Appendix

A Datasheet

A.1 Motivation

Q: For what purpose was the dataset created? This dataset is designed as a test-bed to investigate the behavior of Multimodal Large Language Models in continual instruction tuning. It specifically aims to address the lack of appropriate and diverse tasks for the instruction tuning of MLLMs.

Q: Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The dataset was created by the authors, who are affiliated with the Center for Future Media Lab (CFM) located in the Computer Science and Engineering department at the University of Electronic Science and Technology of China (UESTC).

Q: Who funded the creation of the dataset? This work is supported by grants from the National Key Research and Development Program of China (2022YFC2009903/2022YFC2009900) and the National Natural Science Foundation of China (Grant No. 62122018, No. 62020106008, No. 61772116, No. 61872064).

Q: Any other comments? No.

A.2 Composition

Q: What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Each instance represents a dialog between a human and an assistant, where the human asks a question based on the image, and the assistant answers the question based on its knowledge.

Q: How many instances are there in total (of each type, if appropriate)? As shown in Table 1, the dataset statistics are as follows:

- Grounding Task: 111,770 samples for training, 21,616 samples for testing.
- Classification Task: 117,715 samples for training, 4,600 samples for testing.
- VQAv2: 82,783 samples for training, 44,793 samples for testing.
- ScienceQA: 12,726 samples for training, 4,241 samples for testing.
- TextVQA: 34,602 samples for training, 5,000 samples for testing.
- GQA: 72,140 samples for training, 12,578 samples for testing.
- VizWiz: 20,523 samples for training, 8,000 samples for testing.
- OCR-VQA: 166,043 samples for training, 20,797 samples for testing.

Q: Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? Most tuning datasets use the complete data from the original datasets, except for grounding and ImageNet. For grounding, we use only one annotation per image. For ImageNet, we randomly select 100 categories from the total 1000.

Q: What data does each instance consist of? Each instance consists of an image, an identifier, and a conversation list that includes instructions from the user and responses from the assistant.

Q: Is there a label or target associated with each instance? Each instance includes a value from the assistant that describes the ground truth of the output.

Q: Is any information missing from individual instances? No.

Q: Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? No.

Q: Are there recommended data splits (e.g., training, development/validation, testing)? Yes, we have constructed the training and testing data. For validation, you can partition the training data as desired.

Q: Are there any errors, sources of noise, or redundancies in the dataset? No.

Q: Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? Since we construct the instruction from commonly used vision-language datasets, you need to download the images within these datasets, including GQA, TextVQA train,TextVQA test, ScienceQA, VizWiz train, VizWiz val, VizWiz test, OCR-VQA,ImageNet, COCO and the annotations for grounding RefCOCO,RefCOCO+,RefCOCOg.

Q: Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? No.

Q: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

Q: Does the dataset relate to people? No.

A.3 Collection Process

Q: How was the data associated with each instance acquired? Our datasets are derived from publicly available and widely used vision-language datasets, which we transform into an instruction style using commonly employed templates.

Q: What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? We use a Python script for auto-labeling to generate instruction-style data.

Q: If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Most tuning datasets use the complete data from the original datasets, except for grounding and ImageNet. For grounding, we use only one annotation per image. For ImageNet, we randomly select 100 categories from the total 1000.

Q: Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? No crowdworkers were involved in the curation of the dataset. Open-source researchers and developers enabled its creation for no payment.

Q: Over what timeframe was the data collected? The whole instruction tuning data was generated in 2023.

Q: Were any ethical review processes conducted (e.g., by an institutional review board)? The source data of each task was collected through ethical review processes.

A.4 Preprocessing/cleaning/labeling

Q: Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? We utilize an auto-labeling preprocessing script to generate the instruction labels for the dataset. Apart from this, no additional preprocessing or labeling is performed.

Q: Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? Yes, we need to download the original images from the datasets upon which ours is based. The URLs for these datasets have been provided above.

Q: Is the software that was used to preprocess/clean/label the data available? Yes, we construct the tuning data by scripts.

A.5 Uses

Q: Has the dataset been used for any tasks already? No.

Q: Is there a repository that links to any or all papers or systems that use the dataset? No.

Q: What (other) tasks could the dataset be used for? Although this dataset is created for continual instruction tuning, we can utilize the entire dataset for training a powerful assistant.

Q: Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? No.

Q: Are there tasks for which the dataset should not be used? This dataset should not be used for commercial.

A.6 Distribution

Q: Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? Yes, the dataset will be open-source.

