P2P: AUTOMATED PAPER-TO-POSTER GENERATION AND FINE-GRAINED BENCHMARK

Tao Sun[1,2], Enhao Pan[1], Zhengkai Yang[1], Kaixin Sui[1], Jiajun Shi[2], Xianfu Cheng[2], Tongliang Li[4], Wenhao Huang[1], Ge Zhang[1,2], Jian Yang[*], Zhoujun Li[3]*

[1] ByteDance, China [2] M-A-P [3] Shenzhen Intelligent Strong Technology Co.,Ltd.
[4] College of Computer Science, Beijing Information Science and Technology University
{buaast,jiaya,lizj}@buaa.edu.cn

ABSTRACT

Academic posters are vital for scholarly communication, yet their manual creation is time-consuming. However, automated academic poster generation faces significant challenges in preserving intricate scientific details and achieving effective visual-textual integration. Existing approaches often struggle with semantic richness, structural nuances, and lack standardized benchmarks for evaluating generated academic posters comprehensively. To address these limitations, we introduce **P2P**, the first flexible, LLM-based multi-agent framework that generates high-quality, HTML-rendered academic posters directly from research papers. P2P employs three specialized agents—for visual element processing, content generation, and final poster assembly—each integrated with dedicated checker modules to enable iterative refinement and ensure output quality. To foster advancements and rigorous evaluation in this domain, we argue that generated posters must be assessed from two complementary perspectives: **objective fidelity and subjective quality**. So we establish **P2PEVAL**, a comprehensive benchmark featuring 1738 checklist items and a dual evaluation methodology (Fine-Grained and Universal). Our Fine-Grained Evaluation uses human-annotated checklists to objectively measure the faithful preservation of verifiable content from the source paper. Concurrently, our Universal Evaluation captures subjective, holistic quality by training a model to align with human aesthetic preferences across key design principles. We evaluate a total of 35 models. To power these advancements, we also release **P2PINSTRUCT**, the first large-scale instruction dataset comprising over 30,000 high-quality examples tailored for the academic paper-to-poster generation task. Furthermore, our contributions aim to streamline research dissemination while offering a principled blueprint for evaluating complex, creative AI-generated artifacts. The code is on the https://github.com/multimodal-art-projection/P2P.

1 INTRODUCTION

Academic posters serve as a vital tool in scholarly communication, distilling complex research into visually accessible formats to foster knowledge dissemination. However, manually creating these posters is a time-consuming and skill-intensive process, demanding a blend of content refinement and design proficiency. Automating academic poster generation thus offers a significant opportunity to streamline research dissemination, but it is a complex challenge that goes far beyond simple summarization or image placement.

The difficulty is twofold. First, the generation task itself requires a system to simultaneously master multiple sub-tasks (Qiang et al., 2016): (1) content distillation to identify and synthesize the most critical information; (2) visual textual integration to understand and contextually place figures and tables; and (3) structural reorganization to transform a linear document into a coherent, two-dimensional layout. Second, and equally challenging, is *evaluation*. A poster's quality is difficult to assess directly, as it requires balancing two distinct axes: **objective fidelity** and **subjective quality**.

*Corresponding Authors

Objective fidelity, the faithful preservation of scientific claims and data, can be verified against the source paper. Conversely, subjective quality, which encompasses layout coherence and visual appeal, is inherently a matter of human aesthetic judgment. A robust evaluation must therefore address both (Que et al., 2024; Viswanathan et al., 2025).

Existing approaches to poster generation primarily rely on template-based or rule-driven methods (Xu & Wan, 2021), which often struggle to capture the semantic richness and structural nuances of academic documents (Qiang et al., 2019), typically decomposing the task into isolated subtasks like content extraction (Cheng et al., 2024b), panel attribute inference (Huang et al., 2022), and layout generation (Lin et al., 2024). Although recent advances in multimodal large language models (MLLMs) and large language models (LLMs) show promise in understanding document structures and visual-textual relationships (Jaisankar et al., 2024), their application to academic poster generation remains limited due to insufficient quality control mechanisms and the absence of standardized benchmarks for systematic evaluation.

To overcome these limitations and explore an underexplored research direction, we introduce a foundational ecosystem for automated academic poster generation, built on three synergistic contributions. The first is **P2P**, a novel and flexible multi-agent framework. It employs three specialized agents for visual processing, content generation, and layout orchestration, each integrated with a **checker reflection loop**. This iterative refinement mechanism mimics the human design process of drafting and revision, ensuring both scientific accuracy and structural integrity. The second contribution is **P2PINSTRUCT**, the first large-scale (30,000+ examples) instruction dataset for this underexplored domain, designed to train end-to-end models. The third is **P2PEVAL**, a comprehensive benchmark featuring 1738 checklist items, which implements our dual evaluation philosophy. **Fine-Grained Evaluation** measures objective fidelity by using LLM-as-a-Judge (Gu et al., 2024) to score generated posters against detailed, human-annotated checklists of verifiable content. While its **Universal Evaluation** captures subjective quality by training XGBoost (Chen & Guestrin, 2016) to emulate human annotators' holistic scores across 10 universal design and content criteria. We evaluate a total of 35 LLMs and MLLMs.

Our contributions are as follows:

- We propose **P2P**, the first multi-agent architecture for academic poster generation featuring an innovative checker-reflection mechanism for iterative refinement, providing a robust design pattern for complex document transformation tasks.
- We construct and release **P2PINSTRUCT**, the first large-scale (30K+) instruction dataset specifically designed to train models for the complete paper-to-poster generation workflow.
- We establish **P2PEVAL**, a new and comprehensive benchmark (1738 checklist items and 121 pairs) with a novel dual-evaluation framework that integrates human-annotated checklists and a predictive scoring model for robust, multifaceted analysis of poster quality.

2 METHODOLOGY

Our method, P2P, is a multi-agent framework that automates making posters. As a model-agnostic orchestration pipeline, it can be instantiated with various LLMs. It breaks down the complex job into small, manageable steps, with a specialized agent handling each one. A core innovation is the integration of a **checker-reflection mechanism** at each stage, enabling iterative refinement and ensuring high-quality output. The robustness of P2P also facilitates the creation of P2PINSTRUCT, a large-scale instruction dataset derived from the intermediate outputs of P2P.

2.1 P2P: MULTI-AGENT FOR PAPER-TO-POSTER GENERATION

As illustrated in Figure 1, the P2P workflow is orchestrated by three collaborative agents: the Figure Agent, the Section Agent, and the Orchestrate Agent. Each agent operates in conjunction with a dedicated checker module that triggers a reflection loop if its output fails to meet quality standards.

Problem Formulation. Given a research paper D in a digital format (e.g., PDF), the task of P2P generation is to automatically synthesize an academic poster P in a web-native format (HTML and CSS). We formalize this task as a sequential operation orchestrated by our specialized agents.

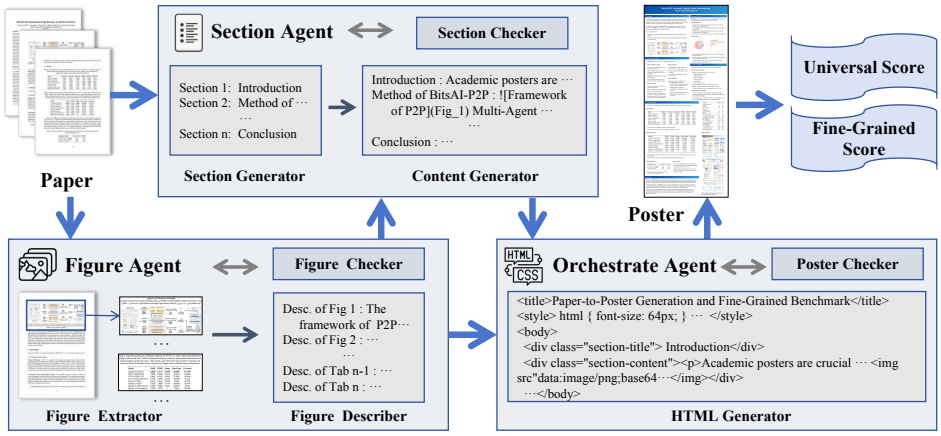


Figure 1: The multi-agent architecture of P2P: papers are processed by the Figure Agent for extraction and description of visual elements, the Section Agent for structural and content generation, and the Orchestrate Agent for poster assembly and HTML rendering. Each agent employs checker modules and reflection loops for iterative enhancement.

The overall process can be expressed as: $P = \mathcal{A}_{\text{Orch}}(\mathcal{A}_{\text{Sec}}(D, F), F)$, where $F = \mathcal{A}_{\text{Fig}}(D)$. Here, the process begins with the **Figure Agent**, $\mathcal{A}_{\text{Fig}}$, which processes the input paper D to extract and describe its visual elements, producing an intermediate set of figures and tables F. Subsequently, the **Section Agent**, $\mathcal{A}_{\text{Sec}}$, generates the poster content $P_{\text{poster\_text}} = \mathcal{A}_{\text{Sec}}(D, F)$ by summarizing the paper D while strategically referencing the visuals in F. Finally, the **Orchestrate Agent**, $\mathcal{A}_{\text{Orch}}$, assembles the textual content $P_{\text{poster\_text}}$ and visual elements F into the final designed poster P.

Figure Agent. The Figure Agent is responsible for processing all visual elements within the input research paper. Its *Figure Extractor* component employs DocLayout-YOLO (Zhao et al., 2024), a state-of-the-art document layout detection model, to extract figures and tables. Concurrently, the *Figure Describer* identifies corresponding captions via spatial relation analysis. These components collaborate to synthesize semantic visual units by combining each extracted graphical component with its associated caption, yielding a set of described visual elements $\mathcal{F}_d = \{(v_1, c_1, \text{desc}_1), \ldots, (v_n, c_n, \text{desc}_n)\}$. Here, v_i denotes the raw visual element (the cropped image file and its metadata), c_i its original caption, and desc_i a detailed description generated by an MLLM, M_{figure}. The *Figure Checker* then validates this output by: (1) preventing duplicate extractions, (2) verifying the capture of all significant visual elements, and (3) confirming accurate visual-caption pairings. To ensure reliable pairings, an initial confidence threshold is applied to detected elements; this threshold is incrementally lowered if discrepancies arise between the counts of identified figures and captions, an iterative process repeated until sufficient alignment is achieved.

Section Agent. The Section Agent focuses on generating the textual content of the poster. Initially, the *Section Generator* analyses the input paper (D) to dynamically infer a detailed structural schema (S) for the target poster. This schema, represented as a JSON object, delineates crucial sections (e.g., Introduction, Methods, Results) and their intended content focus. Subsequently, the *Content Generator* synthesizes semantically coherent textual content for the poster, $P_{\text{poster\_text}}$, by utilizing the structural schema S, the original input paper D, and the detailed descriptions and indices of visual elements $\mathcal{F}_d$ provided by the Figure Agent. This textual content generation can be formally described as: $P_{\text{poster\_text}} = \mathcal{M}_{\text{text}}(D, S, \mathcal{F}_d)$, where $\mathcal{M}_{\text{text}}$ is a LLM specialized in text generation. $\mathcal{M}_{\text{text}}$ employs prompts not only to generate text but also to strategically integrate Markdown-style references to figure indices from $\mathcal{F}_d$ at optimal textual positions, ensuring contextual relevance and visual-textual alignment. The *Section Checker* scrutinizes the generated $P_{\text{poster\_text}}$ for: (1) coherence and logical flow, (2) completeness in covering core contributions, (3) faithfulness to the original paper's findings, and (4) correct and relevant referencing of visual elements. If inadequacies are detected, a reflection loop initiates a revision of the section structure or content by the Section Agent.

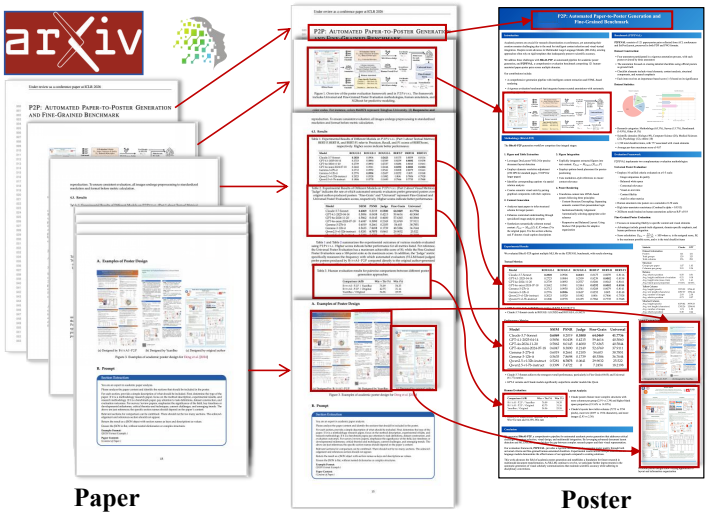


Figure 2: An example of the paper-to-poster transformation achieved by P2P, showing direct correspondences between elements in the input paper (left) and the generated academic poster (right).

Orchestrate Agent. The Orchestrate Agent integrates the visual and textual components into a cohesive and professionally formatted poster. The *HTML Generator* utilizes the Markdown-formatted text $P_{\text{poster\_text}}$ from the Section Agent and the actual visual elements (images/tables $\mathcal{F}_v$, where each figure is additionally provided with its width, height, and aspect ratio as supplementary information) extracted by the Figure Agent, to produce the poster in HTML and CSS. The Orchestrate Agent deliberately omits original captions from $\mathcal{F}_d$ in the final embedded visuals to improve visual clarity and maintain a concise academic presentation. The rendering process adheres to three principles: (1) Content-Structure Decoupling: Decouple semantics from presentation via modular CSS. (2) Institutional Identity Alignment: Customize color schemes to align with the logo of the institution or conference. (3) Responsive and Balanced Layout Generation: Use CSS flexbox for adaptive column structures and whitespace optimization. The *Poster Checker* evaluates the rendered poster for layout aesthetics and structural integrity, triggering iterative adjustments (via reflection) to resolve issues like unbalanced spacing or misaligned elements until the design meets professional standards.

