On Token's Dilemma: Dynamic MoE with Drift-Aware Token Assignment for Continual Learning of Large Vision Language Models

Supplementary Material

A. Additional Experiment Details

A.1. More Implementation Details

In our experiments, we utilize the pre-trained, instruction-untuned LLaVA-v1.5 [38] as the backbone model for continual learning. The model employs Vicuna [11] as its language backbone and a pre-trained CLIP ViT-L/14 visual encoder [47] for visual feature extraction. During continual learning, only the newly added modules are trainable, while all other components remain frozen. In the default setting, 16 rank-4 LoRA experts are added when each new task starts, and $K = 16$ is applied for top-K routing over all trained experts. Different configurations are evaluated in ablation studies. The same configurations are also applied to the baseline model IncMoELoRA in experiments. All experiments are conducted on a compute node equipped with four NVIDIA H100 GPUs. Following the official LLaVA-v1.5 configuration, we adopt a global batch size of 128 and a learning rate of 2×10^{-4}. We set the warmup ratio to 0.03 and use the AdamW optimizer for training. The model is trained in PyTorch with BF16 precision and DeepSpeed ZeRO-2. We set the weights of the load balancing, exclusivity, and specialization losses to 1×10^{-3}. For all compared methods, we follow the default configurations from their original papers. Other remaining settings are consistent with those specified for LLaVA-v1.5 [38].

A.2. Details of Datasets

The eight tasks in the CoIN [7] benchmark are as follows:

ScienceQA (SQA) [42] is a multimodal science question-answering dataset designed to assess models' reasoning over integrated visual and textual information. The training set contains 12,726 samples (6,218 image–text and 6,508 text-only), and the test set includes 4,241 samples (2,017 image–text and 2,224 text-only).

TextVQA [58] focuses on text recognition within visual question-answering. It features real-world images with diverse textual content. The training set includes 34,602 image–text samples, and the test set comprises 5,000 image–text samples.

ImageNet [53] is a large-scale benchmark for image classification. The training set contains 129,833 image–text samples, and the test set includes 5,050 image–text samples.

GQA [26] emphasizes real-world visual reasoning, requiring understanding of object relationships and multi-step inference based on both synthetic and real images with scene graphs. The training and test sets include 72,140 and 12,578 image–text samples, respectively.

VizWiz [22] is designed for visual question-answering in assistive contexts for visually impaired users. It provides 20,523 training samples and 4,319 test samples, all in the image–text modality.

Grounding (Ref) [29] evaluates grounding of natural-language expressions in images. It contains image–text pairs requiring models to predict bounding boxes aligned with textual descriptions. The training set includes 55,885 samples, and the test set includes 30,969 samples.

VQAv2 [19], a visual question-answering benchmark, features balanced answer distributions and broad topical coverage. It provides 82,783 training samples and 214,354 test samples, all in the image–text modality.

OCRVQA [44] integrates OCR with visual question-answering to assess models' ability to extract and reason over textual content in images. The dataset includes 165,348 training samples and 99,926 test samples, all image–text.

Table 6. Performance under varying task orders.

Task Order	Aggregate Results (%)		
	MFN↑	MAA↑	BWT↑
Origin	57.03	57.70	-4.67
Reverse	56.67	57.34	-4.71
Alphabet	56.44	56.98	-4.92

B. More Experiments

B.1. Experiments on Different Task Orders

To assess order sensitivity, we trained our method under multiple task orderings on the eight CoIN datasets. We conduct experiments using three task orderings: "Origin", the original task ordering proposed in the CoIN benchmark; "Reverse", the reversed version of the original ordering; and "Alphabet", where tasks are ordered alphabetically. As summarized in Table 6, our method exhibits excellent stability across different task orderings; the amount of forgetting remains consistently low with minimal variation across orderings. Without techniques like experience replay, our proposed token-level expert assignment regularization within this incremental MoE with LoRA approach consistently learns new task knowledge with minimal forgetting, regardless of the task ordering. This is because our method effectively prevents ambiguous tokens from con-

Table 7. Performance under distinct training instruction types.

Instruction	Aggregate Results (%)		
	MFN↑	MAA↑	BWT↑
Origin	57.03	57.70	-4.67
Diverse	56.90	57.44	-5.12
10Type	56.77	57.25	-4.87

tributing to the learning of new routers and experts, ensuring that effective new knowledge is absorbed into the new experts without any assumptions about the incoming order of the data.

