. Data and codes will be released.

1 Introduction

Visual question answering (VQA) is a deeply interlaced task of CV and NLP which requires answering a question given an image. Driven by its wide range of application and thirst for exploring the interaction between both modalities, VQA has attracted a growing number of researches in recent years. Such enthusiasm has thus fertilized the growth in both the diversity and practicality of the task setting. For example, GQA (Hudson and Manning, 2019) demand comprehension of the scene and reasoning over objects, while OKVQA (Marino et al., 2019) emphasizes the capability of utilizing knowledge.

However, researches on these VQA tasks are usually separate from each other. Intuitively, performing well on a VQA dataset does not necessarily guarantee acceptable results on others. As we can expect all sorts of questions from users, a model trained on a single specific domain may not be competent for real-world application. Therefore, to prompt exploration towards such direction, in this paper, we propose the task of **One-to-Many Visual Question Answering**, which mimics the authentic situation in the real world and demands a single model to answer questions requiring assorted skills.

To perform well on such a challenging stage, an ideal model shall master various VQA abilities. There are former works (Goyal et al., 2017; Hudson and Manning, 2019; Kafle and Kanan, 2017) classifying the VQA questions into various classes. However, their categorization mainly focus on the forms or intention of questions within only a single dataset like GQA (Hudson and Manning, 2019) or TDIUC (Kafle and Kanan, 2017), which fails to cover all VQA abilities. For example, TDIUC (Kafle and Kanan, 2017) divides questions into classes like *Object Presence* and *Sport Recognition*. Their motivation and implementation do not fit in our One-to-Many VQA here. Normally, the focus of a VQA dataset is unique and confined to a single VQA ability, like knowledge or reasoning. To the best of our knowledge, no dataset available is able to complete our proposed task on its own, which means accommodating sufficient diverse VQA resources is necessary.

In addition, to provide better generalization for all sorts of VQA questions, we first break the task of VQA down into the integration of basic VQA abilities. Having taken both the commonality and distinction into consideration, with thoughtful de-

*Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 16931–16943

liberation, we come up with three basic VQA abilities, i.e., Knowledge Capability (KC), Visual Attribute Recognition (VAR) and Scene Comprehension (SC). As their names suggest, KC encapsulates the capability to store and apply knowledge, VAR encompasses attribute recognition and basic forms of reasoning, such as counting, and SC denotes the proficiency to infer relationships across objects. The relationship among these abilities is further analyzed in our experiments. Then, ten common VQA datasets are collected. Through analysis of their motivation and styles, we categorize them according to the focused abilities. Additionally, we select three extra datasets as the Held-Out group, which are not used in training but reserved for the zero-shot testing. Ideally, a model having mastered the three VQA abilities will generalize well to them.

To integrate each ability while maintaining the flexibility for dynamically adjusting the model to focus on the required ability of each sample, we propose a Mixture of LoRAs strategy (MoL). Applying trainable adapters to a frozen MLLM (multimodal large languange model) is an ideal solution under the setting, as LoRA adapters are flexible to merge and expand, while MLLMs already contain basic multimodal skills, which aids to generalization. Specifically, three LoRA adapters are individually trained for each ability and then weightedly averaged during inference. The objective is to allow for dynamic weighting and adjustment, tailoring the emphasis to the core of each question, which is captured by a trained smaller language model.

We present a comprehensive study under the One-to-Many setting and verify the effectiveness of the proposed method. Experiments shows that specialist models which trained on single datasets fail to generalize well to other datasets. Compared to the advanced visual-language pretrained models and multimodal LLMs, our method achieves the best One-to-Many performance on most datasets, as shown in Figure 1. Even in comparison with previous specialist models, our method establishes the new state-of-the-art accuracy on four datasets, OKVQA, KRVQA, COCO-QA and DAQUAR. Besides, our method brings significant improvement on zero-shot performance for Held-Out datasets, VQA abs, VizWiz and A-OKVQA. To summary, our contributions are as follows:

- To the best of our knowledge, we are the first to propose One-to-Many VQA, which is a challenging task simulating real-life scenarios, and conduct detailed analyses into the three VQA abilities on the proposed benchmark.

- We propose a novel Mixture of LoRAs strategy (MoL) to dynamically adjust the capability of the model for each sample demanding various VQA abilities. In experiments, MoL demonstrates promising flexibility and embraces evident improvement under the One-to-Many setting across two MLLMs.

- We establish new state-of-the-art performance on four Held-In VQA datasets, OKVQA, KRVQA, COCO-QA, DAQUAR and significantly improve zero-shot performance on three Held-Out datasets, VQA abs, VizWiz and A-OKVQA.

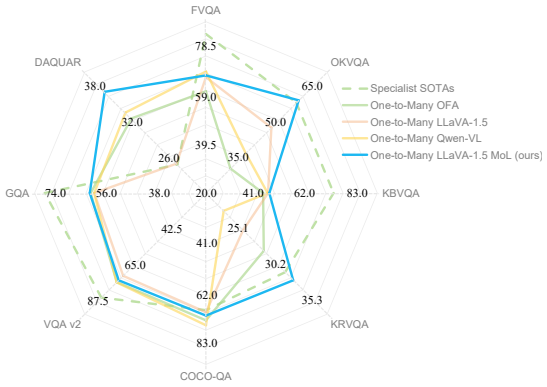

Figure 1: Performance demonstration of our proposed MoL_{LW} applied on LLaVA-1.5-7b (Liu et al., 2023a) against one-to-one specialist SOTAs and strong One-to-Many baselines (one of the most advanced VLP models OFA (Wang et al., 2022b), and MLLMs Qwen-VL and LLaVA-1.5). Our method performs best among the One-to-Many methods and demonstrates competitive performance even compared with specialist models on most datasets. Refer to Appendix C for our One-to-Many baseline settings.

