MSML640 Computer Vision

Project Progress Report

Author: Anirud Mohan, Hengmeng Wang, Hsing-Hao Wang, Javad Baghirov, William Loe

GitHub Repository Link

- https://github.com/willloe/MSML640_Group10

Summary of progress

For the classifier pipeline, text extraction from the SlideVQA dataset was expanded to include all text-related elements such as text boxes, titles, and page numbers, and an OCR system is being integrated to capture text with bounding box positions. A slide-level classifier was developed to detect whether slides contain diagrams, figures, or images, achieving an accuracy of 0.8. Future work is now focused on extending this to detect multiple elements per slide with bounding boxes and multi-label outputs. Metadata for the SlideVQA dataset was generated, and scripts were created to parse the dataset, extract slide images, and annotate them with bounding boxes. A web server is also being implemented to automate the process of parsing presentations into training data.

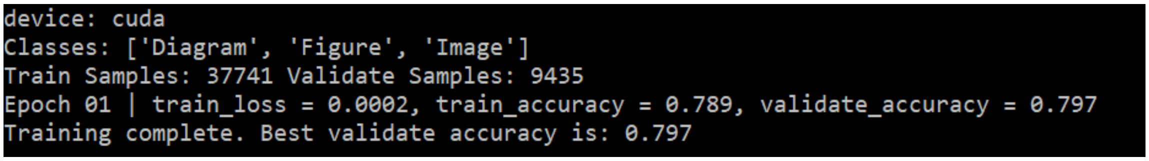
For the diffusion model, the project has a working SDXL-based generation stack with layout awareness and reproducible smoke tests. A preprocessing module converts slide layouts into a 4-channel control tensor and a safe-zone mask, enabling ControlNet-style guidance that respects title/body/image/logo regions while preserving background freedom. Compact JSON Schemas for layouts and palettes are in place, plus a synthetic data module that creates realistic, schema-valid samples for testing without external datasets. The inference pipeline loads SDXL with optional LoRA, applies a readability-focused safe-zone postprocess, and optionally uses ControlNet with a configurable control_from signal (element/safe/edge). Debug artifacts include the exact conditioning image, raw generation, masked output, overlay, and a determinism hash. Smoke scripts run locally (CPU skips heavy gen) and on Colab (T4 GPU), giving a clean, testable foundation for upcoming work on schedulers, ControlNet tuning, and LoRA fine-tuning.

New task division:

- William Loe (Generation stack): ControlNet conditioning MVP, scheduler options (DDIM/DPM++), LoRA training scaffold, 512 -> 1280 upscaler, and optional inpainting hook.
- HengMeng (Data/metrics/integration): WCAG-AA readability metric, layout-safety metric, dataset + manifest, synthetic recipe/themes, CLI + Colab UX, and reporting.

Proof-of-Concept or early results

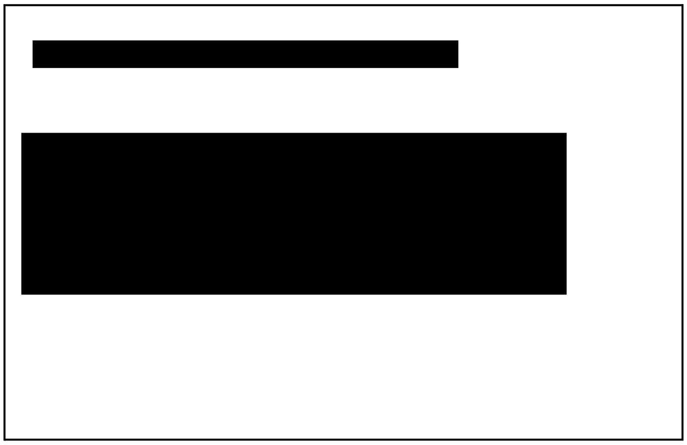
Classifier Model:



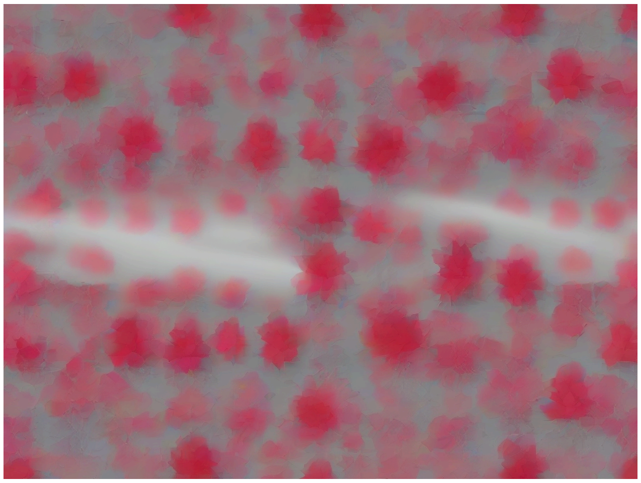
```
device: cuda
Classes: ['Diagram', 'Figure', 'Image']
Train Samples: 37741 Validate Samples: 9435
Epoch 01 | train_loss = 0.0002, train_accuracy = 0.789, validate_accuracy = 0.797
Training complete. Best validate accuracy is: 0.797
```

Diffusion Model:

Generated safe zone for text areas



Generated raw background

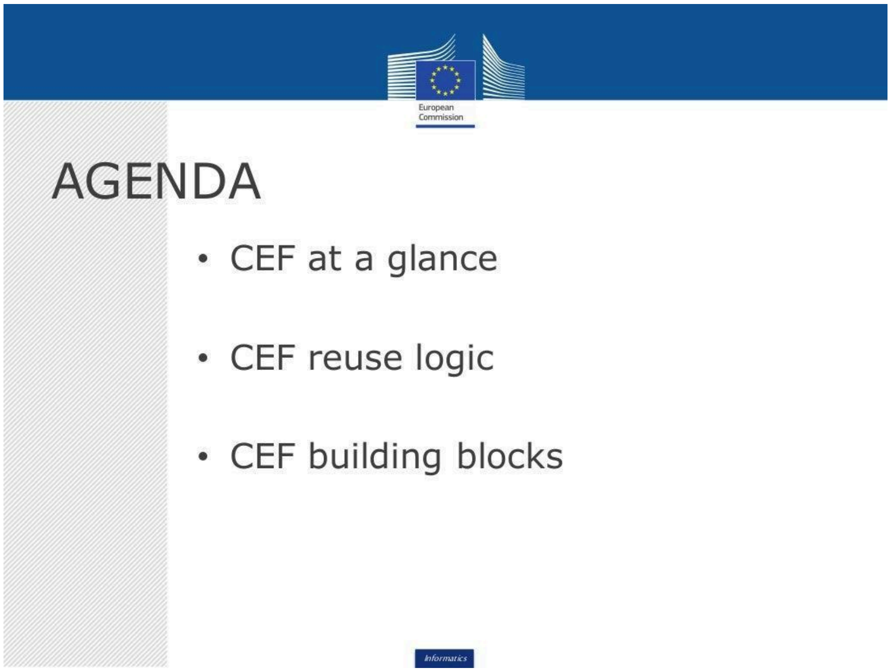


Early result: raw background + safe zone mask

Updates (if applicable)

Classifier Model:

Anirud: The extraction dataset code has been modified to extract all text-related content from the SlideVQA dataset, including TextBoxes, Titles, and Page Numbers. Additionally, an OCR extraction program will be implemented to extract text content along with its corresponding bounding box position on each slide.