B.2. Experiments on Distinct Training Instruction Types

To validate the reliability of our proposed method against distinct instruction templates, we conduct experiments with different template types, reported in Table 7. There are three types of instruction templates in the CoIN benchmark [7]. Following the default setting, the experiments in the main paper are based on the "Origin" type. We further conduct experiments with the other two types of instruction templates, Diverse and 10Type, in [7]: 1) Diverse: Distinct instruction templates tailored to different tasks. 2) 10Type: Randomly sampled from 10 distinct instruction templates. (Details can be found in Table 16.) The results show that forgetting and accuracy on all three metrics are nearly identical across instruction types, indicating the method's stability. This result is significant, as it indicates that our method's token-level routing mechanism is not overfitting to superficial, task-specific prompt formats.

B.3. Ablation Studies on MoE Configurations

In this section, we validate the proposed method on different MoE configurations, including the top-K value, the number of experts, and the expert capacity, and show the results in Table 8, 9, and 10. Experiments with baseline IncMoELoRA are conducted as a reference. The ablation studies demonstrate that the proposed method performs robustly across different MoE configurations and consistently delivers improvements.

Ablations on top-K value. First, we conduct an ablation study on the top-K value, comparing top-8 and top-16 over all experts. As shown in Table 8, the overall performance is relatively insensitive to the choice of top-K, and under both configurations our LLaVA-DyMoE consistently outperforms the baseline method, demonstrating the effectiveness of the proposed approach.

Ablations on expert number. Second, we conduct an ablation study on the number of newly added experts, comparing 8 and 16 experts per task. In the 8-expert setting, we increase the parameters of each expert to maintain a comparable total capacity. As shown in Table 9, LLaVA-DyMoE consistently outperforms the baseline under both configurations, demonstrating its effectiveness.

Table 8. Ablations on top-K value.

Top-K	Method	Aggregate Results (%)		
		MFN↑	MAA↑	BWT↑
8	IncMoELoRA	48.55	49.87	-16.28
	LLaVA-DyMoE	56.89	57.71	-5.12
16	IncMoELoRA	49.68	49.50	-16.67
	LLaVA-DyMoE	57.03	57.70	-4.67

Table 9. Ablations on expert number.

Expert Number	Method	Aggregate Results (%)		
		MFN↑	MAA↑	BWT↑
8	IncMoELoRA	48.39	50.62	-17.93
	LLaVA-DyMoE	56.97	58.37	-5.78
16	IncMoELoRA	49.68	49.50	-16.67
	LLaVA-DyMoE	57.03	57.70	-4.67

Table 10. Ablations on expert capacity.

Expert Capacity	Method	Aggregate Results (%)		
		MFN↑	MAA↑	BWT↑
1	IncMoELoRA	48.23	49.14	-15.79
	LLaVA-DyMoE	56.78	57.34	-4.13
2	IncMoELoRA	49.08	49.25	-16.58
	LLaVA-DyMoE	56.91	57.48	-4.62
4	IncMoELoRA	49.68	49.50	-16.67
	LLaVA-DyMoE	57.03	57.70	-4.67

Ablations on expert capacity. Furthermore, we conduct an ablation study on the impact of expert capacity by varying the LoRA experts' rank to 1, 2, and 4. As shown in Table 10, models with larger expert capacity achieve better performance, and across all capacity settings LLaVA-DyMoE consistently outperforms the baseline, further demonstrating the effectiveness of the proposed approach.

Overall, these ablations show that LLaVA-DyMoE is robust to variations in the MoE configuration: it consistently outperforms the baseline across different choices of top-K, number of experts, and expert capacity.

B.4. Experiments with Different Backbone Sizes

Besides the 7B model, we further validate our method using the larger 13B LLaVA backbone, as shown in Table 11.

Table 11. Performance across different model sizes.

Size	Method	Accuracy on Each Task (%)								Aggregate Results (%)		
		SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA^{v2}	VQA^{OCR}	MFN↑	MAA↑	BWT↑
7B	IncMoELoRA	68.43	50.31	68.42	47.97	39.46	4.56	57.31	60.95	49.68	49.50	-16.67
	LLaVA-DyMoE (Ours)	76.25	53.86	95.80	48.40	52.35	9.25	58.30	62.00	57.03	57.70	-4.67
13B	IncMoELoRA	68.75	51.69	85.80	48.10	40.20	6.55	58.85	64.60	53.07	53.20	-14.23
	LLaVA-DyMoE (Ours)	78.75	56.24	96.05	55.85	53.20	13.85	64.05	65.15	60.39	61.25	-4.64

Table 12. LLaVA-DyMoE is compatible with task-level routing methods.

