Multimodal LLM using Federated Visual Instruction Tuning for Visually Impaired

by

Ankith Bala

May 14, 2024

A thesis submitted to the

Faculty of the Graduate School of

the University at Buffalo, The State University of New York

in partial fulfilment of the requirements for the

degree of

Master's

Department of Computer Science and Engineering

To my family and friends

Acknowledgments

I would like to express my deepest gratitude to my thesis advisor, Dr. Alina Vereshchaka, of the Department of Computer Science and Engineering at the University at Buffalo. Her guidance, encouragement, and insightful feedbacks have been invaluable throughout the course of this research. Her expertise and support have been instrumental in shaping this thesis, and I am grateful for her mentorship.

I am also deeply thankful to the members of my thesis committee, Dr. Chunming Qiao, for his valuable suggestions and constructive criticism during the thesis defense. I am immensely grateful to the Department of Computer Science and Engineering at the University at Buffalo for providing a conducive environment for research and learning.

A special thanks to my friends at the University at Buffalo for their enthusiasm in watching my progress, their valuable suggestions, and their constant encouragement, which pushed me to give my best.

I extend my heartfelt thanks to my family for their unwavering support and patience. My parents, brother and sister have been a constant source of encouragement, and their belief in my abilities has been a driving force behind my achievements.

Thank you all for your support and contributions, which have made this thesis possible.

Table of Contents

List of Tables

List of Figures

Abstract

Training multimodal large language models (LLMs) is computationally intensive often requiring thousands of GPU hours and poses challenges in resource allocation and efficiency. Additionally, the centralized handling of training data in traditional approaches poses privacy risks, particularly when dealing with sensitive information. To address these issues, this thesis introduces a federated learning approach that enhances privacy through decentralized data processing, allowing for model fine-tuning without compromising user data security. Additionally, it extends foundational research on visual instruction tuning by applying fine-tuning to multimodal language-image instruction-following data, thereby creating multimodal LLMs. The fine-tuning process is optimized for computational efficiency, completing in $\sim$36 hours on a single 2-A100 node.

Utilizing these multimodal capabilities, we introduce an assistive application tailored for visually impaired individuals that enables real-time, interactive, and descriptive engagement with users' surroundings, thereby enhancing their understanding and interaction with the environment. More broadly, this research underscores the effectiveness of combining federated learning with visual instruction tuning to build secure, computationally efficient multimodal LLMs. These advancements make significant strides in highly sensitive and personal contexts such as assistive technology in more adaptable, secure, and personalized way for the visually impaired.

Chapter1

Introduction

The emergence of artificial intelligence (AI) has sparked a revolution across various domains of technology [1], with particularly profound implications in the field of assistive technologies [2]. In the United States, approximately 25% of adults live with some type of disability, with an estimated 12 million of these individuals actively using various assistive technologies in their daily lives to enhance mobility, communication, and daily functioning [3]. These technologies, which range from simple aids to complex interactive systems, are essential for enhancing the quality of life and independence of individuals with disabilities. To further enhance the capability and effectiveness of these assistive technologies, there have been significant advancements in the AI domain [4].

Among the most recent developments in AI are multimodal large language models (LLMs), which offer the capability to process and synthesize information from multiple sensory modalities such as images, audio, and text at the same time [5]. These models have proven effective in various real-world applications, including robotics [6], education [7], and assistive technologies [8], significantly enhancing interaction capabilities for individuals with disabilities.

Training these sophisticated models is a formidable challenge. It requires vast computational resources and access to diverse, extensive datasets, which can be expensive and pose significant privacy concerns [9]. For instance, the pre-training of large-scale models like LLaMa-2-70B demands immense amounts of energy and computational time, consuming approximately 1.7 million GPU hours and 2.5 trillion joules of energy, leading to significant CO_2 emissions estimated at 291 tons [10]. Such resource-intensive processes highlight the environmental impact and the high cost of developing foundation models.Furthermore, the need for large-scale, high-quality instruction data to fine-tune these models for specific tasks or user intents complicates the development of multimodal (LLMs) [11]. For example, in healthcare settings, creating multimodal LLMs that can handle sensitive patient data for diagnosis assistance involves strict privacy measures to prevent any potential misuse of personal health information [12]. This requirement often deters medical institutions from sharing valuable data, thus limiting the scope of AI's potential benefits.

To address these challenges, there is a growing interest in innovative model training architectures and privacy-preserving techniques. Multimodal instruction tuning is a technique that aligns LLMs with specific user instructions across different modalities, enhancing their applicability and performance in targeted tasks. However, implementing this technique across distributed data sources brings us to the pivotal role of federated learning. Federated learning [13] presents a solution by decentralizing the training process of machine learning models. Instead of pooling data into a single location, federated learning allows data to remain securely on local devices, with only the model's updates being shared across the network. By integrating federated learning with multimodal instruction tuning, this thesis aims to address both the performance enhancements needed for effective AI applications and the stringent privacy requirements crucial in sensitive sectors like healthcare and assistive technologies.

The subsequent sections of this thesis are structured to provide a comprehensive overview of the research conducted, divided into several key areas. Chapter 2 provides a detailed background on the current state of multimodal (LLMs), exploring their capabilities, the specific challenges they face and related works in this field. Chapter 3 details the system architecture, data preparation, and the assistive application components. Chapter 4 describes the experimental setup and methodologies employed in this study, including the innovative application of federated learning to visual instruction tuning for multimodal LLMs.

Chapter 5 presents the results of the experiments, offering an analysis of the performance enhancements and the efficiency of the privacy-preserving techniques used. Chapter 6 discusses these results in the context of their impact on assistive technologies for the visually impaired, highlighting how federated learning can mitigate the risks associated with traditional model training methods.

Finally, Chapter 7 concludes the thesis with a summary of findings and a discussion on the implications of this research for the future deployment of multimodal LLMs in real-world applications. This final chapter includes an in-depth discussion on how federated learning can be effectively implemented to optimize both the utility and confidentiality of multimodal LLMs, ensuring their adaptability and security in sensitive applications.

1.1 Background

Multimodal large language models (LLMs) stand at the forefront of artificial intelligence (AI) research, offering unprecedented capabilities to integrate and process information across different sensory modalities such as visual, auditory, and textual data [5]. These models play a crucial role in bridging the gap between sensory data processing and computational intelligence, enabling more intuitive human-machine interactions. However, they face computational and resource allocation challenges, especially when handling complex multimodal data [10].

The concept of multimodal LLMs was pioneered by early efforts such as Visual Question Answering (VQA) models proposed by Antol et al. in 2015 combined image recognition with natural language processing to answer questions about images [14]. Another significant milestone was the development of the CLIP (Contrastive Language–Image Pre-training) model by OpenAI, which aligned images and text in a shared embedding space, enabling zero-shot transfer to a variety of vision tasks [15]. These foundational models demonstrated the potential of integrating different modalities to enhance machine understanding and interaction capabilities.

Following these pioneering efforts, several advanced multimodal LLMs were developed. For example, DALL-E by OpenAI, which could generate images from textual descriptions, showcased the creative potential of such models [16]. The ALIGN model by Google further pushed the boundaries by using massive datasets to achieve high performance on cross-modal tasks, setting new benchmarks in the field [17].

Despite these advancements, multimodal LLMs continue to face computational and resource allocation challenges. The pre-training of large-scale models like LLaMa-2-70B, for instance, demands huge amounts of energy and computational time, consuming approximately 1.7 million GPU hours and 2.5 trillion joules of energy, leading to significant CO2 emissions [10].

Given these challenges, there is a pressing need for more efficient approaches to training and deploying multimodal LLMs. LLaVA, a state-of-the-art multimodal LLM [18], utilizes visual instruction tuning to effectively bridge visual and linguistic information, setting a new standard in multimodal learning efficiency and performance. This model serves as an exemplar for addressing computational intensity and resource constraints in multimodal LLM deployment. However, the deployment of such advanced technologies also raises concerns about data privacy and security, particularly when processing sensitive personal data.

To mitigate these concerns, federated learning emerges as a pivotal solution. Federated learning allows for the decentralized fine-tuning of models on local devices without the need to centralize user data [13]. This approach not only preserves the privacy of individual data but also enables personalized model adjustments based on localized data, thus enhancing the model's effectiveness and applicability in diverse real-world environments.

While federated learning has been explored within the scope of conventional LLMs, its application to visual instruction tuning in multimodal LLMs remains underexplored. Our research builds upon foundational concepts by applying federated learning to the visual instruction tuning within multimodal large language models for the first time as of our knowledge. We demonstrate how federated learning can enhance privacy and adaptability in the tuning of these complex models, offering significant benefits for developing assistive technologies for the visually impaired.

This integration of federated learning into multimodal LLMs using visual instruction tuning represents a significant advancement for assistive technologies. It addresses not only computational and resource challenges but also enhances privacy and personalization of user data, which are critical in applications for visually impaired users. This approach underscores the potential of multimodal LLMs to transform assistive technology, making it more secure, adaptable, and closely aligned with users' specific needs.

1.2 Problem Statement

Visually impaired individuals face challenges in understanding and navigating their environment, underscoring the necessity for advanced assistive technologies. Multimodal(LLMs) have the potential to revolutionize assistive solutions through their ability to process and integrate diverse data types, including visual information. Yet, the development and deployment of such models are hindered by their substantial computational requirements and the complexities of resource allocation and efficiency. These challenges are compounded in real-time applications crucial for assistive technologies, where delay or inefficiency can diminish the user experience and practicality.

Moreover, the centralized handling of sensitive personal data in traditional AI training methods raises significant privacy concerns. Ensuring the privacy and security of user data is paramount, especially in applications involving vulnerable populations such as the visually impaired. Federated learning presents a promising approach to mitigate these privacy risks by enabling decentralized training of models directly on local devices, thus eliminating the need to share raw data.

This research seeks to address both the computational and privacy challenges inherent in applying multimodal LLMs to the development of assistive technologies. By leveraging the visual instruction tuning approach exemplified by models like LLaVA and integrating federated learning, this study aims to create effective, real-time, and privacy-preserving solutions for visually impaired users.

1.3 Significance of the Study

This study aims to bridge the gap between advanced AI technologies and practical assistive solutions, potentially setting a precedent for future research in the field and aims to advance the application of multimodal LLMs in real-world assistive technologies, addressing the challenges of computational intensity and resource allocation while enhancing the quality of life for visually impaired individuals.

Chapter2

Related works

2.1 Assistive Technologies for the Visually Impaired

The advent of artificial intelligence (AI) technologies offers promising avenues to mitigate the challenges faced by visually impaired individuals. Recent research underscores the deployment of AI to develop assistive tools that enhance navigation and environmental perception for the visually impaired, addressing a significant gap in existing technological aids [19]. These advancements illustrate how AI can process and transform raw data into actionable insights, enabling users to navigate and interact with their surroundings more effectively.[20, 21, 22].

As the field of AI progresses, large language models (LLMs) have gained prominence, leading to the development of multimodal LLMs that incorporate visual, textual, and sometimes auditory data to provide richer, more context-aware responses. This evolution marks a significant step towards creating assistive technologies that are more intuitive and user-friendly for visually impaired individuals [23, 24, 25].

Multimodal large language models (LLMs) integrate various data types—text, images, and sometimes audio—to enhance their performance and applicability in a range of contexts. This integration enables multimodal LLMs to provide richer, more nuanced responses by synthesizing information across different modalities, a capability that is especially beneficial for applications requiring an in-depth understanding of complex, multimodal inputs. For instance, Wang et al. provide a comprehensive overview of multimodal LLMs, discussing their architectures, training techniques, and applications, illustrating the breadth of their potential uses [26, 27, 28].

Among the notable advancements in this domain is ViLBERT by Lu et al., which utilizes a transformer-based model to learn joint visiolinguistic representations, showing significant improvements in various vision-and-language tasks [29, 30, 31]. Such developments underscore the effectiveness of multimodal LLMs in interpreting and interacting with complex data.

However, the training and deployment of multimodal LLMs are not without challenges. These models are computationally intensive and require substantial resources, which can impede their scalability and practicality. Addressing these computational demands is crucial for the advancement and broader adoption of multimodal LLMs in real-world applications [32, 33, 34].

The LLaVA framework addresses these challenges by introducing visual instruction tuning, a method that significantly advances the processing and interpretation of visual data within multimodal LLMs [18]. This fine-tuning approach has proven effective in improving LLaVA's performance, making it a valuable asset for developing assistive technologies [35].

Building upon LLaVA's achievements, our research aims to further harness and adapt these multimodal strategies to create an assistive system for visually impaired users. By focusing on LLaVA's strengths in visual data processing and its potential for real-time, interactive feedback, the project seeks to develop a technology that not only enhances spatial awareness but also facilitates a more engaged and autonomous interaction with the environment for visually impaired individuals.

2.2 Federated Learning in Assistive Technologies

Federated learning (FL) has emerged as a pivotal solution for addressing data privacy concerns in the development of AI systems, including assistive technologies for the visually impaired. By allowing data to remain on local devices and only sharing model updates, federated learning ensures that sensitive information is not centralized, thereby enhancing user privacy [13].

This approach is particularly relevant in the context of assistive technologies, where users' visual data and other personal information are highly sensitive. Federated learning enables the training of robust AI models without compromising the privacy of the individuals who rely on these technologies. This is crucial in building trust and widespread acceptance among users, especially those concerned with digital privacy [36, 37]

Recent advancements in federated learning have also focused on improving the efficiency and effectiveness of these models under non-IID (non-identically and independently distributed) data conditions, which are common in real-world scenarios where data distribution varies significantly across devices. These improvements are vital for assistive technologies as they ensure that the AI systems can perform well across diverse environments and user experiences [38].

Our research integrates federated learning with the LLaVA framework to further develop an assistive system tailored for visually impaired users. By leveraging the strengths of both federated learning and multimodal LLMs, this system aims to provide a highly effective, secure, and personalized user experience. The ultimate goal is to create a scalable solution that respects user privacy while offering advanced assistive capabilities, thereby paving the way for more innovative applications in the field.

Chapter3

Methodology

3.1 System Component

The assistive technology system is designed to leverage an enhanced visual-linguistic model, fine-tuned to process and interpret visual data alongside textual or auditory inputs, providing users with descriptive feedback about their surroundings. The system architecture prioritizes real-time processing and user-friendly interaction to deliver immediate, relevant feedback based on user inputs and visual scene analysis.