2 One-to-Many VQA

Assorted VQA datasets mainly differ in their scales and required VQA abilities for solution. As shown in Table 1, we investigate into ten common VQA datasets. Taking into account the motivation behind these datasets and the actual cognitive processes involved when humans perform VQA, we categorize them into three groups according to three proposed VQA abilities, i.e., **Knowledge Capability**, **Visual Attributes** and **Scene Comprehension**. Datasets clustered under each of them tend to focus on and

Figure 2: Examples of the ten VQA datasets. Different colors stand for different focus of VQA abilities. Blue for Knowledge Capability, green for Visual Attribute Recognition, and yellow for Scene Comprehension. The styles of these datasets and their motivation make it easy to cluster.

benefit from (but not solely relied on) the same corresponding VQA ability. Note that these abilities are not completely independent from each other. For example, to apply knowledge, basic recognition ability is indispensable.

Knowledge Capability Abbreviated as KC, the ability of Knowledge Capability aims at storing and utilizing knowledge. This group contains FVQA (Wang et al., 2017), OKVQA (Marino et al., 2019), KBVQA (Wang et al., 2015) and KRVQA (Cao et al., 2021). FVQA and KBVQA both provide extra knowledge for solution. The former provides a sentence of fact for each sample, while the latter utilizes DB-pedia (Auer et al., 2007) to consult for knowledge. OKVQA comes from the most open setting of VQA, that models are allowed to use any form of external knowledge, from knowledge bases to even the Internet or GPTs. KRVQA aims to avoid the language-prior shortcut by erasing the mentioned entity in a question and replacing it with a knowledge-based description.

Visual Attribute Recognition Abbreviated as VAR, the ability of visual attribute recognition aims at recognizing attributes and simple reasoning like counting. This group contains TDIUC (Kafle and Kanan, 2017), COCO-QA (Ren et al., 2015), VQA v2 (Goyal et al., 2017) and VG-QA (Krishna et al., 2017). TDIUC generates questions using annotation in MS-COCO images (Lin et al., 2014) and collects filtered samples from VQA v1 (Antol et al., 2015) as well as VG-QA, dividing them into 12 fine-grained tasks. COCO-QA provides a larger and more diverse dataset than DAQUAR as an early work. VQA v2 is proposed to reduce the bias in VQA v1 by collecting complementary data towards existing ones, and has been a widely used dataset for testing on VQA. Visual Genome dataset (Krishna et al., 2017) provides detailed annotations about the objects in images for future analysis. VG-QA is a corresponding VQA dataset provided along with it, which is relatively tough due to its assorted styles of answers.

Scene Comprehension Abbreviated as SC, the ability of Scene Comprehension aims at reasoning over objects to catch relations. This group contains GQA (Hudson and Manning, 2019) and DAQUAR (Malinowski and Fritz, 2014). GQA leverages the annotations of scene graphs to automatically construct questions with a question engine. Questions in GQA usually involve reasoning over objects. Following previous works (Tan and Bansal, 2019; Wang et al., 2022b), we combine the training and validation sets of GQA balanced for training, and use the testdev set for testing. DAQUAR is the first VQA dataset available, focusing on in-door scenarios. However, questions in DAQUAR also require frequent reasoning.

Held-Out Datasets These datasets are selected to evaluate the zero-shot performance and thus not involved in training. This group contains VizWiz

Usage	Group	Datasets	# Images	# Samples	Avg. Len
Held-In	KC	FVQA	2,190	5,826	9.5
		OKVQA	14,031	14,055	8.1
		KBVQA	700	2,741	6.9
		KRVQA	19,739	126,525	11.7
	VAR	TDIUC	167,437	1,654,167	6.9
		COCO-QA	69.172	117.684	8.7
		VQA v2	204,721	658,111	6.2
		VG-QA	108,077	1,445,316	5.7
	GQA	GQA	82,772	1,087,640	8.8
		DAQUAR	1,447	12,468	10.6
Held-Out	Held-Out	VizWiz VQA	17,925	17,925	6.3
		VQA abstract	30,000	90,000	6.2
		A-OKVQA	17,652	18,201	8.8

Table 1: Datasets statistics. The name of each group denotes the focused VQA ability in it. # Images and # Samples stand for the numbers of questions and samples in each dataset. The average length of questions is denoted by Avg. Len.

VQA (Gurari et al., 2018), VQA v1 abstract (Antol et al., 2015) and A-OKVQA (Schwenk et al., 2022). VizWiz VQA aims to help the blind by answering their questions about what they are photoing at, which is why it is poor in image quality and questions are sometimes informal or even unanswerable. To use it as a zero-shot test set, we only pick the answerable questions. VQA v1 abstract comes from the abstract scenes in VQA v1. Its cartoon-style images presents challenge but are not too abstract like CLEVR. A-OKVQA is a newer version of OKVQA, as introduced before. Though similar in goals, A-OKVQA shares no overlap with OKVQA, possessing no risk of information leak. To be consistent with other datasets, we utilize the open-ended answers from A-OKVQA.

Dataset Statistics Table 1 shows the statistics of our collected datasets. As mentioned above, these datasets share large differences against each other, like the focused VQA abilities, sizes and the sparsity of answers. Such properties make it unrealistic to analyze dataset by dataset. Not only will it cost unnecessary effort, but also makes it hard to capture the commonality among them, which is why we analyze by groups. [1]

3 Method

Our method is proposed to address all three abilities mentioned above and dynamically adjust to focus on the required core ability of each sample. This section introduces our design.