Method	Accuracy on Each Task (%)								Aggregate Results (%)		
	SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA^{v2}	VQA^{OCR}	MFN↑	MAA↑	BWT↑
LLaVA-DyMoE	76.25	53.86	95.80	48.40	52.35	9.25	58.30	62.00	57.03	57.70	-4.67
+ Task Router	78.18	53.36	95.63	54.63	53.92	24.46	59.54	60.40	60.02	60.78	-1.73

Table 13. LLaVA-DyMoE is compatible with data-based continual learning strategies. Table 5 in the main paper shows that the proposed token assignment regularization can work with replay techniques. This table shows the performance across different replay buffer sizes, with ProgLoRA [73] (containing replay) as a reference.

Replay Size	Method	Aggregate Results (%)		
		MFN↑	MAA↑	BWT↑
200	ProgLoRA	59.09	62.38	-6.59
	LLaVA-DyMoE	62.08	61.93	-1.55
500	ProgLoRA	59.14	62.74	-6.47
	LLaVA-DyMoE	62.55	62.17	-1.00
1000	ProgLoRA	59.66	63.23	-6.21
	LLaVA-DyMoE	63.19	62.95	-0.64

Scaling to a larger and stronger backbone model yields improved continual learning performance while preserving a low forgetting rate. LLaVA-DyMoE demonstrates robust scalability, effectively leveraging the increased capacity to achieve a higher MFN of 60.39% while maintaining a consistently low forgetting rate (-4.64%). This confirms that our drift-aware token assignment mechanism remains effective regardless of the underlying model size.

B.5. Additional Results on Data-based Strategies

The proposed drift-aware token assignment regularization in LLaVA-DyMoE is orthogonal and compatible with data-based strategies such as replay and data augmentation. By focusing on the core router training, our method can improve performance when combined with these techniques. In the main paper, we have provided the experiments in Table 5. In this section, we provide additional details on the results of replay techniques under different replay buffer sizes. We compare our method, LLaVA-DyMoE equipped with a standard replay buffer [51], against ProgLoRA [73]. This configuration serves as a basic replay-based variant of our dynamic MoE architecture. Following ProgLoRA, we vary the buffer size (200, 500, 1000) to match comparable replay budgets. As shown in Table 13, LLaVA-DyMoE consistently achieves competitive or better performance across different replay buffer sizes.

B.6. Compatibility with Additional Task-specific Router

Our LLaVA-DyMoE, which focuses on rectifying micro-level token routing drifts, is inherently orthogonal to and compatible with macro-level MCIT paradigms based on task-specific routing. In particular, our method can be seamlessly integrated into architectures that employ task-level routing strategies [20, 72, 73, 81]. These approaches first decide which group of experts to activate at the task level, while LLaVA-DyMoE then optimizes the token assignments within the selected group, mitigating the intra-group routing drifts we identified and thus providing complementary benefits.

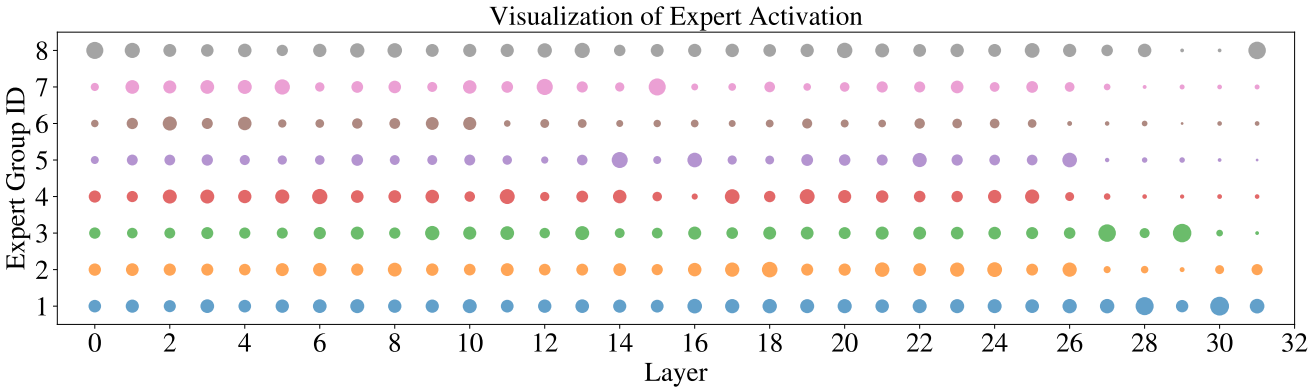
To verify this compatibility, we equip LLaVA-DyMoE with a task-specific router. In this setup, the task router determines which experts are activated for each task, while our dynamic MoE component can further regularize token-level routing. As shown in Table 12, this combination yields improved performance over vanilla LLaVA-DyMoE, demonstrating that LLaVA-DyMoE can provide additive gains when integrated with task-specific routing methods.

