Paper2Poster: Towards Multimodal Poster Automation from Scientific Papers

Wei Pang*, Kevin Qinghong Lin*, Xiangru Jian*, Xi He, Philip Torr
University of Waterloo, University of Oxford, Vector Institute

Project Page: `https://paper2poster.github.io`

Motivation

- **Academic posters** condense a full scientific paper into a single visual page for rapid communication at conferences.
- **Challenge**: Unlike slide generation, poster automation must compress long, interleaved multi-modal documents (text, figures) — preserving both content and visual appeal.
- Existing methods for slides/templates fail due to:
  - Loss of logical structure or visual coherence.
  - Lack of optimal content compression.
  - Poor joint reasoning over text, vision, and layout.
- **Our aim**: Enable fully automated, visually coherent, editable (`.pptx`) academic poster generation.

Related Work

Text-rich Image Generation

- Poster generation and document design by diffusion models struggle with unreadable, blurry text outputs.

Complex Visual Layouts

- Slide/website generation use tool-based or multi-agent workflows for assembly, but not optimized for single-page, dense posters.

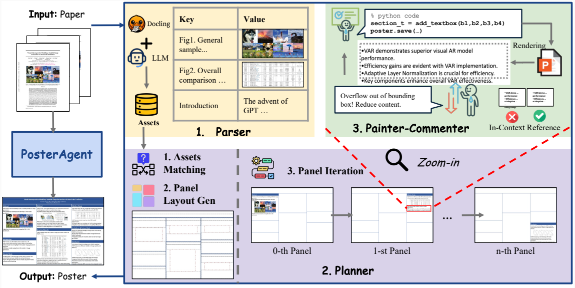
Vision-Language Agents

- Recent LLM "agents" (ReAct) show promise in automation, but lack robust visual feedback and fail on long-context/multi-modal tasks.
- **Our bench:** First to address poster-level multi-modal context compression.

Method: Overview of PosterAgent

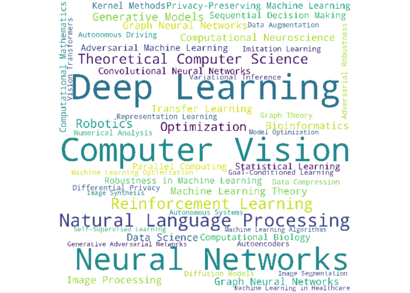
Formulate poster generation as a **multimodal context compression** problem.
Propose **PosterAgent**, a top-down, visual-in-the-loop multi-agent pipeline:

1. **Parser**: Extracts text/visual assets from the paper.
2. **Planner**: Aligns/arranges assets, generates hierarchical layout.
3. **Painter–Commenter**: Refines and checks panel rendering via visual feedback.



Method: Parser (Global Organization)

- **Purpose**: Converts raw PDF into a structured asset library for later stages.
- **Text assets**: Section-wise synopses by LLM + tools (`Marker`, `Docling`) — mimics how humans skim for key content.
- **Visual assets**: Extract significant figures/tables via captions.
- Reduces each document to a concise, machine-usable set of key elements.



Method: Planner (Local Organization)

- **Asset matching:** For each textual section, select most relevant figure/table via semantic alignment.
- **Layout Generation:** Use a binary-tree layout strategy to allocate panels by estimated text length and asset size.
- **Panel iteration:** Fill each poster section iteratively; produce concise bullet points.

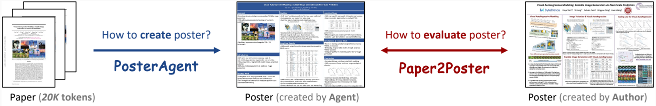
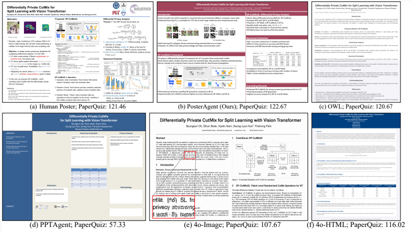


Figure 3: Poster structure and layout organization

Method: Painter–Commenter Loop (Local Refinement)

- **Painter:** Generates panel-level bullet points; renders visuals via executable pptx code.
- **Commenter (VLM as critic):** Examines if text fits, figures are clear, and layout is balanced.
- Uses *in-context references* for robust visual feedback; corrects ”overflow”, ”too blank”, or ”good to go”.
- Iterative loop continues until all panels are visually optimal.