Q: How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The data is available through `https://huggingface.co/datasets/Zacks-Chen/CoIN`.

Q: Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? CC-4.0.

Q: Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

Q: Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

A.7 Maintenance

Q: Who will be supporting/hosting/maintaining the dataset? We will be hosting the dataset on huggingface.

Q: How can the owner/curator/manager of the dataset be contacted (e.g., email address)? The authors can be contacted via their emails mentioned in the paper. Issues can also be opened on our public GitHub repo.

Q: Is there an erratum? Not to the best of our knowledge.

Q: Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Maybe, we will add more diverse task into the CoIN.

Q: Will older versions of the dataset continue to be supported/hosted/maintained? Yes.

Q: If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Not officially, but our benchmark code is open source and pull requests are welcome.

B Dataset details

The curated datasets are kept in JSON-files with the following keys:

- **id**: Identification for each instruction tuning sample.
- **image**: Image path.
- **conversation**: List of instructions and the answers from user and assistant.
  - **from**: Role of the instruction, human or gpt.
  - **value**: Instruction or answer details.

The constructed instruction training samples for each task are placed in **Instruction/Task_Name/train.json**, and testing samples are placed in **Instruction/Task_Name/test.json**. All code is accessible via the repository at `https://github.com/zackschen/CoIN`. In addition, all the training and testing instruction samples can be download from `https://huggingface.co/datasets/Zacks-Chen/CoIN`.

C Additional Experiments

Impact of Backbone Size To evaluate the influence of different model sizes of backbone, we add a larger architecture to evaluate performance across different model sizes. We choose LLaVA-13B as the new backbone to conduct experiments on our proposed benchmark. The comparison of *Truth Alignment* and *Reasoning Capability* between the 13B and 7B models is presented in the table below.

Table 10: The results evaluating the *Truth Alignment* of LLaVA about **different model size** are presented below.

Size	**Accuracy on Each Task**								**Overall Results**	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
7B	82.45	49.99	96.05	56.40	55.45	31.27	62.20	57.08	32.97	-32.62
	21.26	28.74	10.25	36.78	32.45	0.83	42.50	57.08		
13B	82.95	54.25	97.28	52.45	59.40	40.35	68.10	61.00	39.43	-28.79
	60.03	41.19	10.62	31.03	32.67	2.60	46.33	61.00		

Table 11: The results evaluating the *Reasoning Capability* of LLaVA about **different model size** are presented below.

Size	**Accuracy on Each Task**								**Overall Results**	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
7B	92	75	97	72	42	58	75	78	71.28	-10.88
		82	74	55	56	47	52	58	78	
13B	94	77	98	77	46	76	80	79	75.98	-11.00
	89	77	58	59	53	62	62	79		

From these tables, we have the following observations: The learning ability increases with model size, evident in both Truth Alignment and Reasoning Capability, resulting in 39.43% and 75.98% in terms of MAA, respectively. In addition, the increase in model size mitigates catastrophic forgetting in Truth Alignment, resulting in a 3.83% improvement in terms of BWT. We believe this occurs because, with the increase in size, the model maintains a larger optimization space to learn new knowledge, allowing it to avoid overlapping with old knowledge. Finally, the observed decrease in forgetting for Truth Alignment and the increase in forgetting for Reasoning Capability suggest that the forgetting of Instruction Following is mitigating. This phenomenon indicates that increasing the architecture's size effectively mitigates the forgetting of the Instruction Following ability, which is valuable for the practical applications of MLLMs.