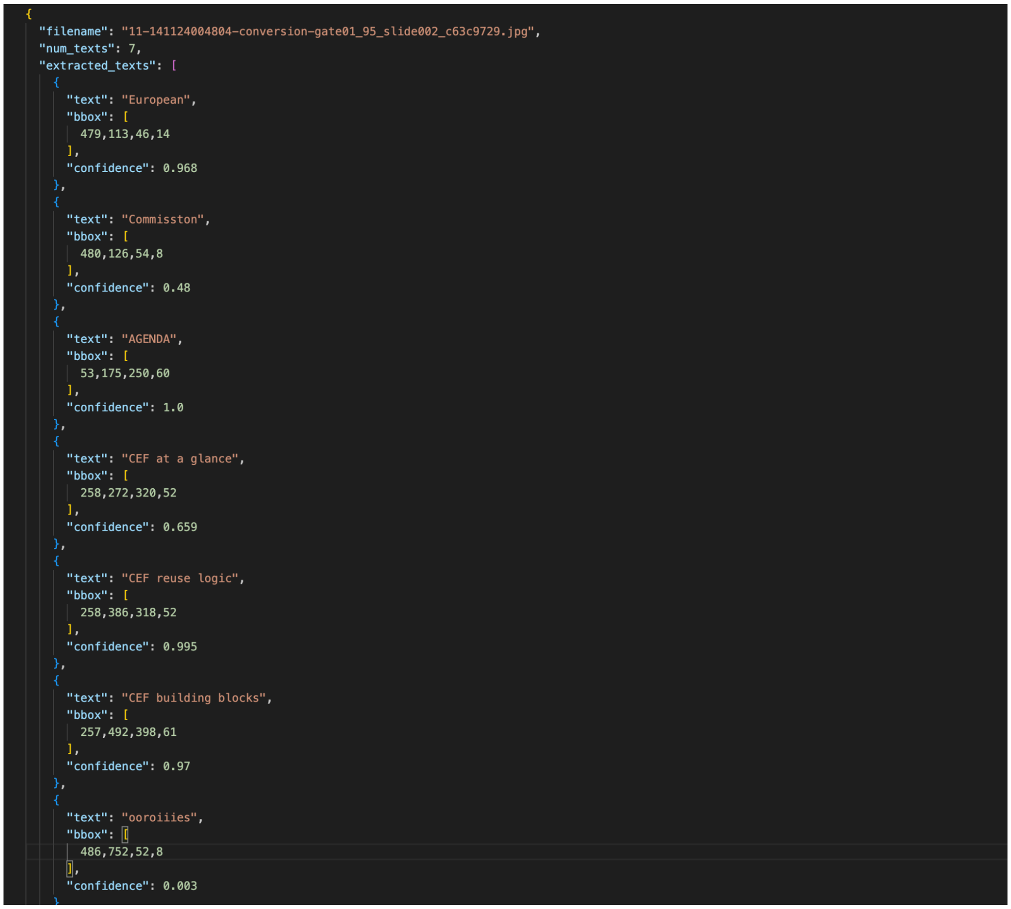


European
Commission

AGENDA

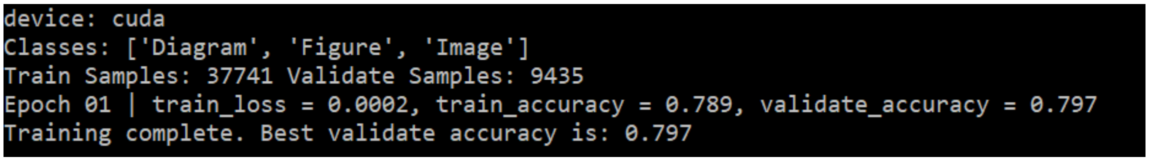
- CEF at a glance
- CEF reuse logic
- CEF building blocks

Informatics



```
{
  "filename": "11-141124004804-conversion-gate01_95_slide002_c63c9729.jpg",
  "num_texts": 7,
  "extracted_texts": [
    {
      "text": "European",
      "bbox": [
        479,113,46,14
      ],
      "confidence": 0.968
    },
    {
      "text": "Commisston",
      "bbox": [
        480,126,54,8
      ],
      "confidence": 0.48
    },
    {
      "text": "AGENDA",
      "bbox": [
        53,175,250,60
      ],
      "confidence": 1.0
    },
    {
      "text": "CEF at a glance",
      "bbox": [
        258,272,320,52
      ],
      "confidence": 0.659
    },
    {
      "text": "CEF reuse logic",
      "bbox": [
        258,386,318,52
      ],
      "confidence": 0.995
    },
    {
      "text": "CEF building blocks",
      "bbox": [
        257,492,398,61
      ],
      "confidence": 0.97
    },
    {
      "text": "ooroiiies",
      "bbox": [
        486,752,52,8
      ],
      "confidence": 0.003
    }
```

