Data Acquisition: Conference Poster Crawlers

Wenqing Zhang Xiaotang Sun Jixuan Wang

Group: NoEyeDeer

December 19, 2025

Project Overview

Goal: Build structured database of AI conference posters
Conferences: ICLR 2025 (session 1) + ICML 2025 Total: 600+ posters

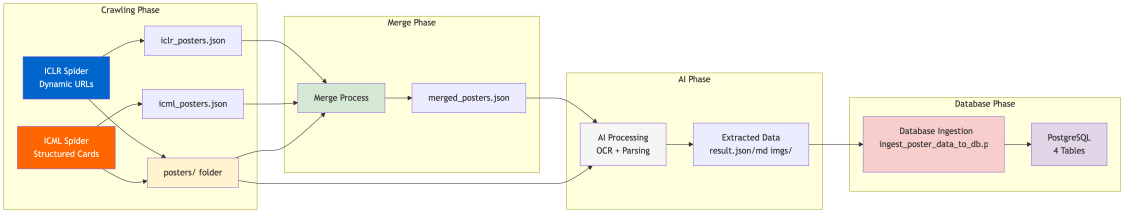


Figure: Dual Spider → Merge → Extract Pipeline

Dual Spider Architecture: ICLR vs ICML

ICLR Spider

`ICLRPosterSpider`

- **Start URL**: Dynamic session-based
- **Target**: Poster Session 1
- **Challenge**: Session filtering

ICML Spider

`ICMLPosterSpider`

- **Start URL**: Fixed events page
- **Target**: Spotlight Posters
- **Filter**: Poster type checking

Aspect	**ICLR**	**ICML**
URL Pattern	`?filter=session&search=Poster+Session+1`	`/virtual/2025/events/2025SpotlightPosters`
Navigation	Extract from `li a::attr(href)`	Loop through `div.displaycards`
Session Type	Poster Session 1 (dynamic)	Spotlight Posters (fixed)

Table: Different URL handling strategies

Code Snippet: ICLR Parsing (Dynamic)

Listing 1: ICLR Dynamic Session Handling

```
start_urls = [
    "https://iclr.cc/virtual/2025/papers.html?filter=session&search=Poster+
        Session+1"
]

def parse(self, response):
    poster_links = response.css("li␣a::attr(href)").getall()
    for link in poster_links:
        url = response.urljoin(link)
        yield scrapy.Request(url, callback=self.parse_poster)
```

Key: Handles dynamic session-based listing

Code Snippet: ICML Parsing (Structured)

Listing 2: ICML Structured Card Parsing

```
def parse(self, response):
    for card in response.css("div.displaycards.touchup-date"):
        title = card.css("a.small-title.text-underline-hover::text").get()
        authors = card.css("div.author-str::text").get()
        abstract_lines = card.css("details > div.text-start.p-4 *::text").
            getall()

        # Filter for Spotlight Posters only
        if "Spotlight Poster" in poster_type:
            yield scrapy.Request(...)
```

Key: Structured card layout with type filtering

Output Files and Merge Process

Raw Output:

- `iclr_posters.json`
- `icml_posters.json`
- `posters/` folder

Merge Step (`merged_posters.py`)

- Combine two JSON files
- Add conference prefix:
  - `ICLR2025_`
  - `ICML2025_`
- Standardize fields
- Preserve image references

Key Design Decision

Both spiders save images to same `posters/` folder, requiring JSON merge to maintain proper metadata-image mapping.

