Welcome to the supplementary repository for the paper

Evaluating Vision Language Models in Microscopy: A Benchmark for Scientific Image Analysis

Author Names
Domain Expert: Dr. X

`Link to the paper: will be added later`

This repository contains instructions for accessing datasets and creating subsets, and ground truths for counting and VQA created by the domain expert among the authors, code files, and some extra evaluations and results.

Data

Certain subsets were created from the publicly available datasets - NFFA, BBBC005, BBBC006, and cell division images - mentioned in the paper. The filenames and associated attributes (like labels) for these created subsets are provided here. The ground truths for counting created by the domain expert among the authors are also provided. Note that the datasets and the links to access them, and the subsets too have been already described in the manuscript.

NFFA-RS

The filenames and class labels for the 250 images in the NFFA-RS dataset are provided in the file `data_NFFA-RS.csv`. This was a randomly sampled subset of the NFFA dataset. Note that the prefixes "L1_, L2_, L3_" in the filenames were removed so that the models could not use this information to guess the class name.

NFFA-MS

The following two files contain the filenames and the ground truths for counts that our domain expert created for the fibers and the particles classes.
`data_ground_truth_NFFA-MS-fibers_counts.csv`
`data_ground_truth_NFFA-MS-particles_counts.csv`

BBBC005-S

The following two files contain the filenames and the ground truths for counting for the w1 and w2 subsets.
`data_ground_truth_BBBC005-S-w1_counts.csv`
`data_ground_truth_BBBC005-S-w2_counts.csv`

Cropped images

For the counting task on NFFA-MS, the images needed to be cropped in order to remove the instrument labels. The labels were interfering with SAM's ability to segment the desired artefacts. The cropped images are made available in the `data_images_NFFA-MS_cropped` folder.

VQA images

Some images were manually selected by our domain expert for VQA tasks from the NFFA dataset. These files are made available in the `data_images_vqa` folder. File names and file sizes have been retained from the original dataset.

VQA ground truths created by our domain expert

The filenames for images that were selected for VQA-T1 (describe the image) task from the NFFA dataset are listed in the file below along with the ground truths for the questions.
`vqa_t1_ground_truths.xlsx`
Note that the files and ground truths used for VQA-T2 (count number of fibers and particles) task were the same as those in NFFA-MS description above. For consistency, they have been copied over to `vqa_t2_counts_ground_truths.xlsx`.
Further, some images were manually selected by our domain expert for VQA-T2 (measure size of fibers and particles) task from the NFFA dataset. The filenames and the ground truths for the average size are listed in the file below. **The average size was calculated by our domain expert manually using the scale bar and ImageJ.**
`vqa_t2_size_ground_truths.xlsx`

Code

Readers and researchers are advised follow the instructions on the VLMs' parent websites to learn how to use and run the models. The specific models that we used, along with the links to their websites, are provided below:

- LLaVA (llava-v1.5-13b)
  https://github.com/haotian-liu/LLaVA
- Gemini (Gemini 1.0 Pro Vision)
  https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-pro-vision
- SAM (ViT-H SAM model)
  https://github.com/facebookresearch/segment-anything

We have provided the code we used to generate our data in the following files:

- `code_llava.py`
  Demonstrates how to run a prompt on multiple images and collect results using downloaded LLaVA model. This python file should be run through the terminal.
- `code_gemini.ipynb`
  This Jupyter notebook demonstrates how to run a prompt on multiple images and collect results using a Gemini model using an API key.
- `code_sam_dice_diff.ipynb`
  This Jupyter notebook demonstrates how to use a downloaded SAM model to segment a directory of images. It also reads the corresponding segmentation ground truth files and calculates Dice and F-1 scores, Jaccard index, and Hausdorff distance for each image. Additionally, difference images are also calculated and can be saved for each image.

Results

The results for classification, segmentation, and counting are fully visualized in the paper and its appendix. Interested users and researchers should be able to reproduce the results by using the code we provided above. Still, the following five files include the numerical data for the results that have been plotted in the manuscript and the appendix. The filenames are (hopefully) self-explanatory.

```
results_classification_NFFA-MS_chatgpt.csv
results_counting_BBBC005-S_w1.csv
results_counting_BBBC005-S_w2.csv
results_counting_NFFA-MS_fibers.csv
results_counting_NFFA-MS_particles.csv
```

Additional VQA results that could not be included in the manuscript or the appendix due to their expansive or supportive nature are provided below.

VQA

The performance of ChatGPT, LLaVA, and Gemini on the VQA tasks on select images from NFFA dataset was evaluated by our domain expert. Quality scores from 0-5 were assigned to each model for each question. The results are provided in the following files:

- `vqa_t1_evaluation.pdf`
  This file contains the filenames used for VQA-T1, the outputs of the models, and the quality scores assigned by our domain expert.
- `vqa_t2_size_results.xlsx`
  This file contains the filenames used for VQA-T2 (measure size of fibers and particles) and the outputs from the models. Note that since the experiments to measure sizes were less comprehensive than those for measuring counts (due to lack of ability to collect ground truths in large numbers), these were not included in the manuscript.