3.1 Architecture

One-to-Many VQA requires a single model to be capable of answering assorted questions, which brings challenge towards the generalization and capacity to the model. This paper proposes to use LoRA (Low-Rank Adaptation) adapters to train MLLMs for each focused VQA ability and merge the adapters as experts. Intuitively, MLLMs (Zhu et al., 2023; Liu et al., 2023b; Peng et al., 2023; Dai et al., 2023) generalize well to assorted instructions and contain rich multimodal knowledge, which aligns well with the proposed One-to-Many VQA task. Meanwhile, not only can LoRA save training resources, but most importantly, it contains the potential to be merged as different experts. Rather than scaling up the parameters by routing among FFNs or models, like the traditional mixture of experts (MoE), we hope to explore using each fine-tuned LoRA as an expert and combine different experts with weighted averaging within a single model. Further, this paper extends to a dynamical weighted averaging strategy that captures the required abilities of each sample and adjusts the focus of the model accordingly.

3.1.1 Overview

The proposed framework is shown in Figure 3. QwenLM (Bai et al., 2023) is a decoder-only large language model trained on 2.2T tokens, containing a vision transformer (ViT)(Dosovitskiy et al., 2020), a VL adapter and a LLM. In addition, under the same paradigm, this paper experiment with LLaVA-1.5-7b (Liu et al., 2023a) as well and our strategy generalizes well to it.

3.2 LoRA Expert Training

This paper trains three LoRAs adapters to learn the focused VQA ability of each group respectively. The model is optimized with cross entropy loss. Assume a sample s, with an input question s_q, an image s_v and an output s_y containing $|s_y|$ tokens. Original model parameter and LoRA parameter are represented by θ_0 and θ_{lora}, respectively, Taking y_m as the mth token and $y_{<m}$ as the tokens ahead of y_m, then the language modeling loss for the sample s is computed as:

[1]There were actually more datasets we considered, like CLEVR (Johnson et al., 2017), KVQA (Shah et al., 2019) and so on. But those datasets based on too abstract scenes like CLEVR are not compatible with our tendency for the real-life application. KVQA requires face recognition and is confined to a limited scene, which is too specific. For reasons similar to the above, we finally reduce them to a total number of thirteen Held-In and Held-Out datasets while maintaining generality.

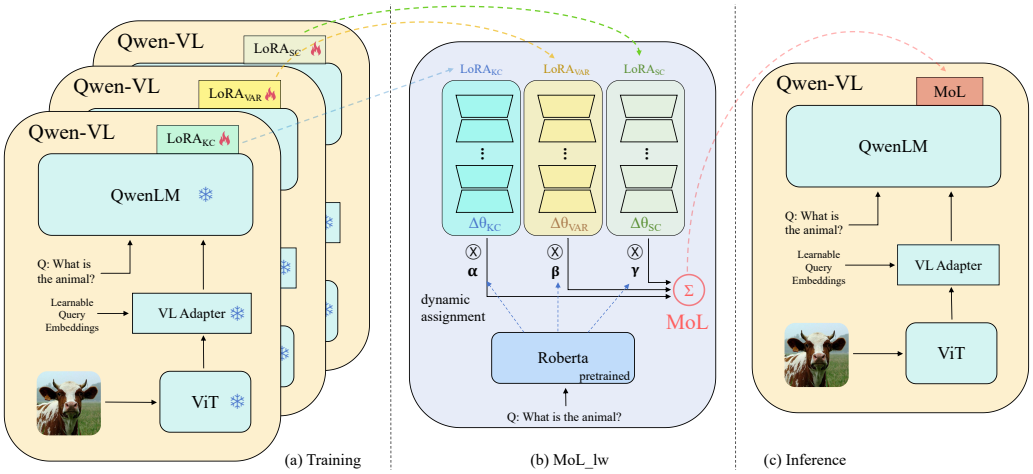

Figure 3: Overview of the proposed method. We experiment with Qwen-VL and LLaVA-1.5 as the backbone in turn, and train a LoRA on each group to learn the corresponding VQA ability. Subsequently, the three individually trained LoRA adapters are merged by weights for integration of VQA abilities. This paper proposes a dynamic weighting method, MoL_{LW}, that generates weights for each sample using a small language model, Roberta, to tailor the focus on the specific ability required for each question. Finally, the merged LoRA is employed for inference.

$$L_s = -\sum_{m=1}^{|s_y|} \log P_{\theta_0+\theta_{lora}}(y_m|y_{<m}, s_v, s_q) \quad (1)$$

We adopt the loss function above to train and obtain three LoRA checkpoints, i.e., $\theta_{KC} = \{B^l_{KC}, A^l_{KC}\}^L_{l=1}$, $\theta_{SAR} = \{B^l_{SAR}, A^l_{SAR}\}^L_{l=1}$ and $\theta_{SC} = \{B^l_{SC}, A^l_{SC}\}^L_{l=1}$, where L presents the number of weight matrices of the LoRA we apply to Qwen-VL and LLaVA-1.5.

3.3 Mixture of LoRAs

The three LoRAs B_iA_i individually trained above can be assumed to have learned the corresponding VQA ability, and in order to integrate their respective wisdom for inference, we mix them together linearly by weights:

$$B_{MoL} = \alpha B_{KC} + \beta B_{VAR} + \gamma B_{SC} \quad (2)$$

$$A_{MoL} = \alpha A_{KC} + \beta A_{VAR} + \gamma A_{SC} \quad (3)$$

$$\theta_{MoL} = \{B_{MoL}, A_{MoL}\}^L_{l=1} \quad (4)$$

where α, β, γ are the weights of adapters, and θ_{MoL} is the parameter of the weightedly mixed adapter.

