An Image-to-Text Dataset for Quantum Circuits

Adarsh Koshiya

OTH Amberg | a.koshiya@oth-aw.de | Winter Semester 2025

Abstract

Image-to-text models typically perform well on natural images but struggle with schematic and symbolic diagrams commonly found in scientific literature. This project investigates the **feasibility of compiling a domain-specific image-to-text dataset for quantum circuit diagrams** from arXiv publications. A fully automatic and reproducible pipeline is developed to extract quantum circuit images, align them with descriptive text, and enrich them with structured metadata. The resulting dataset consists of **250 quantum circuit images** sourced from recent quant-ph papers. This report describes the pipeline architecture, discusses challenges encountered during dataset compilation, analyzes dataset quality, and evaluates the feasibility of scaling the approach to larger collections of scientific documents.

Motivation

Image-to-text models are mainly trained on natural images and therefore struggle with scientific diagrams, which are **symbolic, structured, and sparse**. Unlike photographs, such figures convey meaning through conventions and spatial relationships rather than visual appearance. Quantum circuit diagrams are widely used in quantum computing research to describe algorithms and experimental procedures, yet they are **largely absent from existing image–text datasets**. Standard vision–language models generalize poorly to this domain. This project explores whether a **domain-specific image-to-text dataset for quantum circuits** can be created automatically and at scale. Using a reproducible pipeline, quantum circuit figures and their associated textual context are extracted from recent quant-ph arXiv papers. The resulting dataset of **250 images** serves as a basis for analyzing practical challenges and assessing the feasibility of large-scale dataset construction.

Dataset Requirements and Constraints

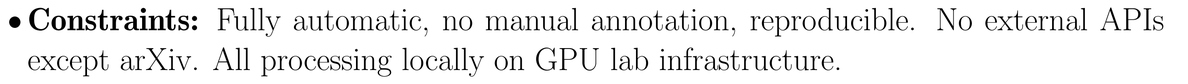
The dataset compilation is governed by **strict requirements** from the project specification.

- **Source:** arXiv papers in the quant-ph category. A fixed list in `paper_list_20.txt` is processed **strictly in order**, without skipping or reordering.
- **Termination:** Pipeline stops as soon as **250 valid** quantum circuit images have been collected.
- **Content:** Each image in PNG format depicting a **schematic quantum circuit**. Other figure types (plots, tables, experimental setups) are excluded.
- **Metadata per image:** arXiv identifier, page number (best-effort), figure number, quantum gates, quantum problem label, descriptions, character-level text positions.
- **Constraints:** Fully automatic, no manual annotation, reproducible. No external APIs except arXiv. All processing locally on GPU lab infrastructure.

Method – Pipeline Architecture (Fig. 1)

An end-to-end, **modular** processing pipeline; each stage performs a well-defined task and communicates through structured intermediate representations. **Stages (as in documentation):**

1. **Paper list** `paper_list_20.txt` →
2. **arXiv Source & PDF Download** (`arxiv_downloader.py`): LaTeX archive and PDF from arXiv.org; deterministic, cached, fixed delay.
3. **LaTeX Extraction** (`latex_extractor.py`): Safe extraction; all `.tex` files collected.
4. **Figure Identification** (`figure_finder.py`): Parse `figure` and `figure*` environments; caption, labels, image paths; only existing files retained.
5. **Quantum Circuit Filtering** (`quantum_circuit_filter.py`): Multi-stage heuristic (LaTeX context + filename; then visual wire detection).
6. **Text Alignment & Extraction** (`text_alignment.py`): Global text with character offsets; `\ref`/"Fig. N"; TF–IDF relevance; record text positions.
7. **Dataset Assembly** (`dataset_builder.py`): Export PNG, build metadata; `dataset_20.json`, `images_20/`, `paper_list_counts_20.csv`.