Hsing-Hao: Created a classifier using a dataset from SlideVQA to classify if a slide is containing diagrams, figures, and images. Classification successful rate is at 0.8. Further implementation is needed to create bounding boxes to classify multiple elements and create multiple labels on one slide.



```
device: cuda
Classes: ['Diagram', 'Figure', 'Image']
Train Samples: 37741 Validate Samples: 9435
Epoch 01 | train_loss = 0.0002, train_accuracy = 0.789, validate_accuracy = 0.797
Training complete. Best validate accuracy is: 0.797
```

Javad: Generated metadata for the SlideVQA dataset using a script, which will later be used for parsing. Created a script to parse the dataset and extract and annotate images from the dataset with bounding boxes which are shown on each slide. Started implementing a web server that parses presentations to compile training data.

Diffusion Model:

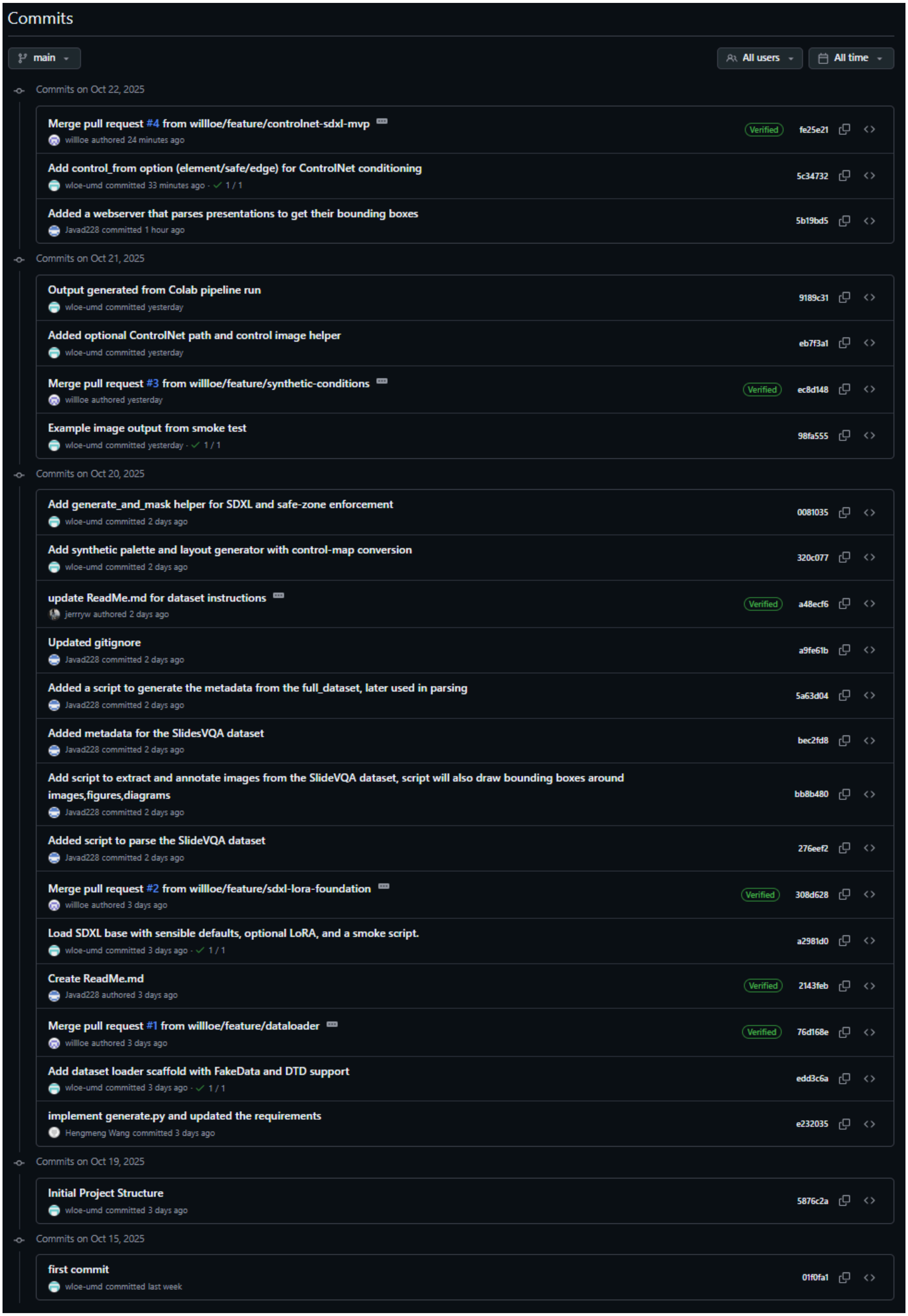
Hengmeng: Implemented structured layout data (like positions and types of titles, images, logos) into a 4-channel control tensor that guides diffusion models during image generation. It encodes element positions, class types, layering (z-order), and reading order into spatial maps that condition the model to respect layout constraints, while also creating "safe zone" masks for areas where the model can generate freely without layout restrictions. Essentially, the “generate.py” is a preprocessing pipeline that translates human-readable design specifications into the continuous tensor format needed by ControlNet-style diffusion models to generate images that follow specific layout structures.