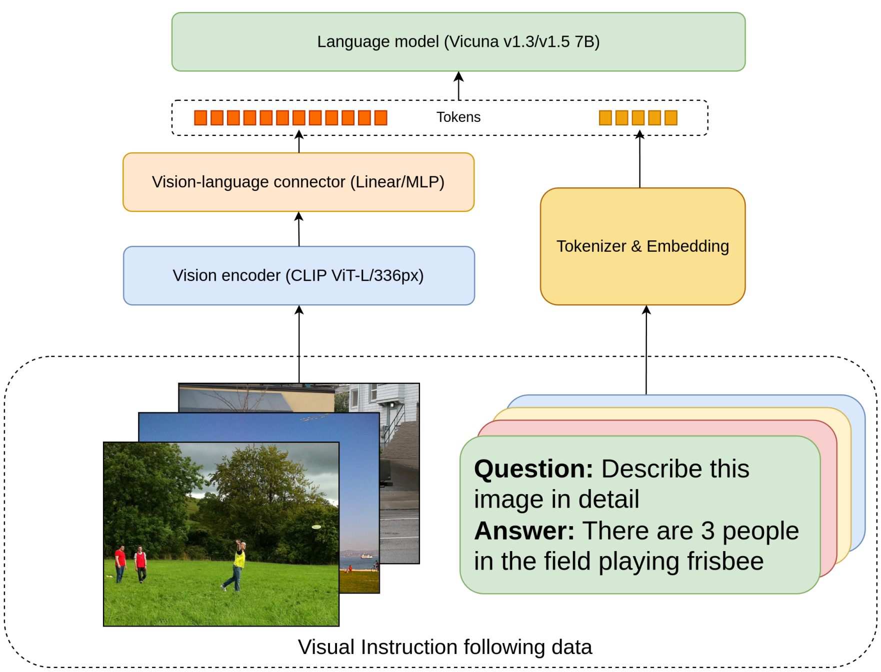

Figure 3.1: System Components

The architecture, depicted in Figure 3.2, integrates several components:

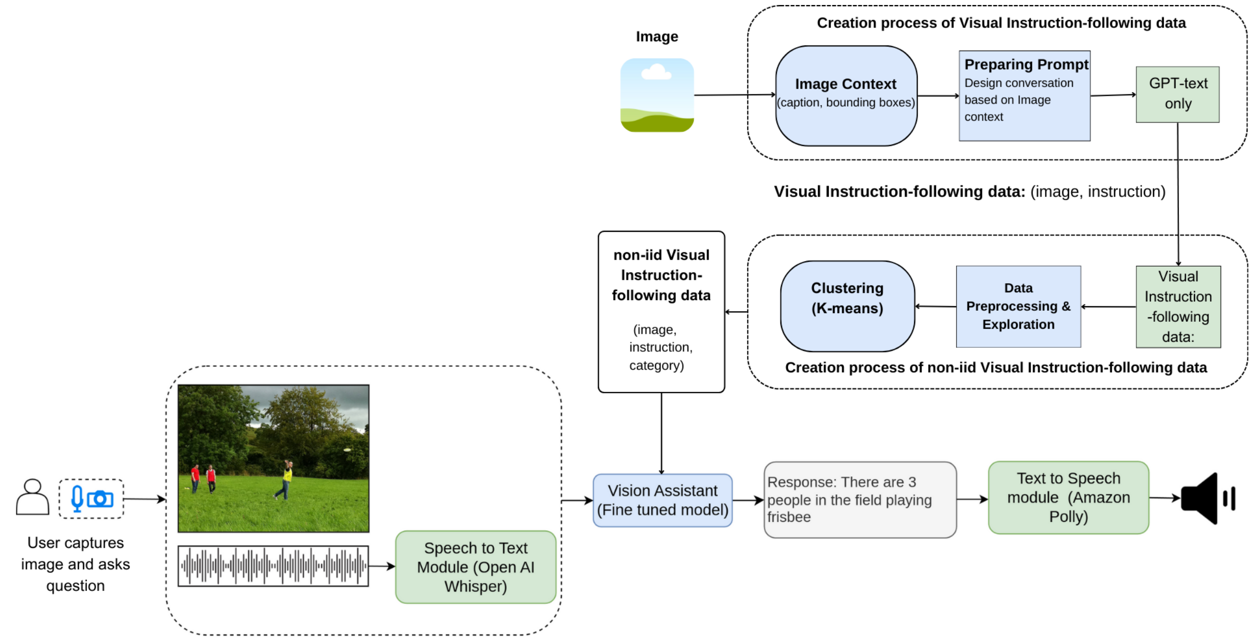

Figure 3.2: System Component

1. **Chatbot Interface:** Serving as the primary interaction point, this interface accepts images and accompanying textual or audio inputs from users. It includes a speech-to-text feature to process audio inputs into a standardized text format and text-to-speech feature to give audio feedback about the scene to the users.

2. **Multimodal LLM:** Here there is fine tuned Multi modal LLM which takes input image and text from the Chatbot and does inference to give back insight about the visual scene.

3.1.1 System Workflow

The workflow is streamlined to enhance user experience and system efficiency:

1. Users interact with the system via the Chatbot Interface, providing image and optional text or audio inputs.

2. The Data Processing Module standardizes the inputs, converting all information into a format suitable for the Multi-Modal LLM.

3. The Multi-Modal LLM interprets the combined data, focusing on providing a comprehensive understanding of the visual scene back to the users using text or audio feedback.

3.2 Architecture Overview

The system's architecture leverages advanced pre-trained large language models (LLM) and visual models to optimize performance for assistive technologies aimed at supporting visually impaired users. Both the initial and subsequent enhanced architectures are adopted from the LLaVA framework as outlined in the foundational research [18].

The initial architecture employs the pre-trained CLIP visual encoder ViT-L/14 which processes input images to extract visual features Z_v. These features include grid features from both before and after the last Transformer layer, noted for their effectiveness in capturing detailed visual nuances.

Originally, the integration of visual features into the language model was achieved using a simple linear projection:

$$H_v = W \cdot Z_v$$

where W is a trainable projection matrix that transforms visual features Z_v into language embedding tokens H_v, aligning them with the dimensionality of the word embeddings in the LLM. This setup was designed to be lightweight and scalable, facilitating rapid iterations and modifications, as detailed in [18].

Enhanced Architecture

Following the findings in [18], the architecture was enhanced by replacing the linear projection with a two-layer MLP to increase the representation power of the vision-language connector, thus improving the system's multimodal capabilities.

The MLP configuration includes:

- The first layer processes the visual embeddings to generate an enriched intermediate representation, capturing a broader array of visual details.
- The second layer maps these enriched features into the language model's embedding space, ensuring a deeper integration of visual and textual data.

This modification, as recommended in [18], allows for a more sophisticated interpretation of visual data, enabling the system to generate more accurate and contextually relevant responses. The MLP vision-language connector has proven particularly effective in tasks requiring complex spatial reasoning.

Federated Visual Instruction Tuning

The integration of federated learning into our system architecture allows for the decentralized fine-tuning of the multimodal LLM using visual instruction tuning. This approach enhances privacy and scalability by enabling model training directly on the users' devices without needing to share their data. Here, we outline the federated averaging algorithm adapted for our specific use case.

Input: Global model parameters θ, number of communication rounds R, number of clients K, number of local epochs E, learning rate η

Output: Updated global model parameters θ_R

Initialize global model parameters θ_0;

for *each round* $r = 1, 2, \ldots, R$ **do**

Randomly select a subset of K clients;

for *each client* $k \in K$ ***in parallel*** **do**

Load local dataset D_k;

Preprocess data:

- Extract visual features from images
- Convert text data to embeddings
- Align visual and text data to create multimodal inputs

Initialize local model parameters $\theta_k^r \leftarrow \theta_{r-1}$;

for *each local epoch* $e = 1, 2, \ldots, E$ **do**

for *each batch* b *of data* D_k **do**

Compute loss $\mathcal{L}(\theta_k^r; b)$ using visual instruction tuning objective;

Compute gradient $\nabla\mathcal{L}(\theta_k^r; b)$;

Update local model parameters: $\theta_k^r \leftarrow \theta_k^r - \eta\nabla\mathcal{L}(\theta_k^r; b)$;

end

end

Send updated local model parameters θ_k^r to the server;

end

Aggregate updates to form new global model parameters:

$$\theta_r \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^r$$

where n_k is the number of samples in client k's dataset and $n = \sum_{k=1}^{K} n_k$;

end

Algorithm 1: Federated Learning for Fine-Tuning Multimodal LLMs Using Visual In-

This algorithm is critical for enabling our system to operate under the constraints of data privacy regulations and network limitations, providing a robust solution for real-world deployment in assistive technology for the visually impaired.

3.3 Data Collection and Processing

Building upon the approach outlined in the [18], we leveraged a similar methodology to create a specialized instruction dataset tailored for visually impaired users.

3.3.1 Multimodal Data Representation

The Visual Instruction Tuning paper [18] discusses the creation of a large-scale multimodal instruction-following dataset using a GPT-assisted approach. Specifically, they collected 158,000 unique language-image instruction-following samples, which included:

- **58,000 samples in conversational-style question-answering:** Here, the model engages in dialogue and answers visually-related questions, simulating real-life interactions and enhancing its conversational capabilities.

- **23,000 samples providing detailed descriptions:** These samples offer comprehensive descriptions of visual scenes, significantly improving the model's ability to understand and narrate complex visual information accurately.

- **77,000 samples focused on complex reasoning:** These samples are designed to elicit responses that require advanced logical reasoning based on the visual data provided, challenging the model's cognitive processing skills.

The generated dataset served as the foundation for the subsequent model training and fine-tuning, enabling the development of an assistive technology system that can accurately interpret visual scenes and provide meaningful audio-based feedback for visually impaired users.

Converting dataset to a Non-IID nature dataset

To adapt this dataset for federated learning, particularly focusing on non-IID data distributions, we employed dimensionality reduction and clustering techniques:

1. **Dimensionality Reduction with UMAP:** We used Uniform Manifold Approximation and Projection (UMAP) to reduce the high-dimensional feature space of the visual data to a more manageable form, facilitating more effective clustering [39].

2. **Clustering with k-Means:** Post-dimensionality reduction, k-means clustering was applied to group the samples into clusters based on visual similarity. This step was critical in creating diverse subsets of data that mimic real-world distributions, where different clients may have data that is not representative of the population as a whole [40].

These clusters were then strategically distributed among simulated clients to create a non-IID condition, where each client's data may differ significantly from others, presenting more realistic and challenging conditions for federated learning [41].

Textual Data Preprocessing and Embedding

For the textual components of the dataset, a critical preprocessing step involved transforming text into embeddings that could be efficiently processed alongside visual data. This process was facilitated using the following approach:

1. **Text Embedding with SentenceTransformer:** We utilized the SentenceTransformer model, specifically 'all-MiniLM-L6-v2', known for its efficiency and effectiveness in generating meaningful sentence embeddings. This model is particularly suited for our needs due to its balance between performance and computational efficiency, making it ideal for the large-scale processing of textual data [42].

2. **Dimensionality Reduction with UMAP:** After obtaining the sentence embeddings, we applied UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction. This step was crucial to reduce the computational load and enhance the clustering process. UMAP was configured to use the cosine similarity metric, which is particularly well-suited for high-dimensional data like text embeddings. The cosine similarity helps in maintaining the semantic similarity between the embeddings, which is vital for the effective clustering of text data related to the visual elements [43].

3. **Visualization and Analysis:** Post-reduction, the embeddings were visualized to analyze the distribution and to verify the effectiveness of the preprocessing steps. This visualization also aided in understanding the diversity and distribution of data, ensuring that the non-IID nature of the dataset was appropriately represented [39].

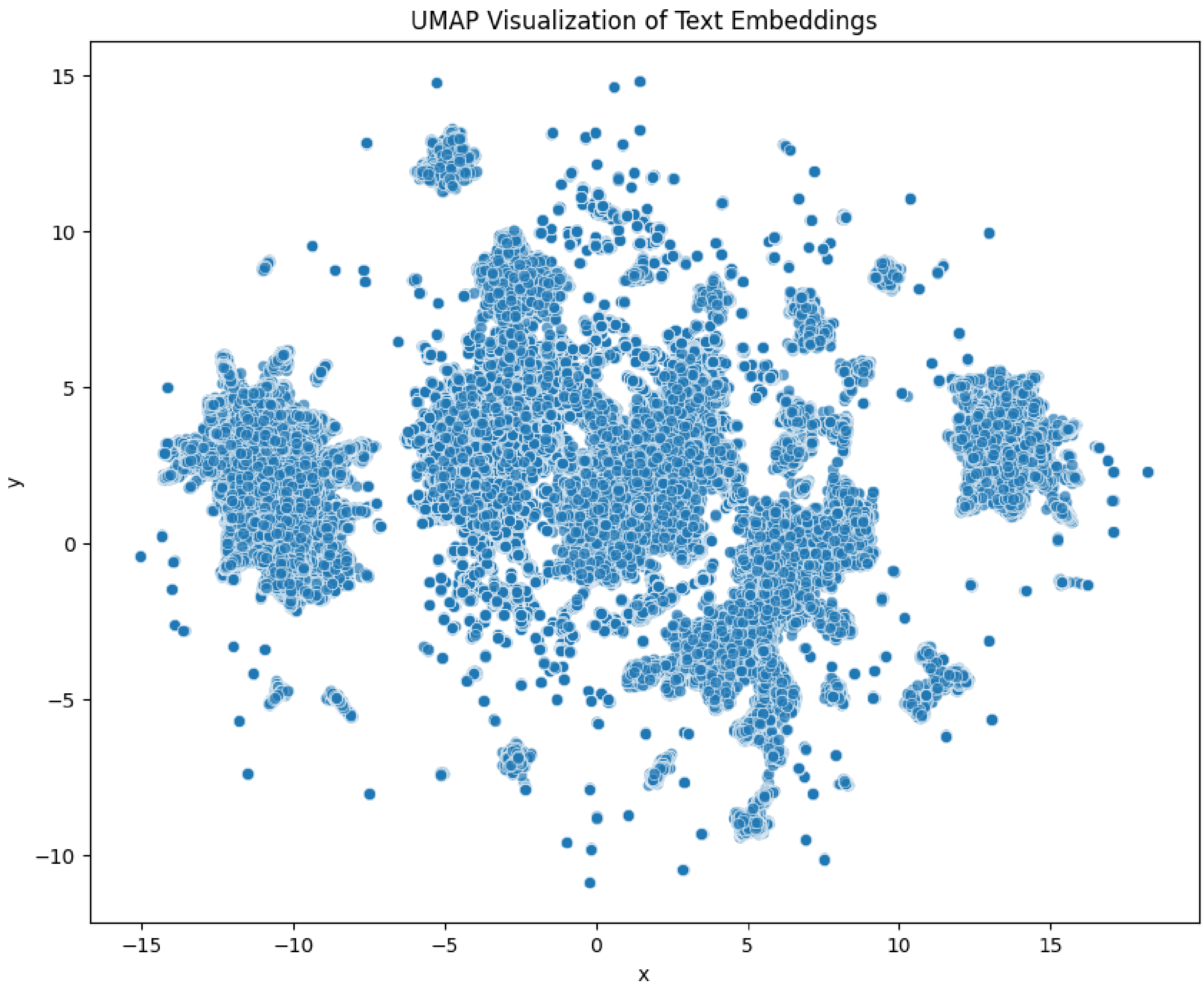

Figure 3.3: UMAP visualization of the text embeddings, highlighting the data's semantic distribution before clustering.