After mixture, assuming $\theta = \theta_0 + \theta_{MoL}$, the model f_θ is evaluated by the average score S_{avg} calculated on all groups:

$$S_{avg} = \sum_{k=1}^{3} \frac{1}{|G_k|} \sum_{j=1}^{|G_k|} \frac{1}{|G_{k_j}|} \sum_{s=1}^{|G_{k_j}|} Score(s, f_\theta(s)) \quad (5)$$

where $|G_k|$ and $|G_{k_j}|$ are the number of datasets in group G_k and number of samples in dataset G_{k_j}. The score $Score(s, f_\theta)$ is computed by the evaluation metric (refer to Appendix D) on sample s. This paper explores several methods for generating weights to achieve best average performance on the three groups with trained LoRAs, which is to find a set of α, β, γ that:

$$\alpha, \beta, \gamma = \underset{\alpha,\beta,\gamma\in[0,1]}{\text{argmax}}\ S_{avg} \quad (6)$$
$$s.t.\ \ \alpha + \beta + \gamma = 1$$

This paper explores three methods for mixture: Simple Average, Empirical Weights and Learned Weights.

Simple Average (MoL_{SA}) MoL_{SA} merges by simply averaging all LoRA adapters, i.e., all weights (α, β, γ in Equ. 6) are set to $1/3$. MoL_{SA} treats all VQA abilities with a same weight, regardless of the sizes of the corresponding groups in training and the overall priority of each ability.

Empirical Weights (MoL_{EW}) To catch an overall priority among the VQA abilities, MoL_{EW} merges by assigning a set of manually decided weights empirically, i.e., conducting a grid search on Equ. 6 for a best set of (α, β, γ).

MoL_{EW} surpasses MoL_{SA} in design by allowing for tendency towards different VQA abilities. However, it still ignores the fact that the multifarious questions coming from all datasets under

our setting depend on varied VQA abilities. For example, for a knowledge-focused question, asking about the function of a building in the image, Scene Comprehension seems much less useful than Knowledge Capability and Visual Attribute Recognition.

Such feature inspires us to dynamically adjust the weights α, β, γ of LoRA adapters to incline according to the focused VQA ability of each sample.

Learned Weights (MoL_{LW}) In order to dynamically identify and incline the model towards the required abilities of each sample, MoL_{LW} trains a small language model, Roberta-large (Liu et al., 2019), to generate a set of weights α, β, γ in Equ. 6. Assuming g for a Roberta model, the weights are generated by: $\alpha, \beta, \gamma = g(s)$.

Take the group VAR, for example. Given samples s^{VAR} from the group G_{VAR}, to train a g to analyze and allocate the weights according to s^{VAR}, we treat it as a three-label regression problem. The target is a set of weights, e.g., $(0.07, 0.82, 0.11)$ for $\alpha_{VAR}, \beta_{VAR}, \gamma_{VAR}$, which are grid-searched for a best performance on VAR. Loss function is Mean Squared Error (MSE):

$$L_s = \frac{1}{3}[(\alpha_{VAR}-\hat{\alpha})^2+(\beta_{VAR}-\hat{\beta})^2+(\gamma_{VAR}-\hat{\gamma})^2] \tag{7}$$

where L_s is the loss of sample s^{VAR}, and $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ are prediction results. The same applies for KC and SC as well.

During the inference stage of MoL_{LW}, since the generated weights are continuous and different by samples, the model needs re-initialization for each sample to assign precise weights to merge adapters, which is inefficient and unnecessary. Therefore, in order to reduce the extra cost from initialization and to initialize by batches instead of by samples, we use k-means clustering to cluster samples with similar weights together. The clustering centers are then used as weights for merging the adapters. The k is set to 20 with random initial centers.

4 Experiments and Analyses

This section presents the results from experiments and corresponding analyses towards different components in our method and the two MLLMs used as backbones. Implementation details are introduced in Appendix A.

4.1 Pilot Experiments

First of all, we wish to verify whether a single dataset is capable of enabling a model to master all VQA abilities (acquiring sound results on all datasets). Unfortunately, but also expectedly, the results are quite poor (refer to Appendix B for detailed pilot experimental results). It is clear that when trained on a single dataset, the model merely acquires acceptable results on its own test set, and its results on test sets from other datasets are generally quite poor. When trained on a mixture of all data simultaneously, although the generalization appears to be better, we still witness evident performance degradation on each test set.

Therefore, as there appears to be an inevitable trade-off between the generalization and specificity, inspired by MoE (Mixture of Experts) (Jacobs et al., 1991), our strategy is to ensure both of them simultaneously by the dynamic allocation and integration of LoRA adapters focusing on different VQA abilities.

4.2 Comparison of LoRA Mixture Methods

In order to explore the performance of the proposed methods in Section 3.3, we provide results in Table 2. According to the results, the proposed MoL_{LW} is the most effective mixing method across both Qwen-VL and LLaVA-1.5, surpassing MoL_{EW} by a notable average margin of 5.0% and 5.7% on Held-In, respectively. Compared with simply training on all data together (the first row), MoL_{LW} improves the Held-In results by 3.4% and 2.6%. Meanwhile, it is worthy to note that neither MoL_{SA} or MoL_{EW} obtains comparable Held-In results with simply training on all data together. We believe simply training on all data enables an automatic trade-off for required general VQA abilities, and thus performs better than grid-searched weights in MoL_{EW}. Since MoL_{LW} is capable of dynamically accommodating to different demands for VQA abilities, rather than fixing to a static allocation of focused abilities, it is much more flexible and versatile, achieving the best performance with clear margin.