Merged JSON Structure Example

Listing 3: merged_posters.json excerpt

```
[
    {
    "poster_id": "ICML_44337",
    "conference": "ICML2025",
    "title": "BaxBench:␣Can␣LLMs␣Generate␣Correct␣and␣Secure␣Backends?",
    "authors": "Mark␣Vero␣,␣Niels␣M ndler␣,␣Viktor␣Chibotaru␣,␣Veselin␣
        Raychev␣,␣Maximilian␣Baader␣,␣Nikola␣Jovanovi ␣,␣Jingxuan␣He␣,␣
        Martin␣Vechev",
    "source_url": "https://icml.cc/media/PosterPDFs/ICML%202025/44337.png?t
        =1752426178.286824",
    "page_url": "https://icml.cc/virtual/2025/poster/44337",
    "local_png_path": "posters/44337.png"
  },
  ...
]
```

Layout Analysis: Challenges & Results

Initial Method:

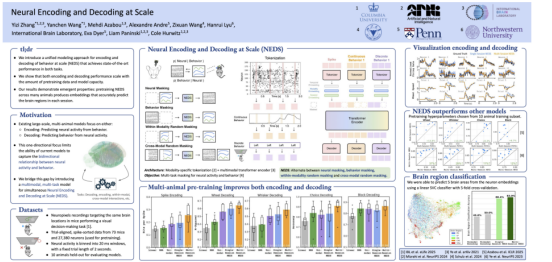
- Detectron2 + Tesseract OCR
- Block detection + text extraction

Problems Encountered:

- Over-segmentation
- Poor OCR accuracy
- Math formulas lost

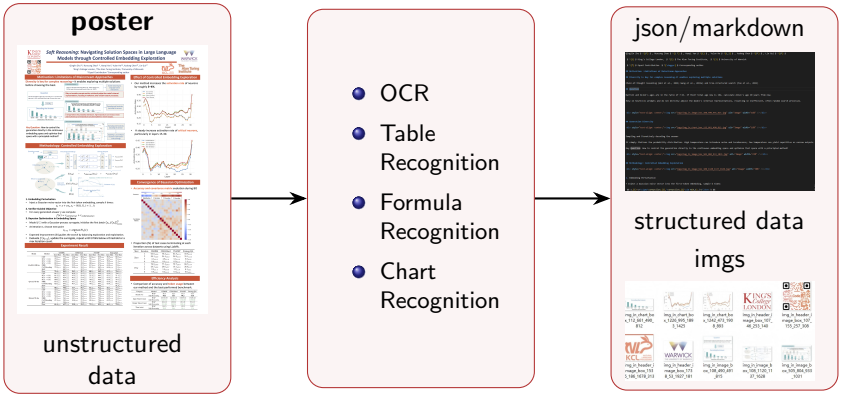
Advanced Attempts:

- KMeans column detection
- Parameter tuning
- Heuristic merging



Poster: ICML_43679

Solution: PaddleOCR-VL



For each poster after process:

- `json` file
- `markdown` file
- `imgs` folder

Extraction to Database

Processing Pipeline

`ingest_poster_data_to_db.py`

1. Read AI-processed data (`result.json`, `result.md`, `imgs/`)
2. Combine with original metadata
3. Insert into PostgreSQL (4 tables)

Tables:

1. `poster_info`: global metadata (title, authors, year)
2. `poster_blocks`: textual blocks + bounding boxes
3. `poster_figure`: cropped figure image paths
4. `poster_table`: parsed table structure (rows, cells)