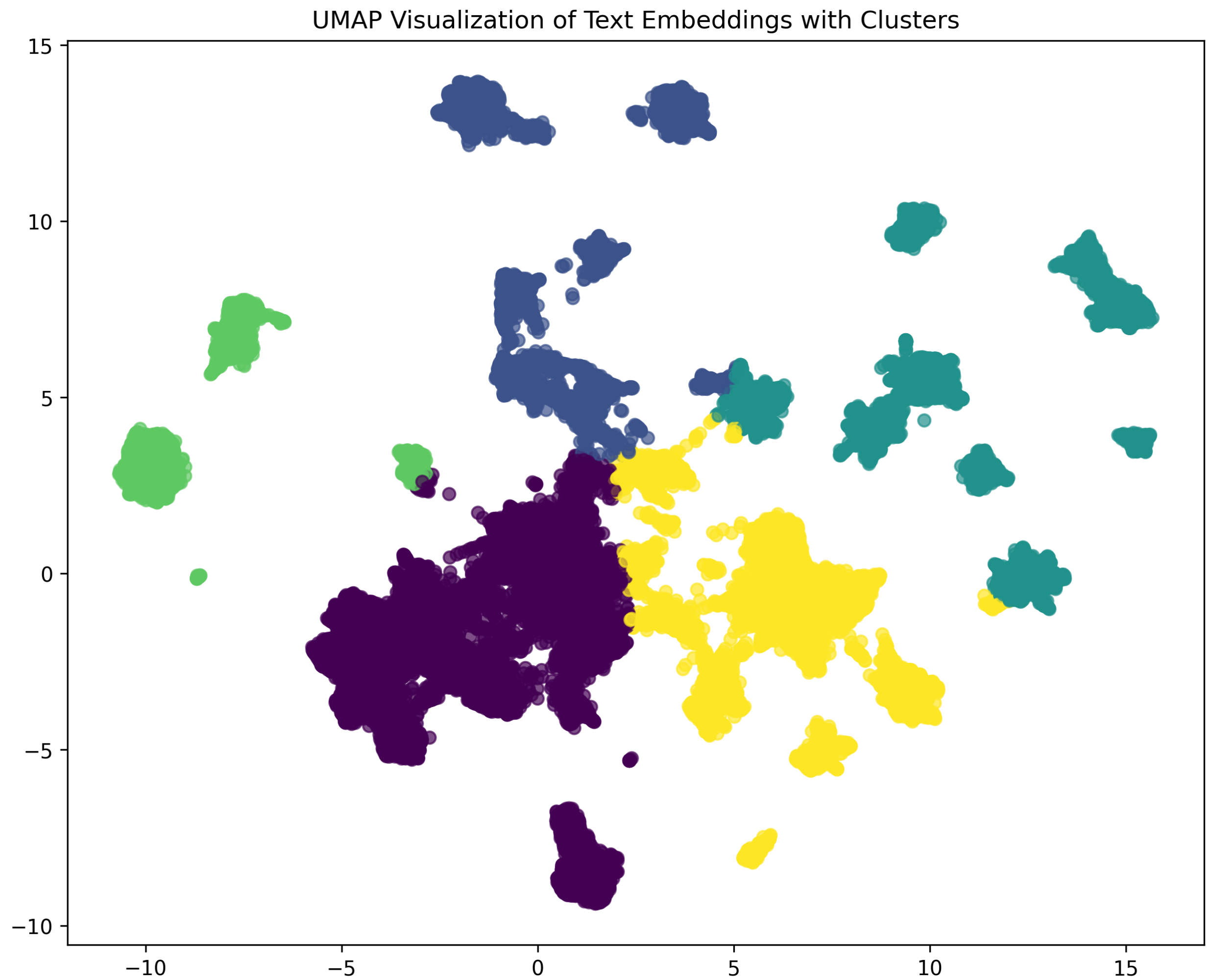

Figure 3.4: K-means clustering on the UMAP-reduced text embeddings, illustrating the formation of distinct data clusters.

Labeling and Categorization of Clusters

After applying UMAP for dimensionality reduction and k-means clustering to identify distinct groups within our dataset, we proceeded to label each cluster with specific categories that reflect the visual content's nature. This step was essential to simulate a non-IID data distribution typically seen in real-world scenarios, where data heterogeneity can significantly affect the learning process in a federated setting [38].

1. **Cluster Labeling:** Each cluster was assigned a label based on the predominant theme or activity represented in the images. The labels and their corresponding cluster indices

are as follows:

- **Everyday** (0): Images depicting daily activities and routines.
- **Sports** (1): Images related to various sports activities.
- **Nature** (2): Images showcasing natural landscapes and wildlife.
- **Adventure** (3): Images depicting adventurous activities and exotic locations.
- **Transport** (4): Images related to different modes of transportation and vehicular movement.

2. **Diversity and Non-IID Nature:** The categorization into diverse themes ensures that the dataset mirrors the variety of visual contexts individuals might encounter in different settings. This diversity is crucial for testing and enhancing the robustness of our federated learning model, as it prepares the model to handle a wide range of visual information. Moreover, the non-IID distribution of the dataset—where different clients may have data that is not representative of the population as a whole—introduces realistic challenges that are often encountered in federated learning applications [41].

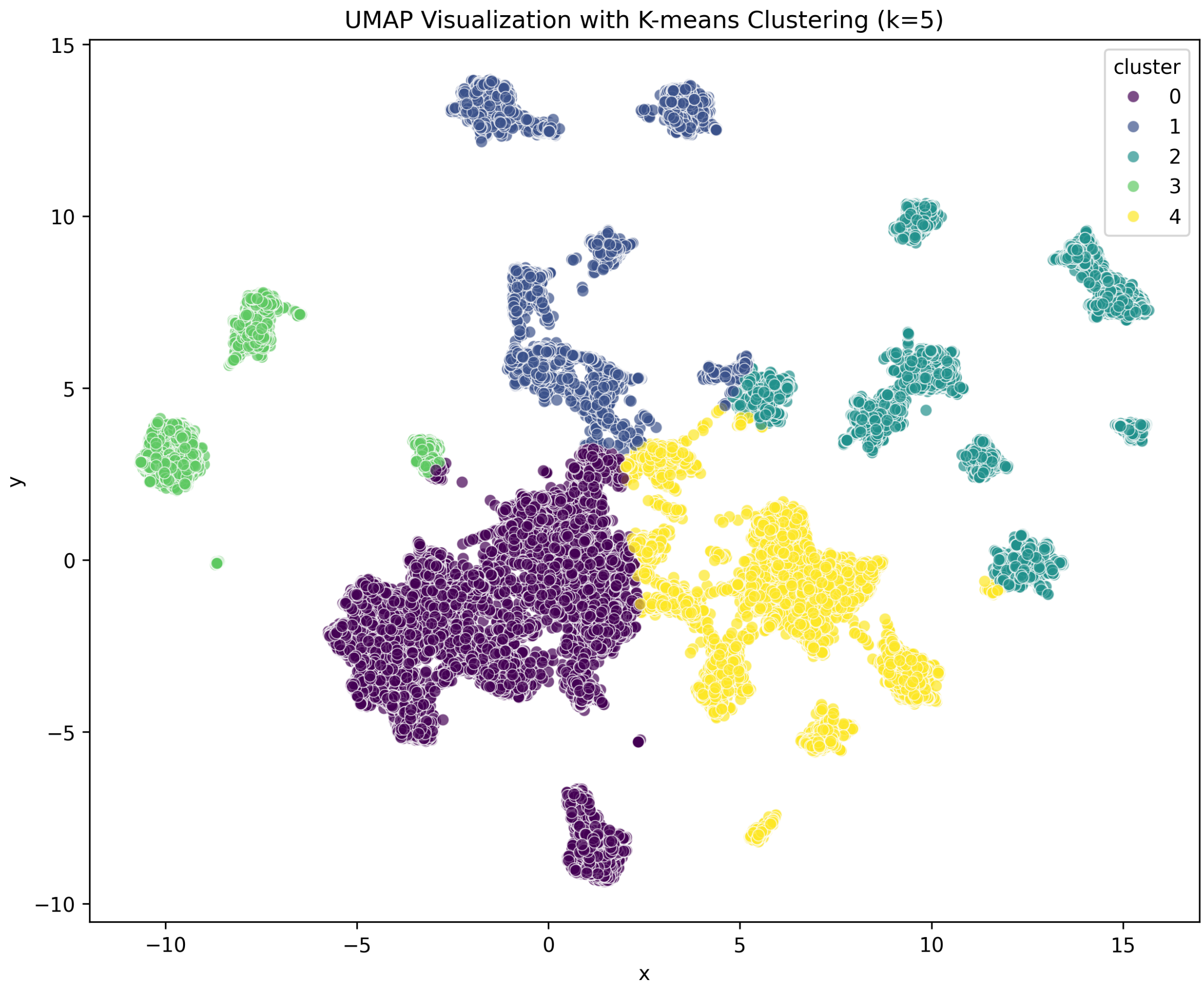

Figure 3.5: Distribution of labeled clusters illustrating the diversity and non-IID nature of the 80k dataset.

Converting dataset to a Non-IID Dataset for Llava_Instruct_150

Similarly, for the Llava_Instruct_150 dataset, we followed the same process of dimensionality reduction and clustering:

1. **Dimensionality Reduction with UMAP:** We used UMAP to reduce the high-dimensional feature space of the visual data.

2. **Clustering with k-Means:** K-means clustering was applied to group the samples into clusters based on visual similarity.

These clusters were then distributed among simulated clients to create a non-IID condition.

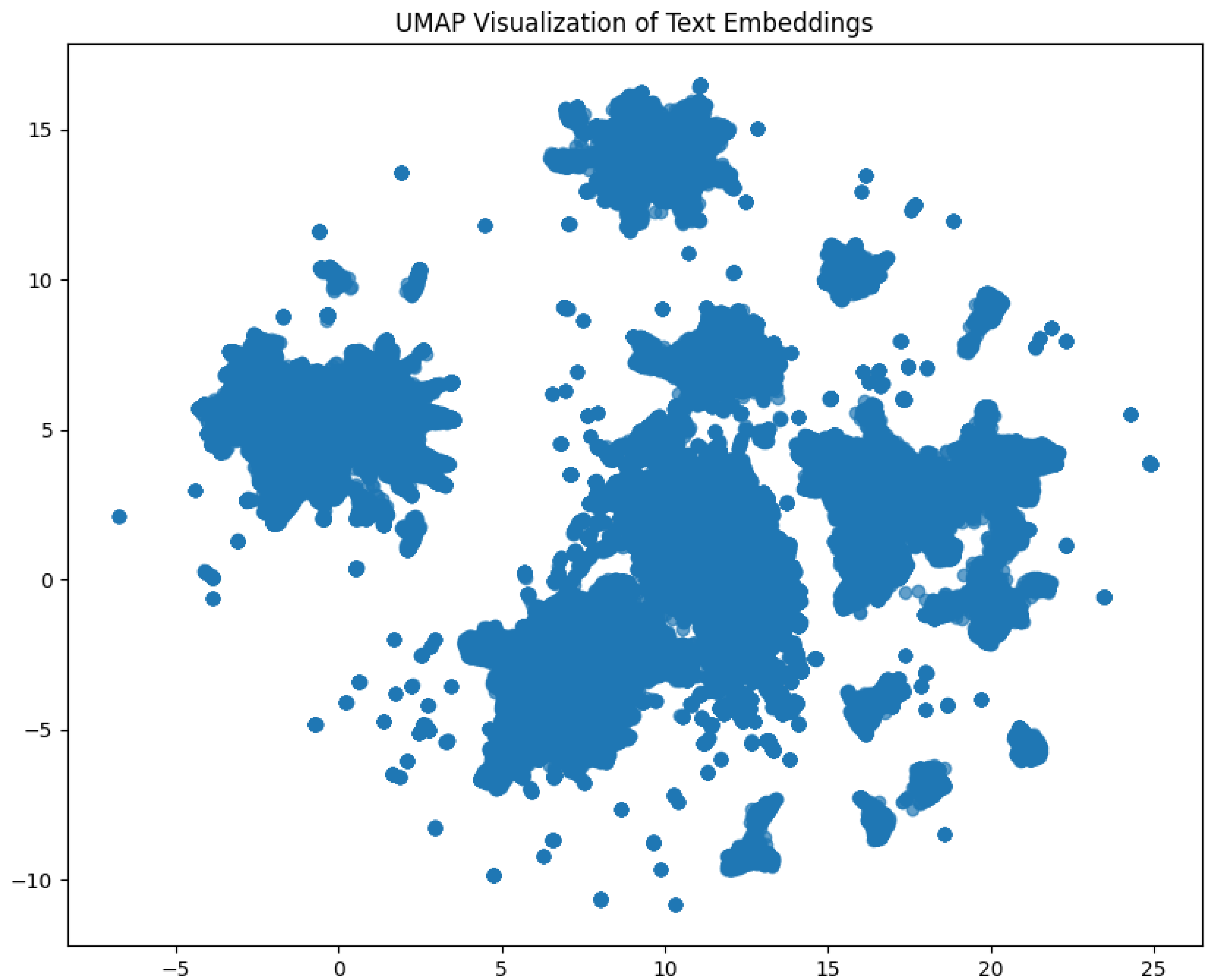

Figure 3.6: UMAP visualization of the text embeddings for llava_instruct_150 dataset, highlighting the data's semantic distribution before clustering.