William: So far, I have set up a working SDXL-based generation stack with layout awareness and reproducible smoke tests. I defined and validated compact JSON Schemas for slide layouts and color palettes. I also built a synthetic data module that produces realistic, schema-valid layouts/palettes so I can test without external datasets. On the model side, I wired up sdxl.py to load Stable Diffusion XL with optional LoRA weights and sensible CPU/GPU defaults. I implemented infer.py to run generation and apply a safe-zone postprocess that keeps text regions neutral for readability. I also added an optional ControlNet path with a control_from switch (element/safe/edge) so I can experiment with which conditioning signal works best. The pipeline now saves the exact control image, raw generation, masked output, and an overlay for quick inspection.

I created lightweight smoke scripts (smoke_sdxl_load.py, smoke_synthetic.py, smoke_infer.py) that run locally (CPU skips heavy gen) and on Colab (GPU) using the same code.

GitHub Activity Snapshot

Main branch:



Commits

main | All users | All time

Commits on Oct 22, 2025

Merge pull request #4 from willloe/feature/controlnet-sdxl-mvp
willloe authored 24 minutes ago — Verified — fe25e21

Add control_from option (element/safe/edge) for ControlNet conditioning
wloe-umd committed 33 minutes ago · ✓ 1 / 1 — 5c34732

Added a webserver that parses presentations to get their bounding boxes
Javad228 committed 1 hour ago — 5b19bd5

Commits on Oct 21, 2025

Output generated from Colab pipeline run
wloe-umd committed yesterday — 9189c31

Added optional ControlNet path and control image helper
wloe-umd committed yesterday — eb7f3a1

Merge pull request #3 from willloe/feature/synthetic-conditions
willloe authored yesterday — Verified — ec8d148

Example image output from smoke test
wloe-umd committed yesterday · ✓ 1 / 1 — 98fa555

Commits on Oct 20, 2025

Add generate_and_mask helper for SDXL and safe-zone enforcement
wloe-umd committed 2 days ago — 0081035

Add synthetic palette and layout generator with control-map conversion
wloe-umd committed 2 days ago — 320c077

update ReadMe.md for dataset instructions
jerryw authored 2 days ago — Verified — a48ecf6

Updated gitignore
Javad228 committed 2 days ago — a9fe61b

Added a script to generate the metadata from the full_dataset, later used in parsing
Javad228 committed 2 days ago — 5a63d04

Added metadata for the SlidesVQA dataset
Javad228 committed 2 days ago — bec2fd8

Add script to extract and annotate images from the SlideVQA dataset, script will also draw bounding boxes around images,figures,diagrams
Javad228 committed 2 days ago — bb8b480

Added script to parse the SlideVQA dataset
Javad228 committed 2 days ago — 276eef2

Merge pull request #2 from willloe/feature/sdxl-lora-foundation
willloe authored 3 days ago — Verified — 308d628

Load SDXL base with sensible defaults, optional LoRA, and a smoke script.
wloe-umd committed 3 days ago · ✓ 1 / 1 — a2981d0