Figure 2 illustrates the core transformation process facilitated by P2P. On the left, a multi-page academic research paper, sourced from repositories such as arXiv or conference proceedings like ICLR, serves as the input. On the right, the corresponding academic poster, generated by P2P, is displayed. The red arrows explicitly map key elements from the original paper, such as the title, specific figures, and sections, to their respective locations and representations in the final poster.

2.2 P2PINSTRUCT: A LARGE-SCALE INSTRUCTION DATASET

The P2PINSTRUCT dataset is derived from the P2P to support training of models for poster generation. Following P2P, we collect 30,460 high-quality instruction-response pairs spanning the complete poster generation workflow. For visual element processing, we prompt Claude to generate 16,848 figure-description pairs through the Figure Describer component, yielding descriptive texts averaging 192 tokens per visual element. For textual content generation, we collect 13,612 instruction-response pairs from the Section Generator, Content Generator, and HTML Generator components. These examples average over 3,300 tokens per response, demonstrating the complexity and richness of the generated content. A detailed account of its generation process, quality validation, and mitigation of potential biases is provided in Appendix D.

3 P2PEVAL: A FINE-GRAINED BENCHMARK FOR POSTER EVALUATION

As shown in Fig 3, we present a benchmark called P2PEVAL for evaluating academic posters.

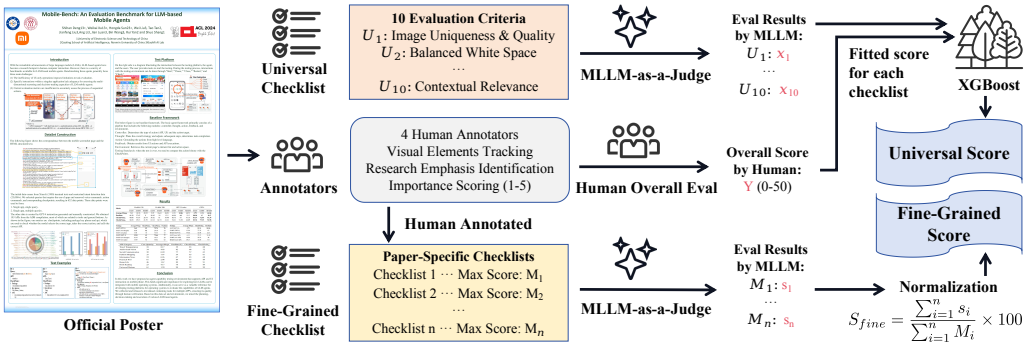


Figure 3: Overview of the poster evaluation framework used in P2PEVAL. P2PEVAL includes Universal and Fine-Grained Poster Evaluation, human annotators, and XGBoost for scoring.

3.1 BENCHMARK CONSTRUCTING

Checklist Design. The core of P2PEVAL lies in its detailed, human-curated and fine-grained checklists, designed to capture the essential elements of a poster. The annotation focuses on creating detailed checklists using official posters as ground truth. Our checklist design incorporates the following elements: **(1) Visual Elements**: Each figure or table present in the official poster constitutes an individual checklist item, evaluated based upon its presence and accurate representation. **(2) Content Analysis**: Each visual element is assessed regarding its textual consistency with the original poster and its visual prominence within the poster layout. **(3) Structural Components**: Annotators identify critical sections such as task definitions, experimental methodologies, and research conclusions within each poster panel. **(4) Research Emphasis**: Essential research findings, methodological details, and explicitly highlighted motivations (often noted by bold or prominent placement) form individual checklist items. **(5) Scoring System**: Each checklist item receives an importance-based score ranging from 1 to 5—minor details are rated as 1, core elements as 3, and critical components central to the paper as 5. The checklist items and their importance scores (1-5) are meticulously crafted by human experts with domain knowledge. Checklist format is listed in Appendix E.1.

Annotation Protocol. The annotation process was governed by a strict protocol to ensure quality and consistency. Each paper-poster pair is processed by four members of our annotation team. Three annotators independently create checklists, while a fourth, senior annotator serves as a verifier and integrator. This fourth annotator's role is to reconcile the three independent annotations by creating a superset (a union) of all identified checklist items. They also normalize the importance scores by calculating the average for each item across the initial annotations and rounding it to the nearest integer, ensuring a final, consensus-based score. Posters with major initial disagreements are excluded from the final test set to ensure data quality. For quality assurance in developing our checklists, we consult with researchers who have previously created posters (while maintaining anonymity). Their feedback confirms that our checklist design accurately reflects their poster creation priorities and decision-making process. Further details about our annotation team are provided in Appendix C.

Dataset Collection and Statistics. P2PEVAL consists of 121 paper-poster pairs collected from the ACL conference series (from 2022 to 2024) under CC4.0 license and from SciPostLayout (Wang et al., 2024a), which contains posters from F1000Research under the CC-BY license. For each pair, P2PEVAL preserves the original research paper in PDF format and the corresponding academic poster in both PDF and PNG formats. As shown in the Fig 4, P2PEVAL encompasses a broad range of research categories and disciplines. **The annotation process results in 1738 checklist items, with 775 associated with visual elements.** The scoring system yields an average per-item maximum score of 4.07. To ensure fairness and diversity, the benchmark includes papers from a wide range of scientific fields, as detailed in Appendix E.2. Our analysis in Appendix E.3 shows that model performance varies across topics, indicating the benchmark does not unfairly favor a single domain.

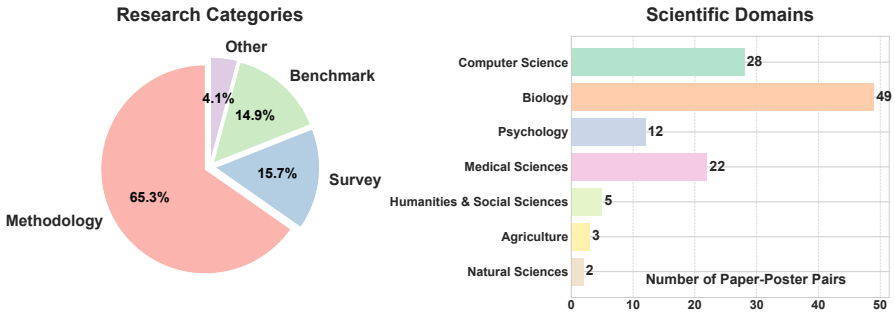


Figure 4: Distribution of P2PEVAL.

3.2 POSTER EVALUATION FRAMEWORK

Our evaluation pipeline consists of two complementary methodologies: Fine-Grained Poster Evaluation and Universal Poster Evaluation. We describe these two evaluation methods in detail below.

3.2.1 FINE-GRAINED POSTER EVALUATION

We design a Fine-Grained Poster Evaluation pipeline to measure a generated poster's fidelity to the key content selected by the original author. This metric focuses on the accurate representation of scientific information, not the imitation of a specific layout. The evaluation is a deterministic process that uses an LLM as an automated verification tool. The process consists of two steps:

Step 1: Automated Verification. For each item on the human-authored checklist (Section 3.1), we prompt an LLM (GPT-4o) to perform a strictly defined check and return relevant score: verify whether that specific, verifiable fact is present and accurately represented in the generated poster. The LLM's role is not subjective judgment. Each checklist item's maximum score is consensus-derived from multiple annotators, ranging from minor visual elements assigned a score of 1 to core research components scored at 5, reflecting their relative importance. This human-centred approach ensures that the scoring system inherently embodies human preferences and domain expertise.

Step 2: Deterministic Scoring. Then we programmatically aggregates these results from LLM. We formally define the final fine-grained evaluation score S_{fine} as $S_{fine} = \frac{\sum_{i=1}^{n} s_i}{\sum_{i=1}^{n} M_i} \times 100$, where S_{fine} is the normalized fine-grained evaluation score on a 0–100 scale, s_i is the assigned score for the i^{th} checklist item represented in the generated poster, M_i denotes the corresponding maximum possible score for that item, and n signifies the total number of checklist items. Consequently, the Fine-Grained Poster Evaluation score comprehensively assesses a generated poster's capability to faithfully preserve the original research's essential content and visual priorities. By emphasizing explicit fidelity to the original author's intended communication goals rather than generic quality alone, the approach enables clear comparative analyses across diverse poster generation methodologies.

3.2.2 UNIVERSAL POSTER EVALUATION

Universal Poster Evaluation employs a unified set of evaluation criteria, each evaluated independently on a discrete scale ranging from 0 to 5. These universal criteria (U_1 through U_{10}) include:

- U_1: Authorship and Title Accuracy
- U_2: Image Uniqueness and Quality
- U_3: Balanced White Space
- U_4: Contextual Relevance
- U_5: Optimal Visual-to-Text Ratio
- U_6: Dimension Appropriateness
- U_7: Visual Consistency
- U_8: Content Fidelity
- U_9: Information Flow Logic
- U_{10}: Self-Contained Explanation

While LLMs can score individual criteria, they often fail to replicate the complex, non-linear weightings humans apply when forming a holistic judgment. To address this, we developed a two-step

Table 1: Experimental results of different models on P2PEVAL. Higher scores indicate better.

Model	Size	ROUGE-1	ROUGE-2	ROUGE-L	BERT[1]	Judge[2]	FineGrain[3]	Universal[4]
Closed-Source Models								
Claude-3.7-Sonnet	🔒	0.2745	0.0830	0.2527	0.8109	0.5537	65.3962	**37.2474**
Claude-3.7-Sonnet[R]	🔒	0.2734	0.0848	0.2516	0.8111	**0.6281**	**65.8848**	35.5062
Claude-3.5-Sonnet	🔒	0.2367	0.0615	0.2185	0.8081	0.2810	47.7385	30.2544
GPT-4.1-2025-04-14	🔒	0.2459	0.0685	0.2281	0.8113	0.4793	60.2879	34.4700
GPT-4.1-mini-2025-04-14	🔒	0.2616	0.0741	0.2407	0.8125	0.3388	55.3493	31.0697
GPT-4.1-nano-2025-04-14	🔒	0.2169	0.0557	0.1990	0.8070	0.2066	41.3446	27.7149
GPT-4o-2024-11-20	🔒	0.2395	0.0668	0.2217	0.8114	0.4959	55.4380	34.3888
GPT-4o-mini-2024-07-18	🔒	0.2362	0.0732	0.2198	0.8167	0.2314	48.8879	30.8409
OpenAI-o1[R]	🔒	0.2385	0.0611	0.2200	0.8088	0.3103	56.8504	34.1659
Seed1.5-VL[R]	🔒	0.2160	0.0539	0.2026	0.8041	0.4050	62.4702	33.9840
Seed-Thinking-v1.5[RT]	🔒	0.2357	0.0701	0.2210	0.8113	0.4711	61.9632	34.6882
Seed-Thinking-v1.5-m[R]	🔒	0.2493	0.0767	0.2315	0.8116	0.3719	57.1457	33.2461
Doubao-1.5-vision-pro	🔒	0.2586	0.0849	0.2409	0.8089	0.0354	45.9282	14.0841
YuanBao[5]	🔒	-	-	-	-	0.0083	57.8677	31.5754
6B+ Models								
InternVL3	8B	0.1980	0.0618	0.1847	0.7994	0.0776	33.3900	22.2245
Qwen3[T]	8B	0.2563	0.0859	0.2373	0.8152	0.1835	45.0272	28.8107
Qwen3[RT]	8B	0.2231	0.0619	0.2082	0.8125	0.2545	53.6611	32.4912
Qwen2.5-VL	7B	0.1090	0.0414	0.1020	0.7645	0.0083	13.7417	13.0597
12B+ Models								
Gemma-3	12B	0.2411	0.0764	0.2250	0.8096	0.0940	46.7903	27.3686
InternVL3	14B	0.2437	0.0736	0.2253	0.8132	0.0756	45.5513	25.6062
27B+ Models								
Gemma-3	27B	0.2500	0.0794	0.2346	0.8133	0.2857	50.8931	28.7410
Gemma-3[T]	27B	0.2536	0.0853	0.2372	0.8132	0.2417	52.1716	28.5901
InternVL3	38B	0.2440	0.0756	0.2258	0.8143	0.2333	52.6634	29.5850
Qwen3[RT]	3/30B	0.2270	0.0637	0.2125	0.8120	0.2562	52.2125	31.1930
Qwen3[RT]	32B	0.2314	0.0659	0.2168	0.8090	0.1736	46.0383	28.9479
Qwen2.5-Coder[T]	32B	0.2666	0.0949	0.2487	0.8167	0.3884	55.9441	32.7935
72B+ Models								
Deepseek-R1[RT]	37/671B	0.1927	0.0461	0.1795	0.8015	0.5333	62.5013	33.9701
Deepseek-V3[T]	37/671B	0.2371	0.0739	0.2232	0.8124	0.5041	59.6805	33.6045
InternVL3	78B	0.2424	0.0789	0.2245	0.8152	0.2773	51.2962	28.9230
Qwen3[RT]	22/235B	0.2278	0.0625	0.2141	0.8077	0.3967	53.7927	31.4551
Qwen2.5-VL	72B	0.2577	0.0909	0.2400	0.8148	0.2833	55.7929	32.3105
Llama-4-Scout	17/109B	0.2806	**0.1208**	0.2625	**0.8172**	0.0413	35.7872	22.9738
Qwen3-P2P[T6]	8B	**0.2882**	0.0955	**0.2675**	0.8135	0.4587	57.6622	32.4996
Qwen2.5-VL-P2P[7]	7B	0.1939	0.0609	0.1797	0.7926	0.3140	37.3078	25.0337
InternVL3-P2P[8]	8B	0.2744	0.0883	0.2551	0.8117	0.3772	51.9670	31.6206

[R] Reasoning/Thinking Mode. [T] Because they are text-only LLMs, we use `Claude-3.7-Sonnet` as the provider of Figure Describer. [1] F1 scores of BERTScore. [2] The rate at which LLM-as-a-Judger prefer generated posters over original author-produced posters. [3] Scores of Fine-Grained Poster Evaluation. [4] Scores of Universal Poster Evaluation. [5] Posters generated by Tencent's AI application called YuanBao. [6-8] Our model, built on `Qwen3-8B`, `Qwen2.5-VL-8B` or `InternVL3-8B`, is fine-tuned using P2PINSTRUCT respectively.

hybrid approach. First, we use an LLM to generate scores for the 10 discrete criteria. Second, we train an XGBoost model to predict a final holistic score from these 10 features, using 1,701 human ratings as ground truth. This hybrid methodology was a principled design choice, as it combines the LLM's strength in stable feature extraction with XGBoost's proven ability to learn complex, non-linear human preference functions. Specifically, the XGBoost model undergoes training with 10-fold cross-validation and utilizes 200 trees. The resulting predictive model exhibits strong performance, achieving an R^2 of 0.92, thus validating the reliability and effectiveness of our Universal Poster Evaluation pipeline. More details about Universal Poster Evaluation can be found in Appendix F.