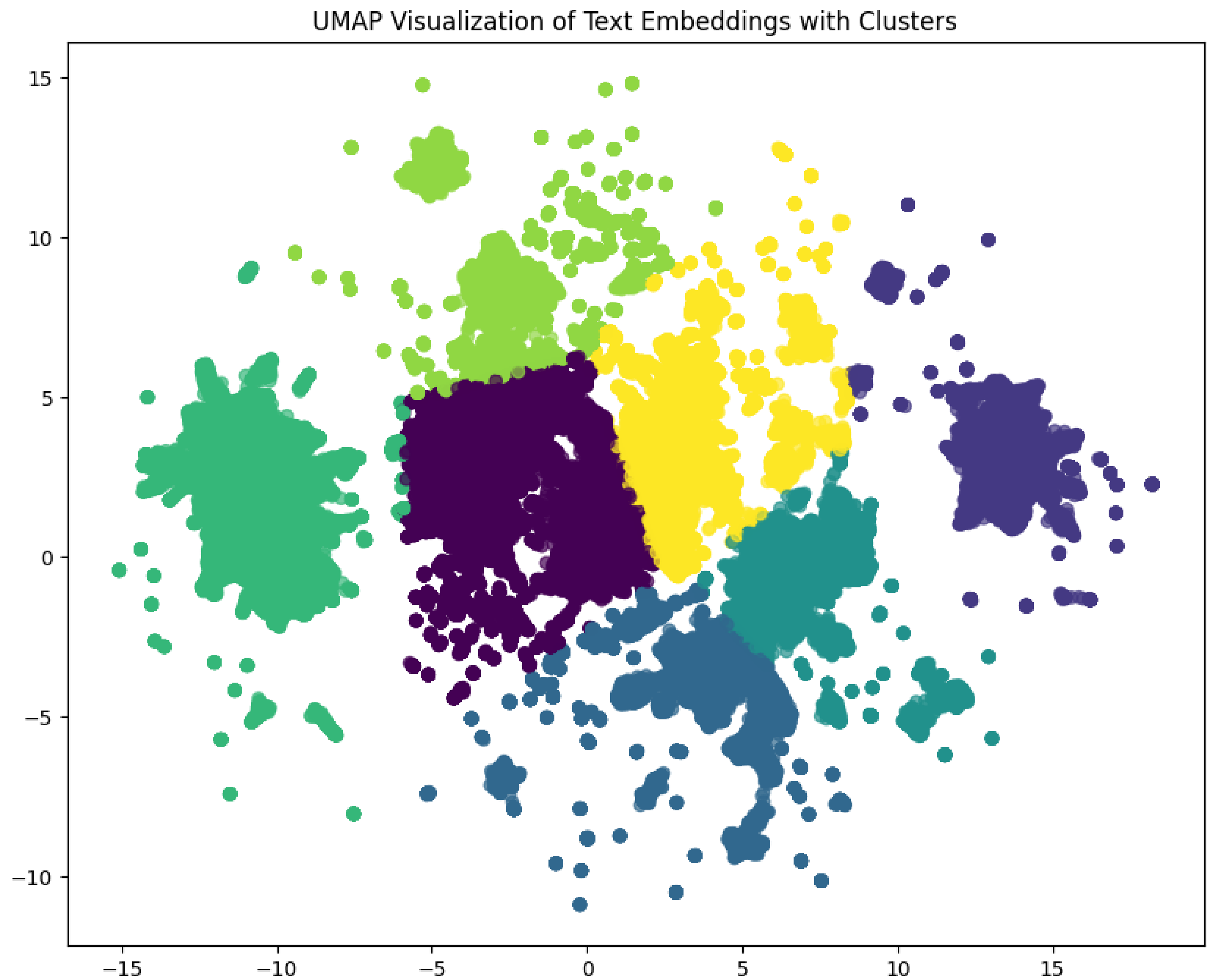

Figure 3.7: K-means clustering on the UMAP-reduced text embeddings for llava_instruct_150 dataset, illustrating the formation of distinct data clusters.

1. **Cluster Labeling:** Each cluster was assigned a label based on the predominant theme or activity represented in the images. The labels and their corresponding cluster indices for the 150k dataset are as follows:

    - **Observations** (0): Images depicting observational activities.
    - **Direct Queries** (1): Images related to direct queries.
    - **Bounding Box Descriptions** (2): Images providing bounding box descriptions.
    - **Books** (3): Images showcasing books and related activities.
    - **Activity Analysis** (4): Images depicting activity analysis.

- **Education** (5): Images related to educational contexts.
- **Quiz** (6): Images depicting quiz-related activities.

2. **Diversity and Non-IID Nature:** The categorization into diverse themes ensures that the dataset mirrors the variety of visual contexts individuals might encounter in different settings. This diversity is crucial for testing and enhancing the robustness of our federated learning model.

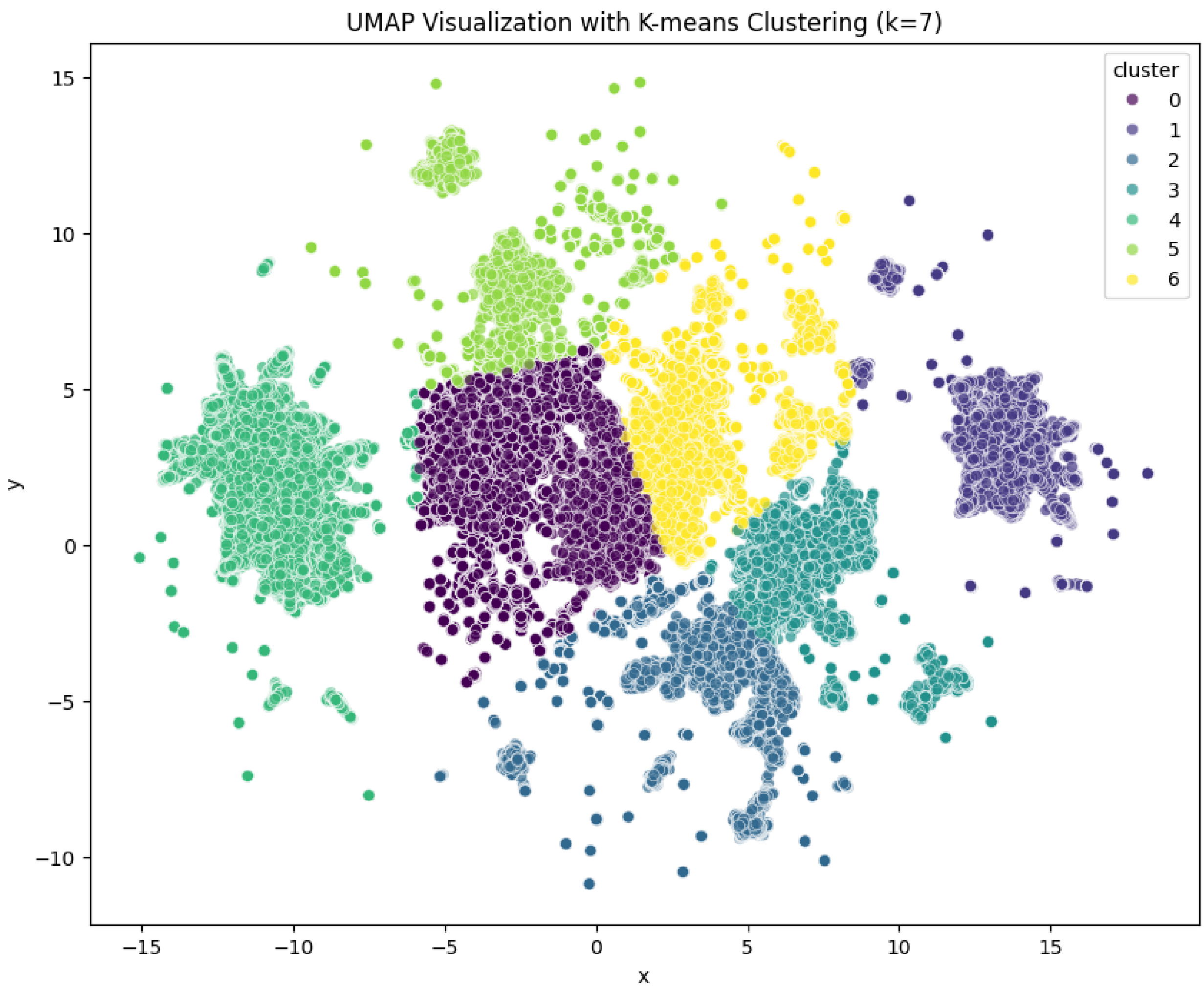

Figure 3.8: Distribution of labeled clusters for llava_instruct_150 dataset illustrating the diversity and non-IID nature of the dataset.

These steps not only ensure that our dataset is well-suited for federated learning but also enhance the model's ability to generalize across different visual domains, which is vital for developing effective assistive technologies for the visually impaired [39, 42, 38].

These preprocessing steps ensured that the textual data were optimally prepared for integration with visual data, facilitating the multimodal learning process. By aligning the text embeddings closely with the visual features, the system is better equipped to handle the complex reasoning tasks required in our application, specifically tailored to assistive technologies for visually impaired users.

The data processing module played a pivotal role in converting raw data into a format suitable for the LLM, ensuring that both visual and textual components were optimally integrated to produce high-quality training and validation sets.

This tailored approach to dataset creation and preprocessing not only enhances the training effectiveness of the multimodal LLM in a federated learning framework but also aligns with the specific needs of visually impaired users, ensuring the assistive technology developed is both effective and efficient.

3.4 Model Training and Fine-tuning

Training our model involved a two-stage process tailored to refine the model's capabilities in handling multimodal data for assistive technologies specifically designed for visually impaired users. This process aligns closely with methodologies from prominent research, particularly the use of a pretrained LLaVA model, which is fine-tuned to meet specific requirements.

Stage 1: Pre-training for Feature Alignment

Initially, we focused on aligning the visual features with textual embeddings efficiently:

- **Training Objectives:** In our adaptation, we utilized a pretrained projection matrix from [18] while keeping the model's visual encoder and LLM weights frozen. This modified approach allowed for the efficient integration of pre-aligned features, focusing on enhancing the compatibility between visual and language modalities.

- **Data Preparation:** The authors curated a subset of the Common Crawl dataset, filtering it down to 595K image-text pairs to balance concept coverage and training efficiency. This subset was used to create instruction-following data resembling single-turn conversations.

- **Visual and Textual Integration:** Each training sample was treated as a single-turn interaction, where a visual query (image) was paired with a textual response (caption). In the original methodology, the primary goal during this stage was to align these multimodal inputs using a trainable projection matrix W designed to match the dimensionality of the visual features with the textual embeddings in the LLM.

This pre-training stage acted as a foundation, setting up the model for more advanced fine-tuning while ensuring that the visual tokenizer was compatible with the frozen LLM.

Stage 2: Fine-tuning End-to-End

After aligning the basic features, we proceeded to an end-to-end fine-tuning stage:

- **Continued Training:** With the visual encoder still frozen, we fine-tuned the projection layer and the LLM together, optimizing both for better integration and response generation.

- **Multimodal Chatbot:** We developed a Chatbot by further fine-tuning the model on 158K language-image instruction-following data, focusing on enhancing its capabilities for multimodal, multi-turn interactions.

- **Use Case Scenarios:** The fine-tuning process was adapted for specific scenarios, including a multimodal chatbot for general interactions and a Science QA application, where the model was tasked with providing detailed responses based on textual or visual prompts.

Federated Learning Integration

In addition to the standard training procedures, we incorporated federated learning techniques to enhance privacy and model robustness:

- **Non-IID Data Handling:** Using the previously mentioned non-IID dataset, we fine-tuned the model across multiple decentralized nodes, simulating real-world scenarios where data distribution is uneven.

- **Local and Global Updates:** Training involved both local updates on client devices and global aggregations to update the central model, ensuring comprehensive learning across diverse data points.

3.5 Parameter-Efficient Fine-Tuning (PEFT)

Parameter-Efficient Fine-Tuning (PEFT) was employed to reduce the computational burden and memory requirements during the fine-tuning process. PEFT focuses on adapting only a small subset of the model parameters, which significantly improves training efficiency while maintaining high performance [44, 45].

Principles of PEFT

PEFT aims to fine-tune large pre-trained models by adjusting only a fraction of their parameters, thereby achieving efficient training with minimal resource utilization. The key principles include:

- **Subset Selection:** Instead of updating all parameters, PEFT selects a subset of parameters that are most influential for the target task [46].

- **Low-Rank Adaptation:** Incorporating low-rank matrices to approximate the full-rank parameter updates, reducing the number of parameters to be fine-tuned [44].
- **Layer-wise Adaptation:** Focusing on specific layers that contribute most to the model's performance, rather than uniformly updating parameters across all layers [45].

Application in Assistive Technologies

For the assistive technology system, PEFT was applied as follows:

- **Visual Encoder:** Fine-tuning involved low-rank adaptation of the visual encoder's final layers, enhancing the model's ability to extract and interpret complex visual features [44].
- **Language Model:** Specific attention layers within the language model were fine-tuned using low-rank matrices to improve the integration of visual and textual data [45].
- **Efficiency and Performance:** The application of PEFT resulted in a significant reduction in computational resources and training time, without compromising the model's ability to generate accurate and contextually relevant descriptions [46].

This comprehensive training approach not only enhanced the model's performance in specific tasks but also ensured its adaptability and effectiveness in real-world applications for the visually impaired.

Chapter4

Practical Application

Assistive technology User Interface (UI)

The user interface (UI) of the assistive technology system is designed with simplicity and accessibility in mind, ensuring that visually impaired users can easily interact with the system. The primary features of the UI include:

1. **Chatbot Interface:** The central element of the UI, where users can input images and optionally provide textual or auditory inputs. This interface includes a speech-to-text feature using OpenAI Whisper to convert audio inputs into text and a text-to-speech feature using Amazon Polly to deliver audio feedback to users.

2. **Voice Commands:** Users can interact with the system using voice commands, enhancing accessibility for those who may find it challenging to use a keyboard or touchscreen.

3. **Screen Reader Compatibility:** The UI is fully compatible with screen readers, ensuring that visually impaired users receive auditory feedback for all interactions.

Gradio Web Server

To provide a seamless and interactive experience, the assistive technology system is hosted on a Gradio web server. Gradio is a Python library that allows developers to create customizable user interfaces for machine learning models. The advantages of using Gradio include:

1. **Ease of Use:** Gradio simplifies the process of creating and deploying web interfaces for machine learning models, making it easy to integrate various input and output modalities.

2. **Real-Time Interaction:** The server enables real-time interaction with the model, providing immediate feedback to users based on their inputs.

3. **Customization:** Gradio offers a high degree of customization, allowing the UI to be tailored to the specific needs of visually impaired users.