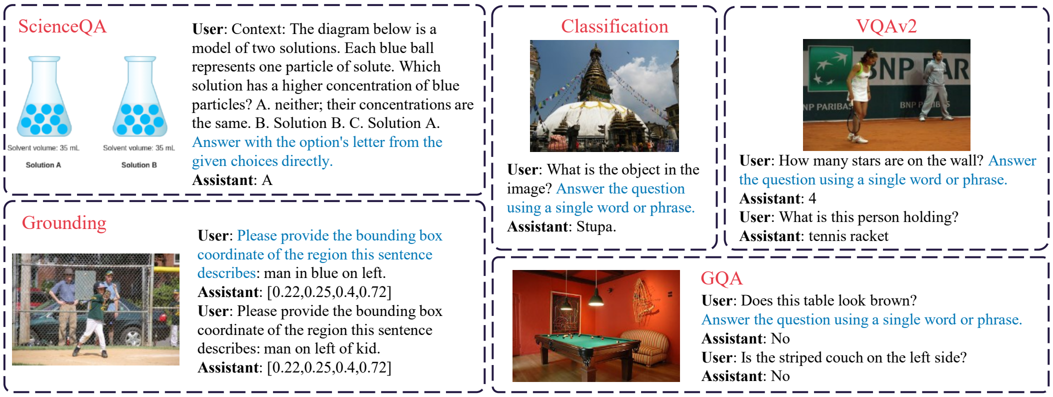
Impact of rank of LoRA The text knowledge always exists when the parameters of the base LLM are frozen, which is consistent with our training setting (Section 3.1.2 in the paper). Therefore, any forgetting primarily occurs in the multimodal knowledge acquired through the additional parameters introduced by LoRA which is very small compared with LLM. To examine this hypothesis further, we conduct additional experiments by increasing the rank of LoRA from 128 to 256. All experiments were conducted with a 40% data volume, as the experiments presented in Table 6 demonstrate that LLaVA achieves superior performance under this setting. The results are shown in the table below.

From Tab. 12, we first observe that performance improves as the rank increases, confirming that a higher number of trainable parameters enhances the model's ability to acquire new multimodal knowledge. Moreover, it is worth noting that knowledge forgetting is also reduced. This is likely because the additional parameters provide the model with sufficient optimization space to learn new multimodal information without overwriting previously utilized space.

Table 12: The results of LLaVA about **different rank of LoRA** are presented below.

Rank	**Accuracy on Each Task**								**Overall Results**	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
128	75.33	47.06	94.95	52.95	50.77	10.25	56.73	55.33	33.18 55.33	-24.85
		49.96	23.60	7.22	36.12	33.05	0.09	39.2		
192	76.30	49.52	97.17	53.87	50.05	7.72	62.90	61.08	38.31 61.08	-21.15
		68.82	40.63	8.72	35.70	30.45	2.95	41.08		
256	76.42	49.21	96.85	51.32	45.75	7.28	63.00	59.02	38.30	-19.8
	69.13	38.51	7.58	36.1	33.83	3.85	41.42	59.02		

D Examples of CoIN



User: Context: The diagram below is a model of two solutions. Each blue ball represents one particle of solute. Which solution has a higher concentration of blue particles? A. neither; their concentrations are the same. B. Solution B. C. Solution A. Answer with the option's letter from the given choices directly.
Assistant: A

User: What is the object in the image? Answer the question using a single word or phrase.
Assistant: Stupa.

User: How many stars are on the wall? Answer the question using a single word or phrase.
Assistant: 4
User: What is this person holding?
Assistant: tennis racket

User: Please provide the bounding box coordinate of the region this sentence describes: man in blue on left.
Assistant: [0.22,0.25,0.4,0.72]
User: Please provide the bounding box coordinate of the region this sentence describes: man on left of kid.
Assistant: [0.22,0.25,0.4,0.72]

User: Does this table look brown?
Answer the question using a single word or phrase.
Assistant: No
User: Is the striped couch on the left side?
Assistant: No

Figure 1: Examples of instruction tuning data in our proposed CoIN, which contains diverse visual understanding and perception tasks, such as classification, referring expression comprehension and image question answering.

To better show the constructed instruction data, we plot some examples with different tasks, as shown in Fig. 1. Blue text represents the instruction templates. We aim for the model to learn the capability of instruction following through these templates.

E *Truth Alignment* Comparison details

For the Image Question Answering task (including VQAv2, ScientQA, TextVQA, GQA, VizWiz, and OCR-VQA), we calculate the accuracy of predicting answers against ground truth, as in LLaVA [35]. In the classification task, the metric is computed by comparing predicted labels with real ones. For the referring expression comprehension task, we employ the widely used Intersection over Union (IoU) as the evaluation criterion to determine the success of the model's predictions. If the IoU of the predicted bounding box and the ground-truth bounding box is greater than 0.5, we consider the prediction to be correct.

F Experiments details

We conduct the experiments on LLaVA and Qwen-VL based on their official code. Following their official hyperparameters, we use the Adam optimizer with no weight decay and a cosine learning rate with a warmup ratio of 3%. During finetuning, gradient checkpointing is used to save GPU memory, and offloading is not used. BF16 and TF32 are enabled to achieve a balance between speed and precision. For MiniGPT-v2, we adapt the official code into $transformer$ training to utilize the $deepspeed$ for offloading frozen parameters to the CPU. In addition, we train all models with 8× 3090s with one epoch.