The columns of Held-Out provide average results on the Held-Out group. Although MoL_{LW} still performs the best, its results from Qwen-VL and LLaVA are different in comparison to w/o MoL. For Qwen-VL, a simple MoL_{SA} is able to surpass w/o MoL by 2.8% on the Held-Out group, while neither MoL_{SA} or MoL_{EW} on LLaVA ob-

Methods	Qwen-VL-Chat					LLaVA-1.5-7b				
	KC	VAR	SC	Held-In	Held-Out	KC	VAR	SC	Held-In	Held-Out
w/o MoL	43.1	72.0	46.9	54.0	53.4	45.1	**68.4**	46.0	53.2	53.5
MoL_{SA}	39.5	66.6	41.2	49.1	56.2	27.7	54.8	40.3	40.9	40.6
MoL_{EW}	41.6	71.3	44.1	52.4	56.8	45.3	64.5	40.7	50.1	52.9
MoL_{LW}	**48.4**	**72.6**	**51.1**	**57.4**	**57.8**	**48.8**	67.9	**50.8**	**55.8**	**56.0**

Table 2: Experimental results from the three methods of mixture on two MLLMs. The tested MLLMs are Qwen-VL-Chat (Bai et al., 2023) and LLaVA-1.5-7b (Liu et al., 2023a). MoL_{SA}, MoL_{EW} and MoL_{LW} denote the three merging methods, Simple Average, Empirical Weights and Learned Weights as introduced in Section 3.3. The row of w/o MoL denotes the results from training a single LoRA adapter on the combination of all groups. KC, VAR, SC, Held-Out represent the average scores of the datasets belong to each group. To avoid direct influence from different numbers of datasets in each group, the column of Held-In is the macro average of scores in Held-In data, which is the average score of KC, VAR and SC, instead of individual datasets. Held-Out is the average results on the three Held-Out datasets.

tains better performance than w/o MoL. Also, there is an evident gap between MoL_{SA} and MoL_{EW} on both the Held-In and Held-Out LLaVA performance. Such phenomenon, as we deduce, is caused by the uneven amount of multimodal training data from Qwen-VL and LLaVA-1.5. The former imports about 1,450M samples for pre-training and instruction-tuning, while that for the latter is merely 1.23M (Liu et al., 2023a). We believe the training on overwhelming amount of data from Qwen-VL is not in vain and empowers Qwen-VL with better generalization ability, which makes LoRA adapters from Qwen-VL more stable and versatile in weighted mixing.

4.3 Mutual Influence Among Abilities

To take a deeper look at the mutual influence of the three VQA abilities, Table 3 provide clues. Results from the first three rows confirm that training for a single VQA ability is far from acquiring an acceptable generalization performance on other groups, which accords with results in pilot experiments that the limited amount and diversity of data in a single dataset or group is not competent for the One-to-Many VQA task. Yet from another perspective, the model trained on a different group is still able to acquire limited scores on the current group, with performance degradation. Thus it lies both commonality and distinction among the proposed three VQA abilities. Further, the commonality among groups may benefit the performance. Both the performance of MoL_{LW} from Qwen-VL and LLaVA-1.5 on the group of KC surpass training on KC itself (the first row) by 0.6% and 0.7% respectively.The boost brought by other groups suggests that biasing the model solely towards the required core ability of each sample does not guarantee best performance. On the contrary, the mixture of the VQA abilities will be more effective in general. Such pattern also applies for low-resourced datasets (like the group of KC here), and importing experts with different focus can be helpful.

4.4 Comparison with SOTAs

Table 4 provides comparison with previous specialist methods on OKVQA, KRVQA, COCO-QA and DAQUAR. Even compared to the specialist models designed for the corresponding dataset, our method surpasses four of them[2]. Especially in OKVQA and KRVQA, our method does not involve external knowledge bases or querying GPT-3 for assistance, which is one of the main sources of improvement for previous methods.

Table 6 reports the comparison of One-to-Many performance on ten Held-In datasets and three Held-Out datasets. We believe that Qwen-VL and LLaVA-1.5 are two of the most advanced multimodal language models with One-to-Many capability, which is why we select them as baselines and use them as backbones. As shown in Table 6 , our method has the best performance on most Held-In dataset with clear margins over OFA, Qwen-VL and LLaVA-1.5 without MoL, bringing much better general performance. As for the Held-Out datasets, our method always performs best in the zero-shot fashion, which verifies the generalization of our One-to-Many models.

[2]The SOTAs of FVQA and KBVQA benefit from utilizing the knowledge bases used to construct these datasets themselves.

Methods	Qwen-VL-Chat					LLaVA-1.5-7b				
	KC	VAR	SC	Held-In	Held-Out	KC	VAR	SC	Held-In	Held-Out
$Group_{KC}$	47.8	60.6	40.4	49.6	51.6	48.1	59.0	40.8	49.3	51.5
$Group_{VAR}$	36.6	72.4	42.3	50.4	55.3	39.3	**71.3**	41.4	50.7	55.6
$Group_{SC}$	34.0	59.4	**52.1**	48.5	50.2	36.4	61.0	**51.6**	49.7	51.6
MoL_{LW} (ours)	**48.4**	**72.6**	51.1	**57.4**	**57.8**	**48.8**	67.9	50.8	**55.8**	**56.0**