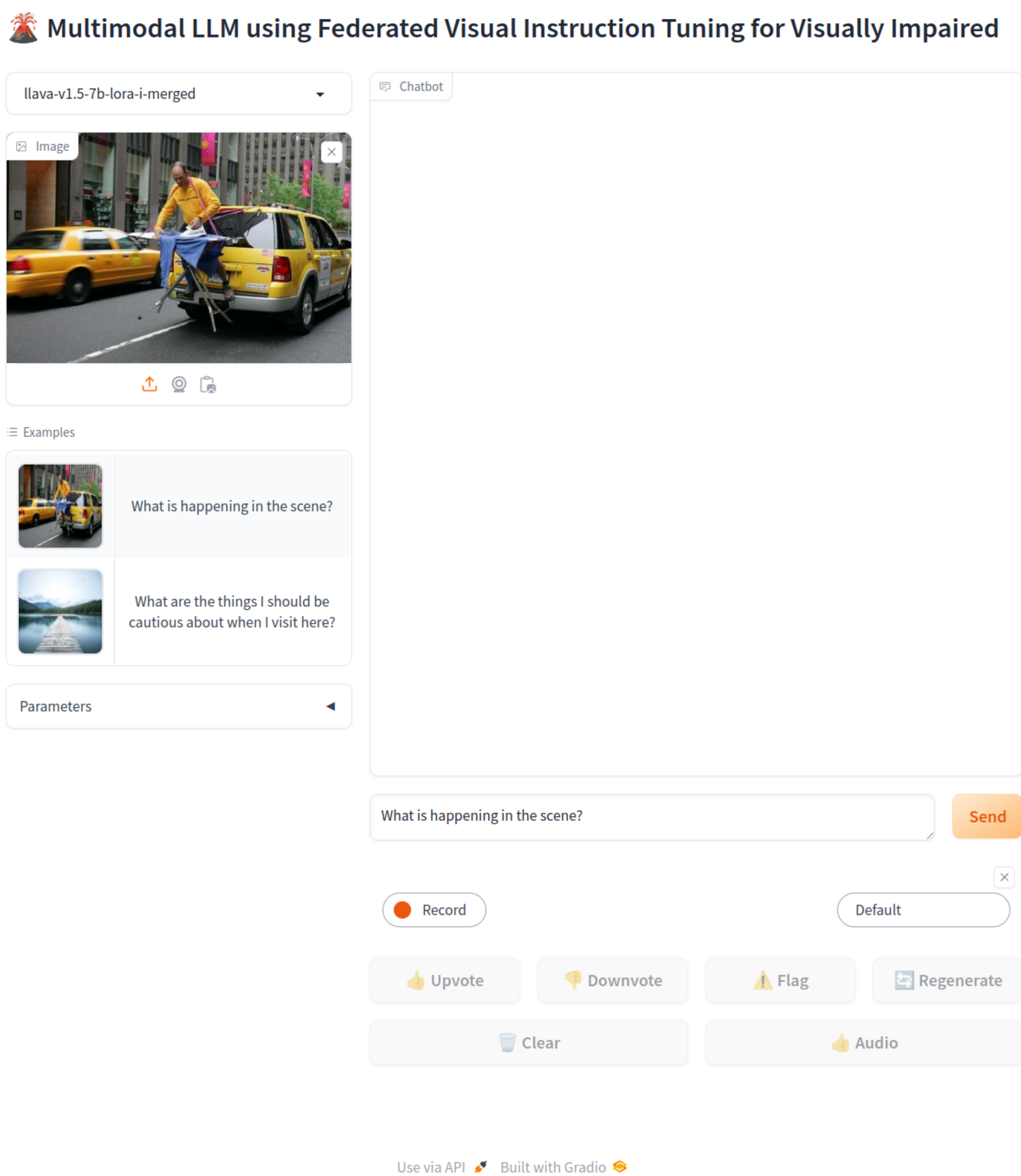

Figure 4.1: Screenshot of the Gradio web interface for the assistive technology system. On the left part of the interface, users can upload an image or take a picture using the camera. The right side of the interface displays the uploaded image along with a question about the scene. Below the image, the system provides textual feedback describing the scene, followed by an audio feedback option. Users can also interact with the system through audio recording, and receive both text and audio responses.

Quantization for Memory Footprint Reduction

To ensure that the assistive technology system can run efficiently on devices with limited computational resources, we employed quantization techniques to reduce the memory footprint of the model. Specifically, we used 4-bit quantization, which involves the following steps:

1. **Model Conversion:** The pre-trained model is converted to a quantized version using a process that reduces the precision of the model's weights from 32-bit floating point to 4-bit integers [47].

2. **Performance Optimization:** Quantization significantly reduces the memory requirements and computational overhead, allowing the model to run on devices with lower processing power without compromising performance [48, 49].

3. **Experimental Results:** The quantized model was tested to ensure that the reduction in precision did not adversely affect the model's accuracy or responsiveness. The results indicated a minimal loss in performance with a substantial gain in efficiency [50].

Application Screenshots

To provide a comprehensive view of the assistive technology system in action, the following screenshots illustrate various functionalities:

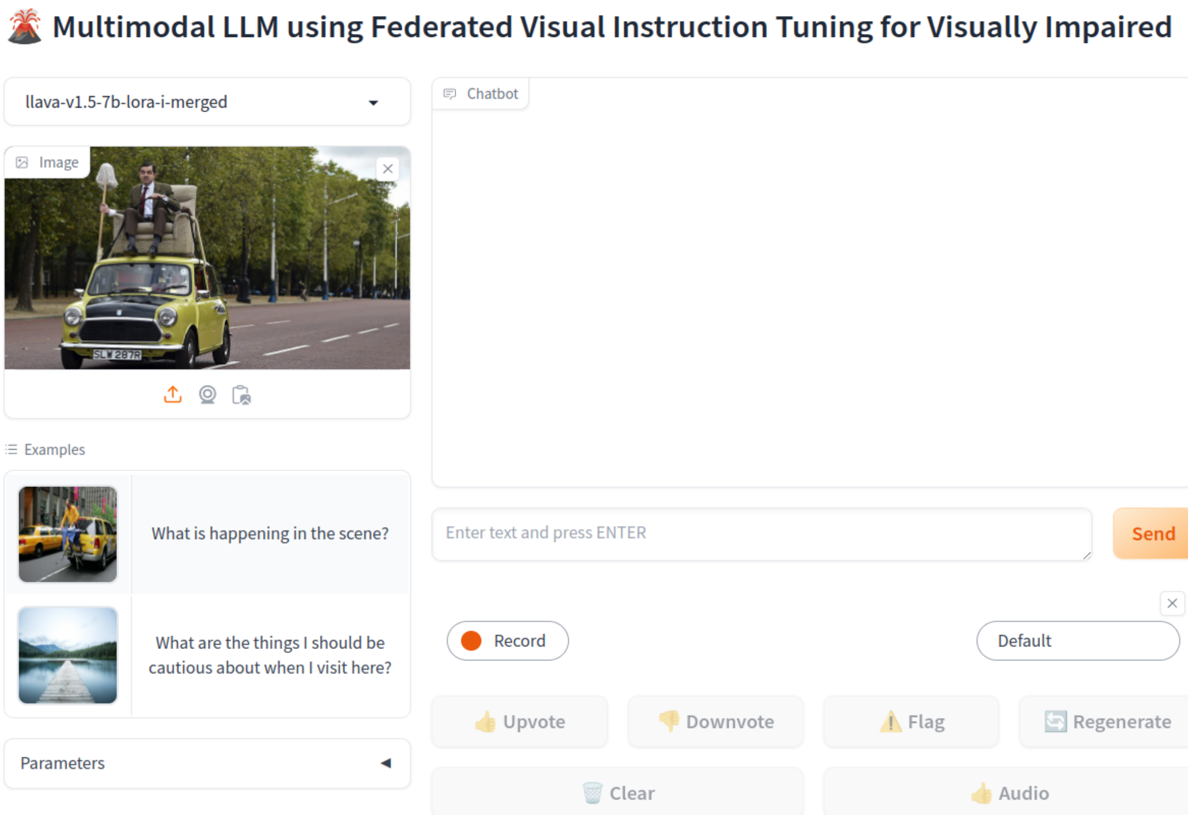

Figure 4.2: User providing an image input

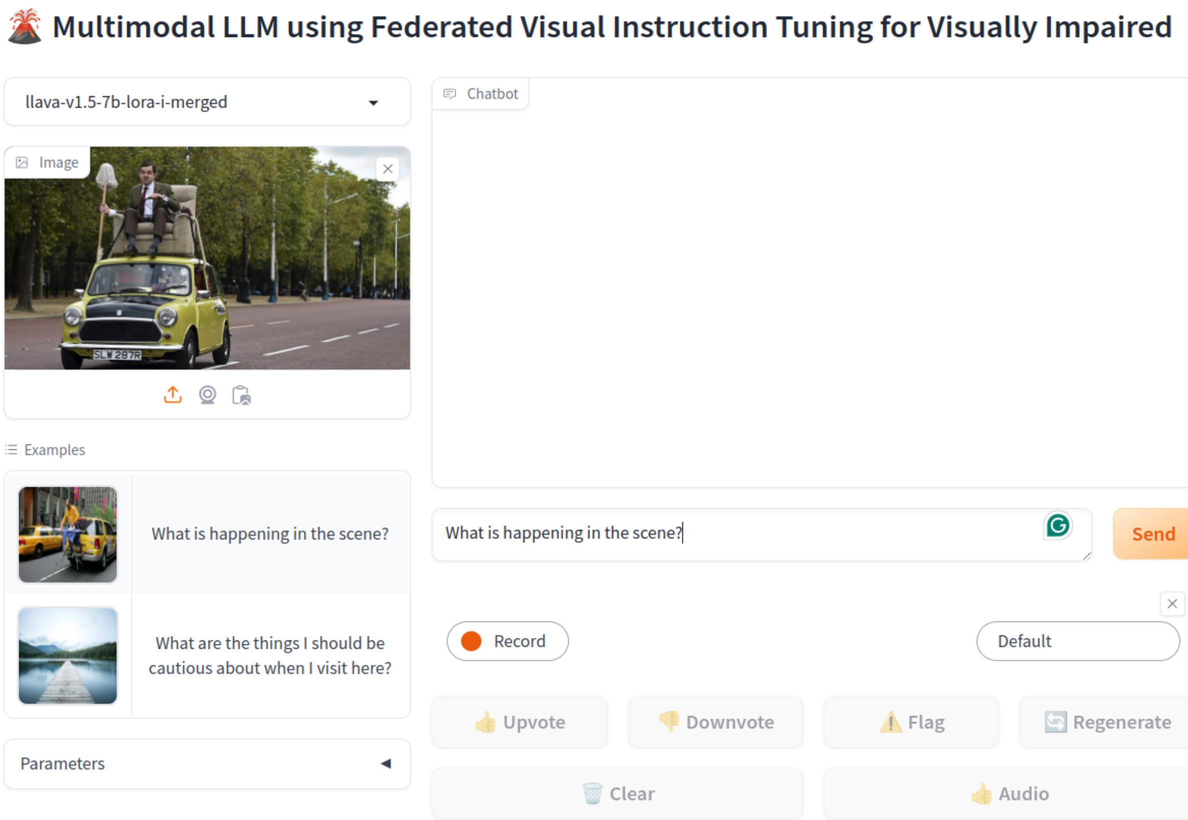

Figure 4.3: User asking question about the image

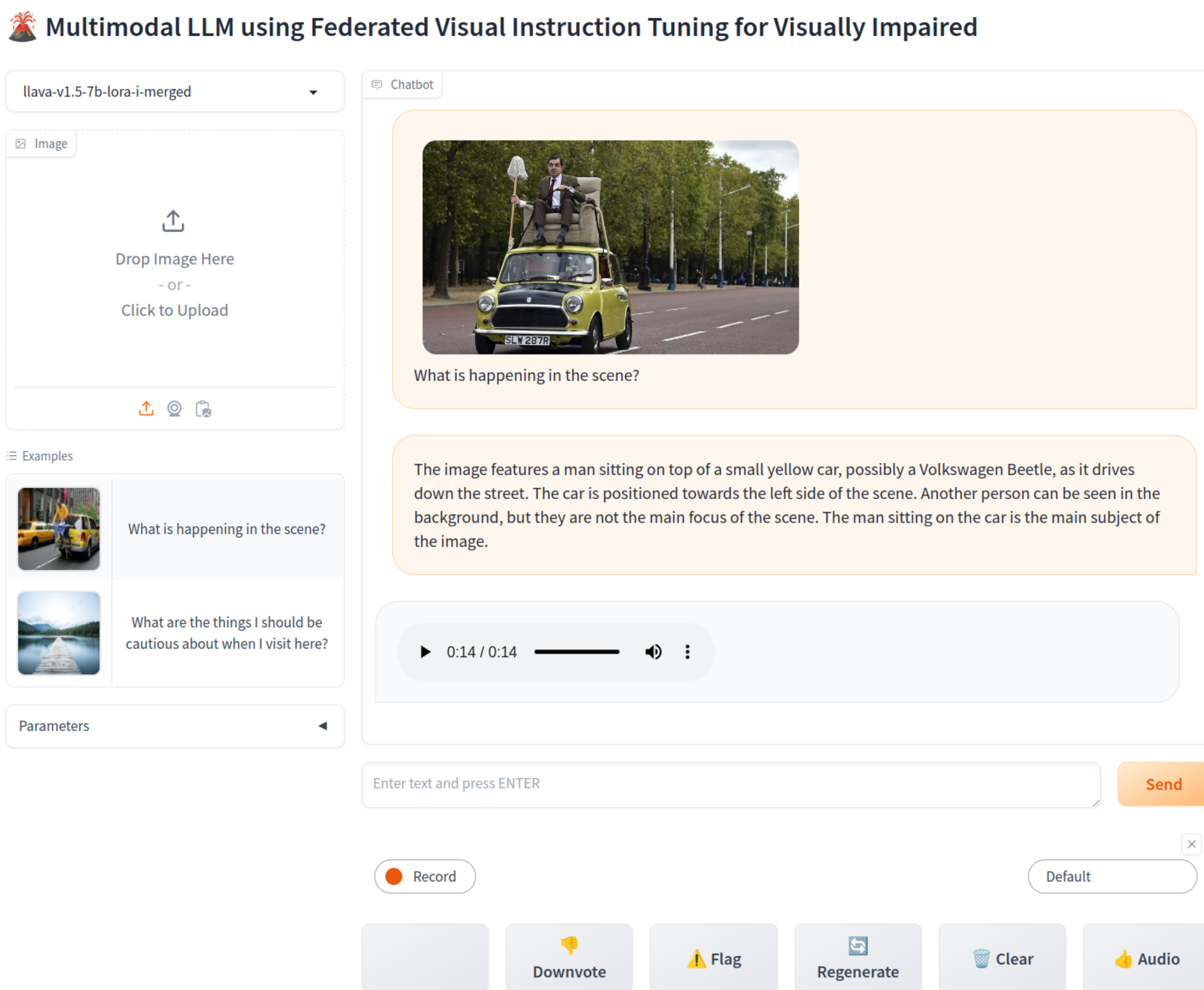

Figure 4.4: Voice command input and audio feedback output.

These screenshots demonstrate the system's capability to provide real-time, contextually relevant feedback to users, enhancing their interaction with and understanding of their surroundings.

Chapter5

Experiments

This chapter explains the experimental procedures undertaken to adapt the LLaVA framework for our specific assistive technology application. The experiments were designed around two experimental settings: utilizing the LLaVA-Instruct-80K and LLaVA-Instruct-150K datasets. Leveraging the Stage 1 Pre-trained projectors from the LLaVA model, Stage 2 fine-tuning was conducted to tailor the model's performance to our use case.

5.1 Experimental Setup

The initial stage of our experiments involved utilizing the LLaVA Stage 1 Pre-trained projectors, which were originally trained on a 558K subset of the LAION-CC-SBU dataset. This foundational training provided a robust starting point for further model refinement.