ER Diagram

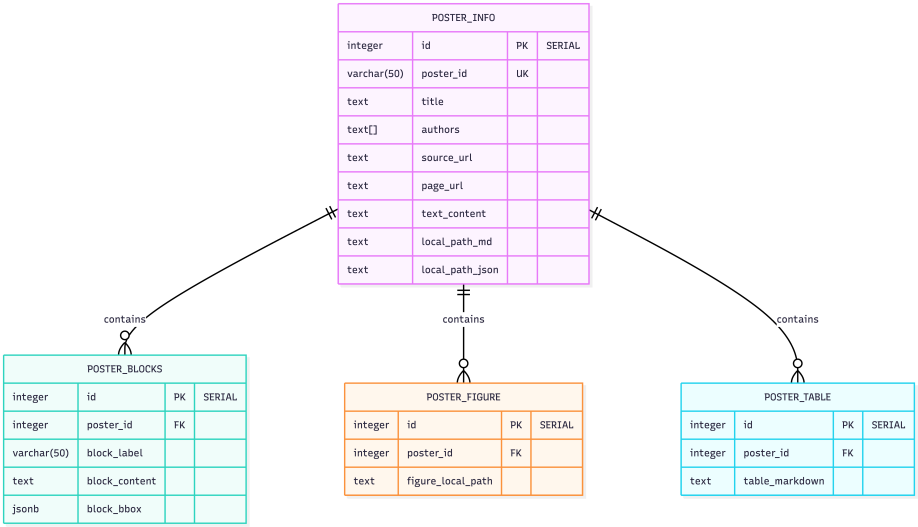

POSTER_INFO
integer id PK SERIAL
varchar(50) poster_id UK
text title
text[] authors
text source_url
text page_url
text text_content
text local_path_md
text local_path_json
contains
contains
contains
POSTER_BLOCKS
integer id PK SERIAL
integer poster_id FK
varchar(50) block_label
text block_content
jsonb block_bbox
POSTER_FIGURE
integer id PK SERIAL
integer poster_id FK
text figure_local_path
POSTER_TABLE
integer id PK SERIAL
integer poster_id FK
text table_markdown

Table 1: poster_info (Core Metadata)

```
SELECT * FROM poster_info LIMIT 1;
```

Field	Value
id	1
poster_id	29917
title	HelpSteer2-Preference: Complementing Ratings with Preferences
authors	Zhilin Wang, Alexander Bukharin, ...
source_url	https://iclr.cc/media/PosterPDFs/ICLR%202025/29917.png
page_url	https://iclr.cc/virtual/2025/poster/29917
text_content	# HelpSteer2-Preference: Complementing Ratings ...
local_path_md	/.../29917/result.md
local_path_json	/.../29917/result.json

Table 2: poster_blocks (AI-Parsed Structure)

```
SELECT * FROM poster_blocks LIMIT 1;
```

Field	Value
id	1
poster_id	1 (references poster_info.id)
block_label	header_image
block_content	*(empty - image block)*
block_bbox	[150, 59, 365, 274]

Bounding Box Explanation

JSONB Format: [x1, y1, x2, y2]

- Coordinates: (150, 59) to (365, 274)
- Width: 215 pixels
- Height: 215 pixels
- Type: `header_image` (poster header image)

Table 3: poster_figure (Extracted Images)

SELECT * FROM poster_figure LIMIT 1;

Field	Value
id	1
poster_id	1 (references poster_info.id)
figure_local_path	/.../29917/imgs/img_in_header_image_box_3341_61_3929_178.jpg

Image Extraction Details

Filename Pattern: img_in_[block_label]_box_[x1]_[y1]_[x2]_[y2].jpg

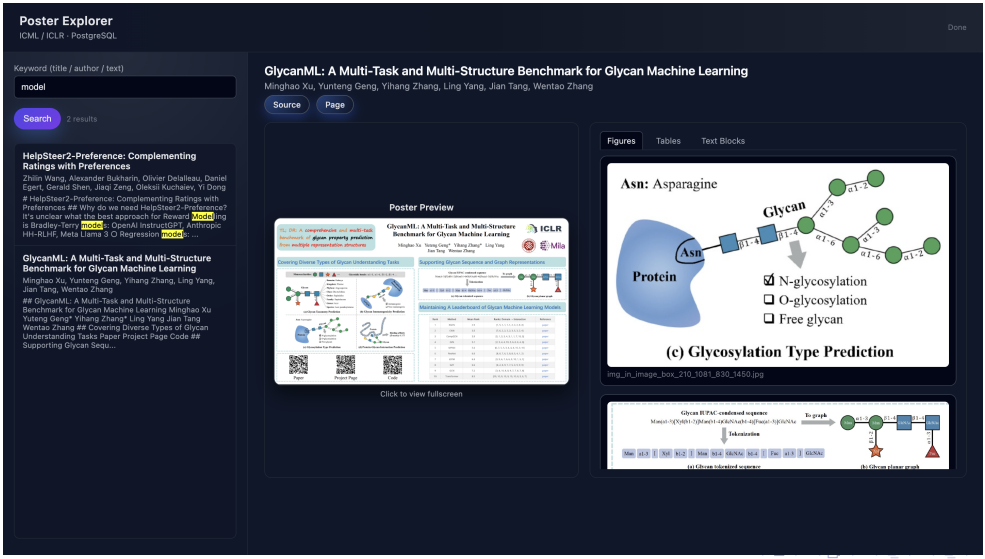
- Source: From header_image block
- Coordinates: matches bounding box
- Format: Cropped JPEG from original poster
- Location: Organized in imgs/ folder