(a) Human Poster; PaperQuiz: 121.46
(b) PosterAgent (Ours); PaperQuiz: 122.67
(c) OWL; PaperQuiz: 120.67
(d) PPTAgent; PaperQuiz: 57.33
(e) 4o-Image; PaperQuiz: 107.67
(f) 4o-HTML; PaperQuiz: 116.02

Figure 4: Panel refinement process (Painter & Commenter loop)

Experimental Method

- **Baselines**:
  - Oracle (Human posters, original paper as upper bounds)
  - End-to-end LLM (GPT-4o) - text/image rendering
  - Multi-agent (OWL-4o, PPTAgent)
  - **Our method: PosterAgent (4o/Qwen)**
- Four major evaluation axes:
  1. Visual Quality
  2. Textual Coherence
  3. Holistic Assessment (VLM-as-Judge)
  4. **PaperQuiz: Knowledge transfer efficacy**

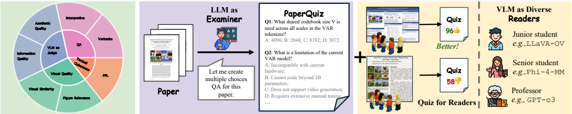


Figure 5: Evaluation framework: multiple axes and VLM–quiz assessment

Experimental Setup

- **Dataset:** `OurBench` (PosterSum) – 100 recent AI conference papers (ICML, NeurIPS, ICLR, 2022–2024).
- **Split:** Diverse sampling – 35 NeurIPS, 37 ICML, 28 ICLR papers.
- **Stats:** Avg. 22.6 pages, 12155 words, 22.6 figures per paper.
- **Environment:** All methods compared using author code and same evaluation protocol.
- **Open-source and Commercial models:** GPT-4o, Qwen-2.5-7B, etc.

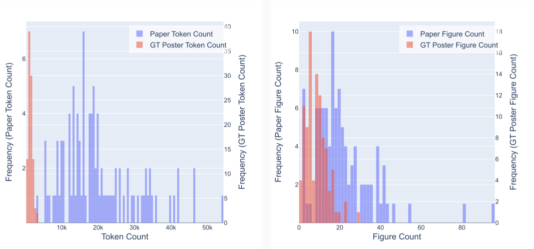
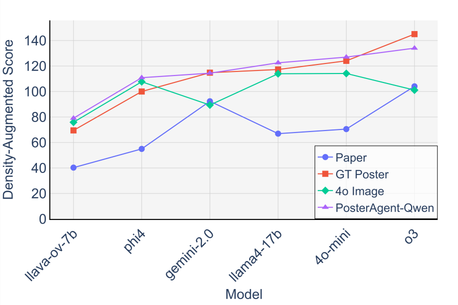


Figure 6: Input statistics: tokens and figures per paper/poster

Experimental Results

- **PosterAgent** outperforms baselines in figure relevance, visual similarity, and informativeness.
- GPT-4o "image mode": visually appealing, but often noisy/incoherent text.
- Human posters favour "visual semantics" – VLM evaluation shows engagement is the hardest aspect.



Ablation Study: Key Module Validation

Study Design

- Key modules examined:
  - Binary-tree layout vs. direct LLM layout
  - Painter–Commenter loop effectiveness
  - In-context reference examples (for visual feedback)
- **Results:** Disabling layout/commenter leads to overflows, blank gaps, or poor structure.

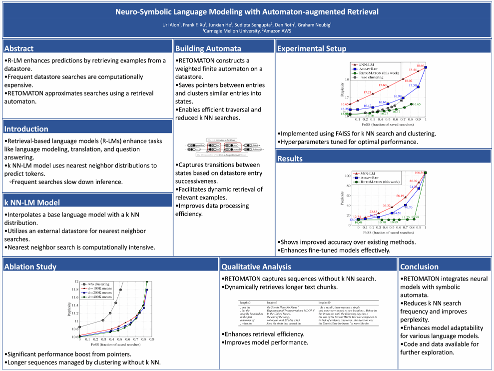


Figure 9: Ablation: Full pipeline preserves clarity; removing commenter/structure causes overflow and clutter.

Limitations

- **Sequential panel refinement** (Painter–Commenter) currently limits efficiency (avg. 4.5min per poster).
- Some categories (e.g., figures with extreme aspect ratios) remain challenging for automated layout.
- Visual engagement still lags behind best human-crafted posters.
- Open-source vision-language models less robust than GPT-4o for visual critique.

Future Work

- **Parallel panel refinement** to boost scalability and speed.
- Integration of external knowledge (e.g., community/openreview comments, logos, branding).
- Human-in-the-loop feedback: combine strong initial agent with interactive editor refinements.
- Enhanced model benchmarking & adaptive layout learning from expanded data.

Thank You!

This video is fully generated by Paper2Video.

Project Page: paper2poster.github.io