G More results about *Reasoning Capability*

Table 13: The results of LLaVA about **different task orders** are presented below.

Order	**Accuracy on Each Task**								**Overall Results**	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
Random	92	75	97	72	42	58	75	78	71.28	-10.88
	82	74	55	56	47	52	58	78		
	GQA	Grounding	ImageNet	OCR-VQA	ScienceQA	TextVQA	VizWiz	VQAV2	MAA	BWT
Alphabet	77	75	98	81	80	75	48	79	68.88	-3.00
	71	94	63	71	90	77	44	79		

Tab. 13 presents the results regarding the *Reasoning Capability* of different task orders. From these comparison results, we observe that the *Reasoning Capability* follows a similar trend to *Truth Alignment*, with the performance of Random being better than Alphabet. Additionally, the forgetting of Alphabet is also milder compared to Random order, resulting in a -3.00% decrease in terms of BWT, indicating that task order impacts *Reasoning Capability*.

Table 14: The results of LLaVA about **different instruction templates** are presented below.

Type	**Accuracy on Each Task**								**Overall Results**	
	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAV2	OCR-VQA	MAA	BWT
Original	92	75	97	72	42	58	75	78	71.28	-10.88
	82	74	55	56	47	52	58	78		
Diverse	92	76	97	71	47	61	71	80	72.54	-9.62
	80	74	53	54	46	73	58	80		
10Type	92	76	98	76	41	74	79	83	74.21	-10.00
	88	77	59	58	45	67	62	83		

Tab. 14 reveals the performance of *Reasoning Capability* on different instruction templates. These comparisons also reveal that the impact of templates on *Reasoning Capability* mirrors that on *Truth Alignment*: simply switching to different templates has minimal effect on overall performance, but increased diversity in templates enhances the robustness of the model.

H Instruction diversity detail

We devise two additional instruction templates to investigate the impact of varying templates. The details of these templates are shown in Tab. 15.

I MoELoRA details

The Mixture-of-Experts (MoE) aims to activate a subset of parameters for each input, enabling a significant increase in model parameters without a corresponding increase in computational efforts. In commonly used transformer-based models, MoE typically transforms the feed-forward layer of each transformer block into an MoE layer [14, 36, 16]. This MoE layer comprises two modules: experts and gate function. The experts are several identical and independent feed-forward neural

Table 15: The list of instructions template for each task.