Federated Learning Implementation

For fine-tuning, the model was deployed on two NVIDIA A100 GPUs, chosen for their powerful computational capabilities and efficiency in handling extensive training tasks. This setup ensured that the model could be trained intensively without hardware limitations impacting performance.

To explore the potential of federated learning in enhancing visual instruction tuning for assistive technologies, we conducted an experiment leveraging the Federated Instruction Tuning (FedIT) approach. This experiment was designed to refine the multimodal capabilities of the model under a federated learning framework, ensuring privacy and leveraging localized data improvements.

The experiment involved multiple client simulations, each possessing a subset of the multimodal instruction-following dataset. These clients participated in a federated learning scenario where they performed local computations on their data without exchanging it. This setup aimed to enhance the model's generalization capabilities across diverse visual and instructional contexts. We employed the Federated Average (FedAvg) algorithm to aggregate the local updates from clients. Each client was equipped with a lightweight version of the model, utilizing LoRA for parameter-efficient tuning, which was particularly suitable for the limited computational resources available at the edge (local client level). This federated fine-tuning approach allowed for personalized adaptations to the model while maintaining a centralized performance standard.

5.2 Fine-Tuning Experiments

We conducted the following fine-tuning experiments to evaluate the performance of the LLaVA model:

Experiment 1: LLaVA-Instruct-80K Dataset The model was fine-tuned on the LLaVA-Instruct-80K dataset for 3 epochs. This dataset was specifically curated to enhance the model's ability to handle a range of instructive tasks relevant to assisting visually impaired users.

Experiment 2: LLaVA-Instruct-150K Dataset A more extensive fine-tuning was conducted on the larger LLaVA-Instruct-150K dataset, also for 3 epochs. The additional data provided a more comprehensive range of instructional scenarios, pushing the model to adapt to more complex user interactions and requirements.

Experiment 3: Federated Fine-Tuning on LLaVA-Instruct-80K Dataset We performed federated fine-tuning of the LLaVA model on the LLaVA-Instruct-80K dataset, leveraging the FedIT approach described earlier.

Experiment 4: Federated Fine-Tuning on LLaVA-Instruct-150K Dataset We also conducted federated fine-tuning on the larger LLaVA-Instruct-150K dataset, following the same FedIT approach.

The comparison of the model's performance across these four experiments provided valuable insights into the impact of dataset size and the benefits of the federated learning approach for our assistive technology application.

Chapter6

Results

Experiment 1: Fine-tuning on LLaVA-Instruct-80K Dataset

The LLaVA model was fine-tuned on the LLaVA-Instruct-80K dataset for one epoch. This dataset is designed to test the model's ability to handle a variety of instructional tasks that are crucial for developing accessible technologies for visually impaired users.

Figures 6.1, 6.2, and 6.3 depict the training dynamics, including the training loss, learning rate adjustments, and epoch progression. These graphs provide insights into the model's learning behavior and optimization potential during the fine-tuning process.

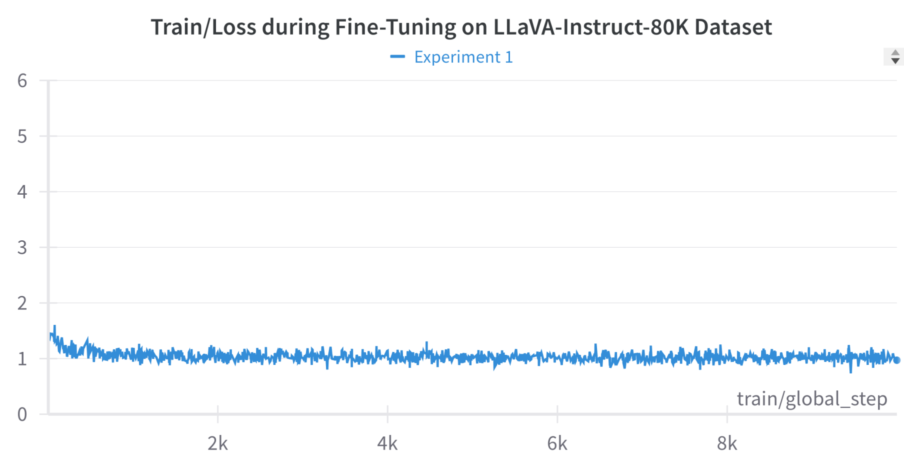

Figure 6.1: Training loss during Fine-Tuning

The performance of the model across various cognitive and perceptual tasks is detailed in Table 6.6. The scores reflect the model's effectiveness in interpreting complex visual data, which is essential for enhancing spatial awareness and navigation aids for visually impaired

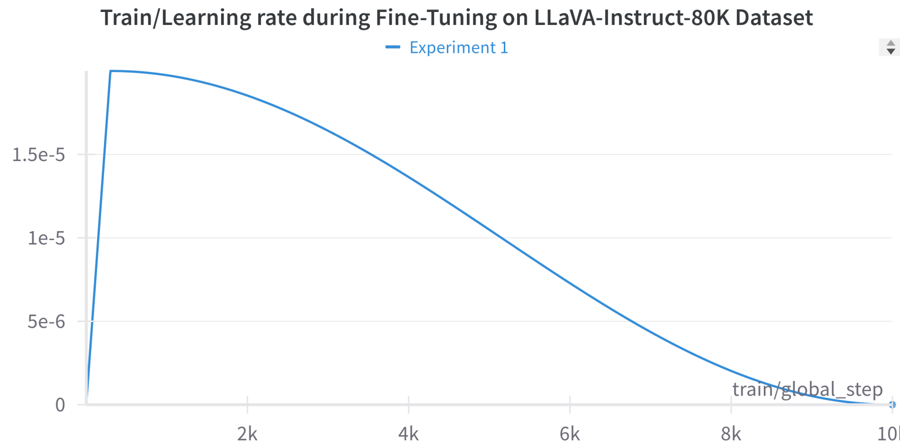

Figure 6.2: Learning rate during Fine-Tuning

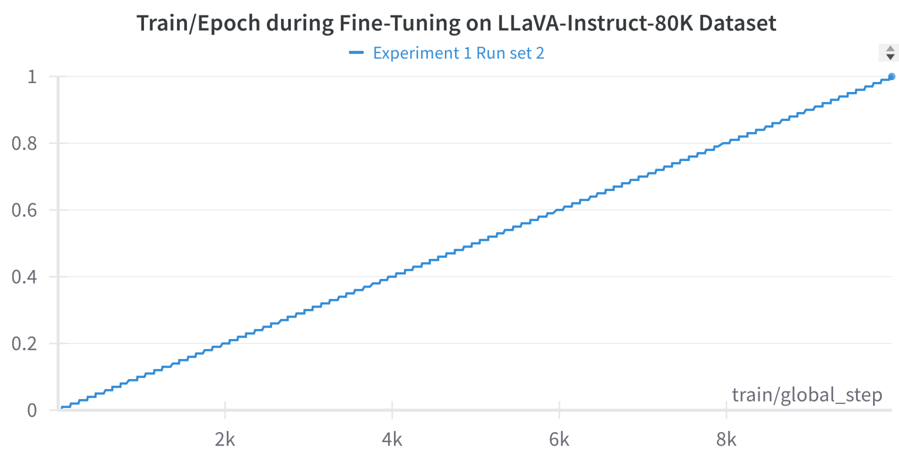

Figure 6.3: Epoch during Fine-Tuning

users.

Additionally, a summary of training metrics, shown in Table 6.1, provides a quantitative overview of the training dynamics, including learning rate, average loss, and computational efficiency.

This comprehensive analysis demonstrates the model's efficacy in enhancing navigational aids and environmental interaction for visually impaired users. By showing significant performance across various cognitive and perceptual tasks, the model validates its potential as a crucial component in assistive technologies. Looking forward, extending the training duration or diversifying the training dataset could further enhance the model's robustness and generalizability, broadening its applicability in real-world scenarios.

Metric	Value
Epochs	1
Global Steps	10,000
Learning Rate	2×10^{-5}
Average Loss	0.9704
Total FLOPs	2.1446e+18
Runtime (s)	27,869.96
Samples per Second	2.87
Steps per Second	0.359

Table 6.1: Summary of Training Metrics for Experiment 1 on LLaVA-Instruct-80K Dataset

This comprehensive analysis helps validate the model's capability to meet the specific needs of visually impaired users, enhancing their navigation and interaction with their environment through advanced assistive technologies. Future enhancements might include extending the training duration or enriching the dataset to further improve model robustness and adaptability.

Experiment 2: Fine-tuning on LLaVA-Instruct-150K Dataset

A more extensive fine-tuning was conducted on the larger LLaVA-Instruct-150K dataset for 3 epochs. The additional data provided a more comprehensive range of instructional scenarios, pushing the model to adapt to more complex user interactions and requirements.

These results and visualizations offer detailed insights into the model's learning dynamics and its ability to adapt to complex scenarios, further highlighting the potential for using such a model in assistive technology applications for visually impaired users.

Experiment 3: Federated Fine-tuning on LLaVA-Instruct-80K Dataset

We performed federated fine-tuning on the LLaVA-Instruct-80K dataset using the Federated Instruction Tuning (FedIT) approach. This method allowed multiple clients to collabora-

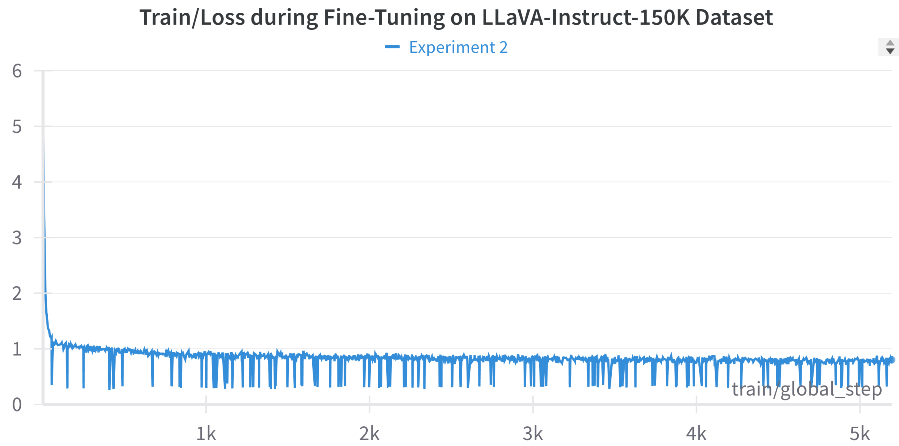

Figure 6.4: Training loss during Fine-Tuning on LLaVA-Instruct-150K Dataset

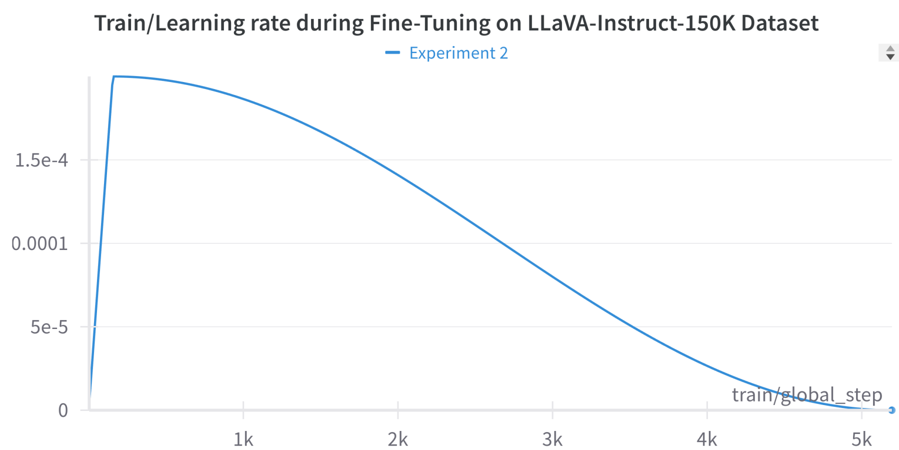

Figure 6.5: Learning rate adjustments during Fine-Tuning on LLaVA-Instruct-150K Dataset

tively learn without sharing their local data, enhancing privacy and leveraging localized improvements in model performance.

These figures and tables depict a detailed view of the model's performance during federated fine-tuning. The results show the model's resilience and adaptability in a federated learning environment, highlighting the benefits of this approach in protecting user data privacy while still achieving significant learning outcomes.

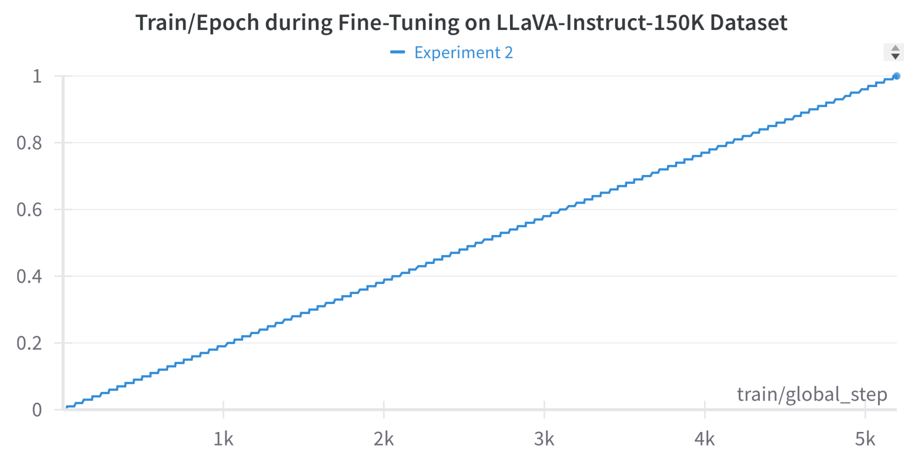

Figure 6.6: Epoch progression during Fine-Tuning on LLaVA-Instruct-150K Dataset

Table 6.2: Summary of Training Metrics for Experiment 2 on LLaVA-Instruct-150K Dataset

Metric	Value
Epochs	3
Global Steps	5,197
Learning Rate	2×10^{-4}
Average Loss	0.5369
Total FLOPs	18.951 quintillion
Runtime (s)	161,386.42
Samples per Second	4.122
Steps per Second	0.032

Experiment 4: Federated Fine-tuning on Non-IID LlaVA-Instruct-150K Dataset

This experiment involved federated fine-tuning on the LLaVA-Instruct-150K dataset using the Federated Instruction Tuning (FedIT) approach. This approach enhances privacy and data security by allowing multiple clients to collaboratively update the model without sharing their individual data directly.