Relevant Methods – Core Pipeline Components

- **QuantumCircuitDatasetPipeline** (`pipeline.py`): Orchestrates pipeline, enforces order, terminates after 250 valid images.
- **ArxivSourceDownloader**, **ArxivPdfDownloader** (`arxiv_downloader.py`): Download LaTeX and PDF with caching and rate limiting.
- **LatexSourceExtractor** (`latex_extractor.py`): Safe extraction of archives.
- **FigureFinder** (`figure_finder.py`): Figure environments, captions, labels, image files.
- **QuantumCircuitFilter** (`quantum_circuit_filter.py`): `text_gate()` (LaTeX context + filename); `wire_ratio_ok()` (horizontal wire detection).
- **TextAligner** (`text_alignment.py`): Descriptive passages aligned with figures via TF–IDF relevance ranking.
- **DatasetBuilder** (`dataset_builder.py`): builds datasate_20.json

Challenges Encountered and Solutions

5.1 Identifying quantum circuits. *Challenge:* High false-positive rate (plots, tables, setups classified as circuits). *Solution:* Hybrid strategy in `quantum_circuit_filter.py`: LaTeX context (circuit commands, gate identifiers), filename heuristics, exclusion list; then OpenCV morphological ops for horizontal wires (`wire_ratio_ok()`). **5.2 Page numbers.** *Challenge:* PDF page numbers often unreliable. *Solution:* Best-effort in `dataset_builder.py`; missing values explicitly marked, no invalid placeholders. **5.3 Empty gate lists.** *Challenge:* Some figures had no extracted gates. *Solution:* Figures with no reliably extracted gates are **excluded**; gate search in captions, LaTeX blocks, descriptive text (`extract_metadata_v3()`). **5.4 Generic "Quantum Circuit" label.** *Challenge:* quantum_problem defaulted to uninformative label. *Solution:* Refined `infer_problem()`: pattern matching for named algorithms (e.g. Shor's, Grover, variational); neutral fallback when unspecified. **5.5 Redundant sub-circuits.** *Challenge:* Multiple small sub-circuits with limited standalone value. *Solution:* Prioritize figures with richer textual context and higher gate density. **5.6 Reproducibility.** *Challenge:* Non-determinism from file traversal, thresholds, text extraction. *Solution:* Sorted file traversal, fixed thresholds, no random sampling; identical inputs ⇒ identical outputs (`pipeline.py`).

Dataset Characteristics and Quality Analysis

Final dataset: **250 quantum circuit images**, each with structured metadata and aligned textual descriptions.

- **Composition:** PNG + bibliographic metadata, figure/page, quantum gates, quantum problem label, one or more descriptive passages. Multimodal structure supports joint reasoning over visual layouts, symbolic operations, and natural language.
- **Gate distribution:** Single- and multi-qubit gates (Hadamard, rotations, CNOT, CZ, etc.); empty gate lists excluded ⇒ every entry has explicit symbolic information.
- **Quantum problem:** Varying specificity; named algorithms when present; neutral fallback for unspecified.
- **Text alignment:** Descriptions from paragraphs referencing the figure; character-level offsets preserved; traceable to source.
- **Metadata:** Missing page numbers represented as absent; schema consistency; all images verified circuits, non-empty gate lists, descriptions traceable.
- **Suitability:** Aligned with image-to-text learning; emphasizes symbolic reasoning, structured layouts, domain terminology.

Feasibility and Conclusion

Feasibility (doc §7.1): Quantum circuit images and textual descriptions can be extracted automatically from arXiv without manual annotation. All stages operate deterministically and rely only on local information; construction is **reproducible and generalizable**. LaTeX-context analysis plus lightweight visual verification proved effective; heuristics follow common conventions and scale beyond the fixed paper list. **Scalability (doc §7.2):** Pipeline exhibits approximately **linear scaling** with number of papers. Resource-intensive steps (LaTeX parsing, image conversion, text alignment) are local and can be parallelized. No fundamental algorithmic limit to scaling; with more resources and minor heuristic adjustments, extension to thousands of papers or related domains is feasible. **Conclusion (doc §7.3):** Constructing a high-quality image-to-text dataset for quantum circuits from scientific literature is **feasible and practical** with fully automatic methods. The project delivers a curated dataset of 250 images suitable for training and evaluating multimodal models, and a **reusable methodological framework** for extracting structured visual–textual data from LaTeX-based publications. Similar approaches can be applied to other scientific domains with schematic diagrams, supporting scientific document understanding and multimodal learning. **References:** Radford et al. (CLIP); arXiv; spaCy; pylatexenc; Salton & Buckley (TF–IDF); OpenCV; Bird et al. (NLP with Python).