Create ReadMe.md
Javad228 authored 3 days ago — Verified — 2143feb

Merge pull request #1 from willloe/feature/dataloader
willloe authored 3 days ago — Verified — 76d168e

Add dataset loader scaffold with FakeData and DTD support
wloe-umd committed 3 days ago · ✓ 1 / 1 — edd3c6a

implement generate.py and updated the requirements
Hengmeng Wang committed 3 days ago — e232035

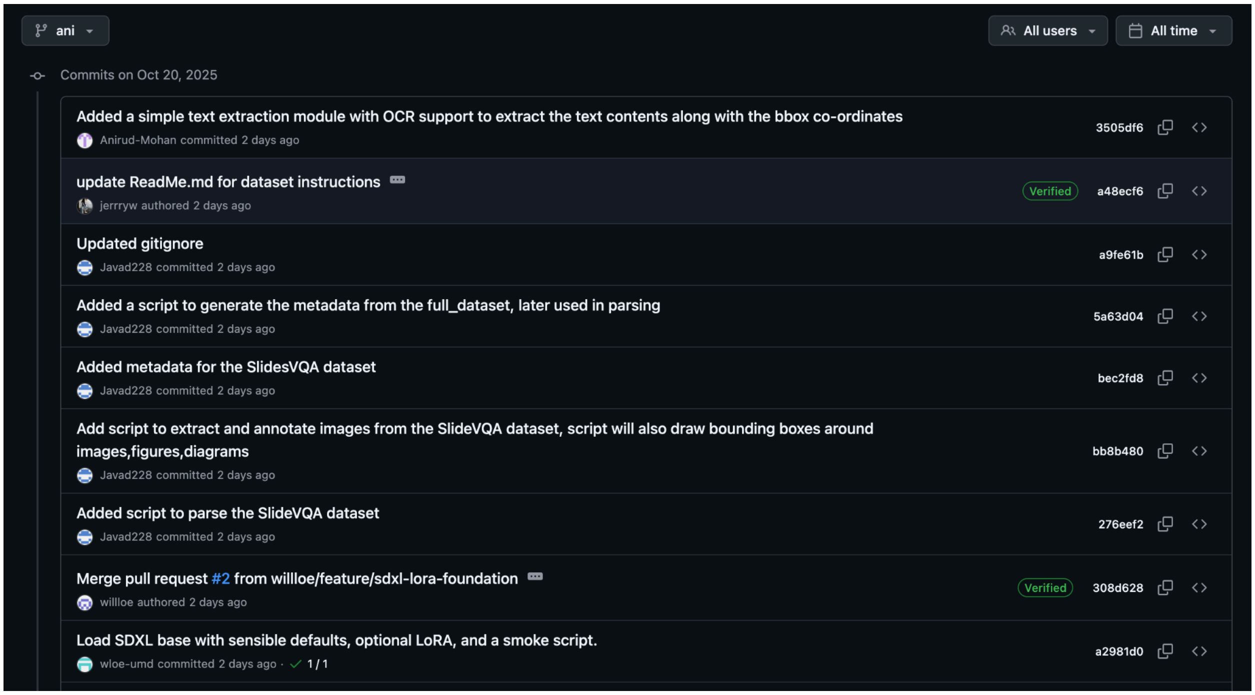
Commits on Oct 19, 2025

Initial Project Structure
wloe-umd committed 3 days ago — 5876c2a

Commits on Oct 15, 2025

first commit
wloe-umd committed last week — 01f0fa1

ani:



Commits on Oct 20, 2025

Added a simple text extraction module with OCR support to extract the text contents along with the bbox co-ordinates
Anirud-Mohan committed 2 days ago — 3505df6

update ReadMe.md for dataset instructions
jerrryw authored 2 days ago — Verified — a48ecf6

Updated gitignore
Javad228 committed 2 days ago — a9fe61b

Added a script to generate the metadata from the full_dataset, later used in parsing
Javad228 committed 2 days ago — 5a63d04

