Scalability Discussion: Vision-Language Model Fine-Tuning

1. Dataset Management

   - Storing datasets in efficient formats like parquet and utilizing cloud storage like AWS S3

   - Using multi-threaded pytorch's DataLoader with n_workers

2. Distributed Training

   - Employ multi-GPU setups with DistributedDataParallel (DDP) or Hugging Face Accelerate.

3. Model Checkpoint

   - Saving and managing checkpoints of models to preventing loss of progress.

4. Deployment Considerations

   - Optimize models using ONNX or TensorRT for reduced inference latency.

   - Use Docker/Kubernetes for scalable deployment on cloud platforms (AWS, GCP, Azure)