Task	Original	Diverse	10Type
ScienceQA	Answer with the option's letter from the given choices directly	Answer with the option's letter from the given choices directly	Answer with the option's letter from the given choices directly Select the correct answer from the given choices and respond with the letter of the chosen option Determine the correct option from the provided choices and reply with its corresponding letter Pick the correct answer from the listed options and provide the letter of the selected option Identify the correct choice from the options below and respond with the letter of the correct option From the given choices, choose the correct answer and respond with the letter of that choice Choose the right answer from the options and respond with its letter Select the correct answer from the provided options and reply with the letter associated with it From the given choices, select the correct answer and reply with the letter of the chosen option Identify the correct option from the choices provided and respond with the letter of the correct option From the given choices, pick the correct answer and respond by indicating the letter of the correct option
Grounding	Please provide the bounding box coordinate of the region this sentence describes	Please provide the bounding box coordinate of the region this sentence describes	Identify and provide the bounding box coordinates that match the description given in this sentence Extract and provide the bounding box coordinates based on the region described in the sentence Please provide the bounding box coordinate of the region this sentence describes Find and provide the bounding box coordinates for the region mentioned in the sentence Provide the coordinates of the bounding box that correspond to the region described in the sentence Give the bounding box coordinates as described in the sentence Determine and provide the bounding box coordinates based on the description in the sentence Identify and provide the coordinates of the bounding box described in the sentence Provide the coordinates for the bounding box based on the region described in the sentence Extract and provide the coordinates for the bounding box described in the sentence Identify and give the coordinates of the bounding box as described by the sentence
GQA	Answer the question using a single word or phrase	Respond to the question briefly, using only one word or a phrase	Respond to the question with a single word or a short phrase Respond to the question using only one word or a concise phrase Answer the question with a single word or a brief phrase Respond with one word or a short phrase Provide your answer in the form of a single word or a concise phrase Respond to the question with just one word or a brief phrase Answer the question using a single word or a concise phrase Provide your response using only one word or a short phrase Respond to the question with a single word or a brief phrase Respond to the question using just one word or a concise phrase Answer the question with one word or a short phrase
ImageNet	Answer the question using a single word or phrase	Express your answer in a single word or a short, descriptive phrase	Express your answer in a single word or a short, descriptive phrase Provide your answer using a single word or a brief phrase Describe the content of the image using one word or a concise phrase Respond to the question with a single word or a short, descriptive phrase Classify the image content using only one word or a brief phrase Give your answer in the form of a single word or a concise phrase Use a single word or a short phrase to categorize the image content Express your answer with one word or a short, descriptive phrase Identify the type of content in the image using one word or a concise phrase Summarize your response in a single word or a brief phrase Use one word or a short phrase to classify the content of the image
OCR-VQA	Answer the question using a single word or phrase	Condense your answer for each question into a single word or concise phrase	Answer with the option's letter from the given choices directly Select the correct answer from the given choices and respond with the letter of the chosen option Determine the correct option from the provided choices and reply with its corresponding letter Pick the correct answer from the listed options and provide the letter of the selected option Identify the correct choice from the options below and respond with the letter of the correct option From the given choices, choose the correct answer and respond with the letter of that choice Choose the right answer from the options and respond with its letter Select the correct answer from the provided options and reply with the letter associated with it From the given choices, select the correct answer and reply with the letter of the chosen option Identify the correct option from the choices provided and respond with the letter of the correct option From the given choices, pick the correct answer and respond by indicating the letter of the correct option
TextVQA	Answer the question using a single word or phrase	Capture the essence of your response in a single word or a concise phrase	Answer the question with just one word or a brief phrase Use one word or a concise phrase to respond to the question Answer using only one word or a short, descriptive phrase Provide your answer in the form of a single word or a brief phrase Use a single word or a short phrase to respond to the question Summarize your response in one word or a concise phrase Respond to the question using a single word or a brief phrase Provide your answer in one word or a short, descriptive phrase Answer the question with a single word or a brief, descriptive phrase Capture the essence of your response in one word or a short phrase Capture the essence of your response in a single word or a concise phrase
VizWiz	Answer the question using a single word or phrase	Provide a succinct response with a single word or phrase	Answer the question using only one word or a concise phrase Respond to the question using only one word or a concise phrase Respond to the question with a single word or a brief phrase Provide your answer using just one word or a short phrase Respond with one word or a concise phrase Answer the question with just one word or a brief phrase Use a single word or a short phrase to answer the question Provide your answer in the form of one word or a brief phrase Reply to the question using one word or a concise phrase Answer with a single word or a short phrase Use one word or a brief phrase to answer the question
VQAv2	Answer the question using a single word or phrase	Answer the question using a single word or phrase	Answer the question using a single word or phrase Answer the question with a single word or a brief phrase Use one word or a short phrase to respond to the question Answer the question using just one word or a concise phrase Provide your answer to the question using only one word or a brief phrase Respond to the question with a single word or a short phrase Use a single word or phrase to answer the question Provide an answer using only one word or a brief phrase Answer the question succinctly with one word or a brief phrase Answer the question with just one word or a short phrase Respond to the question using a single word or a concise phrase

networks, and the gate function models the probability distribution to govern the weights of outputs from these expert networks. Specifically, for an intermediate representation x from the previous attention layer in models, the output of the MoE layer can be mathematically represented as follows:

$$h = \sum_{i=1}^{N} E_i(x)G(x)_i, \tag{3}$$

where the $E_i(\cdot)$ and $G(\cdot)_i$ denote i-th expert and the gate function. In addition, the gate function can be written as follows:

$$G(h) = Softmax(hW_g), \tag{4}$$

where W_g is the trainable weight within gate function $G()$.

Our goal is to tackle the challenge of catastrophic forgetting in the continual instruction tuning of MLLMs. We are inspired by MoE, which employs distinct experts to acquire various types of knowledge, akin to the expansion category of continual learning methods. Therefore, we bring the prevalent method MoELoRA [36, 14] in CoIN to utilize experts to acquire distinct knowledge for different tasks to mitigate forgetting.

The MLLMs in CoIN are fine-tuned in a parameter-effective way, i.e Low-rank Adaptation (LoRA) [25]. LoRA uses two low-rank matrices with rank r to update the knowledge and avoids changing the parameter of the learned model. Specifically, a certain transform feed-forward layer is parameterized with $W \in R^{d_{in} \times d_{out}}$, where d_{in} and d_{out} are the dimension of input and output, respectively. Two low-rank matrix $A \in R^{d_{in} \times r}$ and $B \in R^{r \times d_{out}}$ are used to learn extra knowledge with: $h = Wx + \frac{\alpha}{r}BAx$, where $x \in R^{d_{in}}$ and $h \in R^{d_{out}}$ denote the input and output vector, respectively. The rank r controls the number of trainable matrices. In addition, the constant hyper-parameter α facilitates the tuning of rank r [25].