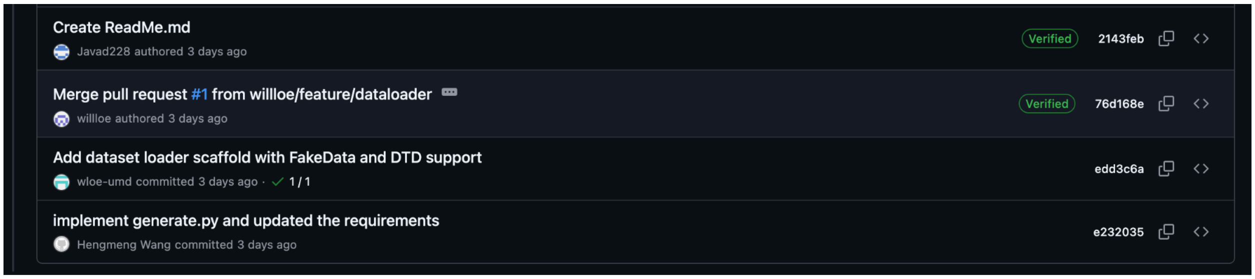
Added metadata for the SlidesVQA dataset
Javad228 committed 2 days ago — bec2fd8

Add script to extract and annotate images from the SlideVQA dataset, script will also draw bounding boxes around images,figures,diagrams
Javad228 committed 2 days ago — bb8b480

Added script to parse the SlideVQA dataset
Javad228 committed 2 days ago — 276eef2

Merge pull request #2 from willloe/feature/sdxl-lora-foundation
willloe authored 2 days ago — Verified — 308d628

Load SDXL base with sensible defaults, optional LoRA, and a smoke script.
wloe-umd committed 2 days ago · ✓ 1 / 1 — a2981d0



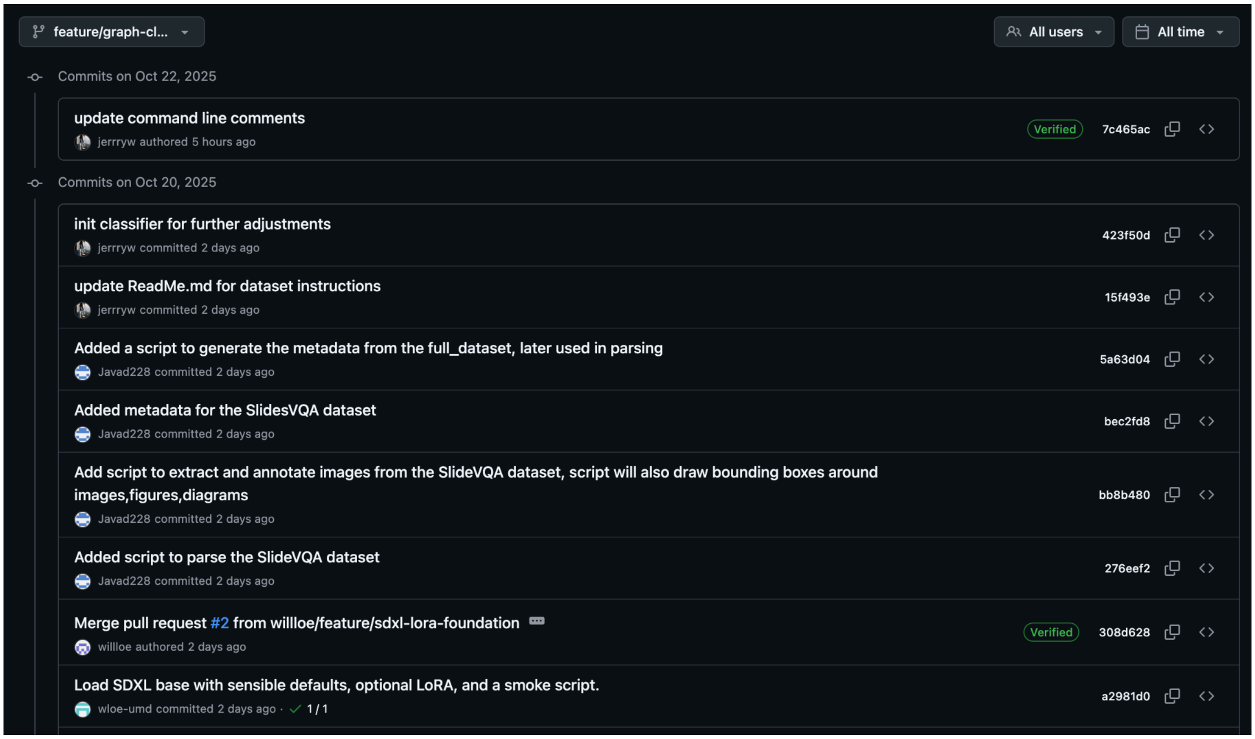
Create ReadMe.md
Javad228 authored 3 days ago — Verified — 2143feb