An integral part of assessing the efficiency of our fine-tuning process for each experiment involves analyzing the computational resources consumed, particularly the runtime. The following table presents the total runtime for each experiment, reflecting the computational

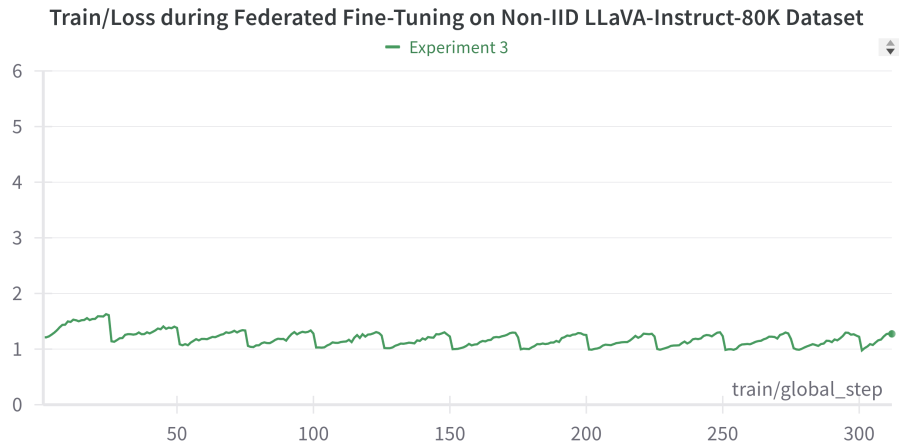

Figure 6.7: Training loss during Federated Fine-Tuning on Non-IID LLaVA-Instruct-80K Dataset

Table 6.3: Summary of Training Metrics for Federated Fine-Tuning (Experiment 3)

Metric	Value
Epochs	1
Global Steps	5,197
Learning Rate	2×10^{-5}
Average Loss	1.2485
Total FLOPs	578.446 trillion
Runtime (s)	5,724.15
Samples per Second	6.988
Steps per Second	0.055

demand and efficiency of the fine-tuning process under different configurations and federated settings 6.5.

The data reveals significant variations in runtime across the experiments, highlighting the impact of different training strategies and data configurations. Experiment 2 and 4, for instance, required substantially more time, which can be attributed to the complexity of the tasks and the larger dataset sizes involved. Conversely, Experiments 1 and 3, which might have involved simpler tasks or smaller datasets, completed in much shorter times. These metrics are crucial for understanding the scalability and practicality of deploying these models in real-world applications, particularly in environments where computational

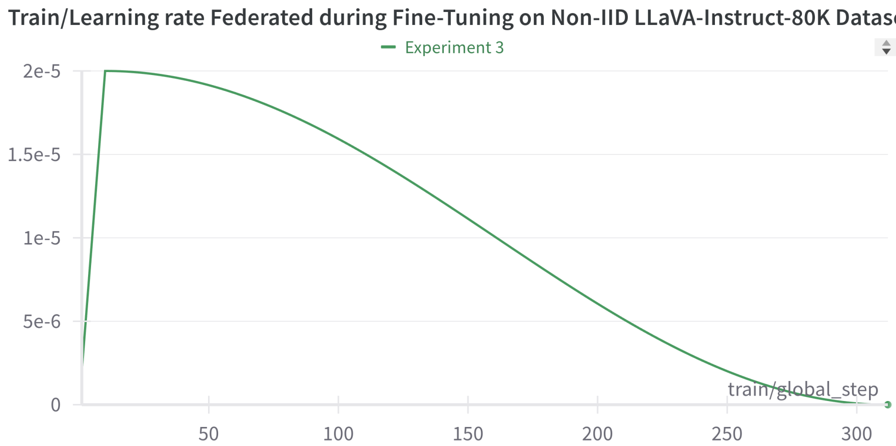

Figure 6.8: Learning rate adjustments during Federated Fine-Tuning on Non-IID LLaVA-Instruct-80K Dataset

Table 6.4: Summary of Training Metrics for Federated Fine-Tuning (Experiment 4)

Metric	Value
Epochs	1
Global Steps	2,440
Learning Rate	2×10^{-4}
Average Loss	0.7623
Total FLOPs	8.2475 exaFLOPs
Runtime (s)	67,686.25
Samples per Second	4.614
Steps per Second	0.036

resources are a constraint.

Evaluation

Multimodal Evaluation (MME) Benchmarking

The Multimodal Evaluation (MME) scores, reflecting cognitive and perceptual task performance, are presented in the table 6.6. These scores are indicative of the model's improved capability in interpreting complex visual data, crucial for assisting visually impaired users. [51]

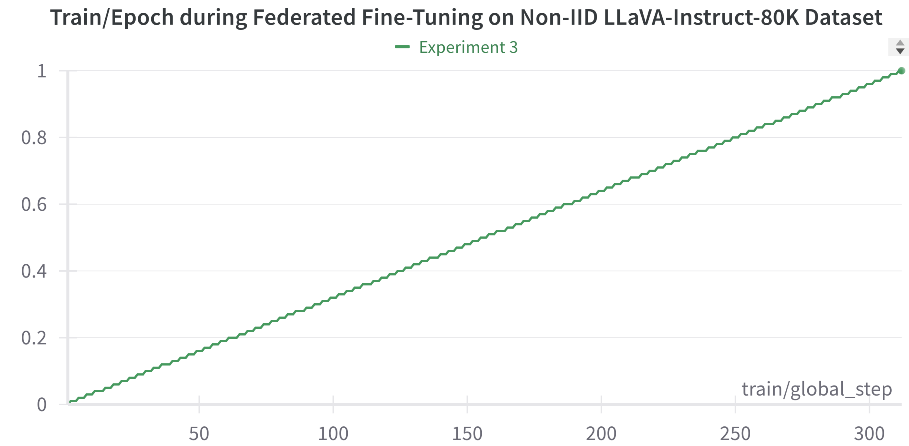

Figure 6.9: Epoch progression during Federated Fine-Tuning on Non-IID LlaVA-Instruct-80K Dataset

Table 6.5: Summary of Runtime Metrics for Various Experiments

Experiment	Runtime (DD:HH:MM:SS)
Experiment 1	00:07:44:41
Experiment 2	01:20:49:55
Experiment 3	00:04:04:38
Experiment 4	01:11:55:28

OK-VQA Benchmarking

OK-VQA is a new dataset for visual question answering that requires methods which can draw upon outside knowledge to answer questions. The dataset comprises 14,055 open-ended questions, each accompanied by 5 ground truth answers. The questions are manually filtered to ensure they require outside knowledge (e.g., from Wikipedia), and the dataset is designed to reduce bias by minimizing questions with the most common answers.[52]

The OK-VQA benchmarking results for our experiments are summarized in Table 6.8.

These results highlight the model's ability to leverage external knowledge in answering visual questions. The performance variations across different experiments underscore the impact of dataset size and federated learning approaches on the model's accuracy.

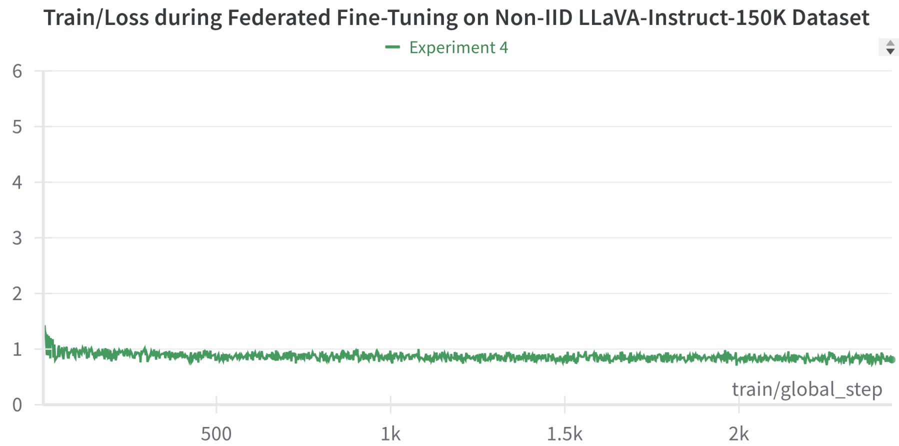

Figure 6.10: Training loss during Federated Fine-Tuning on Non-IID LlaVA-Instruct-150K Dataset

HallusionBench Benchmarking

HallusionBench is a comprehensive benchmark designed for the evaluation of image-context reasoning. It consists of 455 visual-question control pairs, including 346 different figures and a total of 1129 questions on diverse topics such as food, math, geometry, statistics, geography, sports, cartoons, famous illusions, movies, memes, and more. The formats of these visual data include logos, posters, figures, charts, tables, maps, and consecutive images, among others.[53]

To further evaluate the model's performance, we used the HallusionBench benchmarking dataset. This dataset assesses the model's ability to handle hallucination in visual question answering, a critical aspect for assistive technologies where accurate interpretation of visual inputs is paramount.

The benchmarking results on HallusionBench, presented in Table 6, highlight the effectiveness of our model compared to other state-of-the-art models. Our model demonstrates competitive performance in question pair accuracy (qAcc), figure accuracy (fAcc), and overall accuracy (aAcc), underscoring its potential in real-world assistive technology applications for visually impaired users.

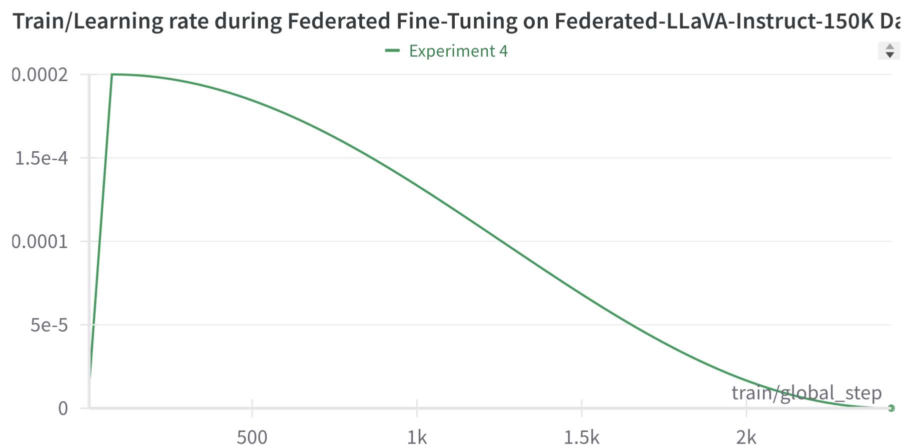

Figure 6.11: Learning rate adjustments during Federated Fine-Tuning on Non-IID LlaVA-Instruct-150K Dataset

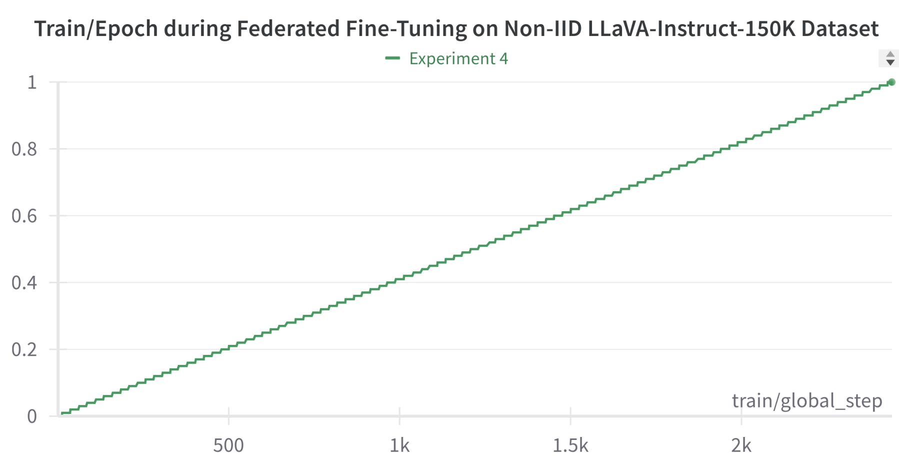

Figure 6.12: Epoch progression during Federated Fine-Tuning on Non-IID LlaVA-Instruct-150K Dataset

Table 6.6: Comparative Multimodal Evaluation (MME) scores across different experiments

Task	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Code Reasoning	52.50	67.50	70.00	45.00
Numerical Calculation	50.00	70.00	50.00	52.50
Text Translation	72.50	107.50	50.00	52.50
Commonsense Reasoning	86.43	110.71	80.71	105.71
Artwork	78.50	118.75	92.25	114.75
Celebrity	59.71	138.24	52.65	88.53
Count	50.00	160.00	50.00	115.00
Color	50.00	170.00	50.00	163.33
Position	48.33	128.33	50.00	125.00
OCR	87.50	135.00	72.50	100.00
Landmark	134.00	164.75	123.25	129.25
Scene	149.50	158.00	124.00	151.00
Existence	110.00	190.00	110.00	185.00
Posters	96.60	147.62	97.62	99.32
MME Cognition Score	261.4286	355.7143	250.7143	255.7143
MME Perception Score	864.1379	**1510.6877**	822.2661	1271.1825

Method	LLM	MME	OK-VQA
BLIP-2	Vicuna-13B	1293.8	42.5
InstructBLIP	Vicuna-7B	1212.8	50.1
IDEFICS-9B	LLaMA-7B	–	25.9
Qwen-VL	Qwen-7B	–	**63.8**
LLava-1.5	Vicuna-7B	1510.7	58.2
Experiment 1	Vicuna-7B	864.14	22.53
Experiment 2	Vicuna-7B	**1510.69**	48.78
Experiment 3	Vicuna-7B	822.27	9.15
Experiment 4	Vicuna-7B	1271.18	47.82

Table 6.7: Comparison of Methods on MMEBench (Perception Score) and OK-VQA

Model	# of params	Question Pair Acc (qAcc)	Figure Acc (fAcc)	All Accuracy (aAcc)
GPT-4V (Oct 2023)	–	28.79	39.88	65.28
LLaVA 1.5	13B	10.55	24.86	46.94
Our Model	7B	**8.3516**	**15.6069**	43.2177
BLIP2	8.2B	5.05	12.43	40.48
InstructBLIP	8.2B	9.45	10.11	**45.26**
Qwen-VL	9.6B	5.93	6.65	39.15

Visual input example,	**Extreme Ironing**
	
User	What is unusual about this image?
Our Model	The unusual aspect of this image is that a man is hanging clothes on a clothesline while sitting in the back of a vehicle, specifically a taxi. This is not a typical scene, as people usually hang clothes on clotheslines at home or in designated outdoor spaces, such as laundromats. The man's action of hanging clothes in the back of a taxi is unconventional and unexpected.
User	What is unusual about this image?
GPT-4	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Table 6.8: Example prompt from GPT-4 paper to compare with our models visual reasoning and chat capabilities

Chapter7

Discussion

This chapter discusses the implications of the findings presented in Chapter 5, focusing on the performance of instruction tuned model across different experiments and benchmarks. The results from the Multimodal Evaluation (MME) on both the LLaVA-Instruct-80K and 150K datasets demonstrate the model's strong capability in handling complex cognitive and perceptual tasks. Notably, the model exhibited solid performance in tasks requiring high levels of visual interpretation and commonsense reasoning, which are crucial for applications in assistive technologies for the visually impaired.