Table 2: Results of pairwise human preference evaluations.

Comparison (A vs. B)	Preferred or Tied (%)	Preferred (%)
P2P / YuanBao	83.05	54.35
P2P / Original	57.63	35.59
YuanBao / Original	20.34	12.40

Table 3: Performance comparison of P2P across different output format.

Output	FineGrain	Universal
HTML	**65.3962**	**37.2474**
SVG	52.7408	30.6648
LaTex	56.8756	25.2585

4 Experiments and Analysis

4.1 Experimental Setup

We conduct comprehensive experiments to evaluate P2P against several MLLMs on P2PEVAL. Specifically, we compare these different model series: GPT (Achiam et al., 2023), Claude (Anthropic, 2024), Doubao (Seed et al., 2025), Qwen (Bai et al., 2025), InternVL (Chen et al., 2024), Gemma (Team et al., 2025), Deepseek (Guo et al., 2025; Liu et al., 2024). Additionally, we include in our evaluation poster images generated by Tencent's AI application, YuanBao (`https://yuanbao.tencent.com/`), which directly produces academic posters in image format and Chinese. We also fine-tune our model `Qwen3-P2P`, `Qwen2-VL-P2P` and `InternVL3-P2P` using P2PINSTRUCT. The detail of training is shown in Appendix G. And input processing is in Appendix I.

4.2 Evaluation Metrics

Beyond the human-validated Universal Poster Evaluation (max 50) and Fine-Grained Poster Evaluation (max 100) using `GPT-4o` in P2PEVAL, we supplement analysis with objective metrics: (1) ROUGE (Lin, 2004), which measures n-gram overlap between generated and reference poster content, thus capturing lexical similarity. (2) BERTScore (Zhang et al., 2019), which leverages contextual embeddings to assess semantic similarity. During evaluation, all image links are removed from the text to ensure fair comparison of purely textual content. And the "Judge" metric reports how frequently VLLM-based automated evaluators prefer P2P's posters over original author-created versions. We also have aesthetic quality evaluation in Appendix L.

4.3 Results and Analysis

Main Results. Table 1 summarizes the experimental outcomes of various models evaluated using P2PEVAL. Our analysis reveals several key findings: **(1) Closed- vs. Open-source Models:** Closed-source models, notably `Claude-3.7-Sonnet`, achieve superior performance in qualitative assessments like the Universal and Fine-Grained evaluation. And leading open-source models such as `Deepseek-R1`(using `Claude` as the provider of Figure Describer), demonstrate strong competitiveness. **(2) Impact of Reasoning Capabilities:** Models employing reasoning or thinking modes such as `Claude-3.7-Sonnet` and `Qwen3` consistently show enhanced performance, especially in the Fine-Grained evaluation. This suggests that advanced reasoning aids in generating outputs that are more aligned with human preferences and detailed content requirements. **(3) Efficacy of P2PINSTRUCT:** Fine-tuning models on P2PINSTRUCT dataset yields substantial and statistically significant improvements (see Appendix H). The `Qwen3-P2P-8B` achieves the highest ROUGE scores across all evaluated models, significantly outperforming its base version and even leading closed-source models in these lexical metrics. It also demonstrates considerable gains in FineGrain and Universal scores over `Qwen3`. The larger gains on lexical metrics like ROUGE reflect the dataset's focus on teaching foundational text generation subtasks, while the consistent improvements on our flagship metrics confirm that these skills translate to higher-level fidelity and quality. These results underscore the value of P2PINSTRUCT. **(4) Supplemental Observations:** A divergence among evaluation criteria is also evident—excellence in lexical overlap (ROUGE) does not uniformly correlate with detailed fidelity (FineGrain), emphasizing the comprehensive nature of P2PEVAL. The strong performance of text-only models utilizing `Claude` for figure description points to the effectiveness of modular, hybrid approaches in this complex generation task.

Table 4: Ablation study results by `Claude-3.7-Sonnet`.

Mutli Agent	Figure Describer	Reflection	FineGrain	Universal
✓	✓	✓	**65.3962**	**37.2474**
✓	✓		64.4556	34.2229
✓		✓	63.7388	35.1107
✓			63.5806	33.1458
			60.7233	34.2554

Analysis of Human Preference Evaluation. To complement our P2PEVAL, we conduct pairwise human preference evaluations, the results of which are presented in Table 2. Participants compare posters generated by P2P using `Claude-3.7-Sonnet`, Tencent's YuanBao, and the original author-created posters. The "Preferred or Tied (%)" and "Strictly Preferred (%)" quantify the proportion of instances where method A is judged superior or equivalent to, and strictly superior to, method B, respectively. The results demonstrate a clear preference for P2P-generated posters over those from YuanBao. Notably, P2P also shows competitive performance against original posters, suggesting its capability to produce posters of superior quality in a significant number of cases.

Analysis of Output Format. Our investigation of different output formats reveals HTML as the optimal medium for academic posters using `Claude-3.7-Sonnet`. As documented in Table 3, HTML-based poster outputs consistently outperform SVG and LaTeX alternatives across both fine-grained and universal metrics. The inherent flexibility of HTML and CSS for layout structuring and content decoupling, coupled with the robust rendering capabilities of modern browsers, contributes to this performance. Furthermore, our experiments suggest that current LLMs exhibit greater proficiency in HTML code generation compared to equivalent SVG or LaTeX implementations, resulting in fewer rendering errors or structural inconsistencies in the final poster artifacts.

Ablation Study. The results of the ablation study of P2P in Table 4 demonstrate that the full system consistently outperforms reduced configurations. When reflection mechanisms (implemented through checker modules) are removed, we observe a moderate decline in universal metrics, suggesting these iterative feedback loops enhance overall poster quality and aesthetic coherence. Similarly, ablating the Figure Describer component, which transforms visual elements into textual descriptions, results in performance degradation. This indicates that directly feeding raw images to MLLMs for content integration can be less effective than providing them with semantically rich textual summaries. These descriptions appear to reduce the interpretative burden on the MLLMs and facilitate a more accurate contextualization of visual information within the poster. Removing all specialized components (resulting in a direct paper-to-poster pipeline without intermediate processing) leads to the greatest performance drop in fine-grained evaluation. This confirms our hypothesis that poster generation benefits significantly from modularized, specialized processing that mimics the distinct cognitive steps humans undertake when creating posters from research papers.

Analysis of Layout without Reflection. A comparative analysis of poster layouts generated by Claude and GPT models, summarized in Table 5, reveals distinct structural tendencies inherent in content segmentation and spatial organization for each model when operating without reflection mechanisms. Claude-generated posters typically exhibit a more fragmented structure, utilizing a greater number of columns. These layouts also demonstrate a tendency towards imbalanced spatial distribution, with taller content often concentrated towards the right and greater variability in column heights. This often results in a higher proportion of blank space, suggesting less efficient spatial utilization. In contrast, GPT-generated posters generally present more uniform and compact layouts. These findings suggest challenges in achieving consistent content allocation across the poster layout, a critical aspect for visual appeal and readability in academic posters.

5 LIMITATIONS AND FUTURE WORK

While P2P demonstrates significant advances, we acknowledge several limitations that offer avenues for future work.

Table 5: Comparison of layout statistics in posters generated by `Claude-3.7-Sonnet` and `GPT-4o-2024-11-20`.

Layout Statistic	Claude	GPT	Layout Statistic	Claude	GPT
General			**Tallest Column**		
Total columns	376	293	Height (px)	7272.22	5794.42
Balance			Text length (char)	2057.37	1554.44
Relative position[1]	0.55	0.51	Number of images	2.93	2.30
Height coefficient of variation[2]	0.21	0.18	**Shortest Column**		
Height ratio (max/min)[3]	1.73	1.61	Height (px)	4379.82	3979.53
Blank space proportion[4]	19.16%	14.92%	Text length (char)	1392.26	1296.84
			Number of images	1.74	1.59

[1] Index of relative column positions within posters; values closer to 0.5 indicate more centered, balanced layouts.
[2] Measure of height consistency across columns; lower values indicate more uniform column heights.
[3] Ratio between tallest and shortest columns; values closer to 1 indicate more even column heights.
[4] Percentage of total poster area occupied by blank space.

- **Output Format:** Our framework currently optimizes for HTML rendering due to its flexibility and the strong capabilities of modern LLMs in generating HTML. As shown in our ablation study (Table 3), generating other formats like LaTeX incurs a performance cost. This presents a practical constraint for users in academic environments where LaTeX or PowerPoint are prevalent. Future work could explore improved code generation for these formats or robust HTML-to-PDF/PPT converters.
- **Computational Cost:** The multi-agent framework with its iterative reflection loops is computationally more intensive than a single-pass generation pipeline. While we demonstrate that the cost is affordable (Appendix K), the latency may be a consideration for real-time applications. The number of reflection iterations serves as a tunable parameter to balance quality and cost.
- **Reasoning Boundaries:** As discussed in Appendix B.2, the checker-reflection mechanism is highly effective at correcting structural and syntactic errors. However, its ability to resolve deep semantic or complex compositional reasoning errors is ultimately bounded by the capabilities of the underlying LLM. For instance, correctly arranging highly intricate, domain-specific multi-panel figures remains a challenge if the base model lacks the necessary visual-spatial reasoning. Our escalation strategy mitigates this but does not eliminate this fundamental boundary.

6 Conclusion

We introduced P2P, a multi-agent framework that effectively transforms research papers into visually coherent and informationally faithful posters. By modularizing the task and integrating checker-reflection loops, our system demonstrates strong performance, producing outputs that often rival human-created examples. We supported this work by creating P2PInstruct, a large-scale instruction dataset, and P2PEval, a comprehensive evaluation benchmark. Crucially, this work argues that evaluating complex, creative artifacts like academic posters requires a dual approach that decouples **objective, verifiable fidelity** from **subjective, holistic quality**. Our P2PEval benchmark embodies this principle, combining fine-grained checklists for content preservation with a universal scoring model trained to emulate human aesthetic judgment. This dual-perspective evaluation, alongside the P2P framework and P2PInstruct dataset, not only provides a robust solution for automated poster generation but also offers a principled methodological blueprint for future research in automated scientific communication and other complex creative AI domains. This foundation promises to enhance research accessibility and dissemination efficiency for the entire academic community.

Ethics Statement

Our work adheres to the principles outlined in the ICLR Code of Ethics. The primary goal of this research is to contribute to societal and human well-being by developing tools that streamline

and accelerate scholarly communication, making research more accessible and saving researchers valuable time.

In line with the principle of upholding high standards of scientific excellence and avoiding harm, we acknowledge the potential risk of automated systems misrepresenting or "hallucinating" scientific content. To mitigate this, our core contribution includes the P2PEVAL benchmark, which is specifically designed to evaluate the factual fidelity of generated posters. Our dual-evaluation framework, combining Fine-Grained checklist-based verification for objective accuracy and a Universal score for holistic quality, represents a direct effort to ensure the trustworthiness and reliability of the output.

To respect the work required to produce new ideas and artefacts, all papers and posters used to construct our P2PEVAL benchmark were sourced from publicly available repositories (ACL, SciPostLayout) under permissive licenses (CC-BY, CC4.0), and we give full credit to the original authors.

In developing our benchmark and datasets, we engaged human annotators. We are committed to fairness and respect for all individuals involved. As detailed in Appendix C, our annotation team consisted of qualified domain experts who were compensated at a rate exceeding the local statutory minimum wage. We established a rigorous annotation protocol, including training and verification steps, to ensure data quality and consistency, while also protecting the privacy and well-being of the participants.

We acknowledge the environmental cost associated with training large models and have reported our resource consumption in Appendix J. We believe the potential benefits of our work in enhancing the efficiency of scientific dissemination for the broader research community justify this computational cost. Finally, we discuss the limitations of our system in Section Q to provide a transparent account of its current capabilities.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we have made comprehensive efforts to provide all necessary components. All code for the P2P multi-agent framework, the P2PEVAL evaluation pipeline, and the fine-tuning process will be made publicly available in the anonymous GitHub repository linked in the paper.

The datasets created for this research will also be released. This includes P2PINSTRUCT, our large-scale instruction-tuning dataset of over 30,000 examples, and P2PEVAL, which comprises 121 paper-poster pairs and 1738 human-annotated checklist items in a structured YAML format.

The paper provides detailed descriptions of our methodology. The architecture of the P2P framework is detailed in Section 2 and illustrated in Figure 1. The construction of the P2PEVAL benchmark and the full evaluation protocol are described in Section 3. For implementation-specific details, Appendix G documents the fine-tuning hyperparameters and setup using the LLaMA-Factory framework. Appendix I describes the input processing pipeline for both text and figures. Furthermore, we include the exact prompts used for each agent and evaluation step in the Appendix to allow for full replication of our generation and evaluation logic.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024. Available at: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Andre R Brunoni, Leandro Valiengo, Alessandra Baccaro, Tamires A Zanão, Janaina F de Oliveira, Alessandra Goulart, Paulo S Boggio, Paulo A Lotufo, Isabela M Benseñor, and Felipe Fregni. The sertraline vs electrical current therapy for treating depression clinical study: results from a factorial, randomized, controlled trial. *JAMA psychiatry*, 70(4):383–391, 2013.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*, 2024.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23550–23558, 2025.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Xianfu Cheng, Hang Zhang, Jian Yang, Xiang Li, Weixiao Zhou, Kui Wu, Fei Liu, Wei Zhang, Tao Sun, Tongliang Li, et al. Xformparser: A simple and effective multimodal multilingual semi-structured form parser. *arXiv preprint arXiv:2405.17336*, 2024a.