Table 3: Mutual influence among the three VQA abilities. $Group_{KC}$, $Group_{VAR}$ and $Group_{SC}$ denote results from training a single LoRA expert on the corresponding group, without mixture of experts. MoL_{LW} represents the results from the proposed method MoL_{LW}.

method	OKVQA
LXMERT (Tan and Bansal, 2019)	37.4
methods with external knowledge base	
TRiG (Gao et al., 2022)	49.4
TwO (Si et al., 2023)	56.7
methods with GPT-3 API	
PICa (Yang et al., 2022)	48.0
Prophet (Shao et al., 2023)	61.1
Qwen-VL MoL_{LW} (ours)	58.6
LLaVA-1.5 MoL_{LW} (ours)	**61.3**

method	KRVQA
Mucko (Yu et al., 2020)	24.0
KM-net (Cao et al., 2019)	25.2
DMMGR (2-steps) (Li and Moens, 2022)	31.8
Qwen-VL MoL_{LW} (ours)	31.3
LLaVA-1.5 MoL_{LW} (ours)	**32.8**

method	COCO-QA
VSE FULL (Ren et al., 2015)	57.8
DPPnet (Noh et al., 2016)	61.2
A+C+Selected (Wu et al., 2017)	71.0
Qwen-VL MoL_{LW} (ours)	**80.4**
LLaVA-1.5 MoL_{LW} (ours)	72.9

method	DAQUAR
DPPnet (Noh et al., 2016)	29.0
A+C+Selected (Wu et al., 2017)	29.2
SANs (Yang et al., 2016)	29.3
Qwen-VL MoL_{LW} (ours)	**38.1**
LLaVA-1.5 MoL_{LW} (ours)	37.7

Table 4: Comparison with previous specialist SOTAs.

5 Related Works

5.1 VQA Datasets

From the first general VQA dataset, DAQUAR (Malinowski and Fritz, 2014), and the much larger VQA v1 and v2, to assorted task-oriented datasets like CLEVR (Johnson et al., 2017), GQA (Hudson and Manning, 2019) and OKVQA (Marino et al., 2019), VQA datasets are becoming larger and more diverse in tasks, requiring various VQA abilities, like knowledge and complicated reasoning. To the best of our knowledge, there is no current VQA dataset aiming for the proposed One-to-Many task.

5.2 VQA Models

Classic VQA models usually follow a two-stage paradigm where image features are extracted by object detection and then interacted with the text feature (Anderson et al., 2018; Tan and Bansal, 2019; Li et al., 2019). During recent years, it is common for to leverage pretrained models for a better performance. Common VQA paradigms includes a pretrained visual-language encoder model with a classifier (Tan and Bansal, 2019; Li et al., 2019) or a transformer-based encoder-decoder generative model (Wang et al., 2022b,a; Lu et al., 2022). MLLMs are also available for VQA tasks (Alayrac et al., 2022), Pali (Chen et al., 2022), as well as the backbones in this paper, Qwen-VL (Bai et al., 2023) and LLaVA-1.5 (Liu et al., 2023a).

5.3 Mixture of Experts

Mixture of Experts, MOE (Jacobs et al., 1991), as a fusion method to integrate multiple FFNs or models, has boosted extensive researches (Zoph et al., 2022; Komatsuzaki et al., 2022; Kudugunta et al., 2021; Zadouri et al., 2023). MOE enables to significantly increase the model capacity as well as the size of a model while causing limited augmentation in inference consumption. Classic methods focus on route enhancing (Zhou et al., 2022; Zuo et al., 2021), of which the basic idea is to select a best expert for the current sample or token. The proposed MoL in this paper differs from previous methods significantly. The model for inference is a single model initialized from the mixture of various LoRA adapters, as opposed to multiple candidate experts, thus maintaining the same amount of parameters and computational cost. Meanwhile, MoL is easy to expand. Given a LoRA adapter trained on another VQA subtask, MoL treats it as an additional adapter and merge it into the previous three.

6 Conclusion

This paper proposes the task of One-to-Many Visual Question Answering, aiming at answering all sorts of common questions in the real world with a single model. To analyze and address the task, we break the task down into the integration of three key VQA abilities and investigate into ten datasets which are categorized into three groups accord-

ing to their emphasized abilities. Then, a Mixture of LoRAs (MoL) strategy is proposed with the aim of dynamically adjusting the capability of MLLMs (Qwen-VL and LLaVA-1.5) towards the focused ability of each sample. Experiments have verified the effectiveness of the proposed method, which significantly improves the One-to-Many performance and generalizes well to zero-shot test datsets. In addition, our method establishes new SOTA results on OKVQA, KRVQA, COCO-QA and DAQUAR, and competitive zero-shot performance on VQA abstract, VizWiz and A-OKVQA.

7 Limitations

Task Compatibility Although the proposed MoL strategy has been verified to be highly effective under the One-to-Many VQA task and has the potential to expand to other VQA subtasks by integrating more LoRA adapters, it is unclear how well it deals with LoRA adapters from another task beyond VQA, e.g., image captioning. In addition, as MoL does not introduce extra parameters or computation, due to the limit of model size and model capacity, merging too many LoRA adapters could cause potential overall performance degradation.

Potential Risk of Hallucination Merging LoRA adapters with weights may cause potential risk of hallucination. As LoRA adapters are individually trained to ensure convenient expansion in a plug-and-use fashion, adapters may import untrue information from other domains.

Limited Available Resources Due to the fact that there are not sufficient accessible and suitable general VQA datasets, results of zero-shot generalization come from the three Held-Out datasets, which might not be diverse enough to cover the three VQA abilities evenly and fairly.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*, pages 722–735. Springer.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Qingxing Cao, Bailin Li, Xiaodan Liang, and Liang Lin. 2019. Explainable high-order visual question reasoning: A new benchmark and knowledge-routed network. *arXiv preprint arXiv:1909.10128*.