B.7. LLaVA-DyMoE with Expert Pruning

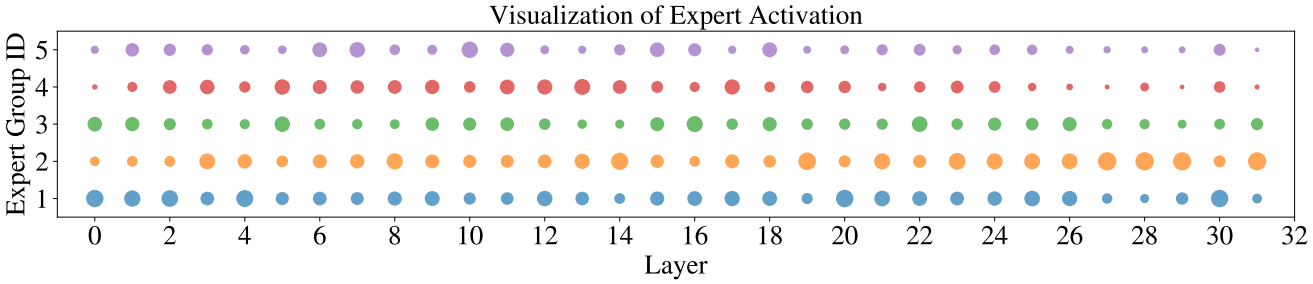
We evaluate the performance of LLaVA-DyMoE under different MoE configurations. In particular, we investigate expert pruning, which removes potentially unnecessary ex-

Table 14. Performance of LLaVA-DyMoE with expert pruning.

Method	Accuracy on Each Task (%)								Aggregate Results (%)		
	SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA^{v2}	VQA^{OCR}	MFN↑	MAA↑	BWT↑
LLaVA-DyMoE	76.25	53.86	95.80	48.40	52.35	9.25	58.30	62.00	57.03	57.70	-4.67
+ Pruning $1/8$	76.04	53.91	96.10	48.16	52.51	9.27	58.21	61.79	57.00	57.62	-4.63
+ Pruning $1/4$	75.59	53.79	95.11	47.78	52.51	9.26	57.94	61.39	56.67	57.37	-4.48



(a) LLaVA-DyMoE after training on the 8-th task.



(b) LLaVA-DyMoE after training on the 5-th task.

Figure 4. Layer-wise expert activation on the CoIN benchmark. Activation frequency is shown for each expert group across layers, and circle size reflects how often an expert is activated.

perts from the MoE. Table 14 reports the performance of LLaVA-DyMoE after pruning either $1/8$ or $1/4$ of the experts with the lowest activation frequencies following the training of each task. The results show that our method remains robust even with expert pruning. Note that we apply only a simple, naive pruning strategy, which leads to a slight performance drop. Although pruning is not the main focus of this work, this experiment demonstrates the potential of our proposed techniques to remain effective under more complex MoE training pipelines.

B.8. Visualization of Expert Activation

In Fig. 4, we present layer-wise expert activation frequencies of LLaVA-DyMoE across the eight tasks in the CoIN benchmark. For clarity, the activation frequencies of newly added experts for each task are merged into one single expert group. The routing scores are aggregated as the expert activation strength. The visualization shows that all experts are activated with varying strengths, exhibiting diverse utilization patterns across layers.

B.9. Qualitative Result Examples

We provide a qualitative comparison in Fig. 5 by randomly sampling data from previous tasks (ScienceQA and ImageNet) after the model has finished training on the final task. As illustrated, LLaVA-DyMoE successfully retains fine-grained knowledge that the baseline often forgets. Specifically, the IncMoELoRA baseline tends to regress to coarse-grained or incorrect labels, such as simplifying a "Bernese mountain dog" to a generic "Dog", or confusing a "Sloth bear" with a visually similar "Otter". In contrast, our method accurately recalls specific species and reason-