To achieve the learning of diverse knowledge from different tasks, MoeLoRA proposes a set of experts to replace the LoRA matrices, denoted as $\{E_i\}_{i=1}^{N}$, where N denotes the number of experts. Therefore, the original computation will change to:

$$h = Wx + \frac{\alpha}{r}\sum_{i=1}^{N} G_i E_i(x) = Wx + \frac{\alpha}{r}\sum_{i=1}^{N} G_i B_i A_i x, \tag{5}$$

where G_i represents the gate function, which we will detail in the following paragraph. The matrices $A_i \in \mathbb{R}^{d_{\text{in}} \times \frac{r}{N}}$ and $B_i \in \mathbb{R}^{\frac{r}{N} \times d_{\text{out}}}$ represent the i-th expert of two low-rank matrices, each with a lower rank of $\frac{r}{N}$. With multiple experts in MoELoRA, the model can learn diverse task knowledge from different experts. Additionally, MoELoRA has the same number of trainable parameters as LoRA, indicating high efficiency.

Since there are many experts in each MoELoRA layer, the key is to create a suitable distribution of each expert to solve each task. As previously emphasized, to mitigate forgetting, the contribution of each expert should be tailored to specific tasks. Therefore, to regulate these contributions, a gate function is introduced. The gate function receives an input similar to the experts and outputs a contribution to choose suitable experts to solve the tasks. This computation is captured by the following equation:

$$G(x) = Softmax(xW_g), \tag{6}$$

where W_g is the trainable weight within gate function $G(\cdot)$. To balance the scale of the output distribution, a softmax operation is applied to normalize the contribution weights. This output distribution is utilized to incorporate the varying percentage contributions of each expert, as outlined in Eq. 5. Ultimately, all the outputs are concatenated to form the final output for the next layer.

J Related Work

J.1 Continual Learning

Recently, numerous methods have been proposed to mitigate catastrophic forgetting in the continual learning paradigm. These methods can be broadly categorized into three groups: *regularization-based*, *memory-based*, and *architecture-based* methods.

Regularization-based methods [52, 66, 1, 32, 7] focus on curing a continual learning network of its catastrophic forgetting by introducing an extra regularization term in the loss function. e.g, EWC[52] penalizes the changes of importance parameters when learning new tasks.

Memory-based methods [5, 49, 4, 39, 6] store previous samples or generate samples for replaying while learning a new task. Some methods [5, 49, 4, 4] use replayed samples from previous tasks to constrain the update of parameters when learning the new task. During training on a new task of EEC [2], reconstructed images from encoded episodes were replayed to avoid catastrophic forgetting.

Architecture-based methods [65, 27, 29, 43, 53] design new architecture modules to each task to prevent any possible forgetting. PNN [50] adds a network to each task and lateral connections to the network of the previous task while freezing previous task parameters. MNTDP [58] provides a learning algorithm to search the modules to combine with, where these modules represent atomic skills that can be composed to perform a certain task.

J.2 Instruction Tuning

Instruction tuning is a promising approach to enable the pre-trained model to follow natural language instructions and improve their generalization performance to unseen tasks. Some methods [64, 10, 59, 63, 11, 34] use the existing vision-language datasets to create instruction tuning data by different templates. At the same time, some methods [35, 60, 24, 62, 56] use the existing vision datasets to generate instructions based on powerful LLMs (e.g GPT-4 [47]). LLaMA [57] observes that a very small amount of instructions improves the performance on MMLU [23], and further improves the ability of the model to follow instructions. LLaVA [35] leverages ChatGPT [46] and GPT-4 [47] for multimodal instruction-following data collection, based on the widely existing image-pair data. InstructBLIP [11] transforms 26 datasets into the instruction tuning format and groups them into 11 task categories for fine-tuning. To further enhance the instruction-following capacity, SPHINX [34] collects instruction data from a wide range of multi-modal tasks, and jointly fine-tune the model to learn a vision generalist, instead of a specialist for specific scenarios.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [No]
   (b) Did you include complete proofs of all theoretical results? [No]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No]