Qingxing Cao, Bailin Li, Xiaodan Liang, Keze Wang, and Liang Lin. 2021. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2758–2767.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

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 preprint arXiv:2305.06500*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*, pages 1965–1973.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. 2022. Sparse upcycling: Training mixture-of-experts from dense checkpoints. *arXiv preprint arXiv:2212.05055*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. 2021. Beyond distillation: Task-level mixture-of-experts for efficient inference. *arXiv preprint arXiv:2110.03742*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Mingxiao Li and Marie-Francine Moens. 2022. Dynamic key-value memory enhanced multi-step graph reasoning for knowledge-based visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10983–10992.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

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*, pages 3195–3204.

Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 30–38.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 146–162. Springer.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

Qingyi Si, Yuchen Mo, Zheng Lin, Huishan Ji, and Weiping Wang. 2023. Combo of thinking and observing for outside-knowledge vqa. *arXiv preprint arXiv:2305.06407*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton Van Den Hengel. 2017. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1367–1381.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2022. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. *arXiv preprint arXiv:2205.11169*.

Jing Yu, Zihao Zhu, Yujing Wang, Weifeng Zhang, Yue Hu, and Jianlong Tan. 2020. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recognition*, 108:107563.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. 2023. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *arXiv preprint arXiv:2309.05444*.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.

Simiao Zuo, Xiaodong Liu, Jian Jiao, Young Jin Kim, Hany Hassan, Ruofei Zhang, Tuo Zhao, and Jianfeng Gao. 2021. Taming sparsely activated transformer with stochastic experts. *arXiv preprint arXiv:2110.04260*.

A Implementation Details

This paper utilizes Qwen-VL-Chat (7B) (Bai et al., 2023), LLaVA-1.5-7b (Liu et al., 2023a) and Roberta-Large (Liu et al., 2019) from hugging-face transformers (Wolf et al., 2020), LoRA (Hu et al., 2021) from hugging-face PEFT(0.6.1) (Mangrulkar et al., 2022), and the code is based on Pytorch(2.1.1) and hugging-face Accelerate(0.24.0). AdamW (Loshchilov and Hutter, 2017) optimizer is used with a peak learning rate 1e-4 for experiments whose training sizes are smaller than a hundred thousand, otherwise 3e-5. The hyper-parameters of AdamW, betas, eps and weight-decay are set to (0.9, 0.95), 1e-8 and 0.1, with a batch size of 4. LoRA rank is 64, with an alpha of 16 and a 0.05 dropout rate. Experiments are conducted on four Tesla A100 gpus. The evaluation metric used in this paper is VQA score (for samples with multiple candidate answers) and Exact Match (for samples with only one answer).

B Pilot Experimental Results

Table 5 provides results from our pilot experiments. It is clear that training on a single dataset is competent for the One-to-Many VQA task, as the performance fails to generalize to other datasets, especially to datasets from another group. Such phenomenon is expected, as different VQA questions share different focus and required abilities, like reasoning or knowledge capability, a single model is incapable of handling every ability simultaneously without significant drop in performance.

C Overall Performance Comparison

Table 6 provides comparison of our methods with previous specialist models and one-to-many baselines. For comparison, we train OFA-large, Qwen-VL-Chat and LLaVA-1.5-7b on all Held-In data together, training to empower them with the capability to handle questions demanding various skills. Yet as shown in the table, there are notable margins on most datasets.

D VQA Evaluation Metrics

VQA evaluation metrics contain Exact Match (Malinowski and Fritz, 2014) and VQA Score (Antol et al., 2015). They apply for different settings in VQA datasets. When only a single correct answer exists in each sample, like DAQUAR (Malinowski and Fritz, 2014), TDIUC (Kafle and Kanan, 2017), GQA (Hudson and Manning, 2019), the Exact Match metric is used. When each sample contains ten candidate answers, like VQA v2 (Goyal et al., 2017), OKVQA (Marino et al., 2019), VizWiz (Gurari et al., 2018), VQA Score is used.

Exact Match Exact Match calculates by judging whether the answer is identical to the annotated ground-truth answer, and if matches, the score will be 1, otherwise 0.

VQA Score VQA Score evaluates how many times the answer appear in the ten candidate answers, and mark the score according to the overlap, which is computed as follows:

$$accuracy = min(\frac{\#\ correct\ hits}{3}, 1)$$

As there are ten candidate answers, # correct hits represents numbers of matched answers. Therefore, as long as there are three or more candidates are the same with the predicted answer, the answer will be considered fully correct, and gets a score of 1.

E Dataset Preprocessing

The preprocessing of datasets affects more under the setting of One-to-Many VQA task than a single-dataset case. The inconsistency among datasets presents challenge to both the training and evaluation by a universal model. For example, the numbers in OKVQA are alphabetic numbers, like 1, 2 and 3, while that in KRVQA are English numbers, like one, two and three, and two-word answers in DAQUAR are all concatenated by a - instead of a space. Therefore, we need to align different formats. In addition, **any adjustment shall not make it unfair for comparison with results from other works**. Specifically, we argue a fair processing shall enable to restore each generated answer back to its original form according merely to its own training set.