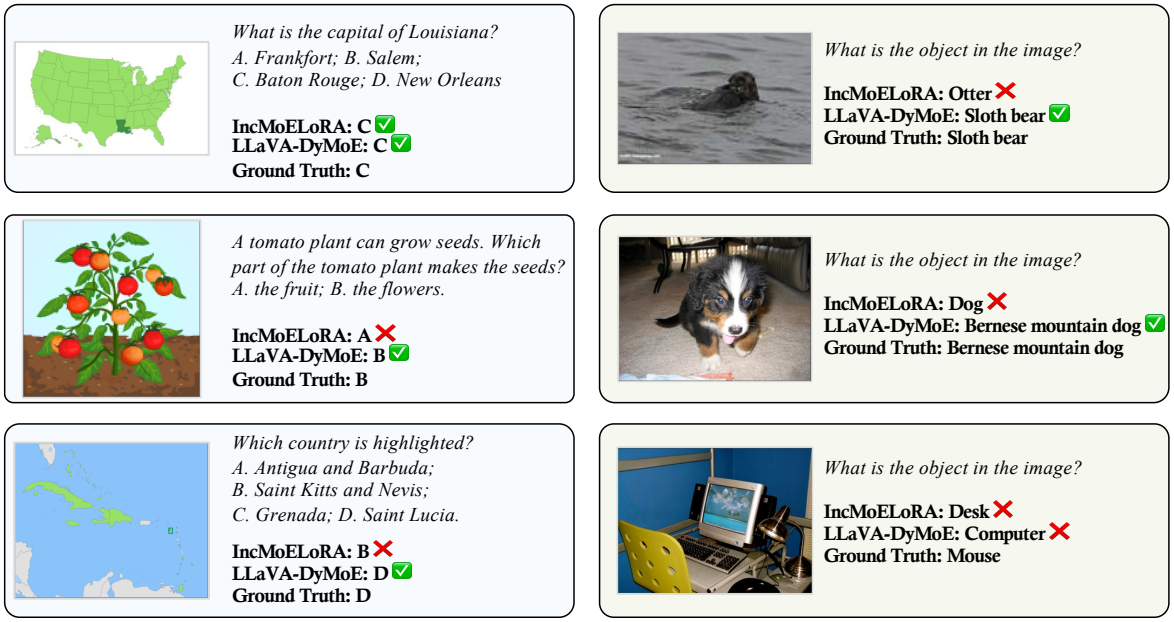


Figure 5. Comparisons between baseline IncMoELoRA and LLaVA-DyMoE on cases after training on the final task. The first column shows cases from ScienceQA, the second column shows cases from ImageNet.

Table 15. Training time of LLaVA-DyMoE during sequential training.

Method	**Training Time on Each Task (min)**								**Average**
	SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA^{v2}	VQA^{OCR}	
IncMoELoRA	7.4	13.6	92.4	123.4	14.9	99.6	105.0	137.1	74.18
LLaVA-DyMoE (Ours)	7.4	14.6	95.5	127.6	15.8	103.2	109.6	145.7	77.43

ing details, such as identifying the correct biological part of a "Tomato" plant. While complex scenes with small objects remain challenging for both models (e.g., the ambiguous Mouse case), our approach exhibits improved knowledge retention compared to the baseline across diverse domains.

B.10. Efficiency

The proposed drift-aware token assignment regularization is applied during training with minor additional computations. To validate the efficiency, we report the training time of the baseline (IncMoELoRA) and our LLaVA-DyMoE in Table 15. Our method incurs only a small training-time overhead of 4.4% (from 74.18 minutes to 77.43 minutes), while leaving inference efficiency unaffected.

C. Ethical and Social Impacts

This work advances MCIT by effectively enabling LVLMs to incrementally perform instruction tuning on new tasks while maintaining proficiency on previously learned ones. A key social benefit of our proposed LLaVA-DyMoE is its emphasis on parameter and inference efficiency. By utilizing a sparse MoE architecture, we minimize the computational energy required for long-term learning compared to dense retraining methods, aligning with the goals of Green AI. Regarding ethical considerations, we note that our model builds upon the pre-trained LLaVA backbone and standard datasets within the open-source CoIN benchmark. As with general data-driven LVLMs, our model naturally reflects the data distributions and characteristics of these foundational resources. While our current work focuses on optimizing knowledge retention and plasticity, we encourage future research to continue exploring safety alignment and fairness as integral components of the continual learning process for real-world applications.

Table 16. The list of instruction templates for each task [7].

Task	Original	Diverse	10Type
SQA	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
VQA^{Text}	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
ImgNet	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
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
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
REF	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
VQA^{v2}	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
VQA^{OCR}	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