Xianfu Cheng, Weixiao Zhou, Xiang Li, Jian Yang, Hang Zhang, Tao Sun, Wei Zhang, Yuying Mai, Tongliang Li, Xiaoming Chen, et al. Sviptr: Fast and efficient scene text recognition with vision permutable extractor. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 365–373, 2024b.

Z Chu et al. A survey of chain of thought reasoning: advances, frontiers and future. comput. res. repository (corr)(2023).

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*, pp. 288–301. Springer, 2006.

Shihan Deng, Weikai Xu, Hongda Sun, Wei Liu, Tao Tan, Jianfeng Liu, Ang Li, Jian Luan, Bin Wang, Rui Yan, et al. Mobile-bench: An evaluation benchmark for llm-based mobile agents. *arXiv preprint arXiv:2407.00993*, 2024.

Yubin Deng, Chen Change Loy, and Xiaoou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017.

Zijian Ding, Qinshi Zhang, Mohan Chi, and Ziyi Wang. Frontend diffusion: Empowering self-representation of junior researchers and designers through agentic workflows. *arXiv preprint arXiv:2502.03788*, 2025.

Nicholas Franzese, Adam Groce, TM Murali, and Anna Ritz. Hypergraph-based connectivity measures for signaling pathway topologies. *PLoS computational biology*, 15(10):e1007384, 2019.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Yi Gui, Zhen Li, Yao Wan, Yemin Shi, Hongyu Zhang, Bohua Chen, Yi Su, Dongping Chen, Siyuan Wu, Xing Zhou, et al. Webcode2m: A real-world dataset for code generation from webpage designs. In *Proceedings of the ACM on Web Conference 2025*, pp. 1834–1845, 2025a.

Yi Gui, Yao Wan, Zhen Li, Zhongyi Zhang, Dongping Chen, Hongyu Zhang, Yi Su, Bohua Chen, Xing Zhou, Wenbin Jiang, et al. Uicopilot: Automating ui synthesis via hierarchical code generation from webpage designs. In *Proceedings of the ACM on Web Conference 2025*, pp. 1846–1855, 2025b.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*, 2024.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4083–4091, 2022.

Md Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. Mapcoder: Multi-agent code generation for competitive problem solving. *arXiv preprint arXiv:2405.11403*, 2024.

Vijay Jaisankar, Sambaran Bandyopadhyay, Kalp Vyas, Varre Chaitanya, and Shwetha Somasundaram. Postdoc: Generating poster from a long multimodal document using deep submodular optimization. *arXiv preprint arXiv:2405.20213*, 2024.

Pia Jeppesen, Rasmus Trap Wolf, Sabrina M Nielsen, Robin Christensen, Kerstin Jessica Plessen, Niels Bilenberg, Per Hove Thomsen, Mikael Thastum, Simon-Peter Neumer, Louise Berg Puggaard, et al. Effectiveness of transdiagnostic cognitive-behavioral psychotherapy compared with management as usual for youth with common mental health problems: a randomized clinical trial. *JAMA psychiatry*, 78(3):250–260, 2021.

Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029*, 2024.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*, 2023a.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*, 2020.

Ryan Li, Yanzhe Zhang, and Diyi Yang. Sketch2code: Evaluating vision-language models for interactive web design prototyping. *arXiv preprint arXiv:2410.16232*, 2024.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models, 2023b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023c.

Shanchao Liang, Nan Jiang, Shangshu Qian, and Lin Tan. Waffle: Multi-modal model for automated front-end development. *arXiv preprint arXiv:2410.18362*, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Jiawei Lin, Jiaqi Guo, Shizhao Sun, Zijiang Yang, Jian-Guang Lou, and Dongmei Zhang. Layoutprompter: Awaken the design ability of large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2(3):9, 2023a.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *CoRR*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023c.

Jiaheng Liu, Zehao Ni, Haoran Que, Tao Sun, Noah Wang, Jian Yang, Hongcheng Guo, Zhongyuan Peng, Ge Zhang, Jiayi Tian, et al. Roleagent: Building, interacting, and benchmarking high-quality role-playing agents from scripts. *Advances in Neural Information Processing Systems*, 37: 49403–49428, 2025.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023d.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034, 2015.

Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *2011 international conference on computer vision*, pp. 1784–1791. IEEE, 2011.

Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 404–411, 2004.

Kanya Paramita and Masayu Leylia Khodra. Tailored summary for automatic poster generator. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pp. 1–6. IEEE, 2016.

Yu-Ting Qiang, Yan-Wei Fu, Xiao Yu, Yan-Wen Guo, Zhi-Hua Zhou, and Leonid Sigal. Learning to generate posters of scientific papers by probabilistic graphical models. *Journal of Computer Science and Technology*, 34:155–169, 2019.

Yuting Qiang, Yanwei Fu, Yanwen Guo, Zhi-Hua Zhou, and Leonid Sigal. Learning to generate posters of scientific papers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, et al. Hellobench: Evaluating long text generation capabilities of large language models. *arXiv preprint arXiv:2409.16191*, 2024.

Rohit Saxena, Pasquale Minervini, and Frank Keller. Postersum: A multimodal benchmark for scientific poster summarization. *arXiv preprint arXiv:2502.17540*, 2025.

Giovanni Scala, Ornella Affinito, Gennaro Miele, Antonella Monticelli, and Sergio Cocozza. Evidence for evolutionary and nonevolutionary forces shaping the distribution of human genetic variants near transcription start sites. *PloS one*, 9(12):e114432, 2014.

Michael Scherer, Almut Nebel, Andre Franke, Jörn Walter, Thomas Lengauer, Christoph Bock, Fabian Müller, and Markus List. Quantitative comparison of within-sample heterogeneity scores for dna methylation data. *Nucleic acids research*, 48(8):e46–e46, 2020.

ByteDance Seed, Yufeng Yuan, Yu Yue, Mingxuan Wang, Xiaochen Zuo, Jiaze Chen, Lin Yan, Wenyuan Xu, Chi Zhang, Xin Liu, et al. Seed-thinking-v1. 5: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*, 2025.

Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: Benchmarking multimodal code generation for automated front-end engineering. *arXiv preprint arXiv:2403.03163*, 2024.

Tao Sun, Yang Yang, Xianfu Cheng, Jian Yang, Yintong Huo, Zhuoren Ye, Rubing Yang, Xiangyuan Guan, Wei Zhang, Hangyuan Ji, et al. Repofixeval: A repository-level program repair benchmark from issue discovering to bug fixing.

Tao Sun, Linzheng Chai, Jian Yang, Yuwei Yin, Hongcheng Guo, Jiaheng Liu, Bing Wang, Liqun Yang, and Zhoujun Li. Unicoder: Scaling code large language model via universal code. *arXiv preprint arXiv:2406.16441*, 2024.

Tao Sun, Jian Xu, Yuanpeng Li, Zhao Yan, Ge Zhang, Lintao Xie, Lu Geng, Zheng Wang, Yueyan Chen, Qin Lin, et al. Bitsai-cr: Automated code review via llm in practice. *arXiv preprint arXiv:2501.15134*, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. Waterbench: Towards holistic evaluation of watermarks for large language models. *arXiv preprint arXiv:2311.07138*, 2023.

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. *arXiv preprint arXiv:2507.18624*, 2025.

Yuxuan Wan, Yi Dong, Jingyu Xiao, Yintong Huo, Wenxuan Wang, and Michael R Lyu. Mrweb: An exploration of generating multi-page resource-aware web code from ui designs. *arXiv preprint arXiv:2412.15310*, 2024a.

Yuxuan Wan, Chaozheng Wang, Yi Dong, Wenxuan Wang, Shuqing Li, Yintong Huo, and Michael R Lyu. Automatically generating ui code from screenshot: A divide-and-conquer-based approach. *arXiv preprint arXiv:2406.16386*, 2024b.

Hao Wang, Shohei Tanaka, and Yoshitaka Ushiku. Scipostlayout: A dataset for layout analysis and layout generation of scientific posters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8136–8141, 2024a.

Zheng Wang, Bingzheng Gan, and Wei Shi. Multimodal query suggestion with multi-agent reinforcement learning from human feedback. In *Proceedings of the ACM Web Conference 2024*, pp. 1374–1385, 2024b.

Ronald L Wasserstein and Nicole A Lazar. The asa statement on p-values: context, process, and purpose, 2016.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Jingyu Xiao, Yuxuan Wan, Yintong Huo, Zhiyao Xu, and Michael R Lyu. Interaction2code: How far are we from automatic interactive webpage generation? *arXiv preprint arXiv:2411.03292*, 2024a.

Shuhong Xiao, Yunnong Chen, Jiazhi Li, Liuqing Chen, Lingyun Sun, and Tingting Zhou. Prototype2code: End-to-end front-end code generation from ui design prototypes. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 88353, pp. V02BT02A038. American Society of Mechanical Engineers, 2024b.

Sheng Xu and Xiaojun Wan. Neural content extraction for poster generation of scientific papers. *arXiv preprint arXiv:2112.08550*, 2021.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, et al. Multilingual machine translation systems from microsoft for wmt21 shared task. *arXiv preprint arXiv:2111.02086*, 2021.

Jian Yang, Jiaxi Yang, Ke Jin, Yibo Miao, Lei Zhang, Liqun Yang, Zeyu Cui, Yichang Zhang, Binyuan Hui, and Junyang Lin. Evaluating and aligning codellms on human preference. *arXiv preprint arXiv:2412.05210*, 2024a.

Jian Yang, Jiajun Zhang, Jiaxi Yang, Ke Jin, Lei Zhang, Qiyao Peng, Ken Deng, Yibo Miao, Tianyu Liu, Zeyu Cui, et al. Execrepobench: Multi-level executable code completion evaluation. *arXiv preprint arXiv:2412.11990*, 2024b.

Jian Yang, Wei Zhang, Jiaxi Yang, Yibo Miao, Shanghaoran Quan, Zhenhe Wu, Qiyao Peng, Liqun Yang, Tianyu Liu, Zeyu Cui, et al. Multi-agent collaboration for multilingual code instruction tuning. *arXiv preprint arXiv:2502.07487*, 2025.

Kaichun Yao, Lan Zeng, Chuan Qin, Hengshu Zhu, Yanjun Wu, and Libo Zhang. Scipg: A new benchmark and approach for layout-aware scientific poster generation.

Sukmin Yun, Haokun Lin, Rusiru Thushara, Mohammad Qazim Bhat, Yongxin Wang, Zutao Jiang, Mingkai Deng, Jinhong Wang, Tianhua Tao, Junbo Li, et al. Web2code: A large-scale webpage-to-code dataset and evaluation framework for multimodal llms. *arXiv preprint arXiv:2406.20098*, 2024.

Junyi Zhang, Jiaqi Guo, Shizhao Sun, Jian-Guang Lou, and Dongmei Zhang. Layoutdiffusion: Improving graphic layout generation by discrete diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7226–7236, 2023.

Tianhao Zhang, Fu Peiguo, Jie Liu, Yihe Zhang, and Xingmei Chen. Nldesign: A ui design tool for natural language interfaces. In *Proceedings of the ACM Turing Award Celebration Conference-China 2024*, pp. 153–158, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *arXiv preprint arXiv:2410.12628*, 2024.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pp. 1015–1022. IEEE, 2019.

Ting Zhou, Yanjie Zhao, Xinyi Hou, Xiaoyu Sun, Kai Chen, and Haoyu Wang. Bridging design and development with automated declarative ui code generation. *arXiv preprint arXiv:2409.11667*, 2024.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*, 2023.

Appendix

CONTENTS

A Related Work **18**

B More Details of P2P and Implementation **18**

B.1 Agent Implementation and Model Agnosticism 18

B.2 Checker-Reflection Mechanisms in Detail 18

C Annotator Team **19**

D More Details of P2PINSTRUCT **20**

E More Details of P2PEVAL **20**

E.1 Checklist Format 20

E.2 Benchmark Statistics 21

E.3 Analysis of Performance Breakdown by Topic 22

F More Details of Universal Poster Evaluation **22**

F.1 Details of the Human Rating Protocol 22

F.2 Details of Evaluation Criteria 23

F.3 Other Methods Performance 24

G Training Details **24**

H Statistical Significance of Results **25**

I Input Processing **25**

J Resource Consumption **25**

K Computational Cost and Latency **25**

L Aesthetic Quality Evaluation **26**

M The Features of Fine-Grained Poster Evaluation **27**

N The Features of HTML Format **28**

O Examples of Poster Generation **28**

P Prompt **30**

Q Limitations **32**

A Related Work

Poster Generation. Academic poster generation involves creating a poster that summarizes the key information from an academic paper. Paramita et al. Paramita & Khodra (2016) develop a model that extracts essential sentences into templates to generate text-based posters. Qiang et al. Qiang et al. (2019) propose a more comprehensive method, decomposing poster generation into three subtasks: content extraction Mihalcea & Tarau (2004); Xu & Wan (2021); Cheng et al. (2024b;a), panel attribute inference Zhong et al. (2019); Li et al. (2020); Huang et al. (2022), and panel layout generation Lin et al. (2024); Zhang et al. (2023). Postdoc Jaisankar et al. (2024) utilizes MLLMs to generate template-based posters but cannot produce flexible layouts with more dynamic integration of figures and text. Additionally, existing academic poster datasets Yao et al.; Xu & Wan (2021); Qiang et al. (2016); Wang et al. (2024a); Saxena et al. (2025) lack fine-grained evaluation metrics necessary for comprehensive quality assessment.

