Multimodal Continual Instruction Tuning with Dynamic Gradient Guidance

Supplementary Material

1. Dataset Distribution Analysis

In the experimental section of this paper, we observed that our algorithm exhibits distinct characteristics on datasets with similar versus disparate image distributions. To provide an intuitive illustration of these distributional differences across datasets, we visualize the three datasets employed in our study—VQAv2, UCIT, and our custom-designed dataset—to visually demonstrate the variations in their image distributions.

1.1. VQAv2 Dataset

The VQAv2 dataset is constructed based on the MS-COCO dataset, which consists of real-world photographs capturing diverse everyday scenes and objects. These images exhibit rich textual information and natural visual characteristics, with all samples in each task residing in a similar distribution space (see Fig. 1). Furthermore, different tasks within VQAv2 often share identical image data across various question-answer pairs (see the Recognition and Judge task in Fig. 1), resulting in minimal distribution shifts between tasks.

1.2. UCIT Dataset

The UCIT benchmark comprises six distinct sub-datasets with substantial distribution discrepancies:

- **ImageNet-R**: Contains various artistic and synthetic renditions of ImageNet classes, including paintings, sketches, and sculptures, representing a significant domain shift from natural images.
- **ArxivQA**: Comprises scientific figures and diagrams extracted from academic papers, featuring schematic representations and specialized visualizations.
- **VizWiz**: Consists of images captured by blind individuals using mobile phones, often containing practical everyday objects with varying quality and unconventional perspectives.
- **IconQA**: Features iconographic images and symbolic representations, characterized by simplified graphics and abstract visual elements.
- **CLEVR**: Utilizes synthetically generated 3D scenes with geometric shapes, exhibiting clean backgrounds and programmed object arrangements.
- **Flickr30k**: Contains natural photographs from the Flickr platform, depicting real-world scenes with diverse contextual elements.

From Fig. 2, we can observe substantial differences in image sources, visual characteristics, and content domains across these six sub-datasets, which result in significant distribution shifts and make UCIT a challenging benchmark.

1.3. Custom Dataset

In the ablation study on gradient approximation, to verify that visual data distribution differences affect our method's dependency on replay data, we construct a custom dataset sequence (VQAv2 $\rightarrow$ VizWiz $\rightarrow$ TextVQA $\rightarrow$ Flickr30k). The TextVQA dataset focuses on visual question answering tasks that require reading and understanding text within images to answer questions about textual content in visual scenes. From Fig. 3, it can be observed that although these four sub-datasets exhibit certain variations in specific visual properties, they primarily consist of real-world photographic data with rich textual information and natural scene representations. Compared to the UCIT benchmark, these datasets share more similar distribution characteristics due to their common origin in photographic imagery and comparable visual texture complexity.

2. More Implementation Details

Model Architecture and Fine-tuning Strategy. Our approach is built upon the LLaVA (Large Language-and-Vision Assistant) model, which represents a pioneering framework for integrating visual and linguistic understanding. LLaVA connects a pre-trained vision encoder with a large language model through a carefully designed projection layer that aligns visual features with the language model's semantic space. This architecture enables the model to process multimodal inputs by first encoding visual information through the vision encoder, projecting these features into the language model's embedding space, and then jointly reasoning about visual and textual information using the language model's transformer blocks.

For parameter-efficient fine-tuning, we employ Low-Rank Adaptation (LoRA), a technique that approximates weight updates through low-rank decomposition. Specifically, for a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA constrains its update by representing it as the product of two low-rank matrices:

$$W = W_0 + \Delta W = W_0 + BA \tag{18}$$

where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and the rank $r \ll \min(d, k)$. During training, only A and B are updated while W_0 remains frozen, significantly reducing the number of trainable parameters.

Training Configuration. All experiments were conducted with a consistent batch size of 32 across both

Figure 1. Illustration of VQAv2 dataset.

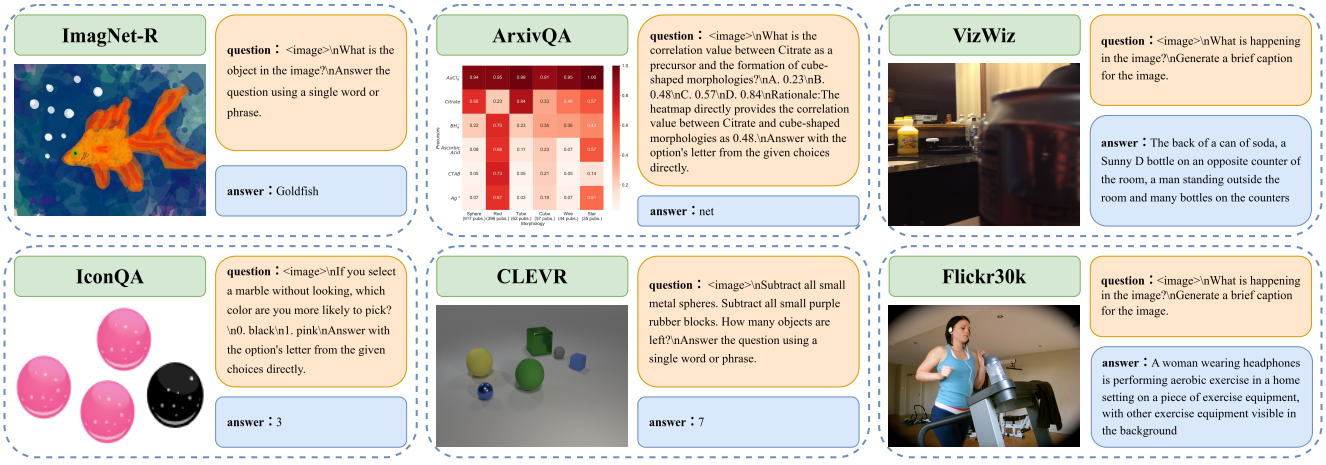

Figure 2. Illustration of UCIT dataset.

datasets and tasks. For the VQAv2 dataset, all subtasks except for the Causal task were trained for a single epoch, as this configuration provided sufficient convergence while minimizing computational overhead. The Causal task, which contains significantly fewer training samples compared to other subtasks, was trained for 4 epochs to ensure adequate learning. Similarly, all tasks in the UCIT dataset were trained for a single epoch to maintain consistency in training strategy across datasets. This differential training strategy ensures balanced optimization across all tasks regardless of their dataset sizes. The learning rate was maintained at 1×10^{-4} throughout the training process, with

Figure 3. Illustration of our custom dataset.

linear warmup and cosine decay scheduling applied for stable convergence. Each experiment is conducted over three trials.