Merge pull request #1 from willloe/feature/dataloader
willloe authored 3 days ago — Verified — 76d168e

Add dataset loader scaffold with FakeData and DTD support
wloe-umd committed 3 days ago · ✓ 1 / 1 — edd3c6a

implement generate.py and updated the requirements
Hengmeng Wang committed 3 days ago — e232035

feature/graph-classification



Commits on Oct 22, 2025

update command line comments
jerrryw authored 5 hours ago — Verified — 7c465ac

Commits on Oct 20, 2025

init classifier for further adjustments
jerrryw committed 2 days ago — 423f50d

update ReadMe.md for dataset instructions
jerrryw committed 2 days ago — 15f493e

Added a script to generate the metadata from the full_dataset, later used in parsing
Javad228 committed 2 days ago — 5a63d04

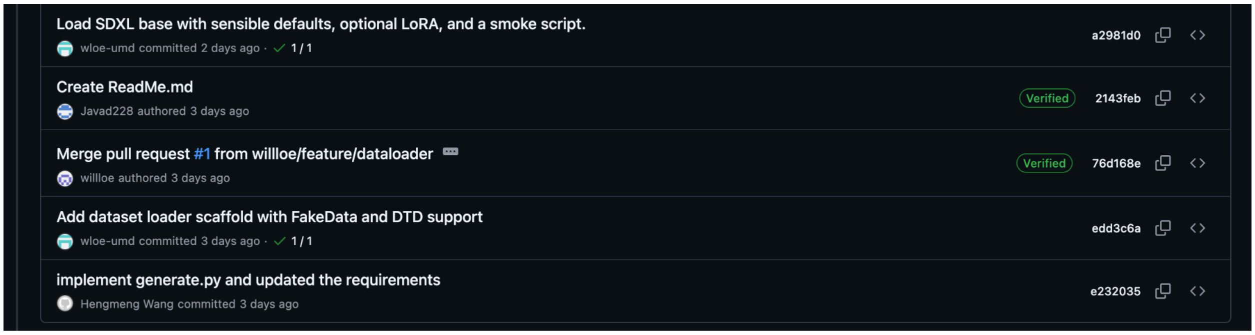
Added metadata for the SlidesVQA dataset
Javad228 committed 2 days ago — bec2fd8

Add script to extract and annotate images from the SlideVQA dataset, script will also draw bounding boxes around images,figures,diagrams
Javad228 committed 2 days ago — bb8b480

Added script to parse the SlideVQA dataset
Javad228 committed 2 days ago — 276eef2

Merge pull request #2 from willloe/feature/sdxl-lora-foundation
willloe authored 2 days ago — Verified — 308d628

Load SDXL base with sensible defaults, optional LoRA, and a smoke script.
wloe-umd committed 2 days ago · ✓ 1 / 1 — a2981d0



Load SDXL base with sensible defaults, optional LoRA, and a smoke script.
wloe-umd committed 2 days ago · ✓ 1 / 1 — a2981d0

Create ReadMe.md
Javad228 authored 3 days ago — Verified — 2143feb

Merge pull request #1 from willloe/feature/dataloader
willloe authored 3 days ago — Verified — 76d168e

Add dataset loader scaffold with FakeData and DTD support
wloe-umd committed 3 days ago · ✓ 1 / 1 — edd3c6a

implement generate.py and updated the requirements
Hengmeng Wang committed 3 days ago — e232035