HTML Code Generation and Multi-Agent. Recent research in automated front-end development focuses on generating HTML from diverse inputs such as screenshots, prototypes and natural language. This has spurred the creation of benchmarks like Design2Code (Si et al., 2024; Yang et al., 2025), Websight (Laurençon et al., 2024), WebCode2M (Gui et al., 2025a), and Web2Code (Yun et al., 2024). Code generation methodologies vary, including direct translation, structured approaches such as DCGen's (Wan et al., 2024b) divide-and-conquer strategy and UICopilot's (Gui et al., 2025b) hierarchical generation. Applications target mobile UIs (Xiao et al., 2024b; Zhou et al., 2024), multi-page websites (Wan et al., 2024a), and web design (Xiao et al., 2024a; Li et al., 2024; Zhang et al., 2024), with model fine-tuning (Liang et al., 2024) enhancing performance. Multi-agent systems are increasingly adopted for complex tasksHan et al. (2024); Liu et al. (2025); for instance, agentic workflows can convert designs to code (Ding et al., 2025; Islam et al., 2024), and some systems employ distinct agents for sub-tasks with iterative human feedback(Wang et al., 2024b).

LLM as a Judge. The use of LLMs as evaluators, termed "LLM-as-a-Judge," is well-studied and has demonstrated high consistency with human judgment, with early work focusing on LLMs evaluating other LLMs, as seen in JudgeLM Zhu et al. (2023); Yang et al. (2024a). Subsequent research introduced systems like AUTO-J Li et al. (2023a), leveraging pairwise and single-response evaluations to achieve strong agreement with human assessments Bai et al. (2023); Li et al. (2023b); Yang et al. (2024b); Li et al. (2023c); Sun et al.; Yang et al. (2021); Sun et al. (2025). With the rise of MLLMs, their potential as evaluators in multimodal tasks is being explored, as traditional metrics often fail to capture the nuances of complex multimodal outputs Antol et al. (2015); Liu et al. (2023a;d;c;b). To enhance LLM evaluation capabilities, techniques such as Chain-of-Thought Wei et al. (2021); Chu et al.; Chai et al. (2025); Sun et al. (2024) and Training-free instruction following Brown et al. (2020); Wei et al. (2021) have been proposed, addressing the need for more robust evaluators in both unimodal and multimodal contexts.

B More Details of P2P and Implementation

B.1 Agent Implementation and Model Agnosticism

As stated in the main paper, P2P is a model-agnostic orchestration framework. For any given experiment in Table 1, all core generative components (the Figure Describer, Section Generator, Content Generator, and HTML Generator) are powered by the same underlying LLM being evaluated. This experimental design ensures a fair, end-to-end benchmark of each model's capabilities on the complex paper-to-poster task. The only fixed components (shown in Appendix I) are programmatic utilities like PyMuPDF for text extraction and DocLayout-YOLO for initial figure detection.

B.2 Checker-Reflection Mechanisms in Detail

The checker-reflection mechanism is central to P2P's quality assurance. As shown in Fig 5, here we detail each checker's implementation, failure triggers, and the nature of the reflection loop.

Figure Checker. This module is rule-based and programmatic for efficiency. Its reflection loop is triggered if the count of detected figures mismatches the count of detected captions (paired using

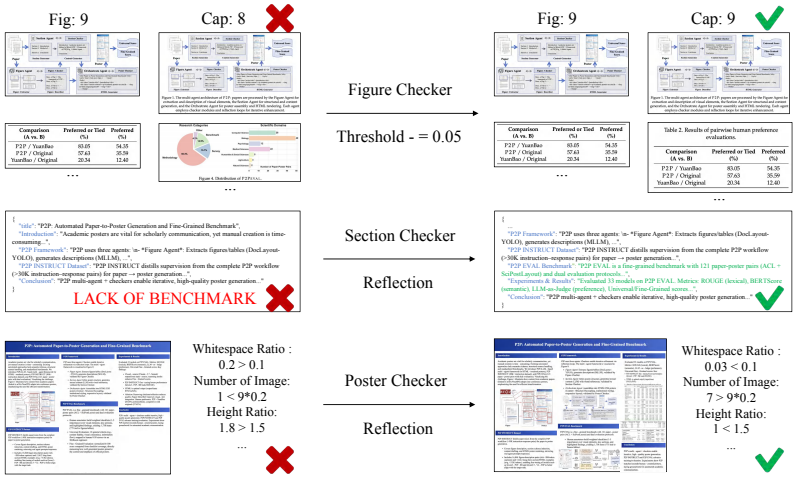


Figure 5: Illustration of the Checker-Reflection Mechanisms.

Manhattan distance). The initial confidence threshold for the DocLayout-YOLO detector is set to a high 0.85 to minimize false positives. If a mismatch occurs, the threshold is iteratively decreased by 0.05. This allows the Figure Extractor to capture elements that may have been initially assigned a lower confidence score (e.g., tables that visually resemble plain text), ensuring all significant visuals are reliably captured.

Section Checker. This module uses an LLM-as-a-Judge to assess semantic qualities. After content generation, the checker prompts the LLM to evaluate the text against the source paper on coherence, completeness, faithfulness, and correct referencing. The LLM returns a binary decision ("OK" or "Problem"). If a problem is detected, it also provides actionable feedback for revision (e.g., "The conclusion section omits the key limitation mentioned in the paper."). The Content Generator then re-runs with this new feedback, ensuring iterative refinement.

Poster Checker. This module is also rule-based. It programmatically parses the generated HTML to check for structural integrity and layout aesthetics. For example, a reflection loop is triggered if the **proportion of blank space exceeds 10%** or if column heights are severely imbalanced. The checker then feeds these failed metrics (e.g., ‘"whitespace_ratio": 0.15‘) and previous HTML code back to the HTML Generator, instructing it to regenerate the layout. On average, this loop runs for 2.32 iterations for GPT-4.1 and 3.47 for Qwen-2.5-VL-7B.

Failure Boundaries and Escalation. The reflection paradigm is highly effective at correcting structural and syntactic errors. However, it can struggle with deep semantic or complex compositional reasoning errors. For instance, if a paper contains numerous, intricate multi-panel figures that require a highly specific arrangement to be coherent, the LLM might struggle to generate a logical HTML structure. The Poster Checker may detect the resulting layout imbalance (e.g., high whitespace), but if the root cause is the LLM’s inability to reason about the complex spatial relationship between sub-figures, reflection alone may not solve it. The LLM, lacking a better strategy, might simply reshuffle the same flawed components. To mitigate this, we employ an **escalation strategy**: if a checker fails after 5 retries, the reflection loop can escalate to an earlier agent, prompting a more fundamental revision. For example, a persistent layout failure in the Orchestrate Agent can trigger the Section Agent to regenerate its content entirely, attempting to provide a simpler or more structured input that the Orchestrate Agent can handle. This provides a multi-level recovery mechanism, though it is ultimately still bounded by the LLM’s core reasoning ability.

C Annotator Team

Our team comprised 12 annotators, all holding university degrees, with six possessing research-based master’s degrees and six holding doctorates. This composition guaranteed that for any given paper, at

least three annotators possessed relevant domain expertise. Before the formal rating process, each annotator completed a one-hour training session.

We have identified the following potential risks and their corresponding mitigation measures:

1. **Privacy and Confidentiality Risks:** The sharing of personal information by participants may lead to privacy breaches. To mitigate this, all data will be anonymized and stored securely, with access restricted to authorized research personnel only.
2. **Psychological Risks:** Participants may experience discomfort or stress during task execution. To address this, we provide detailed task instructions and debriefing sessions to ensure participants feel supported throughout the process. Additionally, participants have the right to withdraw from the study at any stage without penalty.
3. **Physical Risks:** Although our research procedures do not involve significant physical risks, we will closely monitor participants for any signs of distress and provide necessary support promptly.

Furthermore, regarding data annotation, we have paid annotators a wage higher than the statutory minimum wage in the country of the data annotators, as a sign of respect for their labor.

D More Details of P2PInstruct

Data Source and Generation. P2PInstruct was constructed using the training split of the SciPostLayout dataset, which is separate from our test set. We chose this source for its permissive CC-BY license and topic diversity. And we avoided using papers from sources like ICLR after consulting with legal professionals, as their copyrights would require author-by-author permissions. To mitigate the risk of "imprinting bias" from a purely synthetic loop, we used a **"teacher-forcing-like" approach**. At each stage of data generation, the model was guided by the corresponding part of the original, human-created poster. For example, the HTML Generator received the generated Markdown but was also shown the ground-truth poster's image as a visual layout reference. This process anchors the synthetic data to human design principles.

Quality Validation. The "high-quality" claim is supported by three pillars: (1) **Built-in Quality Control** from the checker-reflection loops during generation; (2) **Manual Verification** of a random sample of 20 pairs from each stage; and (3) **Empirical Validation**, where models fine-tuned on P2PInstruct consistently and significantly outperform their base versions (Table 1), providing strong evidence of the dataset's effectiveness.

E More Details of P2PEval

E.1 Checklist Format

All checklist annotations, including unique paper identification, detailed evaluation criteria, reference figures (when applicable), and established maximum scores, are documented in YAML format.

Each item is a distinct, verifiable component with a description, figure, and max score. For clarity, here is an example that is similar to the following simplified version:

```yaml
name: paper_id
checklist:
 - description: Does the introduction section highlight the limitations
    of current methods in code generation tasks and motivate the use of
    xxxxx?
  max_score: 4
 - description: Does the introduction section provide xxxxx examples for
    visual demonstration?
  figure: 0
  max_score: 5
```

Listing 1: A YAML checklist for paper evaluation.

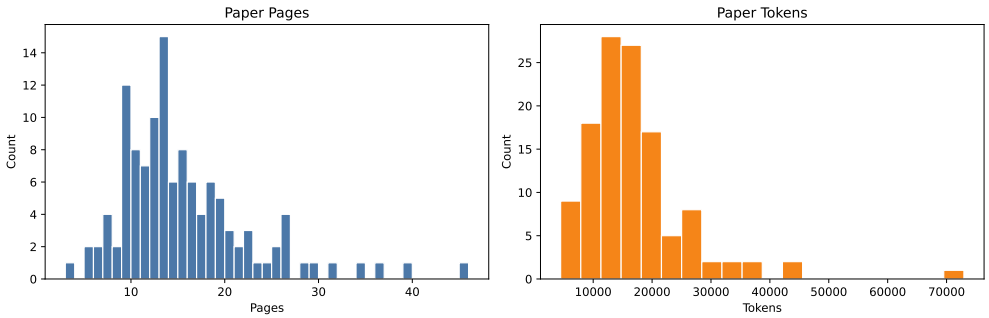


Figure 6: Distributions of Paper

Among them, xxxxx refers to what was proposed in paper_id, which has been anonymized for double-blind review; figure: 0 indicates there is a reference image (the image is taken from the original paper); max_score represents the weight assigned by human annotators. More examples are provided in the anonymous GitHub repository linked to supplementary materials.

E.2 Benchmark Statistics

P2PEVAL is composed of 121 complete paper-poster pairs. Each sample in the benchmark is comprehensive, including the original research paper (‘.pdf‘), the author-created poster (‘.pdf‘ and ‘.png‘), and our detailed, human-annotated ‘checklist.yaml‘ file. The statistics presented in Table 6 highlight the complexity and diversity of the data, underscoring the challenges of the paper-to-poster task.

Table 6: Key statistics of the P2PEVAL benchmark dataset.

Metric	**Value**
Total Paper-Poster Pairs	121
Total Checklist Items	1,738
Total Visual Elements in Checklists	775
Average per Sample	
Paper Length (Pages)	15.13
Paper Content (Tokens)	17,104
Poster Content (Tokens)	2,050
Checklist Content (Tokens)	548
Images per Sample	7.25

The source documents are substantial academic works, with an average length of 15.1 pages and approximately 17,100 tokens per paper. The corresponding posters are significantly more concise, averaging around 2,050 tokens. This reflects a textual summarization ratio of approximately 8.3-to-1, demonstrating the extensive content distillation required for poster generation.

The multimodal nature of the task is evident from the data. On average, each sample contains 7.25 visual elements (figures and tables), with some papers including as many as 28. The richness of our annotations is highlighted by the 1,738 unique checklist items across the dataset, averaging 548 tokens of descriptive metadata per sample. This fine-grained annotation provides a robust foundation for the detailed evaluation of model-generated posters. The distribution histograms 6 show that the

majority of papers range from 8 to 17 pages, and most contain between 5 and 10 images, representing a typical cross-section of academic publications.

E.3 Analysis of Performance Breakdown by Topic

Table 7: Performance breakdown by research category.

Metric	Methodology	Survey	Benchmark	Other
Fine-Grained Score	64.24	65.88	**68.78**	65.92
Universal Score	42.21	40.95	**42.46**	40.63

Table 8: Performance breakdown by scientific field.

Metric	Comp. Sci.	Biology	Psychology	Medical Sci.	Hum. & Soc. Sci.
Fine-Grained Score	66.59	64.17	**73.42**	66.51	62.84
Universal Score	**42.80**	42.21	41.45	39.85	39.64

We analyze the performance of our best-performing model (Claude-3.7-Sonnet) across different research categories and scientific fields from our P2PEVAL.

- **Performance by Research Category**: The tab 7 shows that our P2P performs exceptionally well on Benchmark papers, which typically have a highly regular structure (e.g., Task Definition, Dataset, Results). This structure is more easily parsed and summarized by LLMs. Methodology and Survey papers, which can be more conceptually dense, also yield strong results, demonstrating the model's versatility.
- **Performance by Scientific Field**: The tab 8 shows that P2P achieves the highest Fine-Grained score in Psychology, likely because papers in this field often use standardized charts (bar graphs, scatter plots) that are well-understood by MLLMs. Computer Science papers achieve the highest Universal score, potentially due to the abundance of CS papers in training data. The slightly lower scores in Humanities & Social Sciences may be due to the reliance on denser, narrative text over structured figures, highlighting a potential area for future improvement.

F More Details of Universal Poster Evaluation

F.1 Details of the Human Rating Protocol

As stated in the main paper, our Universal Poster Evaluation score is not directly generated by an LLM. Instead, it is predicted by an XGBoost model trained to replicate human judgment. This was a deliberate design choice, as we found that while LLMs can reliably score discrete, objective criteria, they struggle to capture the complex, non-linear weightings that humans apply when forming a single holistic impression of a poster's quality. The following protocol was used to collect the human scores that serve as ground truth labels for training this XGBoost model.