First, we have verified that for numbers, almost all dataset used in this paper are either fully in alphabetic form or English form. Although FVQA has 5 samples of exception, 10 for VQA v1 abstract, we consider them to be negligible compared to the total amount of samples.[3] However, sub-

[3]There are a few samples in VQA v2 and KRVQA that contains answers of English number one or two. However, we observe that the questions of these samples actually are not about numbers. For example, a question asking What activity is the man doing contains a candidate answer of One, which is quite confusing but does not affect the fairness if we convert it to alphabetic form, because the question does not lead to counting.

Training Set	FVQA	OKVQA	KBVQA	KRVQA	TDIUC	COCO-QA	VQAv2	VG-QA	GQA	DAQUAR	*Avg.*
FVQA	**70.1**	47.6	36.3	9.9	81.8	59.4	73.5	34.2	52.2	25.1	47.3
OKVQA	54.1	**59.5**	34.0	10.1	78.3	55.7	70.9	30.1	53.2	28.0	46.3
KBVQA	55.5.	49.0	**43.4**	6.5	82.9	54.0	74.4	33.0	52.5	27.9	46.6
KRVQA	54.2	49.3	28.7	**31.8**	80.3	54.0	72.4	28.0	52.7	26.9	46.5
TDIUC	55.0	43.5	43.1	7.4	**92.7**	62.5	74.4	36.0	51.7	25.1	47.4
COCO-QA	54.7	42.8	41.0	7.1	83.7	**82.3**	73.8	33.8	52.1	28.1	48.3
VQA v2	53.8	45.6	43.0	9.0	86.0	60.0	**78.9**	34.7	52.7	27.0	47.5
VG-QA	52.5	42.1	40.5	7.0	85.0	59.5	75.9	**45.1**	51.1	27.1	47.0
GQA	50.8	42.9	39.4	8.6	77.7	60.4	72.8	31.1	**65.3**	28.3	47.6
DAQUAR	54.1	46.5	39.0	8.0	77.4	55.5	75.8	32.5	54.8	**39.9**	48.2
ALL	63.0	45.7	36.7	26.8	90.5	77.6	76.9	43.0	60.6	33.1	**54.0**

Table 5: Pilot experiments with specialist models. A LoRA model based on Qwen-VL-Chat is trained on a single dataset from the ordinate each time and tested on each dataset in the abscissa. Bold numbers indicate the best result tested on each dataset, which in this case, are all on the main diagonal. The last row, All, is a generalist model that trained with all datasets together.

Methods	KC				VAR				SC		Held-Out		
	FVQA	OKVQA	KBVQA	KRVQA	TDIUC	COCO-QA	VQAv2	VG-QA	GQA	DAQUAR	VQA abs	VizWiz	A-OKVQA
Specialist SOTAs	81.2	61.1	69.6	31.8	-	71.0	86.1	-	77.0	29.3	-	-	-
OFA-large†	54.5	42.1	36.7	29.2	**91.9**	74.6	76.4	42.2	61.0	32.2	63.7	25.9	49.4
Qwen-VL-Chat†	63.0	45.7	36.7	26.8	90.5	77.6	76.9	43.0	60.6	33.1	67.2	38.5	54.4
LLaVA-1.5-7b†	60.7	53.9	39.0	26.6	86.8	71.7	74.4	40.7	62.6	29.4	62.1	38.1	60.3
Qwen-VL MoL_{LW} (ours)	**64.0**	58.6	**39.6**	31.3	86.2	**80.4**	**79.0**	**45.0**	**64.1**	**38.1**	**70.1**	**44.6**	58.6
LLaVA MoL_{LW} (ours)	61.5	**61.3**	39.5	**32.8**	83.0	72.9	76.8	39.0	63.9	37.7	61.3	43.8	**62.9**

Table 6: Comparison on each dataset with baselines. † denotes results come from our implementation where models are trained on all groups. Specialist SOTAs denotes the SOTAs on each dataset from one-to-one task-specific models: FVQA (Li and Moens, 2022), OKVQA (Shao et al., 2023), KBVQA (Wang et al., 2015), KRVQA (Li and Moens, 2022), COCO-QA (Wu et al., 2017), VQA v2 (Chen et al., 2022), GQA (Yao et al., 2022), DAQUAR (Yang et al., 2016).As we have modified TDIUC and VG-QA to improve consistency (refer to Appendix E), no SOTA results are available. Note that all results of each group on Held-Out datasets come from zero-shot testing.

stantial number answers in VG-QA are either in alphabetic or English forms, which would be unfair to simply convert to a unified form. Due to the metric of evaluating the correctness of answers, answers with similar meaning but different types against the ground truth (like 1 against One) are treated as errors. Considering there are few work about the accuracy of VG-QA recently (perhaps due to the same reason), we do not compare its result with former ones and provide results under our setting as a benchmark, and suggest following work to maintain such setting for rationality and consistency. In addition, we find the first letter of all answers are either all uppercased or not, so it is fair to lower them.

In addition, for TDIUC, there are substantial samples (about 22.35%) with answers of *doesnotapply*, which refer to questions that are unanswerable. No other training dataset or group contains similar features. Consequently, when trained on other datasets or groups that does not contain TDIUC samples, the model is unable to predict *doesnotapply* and thus the performance on TDIUC drops significantly, causing interference for analyses. Therefore, for consistency among datasets, we remove the samples in TDIUC that are labeled with *doesnotapply*.

Therefore, in this paper, our preprocessing can be concluded as follows: 1) Mapping all numbers into alphabetic numbers. 2) Replacing the short dash - in DAQUAR answers with a space (two-word answers with a comma in the middle are not revised). 3) Removing the dot at the end of VG-QA answers. 4) Lowering all texts. 5) Removing samples with answers of *doesnotapply* in TDIUC. 6) Removing same samples that appear across any validation set with training sets to avoid sample leak.