Documentation - Team Members

Project Title: Dataset Development & Annotation Tool

Organization: MealLens

Group 4

I have Assigned tasks to each member in a tabular format to avoid miscommunication on Slack where a channel was assigned for our team.

Team

Task	Name
6.0 - 6.4	@Akash
6.5 - 6.9	@Nishant Patel
Folder 6 a-f	@Aayush Thakur
7.0 - 7.4	@Tushar Sharma
7.5 - 7.9	@Sai Akshay Indla
Folder 7 a-f	@Sai R Kasi

1. **Name: Aayush Thakur**

Data Cleaning Report: Food Image Dataset (25GB Subset)

As part of a collaborative dataset preparation task, our team was assigned the responsibility of curating a comprehensive image repository totalling 25GB, which contained a mix of food and non-food images. After a series of planning discussions with the team lead, we agreed to divide the workload equally, **allocating 5GB of raw data per member** to ensure a structured and parallel approach.

I handled one 5GB segment with the goal of extracting only high-quality, contextually relevant food images for use **in future machine learning or computer vision pipelines**. The initial dataset contained significant noise, including images of garbage, random backgrounds, and unrelated content.

To perform this cleaning task, I followed a manual curation process, leveraging **domain knowledge and visual inspection techniques** to assess each image based on relevance, quality, and clarity. The steps included:

- Conducting a comprehensive scan of each directory and subfolder.
- Applying content-based judgment to assess whether images met the criteria for inclusion.
- Eliminating mislabeled, unclear, or unrelated visuals to enhance dataset consistency.

- Maintaining a clean and categorized directory structure to support scalability and reuse.
- This **refined dataset is now focused, noise-free,** and prepared for integration. I've shared my cleaned files and am ready to proceed with the final merging phase.

2. **Name: Tushar Sharma**

As part of a collaborative effort to curate a 25GB image repository containing both food and non-food images, our team **divided the workload equally,** with each member assigned a 5GB subset. My task was to clean a 5GB segment, isolating high-quality, relevant food images for future machine learning or computer vision applications. **The raw dataset included significant noise**, such as irrelevant images (e.g., garbage, random backgrounds).

To address this, I implemented a **manual curation process**, using domain expertise and visual **inspection to evaluate images based on relevance**, clarity, and quality. The process involved:

- Thoroughly reviewing all directories and subfolders.
- Assessing images against predefined inclusion criteria.
- Removing mislabeled, low-quality, or unrelated images to ensure dataset integrity.
- Organizing the refined dataset into a clear, scalable directory structure.

The resulting dataset is now streamlined, free of noise, and ready for integration. I have shared the cleaned files and am prepared to support the final merging phase.

3. **Name: Akash patil**

For a joint team effort, we were handed the responsibility of assembling a 25GB image dataset that included a variety of food and non-food visuals. To keep things streamlined and manageable, **we strategized with our team lead and settled on a divide-and-conquer approach**. The data was split evenly, giving each of us 5GB of raw images to tackle on our own, allowing for parallel progress without stepping on each other's toes.

I was assigned for curating one 5GB segment of image directories 6.0 to 6.4, with a specific focus on extracting high-quality, relevant food images. The original data **contained a significant amount of noise**, such as images of garbage, random objects, and unrelated backgrounds, which needed to be filtered out to improve the dataset's relevance for downstream use in machine learning and computer vision pipelines.

To carry out this task, I employed a **manual curation process based on visual inspection** and domain-specific judgment. I conducted a thorough scan of each folder and subfolder to review every image. Each image was **evaluated based on clarity, contextual relevance to food**, and overall quality. This hands-on approach allowed me to identify and remove blurry, mislabeled, illustrations, drawings or unrelated content effectively.

Beyond just cleaning the images, I took extra care to organize everything into a neat, well-structured folder system. Think of it as building shelves in a library each image has its rightful place, making it **easier to find, scale, and reuse later without any chaos**. The end result? A polished, clutter-free dataset that's all set to plug into the larger puzzle.

With my part wrapped up and the files handed over, I'm now geared up to join the rest of the team for the final merge and bring the complete dataset together.

4. **Name: Nishant Patel**

As part of **our collaborative effort to curate a 25GB dataset** containing both food and non-food images, I was assigned a 5GB subset for detailed cleaning and refinement. The primary goal was to isolate high-quality, contextually relevant food images suitable for future use in machine learning and computer vision pipelines.

The initial dataset included **a significant amount of noise such as unrelated visuals** (e.g., product packaging, poor lighting, backgrounds, or non-food subjects). To improve the dataset's usability and consistency, I followed a thorough manual curation process, applying both domain knowledge and visual inspection.

Key steps performed:

- **Systematic scan of all directories** and subfolders in the assigned 5GB portion.
- Content-based filtering to assess relevance, clarity, and whether food was the primary subject.
- Removal of irrelevant or low-quality images, including those with:
    - Non-food content
    - Excessive noise or blur
    - Distracting elements or poor framing
    - Misleading labels
- Image cropping was performed when needed to:
    - Focus on the food items
    - Enhance visual clarity without compromising context
    - Directory restructuring to ensure **a clean, scalable, and integration-ready dataset layout.**

5. Sai Akshay Indla

As part of our collaborative industrial internship project, our team was tasked with curating a large 25GB image dataset containing both food and non-food visuals. To handle this efficiently, **we adopted a divide-and-conquer strategy** under the guidance of our team lead. The dataset was split into equal portions, and I was responsible for processing a dedicated 5GB segment of image directories ranging from 6.5 to 6.9.

My focus was on identifying and retaining high-quality, contextually relevant food images while eliminating noisy or irrelevant content. This included filtering out images that featured non-food items, garbage, distorted visuals, or unrelated scenes that didn't contribute to the purpose of our machine learning and computer vision objectives.

To achieve this, I used a manual review process that involved visually inspecting each image in the assigned folders. By applying **domain knowledge and quality benchmarks** such as image clarity, relevance, and framing, I was able to methodically remove unsuitable content like cartoons, blurry shots, or mislabelled files. This approach, although time-intensive, ensured a higher degree of accuracy in the curation.

I also made it a point to maintain an organized folder structure throughout the process, so the cleaned dataset is not just accurate but also **easy to navigate and scalable** for future work. My part of the dataset is now refined and ready to be integrated into the collective dataset, contributing to a more robust and application-ready final output.