Table 4: poster_table (Markdown Tables)

```
SELECT * FROM poster_table LIMIT 3;
```

```
1,"| Selection Method | General Knowledge (3 tasks) | Commonsense Reasoning (4 tasks) | Reading Comprehension (2 tasks
| ---------------- | --------------------------- | ------------------------------- | ------------------------------ |
| Random           | 50.33                       | 36.19                           | 39.09                          |
| DSIR             | 50.37  $ \uparrow $ 0.04    | 34.01  $ \downarrow $ 2.18      | 38.80  $ \downarrow $ 1.29     |
| PPL              | 48.71  $ \downarrow $ 1.62  | 37.72  $ \uparrow $ 1.53        | 38.57  $ \downarrow $ 0.52     |
| Semdedup         | 50.99  $ \uparrow $ 0.66    | 36.11  $ \downarrow $ 0.08      | 39.44  $ \uparrow $ 0.35       |
| Qurating         | 51.56  $ \uparrow $ 1.23    | 35.93  $ \downarrow $ 0.26      | 39.70  $ \uparrow $ 0.61       |
| MATES            | 50.45  $ \uparrow $ 0.12    | 36.06  $ \downarrow $ 0.13      | 39.83  $ \uparrow $ 0.74       |
| Quad(ours)       | 52.08  $ \uparrow $ 1.75    | 37.03  $ \uparrow $ 0.84        | 41.07  $ \uparrow $ 1.98       |
2,"| Defence          | Environment | Action Space | Certification |
| ---------------- | ----------- | ------------ | ------------- |
| Deterministic    | Stochastic  | Discrete     | Continuous    |
| COPA [Wu et al.,] | ✓           | ✗            | ✓             |
| Our Method       | ✓           | ✓            | ✓             |"
2,"| Environment | Method          | Noise          | Avg. Cumulative Reward | Action-level Mean Radii |
| ----------- | --------------- | -------------- | ---------------------- | ----------------------- |
| DQN         | C51             | DQN Transition | Trajectory             | C51 Transition          |
| Freeway     | Proposed (RDP)  | 0.0            | 20.1                   | 21.3                    |
| 1.0         | 16.9            | 16.1           | 128.1                  | 32.6                    |
| 2.0         | 16.6            | 15.3           | 145.5                  | 58.7                    |
| 3.0         | 16.0            | 15.1           | 160.0                  | 102.4                   |
| COPA        | N/A             | 16.4           | 16.4                   | N/A                     |
| Breakout    | Proposed (RDP)  | 0.0            | 385.4                  | 389.3                   |
| 1.0         | 366.6           | 369.0          | 3.4                    | 3.2                     |
| 1.5         | 320.8           | 270.4          | 7.9                    | 7.6                     |
| 2.0         | 268.4           | 102.7          | 17.7                   | 16.9                    |
| COPA        | N/A             | 325.7          | 330.1                  | N/A                     |"
```

Interface Preview



Thank You