Annotator Qualifications and Training. Details of our annotation team are shown in Appendix C. During the training session, we reviewed the 10 universal criteria (detailed in Section F.2) and provided scoring guidelines using concrete examples. For instance, award-winning NeurIPS, ICML and ICLR posters were presented as examples deserving a full score (50/50), while posters with significant clarity issues, poor visual design, or logical gaps were used to demonstrate how deductions should be made.

Rating Procedure and Scale. For Universal Poster Evaluation, annotators produced a single holistic score in [0, 50], referencing U1–U10. We provided explicit anchor points to reduce scale ambiguity:

- **45–50: Outstanding.** Meets or exceeds expectations on most criteria; highly professional.
- **35–44: Strong.** Minor issues on a few criteria; clearly presentation-ready.
- **25–34: Adequate.** Several weaknesses, but overall understandable and usable.
- **15–24: Weak.** Multiple violations (e.g., poor fidelity, cluttered layout, weak flow).
- **0–14: Poor.** Severe issues (e.g., illegible, misleading, not self-contained).

Each poster was rated independently by three annotators. Similar to the annotation of the checklist, any posters with substantial initial scoring disagreements were excluded from the final training set for the XGBoost model to maintain high data quality. We utilize both powerful models like GPT-4o and lighter models such as Qwen-VL-2.5-32B, ensuring the trained annotators are exposed to diverse samples to enhance generalizability. We collected a total of 1,701 human ratings across the dataset. To avoid any leakage when training the scoring model, these ratings were collected on outputs produced by an ablated system variant without the checker-reflection mechanism.

Reliability and Agreement. The rigor of our protocol is validated by strong inter-annotator agreement statistics. We achieved a Krippendorff's Alpha of **0.95**, a Spearman correlation of **0.96** and Cohen's Kappa of 0.71, indicating exceptional reliability and consistency among our raters' judgments. These high-quality human scores formed the ground-truth labels (y) used to train our final XGBoost scoring model. We also observed low divergence between predicted and human score distributions (KL divergence: 0.1093; JS divergence: 0.0239), supporting the reliability of our scoring pipeline.

F.2 Details of Evaluation Criteria

The 10 universal criteria form the input feature vector for our XGBoost scoring model. For each generated poster, an LLM-as-a-Judge (GPT-4o) is prompted to score the poster on each of the following 10 dimensions on a discrete scale from 0 to 5. Below are the detailed descriptions for each criterion as used in the evaluation prompt.

U_1: Authorship and Title Accuracy *Description:* Does the poster clearly and accurately display the complete paper title and the full names of all authors without any spelling errors, omissions, or formatting mistakes? *Scoring Guideline:* A score of 5 requires a perfect match. Deductions are made for typos, missing authors, or an incomplete title.

U_2: Image Uniqueness and Quality *Description:* Are all images in the poster unique (i.e., no unintended duplications)? Is the visual quality of each image (resolution, clarity) sufficient for a professional presentation? *Scoring Guideline:* A score of 5 requires all images to be distinct and high-resolution. Points are deducted for blurry, pixelated, or duplicated figures.

U_3: Balanced White Space *Description:* Is the negative or "white" space distributed effectively across the poster? Does the layout avoid areas that look visually overcrowded or, conversely, excessively empty? *Scoring Guideline:* A high score indicates a layout that feels balanced and guides the eye naturally.

U_4: Contextual Relevance *Description:* Do the visual elements (figures, tables) align logically with the adjacent text? Is their placement and thematic connection clear, enhancing the reader's comprehension? *Scoring Guideline:* A score of 5 means every figure is placed next to the text that describes or references it.

U_5: Optimal Visual-to-Text Ratio *Description:* Does the proportion of the poster's area covered by images effectively serve the research content? Is there a good balance between visual evidence and textual explanation? *Scoring Guideline:* The optimal ratio is context-dependent, but a high score reflects a poster that uses visuals impactfully without overwhelming the text, or vice-versa.

U_6: Dimension Appropriateness *Description:* Are the overall dimensions (width and height) of the poster suitable for its content and a typical presentation environment? Does it avoid extreme aspect ratios that would make it difficult to read or display? *Scoring Guideline:* A score of 5 indicates standard poster dimensions (e.g., portrait A0, landscape 48"x36"). Extreme, banner-like shapes receive lower scores.

U_7: Visual Consistency *Description:* Do the design elements—such as color schemes, typography (fonts, sizes), and section heading styles—maintain a cohesive and consistent identity throughout the poster? *Scoring Guideline:* A high score is given for a poster with a unified visual theme. Using a chaotic mix of fonts or colors results in a low score.

U_8: Content Fidelity *Description:* Are the data representations, mathematical formulas, key terminology, and scientific findings presented on the poster identical to those in the original research paper? *Scoring Guideline:* A score of 5 requires zero "hallucinated" or misrepresented facts. Any deviation from the source paper's content results in a score deduction.

U_9: Information Flow Logic *Description:* Is the content organized in a logical and intuitive sequence (e.g., Introduction $\rightarrow$ Methods $\rightarrow$ Results $\rightarrow$ Conclusion)? Can a viewer easily follow the research narrative from start to finish? *Scoring Guideline:* High scores are for posters with a clear, linear, or grid-based flow. Disorganized or confusing layouts receive low scores.

U_{10}: Self-Contained Explanation *Description:* Can the poster be fully understood on its own, without requiring a verbal explanation from a presenter? Does it provide enough context and detail for a knowledgeable reader to grasp the core concepts and contributions? *Scoring Guideline:* A score of 5 indicates a poster that is fully self-sufficient. A poster that is just a collection of figures with cryptic captions would score poorly.

F.3 Other Methods Performance

Our choice of a hybrid MLLM-Featurizer and XGBoost model for the Universal Score was a principled design decision to create a more robust and human-aligned evaluation. Directly asking an LLM for a single holistic score is often noisy and inconsistent. Our approach combines the strengths of different models:

- **Stable Feature Extraction (LLM):** We use the LLM for what it does best: scoring 10 discrete, well-defined criteria (U1-U10).
- **Learning Human Preferences (XGBoost):** We use XGBoost to learn the complex, non-linear function of how humans weigh these criteria, training it on 1,701 human preference scores.

To empirically validate this, we conducted an ablation study comparing our method against simpler baselines. As shown in Table 9, our hybrid approach is significantly more aligned with human judgment than end-to-end LLM scoring based on Qwen-2.5-VL-32B.

Table 9: Ablation study on the Universal Score methodology, validating the choice of the MLLM-Featurizer and XGBoost approach. Metrics are correlations (R^2) and distributional similarity (KL/JS Divergence, lower is better) against human scores.

Method	R^2	**KL-Divergence**	**JS-Divergence**
Direct LLM Scoring (End-to-End)	0.27	1.60	0.62
LLM as Regressor (Replaces XGBoost)	0.51	0.34	0.22
Fine-Tuned LLM as Aggregator	0.70	**0.06**	0.14
Our MLLM-Featurizer + XGBoost	**0.92**	0.11	**0.02**

Additionally, we experiment with other methods, including Ordinary Least Squares($R^2 = 0.66$), Random Forest($R^2 = 0.83$), and various regularization techniques($R^2 = 0.89$); however, these approaches yield suboptimal performance compared to XGBoost.

G Training Details

We utilized the unified and efficient LLaMAFactory framework (Zheng et al., 2024) for all fine-tuning experiments. We use our P2PInstruct dataset with a learning rate of 5×10^{-5} for 3 epochs and employing AdamW (Loshchilov & Hutter, 2017). To ensure optimal instruction-following, we use

the native chat template for each respective model (e.g., Qwen3's template) . Training is conducted with BF16 mixed-precision to accelerate the process. No sequence packing was used, and we set a maximum sequence length of 8000, truncating longer examples.

H Statistical Significance of Results

We performed statistical significance testing for our key experimental comparisons. Following modern best practices (Wasserstein & Lazar, 2016; Jeppesen et al., 2021; Brunoni et al., 2013), we report **Effect Size (Cohen's d)**, which quantifies the practical magnitude of an observed difference. A large effect size ($|d| > 0.8$) indicates a meaningful, substantial improvement. As shown in Table 10, fine-tuning on P2PINSTRUCT yields a large effect size, and our ablation study confirms that removing key components has a medium-to-large negative effect on quality.

Table 10: Effect size (Cohen's d) for fine-tuning and ablation studies. Positive values indicate improvement over the baseline; negative values indicate degradation.

Comparison	**FineGrain (d)**	**Universal (d)**
Effect of Fine-Tuning on P2PINSTRUCT		
InternVL3-P2P vs. Base	2.13	1.72
Qwen2.5-VL-P2P vs. Base	4.18	3.00
Qwen3-P2P vs. Base	1.21	0.60
Effect of Ablating Components (vs. Full P2P*)*		
w/o Reflection	-0.15	-0.61
w/o Figure Describer	-0.26	-0.43
Multi-Agent Only	-0.28	-0.82
End-to-End Only	-0.72	-0.60

I Input Processing

All models are configured with the temperature of 1 and the maximum output token length of 8000 to ensure fair comparison while maintaining generation diversity. To isolate the benefit of our architecture from backbone capacity, we also include text-only backbones that consume pre-extracted figure descriptions(using `Claude-3.7-Sonnet` as the provider of Figure Describer).

And all models receive the same inputs:

- PDF text extracted with PyMuPDFLoader(https://github.com/pymupdf/PyMuPDF) for Section Agent.
- Figures/tables detected with DocLayout-YOLO (Zhao et al., 2024) for Figure Agent.
- For text-only LLMs, which cannot process images, figure descriptions are generated using `Claude-3.7-Sonnet` as the provider of Figure Describer.

J Resource Consumption

The training phase consumes GPU compute equivalent to 80 A100-GPU hours, whereas evaluating multiple models via the APIs offered by our provider, OpenRouter and Volcano Engine, incurs a cost of more than $500.

K Computational Cost and Latency

To help users assess practical deployability, we provide a detailed cost and latency analysis for generating a poster for an average paper (based on Table 6: 17,104 text tokens and 7.25 images). We report metrics for GPT-4.1 (representing powerful closed-source models) and Qwen-2.5-VL-7B (representing efficient open-source models). The average runtimes, using multi-threading and parallel

optimizations, are 209.32 seconds for GPT-4.1 and 55.61 seconds for Qwen-2.5-VL-7B. Below is an approximate breakdown of token consumption and cost, accounting for the retry factors measured during our experiments. The cost of the Figure Extractor (YOLO detection) is excluded as it is a fixed, local computation.

GPT-4.1 Breakdown

- **Figure Agent:** Avg. input = 391.84 (image tokens) $\times$ 7.25; Avg. output = 206.40×7.25.
- **Section Agent:** Avg. input = $(17,104 \times 2) + 338$ (prompt) $+ (206.40 \times 7.25)$; Avg. output = $2,158.84$. This stage has a retry factor of $1.05\times$.
- **Orchestrate Agent:** Avg. input = $17,104 + 2,158.84 + 1,145$ (prompt); Avg. output = $2,505.84$. This stage has a retry factor of $2.32\times$.
- **Total:** This results in an average total of **88,031 input tokens** and **9,577 output tokens**.

Based on OpenRouter pricing for GPT-4.1 ($2/M input, $8/M output):
Cost = $(88,031/1,000,000 \times \$2) + (9,577/1,000,000 \times \$8) \approx \$\mathbf{0.25}$ per poster.

Qwen-2.5-VL-7B Breakdown

- **Figure Agent:** Avg. input = 391.84 (image tokens) $\times$ 7.25; Avg. output = 148.96×7.25.
- **Section Agent:** Avg. input = $(17,104 \times 2) + 338$ (prompt) $+ (148.96 \times 7.25)$; Avg. output = $1,479.50$. This stage has a retry factor of $1.03\times$.
- **Orchestrate Agent:** Avg. input = $17,104 + 1,479.50 + 1,145$ (prompt); Avg. output = $1,637.47$. This stage has a retry factor of $3.47\times$.
- **Total:** This results in an average total of **107,993 input tokens** and **8,286 output tokens**.

Based on OpenRouter pricing for Qwen-2.5-VL-7B ($0.20/M for both input/output):
Cost = $(107,993/1,000,000 \times \$0.20) + (8,286/1,000,000 \times \$0.20) \approx \$\mathbf{0.02}$ per poster.

This analysis demonstrates that P2P offers a tunable cost-performance trade-off and is practically affordable for users, with costs ranging from a couple of cents to a quarter per poster depending on the chosen backbone model.

L AESTHETIC QUALITY EVALUATION

To complement our existing metrics, we introduce a quantitative evaluation of aesthetic principles standard in graphic layout analysis (Deng et al., 2017; Datta et al., 2006; Lu et al., 2015; Marchesotti et al., 2011). We analyze the generated posters using a suite of metrics measuring alignment, balance, symmetry, and other design fundamentals. As shown in Table 11, our best-performing model produces layouts with strong balance and color harmony, providing an objective measure of aesthetic quality.

Table 11: Quantitative aesthetic evaluation of generated posters. Results show mean $\pm$ 95% CI. Higher is generally better. Best results are in bold.

Metric	Higher is better?	Claude-3.7-Sonnet	GPT-4o-mini-2024-07-18
Alignment	Yes	0.455 ± 0.005	$\mathbf{0.476 \pm 0.007}$
Balance (Left/Right)	Yes	$\mathbf{0.894 \pm 0.014}$	0.744 ± 0.032
Balance (Top/Bottom)	Yes	$\mathbf{0.790 \pm 0.022}$	0.784 ± 0.020
Symmetry	Yes	$\mathbf{0.573 \pm 0.016}$	0.447 ± 0.026
Whitespace Ratio	Moderate (target-dependent)	0.719 ± 0.007	0.684 ± 0.015
Rule of Thirds	Yes	0.666 ± 0.015	$\mathbf{0.668 \pm 0.017}$
Contrast (RMS)	Yes	$\mathbf{0.257 \pm 0.007}$	0.247 ± 0.009
Color Harmony	Yes	$\mathbf{0.908 \pm 0.010}$	0.846 ± 0.014
Simplicity (1 - Clutter)	Yes	$\mathbf{0.784 \pm 0.004}$	0.779 ± 0.007

We define each metric used in our evaluation as follows:

Alignment: This metric quantifies the proportion of strong lines in the image that are aligned with the horizontal and vertical axes. It is calculated by first computing the image gradients using a Sobel filter. Then, it measures the magnitude-weighted sum of pixels whose gradient direction is within a small tolerance ($\pm 10^\circ$) of 0, 90, or 180 degrees. This sum is normalized by the total gradient magnitude across the entire image. A high score indicates a strong presence of structured horizontal and vertical elements, which is a key principle of organized design.