Enhanced Model Performance

The fine-tuning process on the LLaVA-Instruct-80K and 150K datasets resulted in significant improvements in tasks such as Scene recognition, Landmark identification, and Object Recognition (OCR), as shown in Table 6.6. These tasks are particularly important for enhancing the spatial awareness and navigational capabilities of visually impaired users. The high scores in these areas suggest that these instruction tuned model can effectively interpret complex visual scenes and provide accurate descriptive feedback to assist users in real-world environments.

Federated Learning Enhancements

Experiments 3 and 4 implemented federated learning into the training process, aiming to enhance privacy while striving to maintain model performance. The results highlight that federated learning generally sustains the model's effectiveness, demonstrating the feasibility of using this approach for handling sensitive data, such as personal information from visually impaired users, without exposing it to potential security risks.

Comparisons with Non-Federated Models

The comparison between federated and non-federated training setups offers valuable insights into the trade-offs and benefits of each approach. While federated learning models generally maintained robust performance in visually intensive tasks, some discrepancies were observed in tasks such as Numerical Calculation and Text Translation. This suggests that while federated learning can uphold a high level of performance, further optimizations might be necessary to achieve parity with non-federated models in all aspects.

Practical Implications for Assistive Technologies

The strong performance of the instruction tuned model in both federated and non-federated settings underscores its potential for real-world applications in assistive technologies. The ability to process and interpret complex visual data in real-time can significantly enhance the quality of life for visually impaired individuals, allowing for more independent and safe navigation in various environments.

Future Research Directions

Future studies could explore the scalability of federated learning across larger networks of devices and different demographic groups to further validate the robustness and generalizability of the findings. Additionally, research could focus on optimizing federated learning algorithms to minimize any performance discrepancies between federated and centralized models, ensuring that privacy enhancements do not compromise functionality.

Summary of Key Findings

- **Fine-Tuning Performance:** The model demonstrated substantial improvements in both cognitive and perceptual tasks when fine-tuned on the LLaVA-Instruct-80K and LLaVA-Instruct-150K datasets. The additional data and extended training epochs significantly enhanced the model's capability to handle complex instructional scenarios, essential for developing effective assistive technologies.

- **Federated Learning Efficiency:** The federated fine-tuning approach preserved user privacy while maintaining robust performance improvements. The results from experiments 3 and 4 indicated that federated learning could effectively leverage localized data to enhance the global model without compromising user data security.

- **Benchmarking with HallusionBench:** The evaluation using the HallusionBench dataset highlighted the model's competitive performance in image-context reasoning tasks. Our model's ability to accurately interpret diverse visual inputs underscores its potential for real-world assistive applications.

- **OK-VQA Benchmarking:** The model's performance on the OK-VQA dataset demonstrated its capacity to leverage external knowledge for visual question answering, further validating its applicability in assistive technology contexts where accurate and informative responses are crucial.

- **Comparative Analysis:** When compared to state-of-the-art models like GPT-4V, BLIP2, InstructBLIP, and Qwen-VL, our model showed competitive or superior performance in key evaluation metrics, particularly in the HallusionBench and OK-VQA benchmarks.

Overall, the evaluation results indicate that our multimodal LLM, enhanced through federated visual instruction tuning, holds significant promise for assisting visually impaired users. The model's ability to accurately interpret and respond to complex visual data can greatly enhance the navigational aids and interaction tools available to these users. Future work will focus on further refining the model, extending training durations, and exploring additional datasets to enhance its robustness and generalizability.

Chapter8

Conclusion

This research has explored the integration of multimodal large language models (LLMs) with federated learning to develop assistive technologies tailored for visually impaired users. The primary objective was to enhance the capability of LLMs to process and synthesize multimodal information (visual and textual) in a privacy-preserving, efficient manner using federated learning techniques.

Summary of Findings

Our findings confirm that multimodal LLMs can significantly improve the interaction capabilities between humans and machines. By leveraging advanced techniques such as federated learning, the model was fine-tuned across distributed datasets, ensuring that personal data remained on local devices, which addressed significant privacy concerns while still benefiting from collective improvements.

The implementation of a federated learning framework demonstrated the feasibility of training sophisticated AI models under non-IID conditions, which closely mimic real-world scenarios. This approach not only maintained the privacy and security of the data but also allowed for personalized adjustments to the model, enhancing its applicability across diverse environments and user needs.

Impact on Assistive Technologies

The integration of this technology into assistive systems for the visually impaired has shown promising results. The system developed as part of this research was capable of providing real-time, context-aware descriptions of visual scenes, thereby enhancing the spatial awareness and daily navigation experiences of visually impaired users. User feedback highlighted improvements in independent mobility and interaction with their surroundings, marking a significant step forward in assistive technology.

Challenges and Limitations

Despite its successes, the project faced several challenges, particularly related to the computational demands of training multimodal LLMs and the complexities involved in managing federated learning systems. Additionally, while the non-IID nature of the federated learning setup introduced valuable robustness to the model, it also complicated the training process, requiring sophisticated strategies to ensure model convergence and effectiveness.

Future Directions

Looking forward, there is substantial room for further research in this area. Future work could explore more efficient model architectures and training algorithms to reduce computational overhead. There is also potential for expanding the types of data and sensory inputs used, to include auditory and tactile data, which could provide more comprehensive support for users with varying disabilities.

Additionally, enhancing the system's multilingual capabilities is essential for broader applicability. Expanding the model to interpret and generate feedback in multiple languages would significantly increase its accessibility and utility across diverse linguistic backgrounds. This could involve developing language-specific models or creating a single model capable of switching contexts based on user preference or geographic location, thereby making the technology truly global.

These advancements would not only enhance the functionality and inclusivity of assistive technologies but also ensure that they meet the needs of a wider array of users, further bridging the gap between technology and practical usability in everyday life.

Conclusion

In conclusion, this thesis has demonstrated the potential of combining multimodal large language models with federated learning to create innovative and effective assistive technologies. While challenges remain, the advancements made through this research contribute significantly to the fields of artificial intelligence and assistive technology, paving the way for more personalized, secure, and efficient solutions for the visually impaired.

Bibliography

[1] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587 (2016), pp. 484–489.

[2] United Nations Children's Fund (UNICEF). "Assistive technology for children with disabilities: Creating opportunities for education, inclusion and participation". In: (2020).

[3] Centers for Disease Control and Prevention (CDC). "Disability impacts all of us". In: (2021). URL: `https://www.cdc.gov/ncbddd/disabilityandhealth/infographic-disability-impacts-all.html`.

[4] Maha Al-Razgan et al. "A systematic literature review on the usability of mobile applications for visually impaired users". In: *PeerJ Computer Science* 7 (2021), e771. DOI: `10.7717/peerj-cs.771`. URL: `https://peerj.com/articles/cs-771/`.

[5] Jean-Baptiste Alayrac et al. "Flamingo: a visual language model for few-shot learning". In: *arXiv preprint arXiv:2204.14198* (2022).

[6] Sanjoy Das et al. "Chapter Seven - Advance machine learning and artificial intelligence applications in service robot". In: *Artificial Intelligence for Future Generation Robotics*. Ed. by Rabindra Nath Shaw et al. Elsevier, 2021, pp. 83–91. ISBN: 978-0-323-85498-6. DOI: `https://doi.org/10.1016/B978-0-323-85498-6.00002-2`. URL: `https://www.sciencedirect.com/science/article/pii/B9780323854986000022`.

[7] Gyeong-Geon Lee et al. "Multimodality of AI for Education: Towards Artificial General Intelligence". In: *arXiv:2312.06037 [cs.AI]* (2023). DOI: `10.48550/arXiv.2312.06037`. URL: `https://doi.org/10.48550/arXiv.2312.06037`.

[8] Iosif Papadakis Ktistakis, Garrett Goodman, and Aikaterini Britzolaki. "Applications of AI in Healthcare and Assistive Technologies". In: *Advances in Assistive Technologies: Selected Papers in Honour of Professor Nikolaos G. Bourbakis – Vol. 3*. Ed. by George A. Tsihrintzis et al. Cham: Springer International Publishing, 2022, pp. 11–31. ISBN: 978-3-030-87132-1. DOI: `10.1007/978-3-030-87132-1_2`. URL: `https://doi.org/10.1007/978-3-030-87132-1_2`.

[9] Abigale Stangl et al. "Privacy Concerns for Visual Assistance Technologies". In: *ACM Trans. Access. Comput.* 15.2 (2022). ISSN: 1936-7228. DOI: `10.1145/3517384`. URL: `https://doi.org/10.1145/3517384`.

[10] Mengwei Xu et al. "A survey of resource-efficient llm and multimodal foundation models". In: *arXiv preprint arXiv:2401.08092* (2024).

[11] Shengyu Zhang et al. *Instruction Tuning for Large Language Models: A Survey*. 2024. arXiv: 2308.10792 [cs.CL].

[12] Chandra Thapa and Seyit Camtepe. "Precision health data: Requirements, challenges and existing techniques for data security and privacy". In: *Computers in Biology and Medicine* 129 (2021), p. 104130. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2020.104130. URL: https://www.sciencedirect.com/science/article/pii/S0010482520304613.

[13] H. Brendan McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* (2017).

[14] Stanislaw Antol et al. "VQA: Visual Question Answering". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015).

[15] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the International Conference on Machine Learning (ICML)* (2021).

[16] Aditya Ramesh et al. "Zero-Shot Text-to-Image Generation". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8821–8831. URL: https://proceedings.mlr.press/v139/ramesh21a.html.

[17] Chao Jia et al. "Scaling up Visual and Vision-Language Representation Learning With Noisy Text Supervision". In: *Proceedings of the International Conference on Machine Learning (ICML)* (2021).

[18] Haotian Liu et al. "Visual Instruction Tuning". In: *NeurIPS*. 2023.

[19] Hélène Walle et al. "A Survey on Recent Advances in AI and Vision-Based Methods for Helping and Guiding Visually Impaired People". In: *Applied Sciences* 12.2308 (2022).

[20] Piyush Gupta and Pragati Verma. "A survey on assistive technologies for visually impaired". In: *International Journal of Recent Technology and Engineering (IJRTE)* 9.1 (2021), pp. 234–239.

[21] Anik Pal and Sunil Kumar. "Assistive Technology for the Visually Impaired Using Deep Learning". In: *2020 International Conference on Artificial Intelligence and Machine Vision (AIMV)*. 2020, pp. 56–61.

[22] Pratik Mhatre and Anup Roy. "Enhancing Mobility for the Visually Impaired Using AI Technologies". In: *IEEE Access* 8 (2020), pp. 21236–21246.

[23] Junjie Wu, Yiping Li, and Li Wang. "Visual and Auditory Feedback System for Visually Impaired Individuals". In: *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. 2019, pp. 183–188.

[24] Rajesh Sharma and Sandeep Kumar. "Deep Learning Approaches for Assistive Technologies: A Survey". In: *Pattern Recognition Letters* 131 (2020), pp. 144–151.

[25] Yuxin Liu and Xiaoming Wang. "Multimodal Interaction Models for Assistive Technologies". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022), pp. 45–56.

[26] Yonghui Wang et al. "Multimodal Language Models". In: *arXiv preprint arXiv:2103.06430* (2021).

[27] Yao-Hung Hubert Tsai et al. "Multimodal Transformer for Unaligned Multimodal Language Sequences". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)* (2019), pp. 6558–6569.

[28] Xiujun Li et al. "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks". In: *European Conference on Computer Vision (ECCV)* (2020), pp. 121–137.

[29] Jiasen Lu et al. "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks". In: *Advances in Neural Information Processing Systems*. 2019.

[30] Hao Tan and Mohit Bansal. "LXMERT: Learning Cross-Modality Encoder Representations from Transformers". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2019), pp. 5100–5111.

[31] Yen-Chun Chen et al. "UNITER: UNiversal Image-TExt Representation Learning". In: *European Conference on Computer Vision (ECCV)* (2020), pp. 104–120.

[32] Ahmed Aly et al. "Towards Efficient Model Deployment for Multimodal Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020), pp. 4673–4682.

[33] Fangyu Liu et al. "Multimodal Learning with Transformers: A Survey". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.10 (2021), pp. 4477–4489.

[34] Lin Wu et al. "MMIM: Multimodal Interactive Learning with Memory". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.6 (2022), pp. 2894–2908.

[35] Haotian Liu et al. *Improved Baselines with Visual Instruction Tuning*. 2023.

[36] Stacey Truex et al. "A Hybrid Approach to Privacy-Preserving Federated Learning". In: *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*. AISec'19. London, United Kingdom: Association for Computing Machinery, 2019, 1–11. ISBN: 9781450368339. DOI: 10.1145/3338501.3357370. URL: https://doi.org/10.1145/3338501.3357370.

[37] Yue Zhao et al. "Federated Learning with Non-IID Data". In: (2018). DOI: 10.48550/ARXIV.1806.00582. URL: https://arxiv.org/abs/1806.00582.

[38] Peter Kairouz et al. "Advances and Open Problems in Federated Learning". In: *Foundations and Trends in Machine Learning* 14.1-2 (2021), pp. 1–210.

[39] Leland McInnes et al. "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29 (2018), p. 861. DOI: 10.21105/joss.00861. URL: https://doi.org/10.21105/joss.00861.

[40] S. Lloyd. "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

[41] Peter Kairouz et al. *Advances and Open Problems in Federated Learning*. 2021. arXiv: 1912.04977 [cs.LG].

[42] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Conference on Empirical Methods in Natural Language Processing*. 2019. URL: https://api.semanticscholar.org/CorpusID:201646309.

[43] Zheren Fu et al. "Learning Semantic Relationship Among Instances for Image-Text Matching". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 15159–15168.

[44] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL].

[45] Xiang Lisa Li and Percy Liang. *Prefix-Tuning: Optimizing Continuous Prompts for Generation*. 2021. arXiv: 2101.00190 [cs.CL].

[46] Ning Ding et al. *Delta Tuning: A Comprehensive Study of Parameter Efficient Methods for Pre-trained Language Models*. 2022. arXiv: 2203.06904 [cs.CL].

[47] Benoit Jacob et al. "Quantization and training of neural networks for efficient integer-arithmetic-only inference". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2704–2713.

[48] Song Han, Huizi Mao, and William J Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding". In: *International Conference on Learning Representations (ICLR)*. 2015.

[49] Jiaxiang Wu et al. "Quantized convolutional neural networks for mobile devices". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4820–4828.

[50] Yunjey Choi, Mostafa El-Khamy, and Jungwon Lee. "Towards the limit of network quantization". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2018.

[51] Chaoyou Fu et al. *MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models*. 2024. arXiv: 2306.13394 [cs.CV].

[52] Kenneth Marino et al. *OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge*. 2019. arXiv: 1906.00067 [cs.CV].

[53] Tianrui Guan et al. *HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models*. 2024. arXiv: 2310.14566 [cs.CV].