Balance (Left/Right): This measures the distribution of visual weight between the left and right halves of the image. Visual weight is approximated by the density of edges detected by a Canny edge detector. The score is calculated as $1 - \frac{|E_L - E_R|}{E_L + E_R}$, where E_L and E_R are the sum of edge pixel intensities in the left and right halves, respectively. A score of 1.0 signifies perfect left-right balance.

Balance (Top/Bottom): Similar to Left/Right Balance, this measures the equilibrium between the top and bottom halves of the image based on edge distribution. The score is calculated as $1 - \frac{|E_T - E_B|}{E_T + E_B}$, where E_T and E_B are the sum of edge pixel intensities in the top and bottom halves. A score of 1.0 signifies perfect top-bottom balance.

Symmetry: This metric assesses horizontal reflectional symmetry. It is computed by calculating the normalized cross-correlation (cosine similarity) between the grayscale left half of the image and a horizontally flipped version of the right half. The score ranges from 0 to 1, where 1 indicates perfect symmetry between the two halves.

Whitespace Ratio: This metric calculates the proportion of the image area that constitutes the background or "whitespace." The background color is estimated by taking the median color of the image's border pixels in the perceptually uniform CIELAB color space. The score is the fraction of total pixels that are perceptually close (within a fixed threshold) to this estimated background color.

Rule of Thirds: This evaluates the composition's adherence to the Rule of Thirds. It first identifies the center of mass (centroid) of the image's edge content. It then measures the shortest distance from this centroid to one of the four "power points" (intersections of the rule-of-thirds grid). The score is inversely proportional to this distance, normalized so that a centroid located directly on a power point yields a high score.

Contrast (RMS): This measures the global contrast of the image, calculated as the Root Mean Square (RMS) contrast. It is defined as the standard deviation of the pixel intensities in the grayscale version of the image, normalized to a [0, 1] range. A higher value indicates a wider dynamic range and stronger tonal separation.

Color Harmony: This metric, derived from the 'hue_peak_score', assesses the simplicity and harmony of the color palette. It computes a histogram of the image's hue channel and sums the normalized frequencies of the two most dominant hues. A high score suggests the palette is concentrated around a small number of primary hues, often leading to a more cohesive and harmonious visual experience.

Simplicity (1 - Clutter): This metric quantifies the visual simplicity of the poster by scoring the inverse of clutter. Clutter is defined as a weighted sum of two components: 1) structural complexity, measured by the overall density of edges, and 2) color complexity, measured by the entropy of the image's saturation and value (brightness) channels. The final score is $1 -$ clutter, where a higher value corresponds to a cleaner, less cluttered design.

M THE FEATURES OF FINE-GRAINED POSTER EVALUATION

The Fine-Grained Poster Evaluation pipeline offers several distinct advantages:

1. **Ground-Truth Alignment**: Each checklist item references specific elements from the official academic posters and corresponding papers, ensuring accurate evaluation aligned with the original author's intent.
2. **Domain-Specific Emphasis**: The pipeline captures domain-specific expectations and conventions, which universal criteria may overlook, reflecting discipline-specific priorities.

3. **Essential Research Component Verification**: Critical content such as key figures, methodology details, and conclusions is explicitly accounted for using detailed scoring mechanisms, ensuring comprehensive evaluation.
4. **Human Preference Integration**: Carefully calibrated by four human annotators, checklist item scores inherently encode domain expertise and human judgment regarding item significance and presentation quality.

N THE FEATURES OF HTML FORMAT

We compare the advantages of HTML for SVG and LaTeX:

- **Universal Accessibility and Portability**: HTML posters can be easily viewed on any device with a web browser, requiring no specialized software (unlike LaTeX, which needs compilation, or potentially specific viewers for complex SVGs).
- **Rich Interactivity**: HTML, often combined with CSS and JavaScript, allows for the seamless integration of interactive elements such as hyperlinks (to papers, datasets, author profiles), tooltips, expandable sections, or even embedded multimedia. This level of interactivity is more cumbersome to achieve and less natively supported in LaTeX or static SVG.
- **Flexible and Modern Styling**: CSS provides powerful and flexible control over the visual presentation, enabling modern, responsive, and aesthetically engaging designs that can adapt to various screen sizes. This offers more design freedom than typical LaTeX layouts and better structural organization for complex content than a single SVG.
- **Ease of Web Integration**: As the native language of the web, HTML posters can be effortlessly embedded into websites, shared via links, and are inherently well-suited for online conference platforms and digital dissemination.

O EXAMPLES OF POSTER GENERATION

Examples of poster generation are shown in Fig 7, Fig 8 and so on. A wide range of examples features spanning different formats (e.g., landscape, portrait, multi-column, spanning columns) and scientific disciplines, effectively highlighting the P2P's flexibility and versatility.

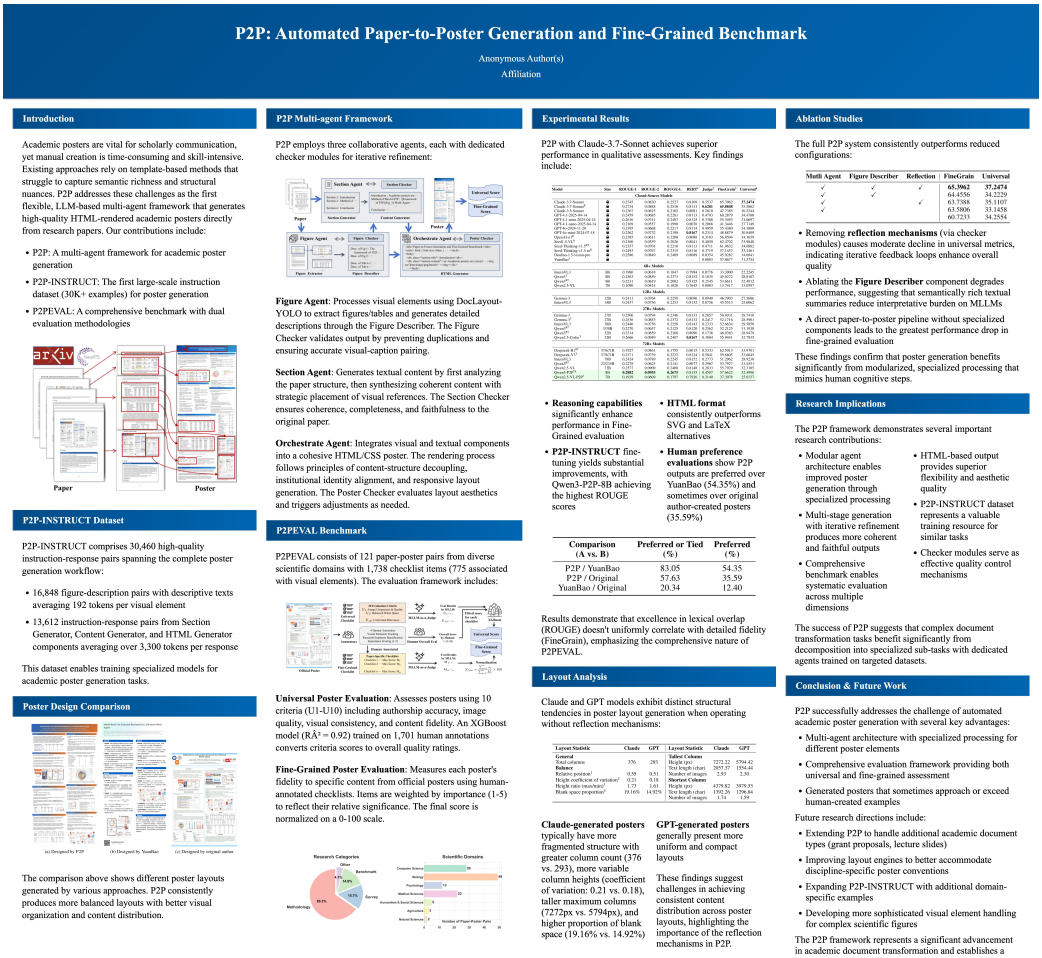


Figure 7: The poster for this paper, powered by P2P.

P PROMPT

Section Extraction

You are an expert in academic paper analysis.

Please analyze the paper content and identify the sections that should be included in the poster.

For each section, provide a simple description of what should be included. First, determine the type of paper. If it is a methodology research paper, focus on the method description, experimental results, and research methodology. If it is a benchmark paper, pay attention to task definitions, dataset construction, and evaluation outcomes. For survey/review papers, emphasize the significance of the field, key timelines or developmental milestones, critical theories and techniques, current challenges, and emerging trends. The above are just references; the specific section names should depend on the paper's content.

Relevant sections for comparison can be combined. There should not be too many sections. The acknowledgement and references section should not appear.

Return the result as a JSON object with section names as keys and descriptions as values.

Ensure the JSON is flat, without nested dictionaries or complex structures.

Example Format:
(JSON Format Example.)

Paper Content:
(Content of Paper.)

Image Description

You are an academic image analysis expert. Your task is to provide detailed descriptions of academic figures, diagrams, charts, or images. Describe what the figure shows, its potential purpose in an academic paper, and any key data or trends visible. The description should be concise and to the point, and should not exceed 100 words.

Image Data:
(Base64 PNG Image Data.)

Text-based Poster Generation

You are a helpful academic expert, who is specialized in generating a text-based paper poster, from given contents.

Figure Description:
(Figures with Description.)

Paper Content:
(Content of Paper.)

If the content of the poster can be described by figures, the relevant text-based content must be simplified to avoid redundancy. Important mathematical formulas can be appropriately placed to assist in understanding.

All sections should be detailed in a markdown format. Do not use headings.

Image-based Poster Generation

You are a helpful academic expert, who is specialized in generating a paper poster, from given contents and figures.

Figure Description:
(Figures with Description.)

Text-based Poster:

(Text-based Poster Content.)

Paper Content:
(Content of Paper.)

Help me inside insert figures into my poster content using my figure index as '![figure_description](figure_index)'

figure_index starts from 0 and MUST be an integer, and don't use any other string in the figure_index.

Each figure can only be used once, and its placement should be precise and accurate.

Use pictures and tables based on their importance.

Poster Rendering

You are a professional academic poster web page creator and your task is to generate the HTML code for a nicely laid out academic poster web page based on the object provided.

Object Description:

- The object contains several fields. Each field represents a section, except for the title, author and affiliation fields. The field name is the title of the section and the field value is the Markdown content of the section.
- The image in Markdown is given in the format ![alt_text, width = original_width, height = original_height, aspect ratio = aspect_ratio](image_index).

HTML Structure:

- Only generate the HTML code inside <body>, without any other things.
- Do not use tags other than , <p>, <ol>, <ul>, <li>, <img>, , .
- Do not create sections that are not in the object.
- Place title, author and affiliation inside . Place title inside , author inside and affiliation inside .
- Place content inside .
- Place each section inside . Place section title inside and section content inside .
- Use <p> for paragraphs.
- Use <ol> and <li> for ordered lists, and <ul> and <li> for unordered lists.
- Use <img src="image_index" alt="alt_text"> for images.

Color Specification:

- Do not add styles other than color, background, border, box-shadow.
- Do not add styles like width, height, padding, margin, font-size, font-weight, border-radius.
- Pick at least 2 colors from the visual identity of the affiliation. If there are multiple affiliations, consider the most well-known affiliation.
- For example, Tsinghua University uses #660874 and #d93379, Beihang University uses #005bac and #003da6, Zhejiang University uses #003f88 and #b01f24. These are just examples, you must pick colors from the visual identity of the affiliation.
- Add text and background color to poster header and section title using inline style. Use gradient to make the poster more beautiful.
- The text and background color of each section title should be the same.

Layout Specification:

- Optionally, inside , group sections into columns using and .
- You must determine the number and flex grow of columns to make the poster more balanced. If the height of one column is too large, move some sections into other columns.

- Optionally, inside , group texts and images into columns using and .
- For example, if there are two images in two columns whose aspect ratios are 1.2 and 2 respectively, the flex grow of two columns should be 1.2 and 2 respectively, to make the columns have the same height.
- Calculate the size of each image based on columns and aspect ratios. Add comment <!– width = display_width, height = display_height –> before each image.
- Rearrange the structure and order of sections, texts and images to make the height of each column in the same group approximately the same.
- For example, if there are too many images in one section that make the height of the column too large, group the images into columns.
- DO NOT LEAVE MORE THAN 5% BLANK SPACE IN THE POSTER.

Existing Style:
(Existing CSS Style.)

Object:
(Poster Object.)

Q LIMITATIONS

While P2P demonstrates significant advances in automated academic poster generation, several limitations warrant acknowledgement. Our approach currently optimizes for HTML rendering, which may present compatibility challenges in academic environments where LaTeX or PowerPoint formats remain prevalent. Additionally, the system's effectiveness is constrained by the visual understanding capabilities of underlying MLLMs, occasionally resulting in misinterpretation of complex scientific visualizations or specialized notations. The multi-agent architecture, while comprehensive, introduces computational overhead that may limit accessibility for resource-constrained researchers.

(a) Designed by P2P (b) Designed by YuanBao (c) Designed by original author

Figure 8: Examples of academic poster design for Deng et al. (2024).

WaterBench: Towards Holistic Evaluation of Watermarks for Large Language Models

Shangqing Tu*, Yuliang Sun*, Yushi Bai, Jifan Yu, Lei Hou, Juanzi Li
Department of Computer Science and Technology, Tsinghua University & School of Computer Science and Engineering, Beihang University

Introduction

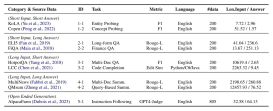
Large language models (LLMs) have raised concerns about potential misuse, necessitating reliable detection methods such as watermarking. However, current watermark evaluations typically assess detection and generation performance separately using inconsistent hyperparameters, often limited to single datasets with perplexity-based metrics.

Prompt	Z-score	Detection results	Generation metric
What are the names of some famous actors that started their careers on Broadway?			
No watermark 1. Hugh Jackman 2. Audra McDonald 3.Idina Menzel...	0.3	TN: 1 FP: 0	58
With watermark There is a very successful list actors who started in Broadway...	7.4	TP: 1 FN: 0	31

WaterBench addresses these limitations as the first comprehensive benchmark for LLM watermarks, enabling fair comparisons while evaluating real-world performance across diverse tasks and metrics.

Task Selection

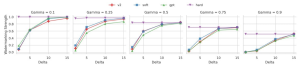
Our diverse benchmark includes nine tasks across five categories based on input-output length patterns: (1) Short Input, Short Answer (entity/concept probing), (2) Short Input, Long Answer (ELI5/finance QA), (3) Long Input, Short Answer (HotpotQA/code completion), (4) Long Input, Long Answer (multi-doc/query-based summarization), and (5) Open-Ended Generation (instruction following).



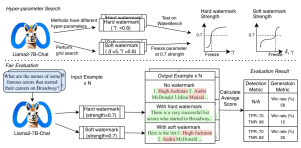
This taxonomy tests watermarking robustness across various language generation scenarios with different computational demands.

Benchmarking Procedure

We introduce a novel approach to ensure fair comparisons between watermarking methods: (1) Define watermarking strength as True Positive Rate (TPR), (2) Perform grid search to determine hyperparameters (γ, δ) that achieve the same watermarking strength across methods,



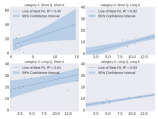
(3) Fix these parameters to jointly evaluate both detection accuracy and generation quality.



This approach prevents unfair advantages from varying watermarking strengths and reveals true trade-offs between detection and quality.

Analysis and Findings

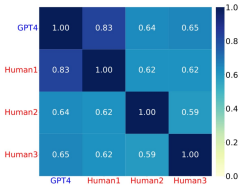
Our analysis reveals: (1) Unifying watermarking strength is crucial for fair comparisons - the same method can appear best or worst depending on strength settings,



(2) Short-text tasks pose significant watermarking challenges with TPR dropping from ~90% to ~0% when strength decreases from 0.95 to 0.7, (3) All watermarks cause substantial quality degradation in open-ended generation (65-99%), (4) Different watermarking methods show varying robustness across task categories with V2 watermark demonstrating the most consistent performance, (5) Correlation analysis between tasks in the same category validates our taxonomy design while highlighting performance differences between short and long-form outputs.

Evaluation Metrics

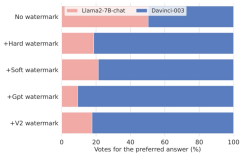
We evaluate watermarks using: (1) Detection metrics - True Positive Rate (TPR) for watermarked text detection and True Negative Rate (TNR) for unwatermarked text classification, (2) Generation metrics - task-specific quality measures (F1, ROUGE-L) for specialized tasks and GPT4-Judge for instruction-following tasks.



Human evaluation confirms GPT4-Judge's effectiveness with Cohen's kappa coefficients >0.6, indicating substantial agreement with human annotators.

Human Evaluation

We verified GPT4-Judge effectiveness by comparing 100 samples against human evaluations. Results show strong alignment between GPT4 and human preferences, with unwatermarked Llama2 achieving ~50% win rate against Davinci-003, while watermarked versions show significantly lower preference rates.



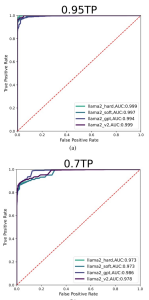
The Cohen's kappa coefficient between GPT4 and human annotators ranged from 0.64-0.83, exceeding the agreement between human annotators (0.59-0.62), confirming GPT4-Judge's reliability as an evaluation metric.

Experimental Results

We evaluated four watermarking techniques (Hard, Soft, GPT, V2) on two LLMs (Llama2-7B-chat, InternLM-7B-8k) at two watermarking strengths (0.7, 0.95). Key findings: (1) At 0.95 strength, all methods show substantial generation quality drops (31-73%),

Model	C1: (Short Q, Short A) *Factual Knowledge*				C2: (Short Q, Long A) *Long-form QA*				C3: (Long Q, Short A) *Reasoning & Coding*			
	TP	TN	GM	Drop	TP	TN	GM	Drop	TP	TN	GM	Drop
Llama2-7B-chat	–	–	17.8	–	–	–	21.3	–	–	–	37.5	–
+ hard watermark	0.0	100.0	13.7	↓ 23.3%	100.0	100.0	19.4	↓ 8.9%	39.2	100.0	21.0	↓ 44.1%
+ soft watermark	0.0	100.0	13.8	↓ 22.6%	100.0	100.0	19.4	↓ 8.7%	41.2	100.0	20.6	↓ 45.1%
+ gpt watermark	11.8	100.0	17.0	↓ 4.4%	99.5	99.8	13.8	↓ 35.0%	25.1	100.0	17.3	↓ 53.9%
+ v2 watermark	0.0	100.0	14.9	↓ 16.6%	99.5	100.0	19.4	↓ 8.8%	39.8	100.0	25.1	↓ 33.2%

Model	C4: (Long Q, Long A) *Summarization*				C5: Open-Ended *Instruction Following*				Overall: (12345) *Detection & Generation*			
	TP	TN	GM	Drop	TP	TN	GM	Drop	TP	TN	GM	Drop
Llama2-7B-chat	–	–	23.3	–	–	–	54.7	–	–	–	28.3	–
+ hard watermark	91.8	100.0	19.9	↓ 14.4%	96.5	99.8	17.3	↓ 68.4%	70.7	99.9	18.4	↓ 35.1%
+ soft watermark	92.0	100.0	20.2	↓ 13.3%	95.4	99.8	19.0	↓ 65.2%	70.7	99.9	18.6	↓ 34.4%
+ gpt watermark	96.0	100.0	15.0	↓ 35.4%	93.4	99.9	4.1	↓ 92.5%	69.9	99.9	14.5	↓ 48.7%
+ v2 watermark	88.8	100.0	19.7	↓ 15.3%	94.0	99.9	17.0	↓ 68.9%	69.4	100.0	19.5	↓ 31.2%



(2) Short-answer tasks are more challenging for watermark detection, (3) V2 watermark achieves the best balance of detection and quality preservation, (4) Open-ended generation suffers the most severe quality degradation (>90%). Test speed analysis shows minimal computational overhead from watermarking (<3%).

Figure 9: Examples of academic poster (Tu et al., 2023), powered by P2P.

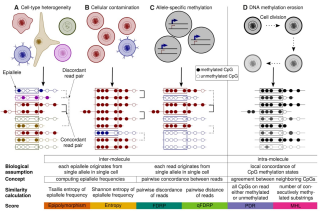
Quantitative comparison of within-sample heterogeneity scores for DNA methylation data

Michael Scherer, Almut Nebel, Andre Franke, Jörn Walter, Thomas Lengauer, Christoph Bock, Fabian Müller, Markus List

Max Planck Institute for Informatics, Saarland University, Kiel University, Austrian Academy of Sciences, Stanford University, Technical University of Munich

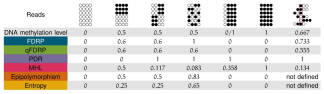
Introduction

DNA methylation heterogeneity within samples provides important information beyond average methylation levels. Within-sample heterogeneity (WSH) can arise from multiple biological phenomena: cell-type heterogeneity (mixed cell populations), cellular contamination (impure samples), allele-specific methylation (ASM), and stochastic DNA methylation erosion during cell division. Bisulfite sequencing enables quantification of this heterogeneity through pattern analysis across sequencing reads. While standard methylation analysis typically reports only average levels at individual CpGs, WSH scores capture the variance in methylation patterns, potentially revealing regulatory regions not apparent from average methylation alone.



Heterogeneity Scores

Six WSH scores were evaluated, divided into two categories: **Intra-molecule scores** quantify agreement between CpG methylation states on individual reads: (1) **PDR** (Proportion of Discordant Reads) measures locally disordered methylation patterns; (2) **MHL** (Methylation Haplotype Load) quantifies stretches of consecutively methylated CpGs. **Inter-molecule scores** capture variance between reads: (3) **Epipolymorphism** and (4) **Entropy** measure epiallele frequency distributions in four-CpG windows; (5) **FDRP** (Fraction of Discordant Read Pairs) and (6) **qFDRP** (quantitative FDRP) are newly proposed scores that quantify pairwise methylation pattern disagreement between reads. qFDRP modifies FDRP using Hamming distance to better balance discordance detection.

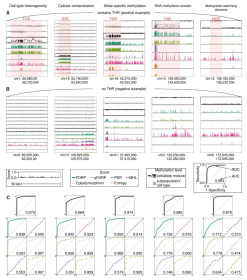


Implementation

All six WSH scores are implemented in an open-source R package, supporting integration into bisulfite sequencing analysis workflows. The package requires aligned reads (BAM files) and genomic annotations as input. Computational optimizations include read subsampling (maximum 40 reads for high-coverage sites) to mitigate combinatorial explosion in pairwise calculations for FDRP/qFDRP. The package implements flexible parameter settings for minimum coverage thresholds, overlap requirements, and window sizes adaptable to various experimental designs. Code is available at GitHub (https://github.com/MPIIComputationalEpigenetics/WSHPackage) with extensive documentation and vignettes demonstrating score applications.

Simulation Framework

Bisulfite sequencing reads were simulated from the human reference genome using Sherman to generate different heterogeneity scenarios: (1) Cell-type heterogeneity: 2-10 cell types with distinct methylation patterns mixed at equal proportions; (2) Cellular contamination: two cell types mixed at varying proportions (50-100%); (3) Allele-specific methylation: two distinct methylation patterns at 50:50 ratio; (4) DNA methylation erosion: stochastic loss of methylation across reads; (5) Methylation switching domains: abrupt changes in methylation states. Technical parameters (coverage: 5-50x, read length: 40-150bp, error rate: 1-10%, CpG density) were systematically varied to assess score robustness.



Recommendations

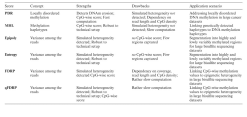
PDR: Best for locally disordered methylation in large cancer datasets; affected by technical parameters.

MHL: Suitable for linking genetic and methylation haplotypes; robust to technical setup but less sensitive to heterogeneity.

Epipolymorphism/Entropy: Effective for detecting heterogeneity in regions with high CpG density; limited by four-CpG window requirement.

FDRP: Good for CpG-wise heterogeneity; sensitive to technical variations.

qFDRP: Recommended for general heterogeneity analysis; offers CpG-wise resolution, detects various heterogeneity types, and shows robustness to technical parameters.



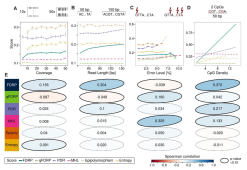
WSH scores provide complementary information to average methylation levels and can identify novel regulatory regions not captured by standard analysis. Score selection should be tailored to specific research questions and data characteristics.

Score Performance

Inter-molecule WSH scores (FDRP, qFDRP, Epipolymorphism, Entropy) effectively detected cell-type heterogeneity, contamination, and ASM, with AUC values often >0.9. These scores showed positive correlation with the number of simulated cell types, capturing the degree of heterogeneity in samples.

Intra-molecule scores (PDR, MHL) were less effective at detecting these phenomena. For DNA methylation erosion, both PDR and inter-molecule scores showed moderate detection ability. All scores correctly showed low values for methylation switching domains, differentiating true heterogeneity from methylation level changes.

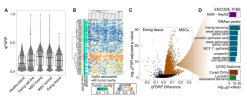
Regarding technical robustness, qFDRP and MHL were most resilient to variations in coverage, read length and CpG density, while FDRP, PDR and Epipolymorphism showed dependencies on these parameters.



Clinical Application

WSH scores were applied to an Ewing sarcoma dataset comprising 140 tissue samples, 16 cell lines, and MSCs (mesenchymal stem cells). qFDRP revealed highest heterogeneity in MSCs and lower heterogeneity in tumor samples, complementary to average methylation analysis.

Using machine learning (LASSO), qFDRP values at 26 specific CpG sites accurately predicted tumor purity levels (correlation: 0.966), outperforming other scores and average methylation. Differential heterogeneity analysis between MSCs and tumor samples identified genomic regions enriched for transcription factor binding sites and DNase-hypersensitive sites linked to Ewing sarcoma.



The patterns found were distinct from those using average methylation, demonstrating the complementary information provided by heterogeneity analysis.

Figure 10: Examples of academic poster (Scherer et al., 2020), powered by P2P.

ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function

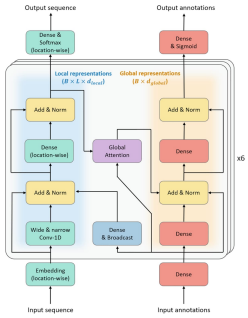
Nadav Brandes†, Dan Ofer†, Yam Peleg, Nadav Rappoport and Michal Linial (* Corresponding author, † Equal contribution)
The Hebrew University of Jerusalem, Israel; Deep Trading Ltd., Haifa, Israel; Ben-Gurion University of the Negev, Israel

Introduction

While protein sequences accumulate exponentially, understanding their functions remains challenging. Self-supervised deep language models have revolutionized natural language processing and have recently been applied to biological sequences. However, existing approaches typically import architectures designed for human language without considering protein-specific characteristics. We present ProteinBERT, a deep language model specifically designed for proteins with novel architectural elements and training methodology that significantly improves efficiency while maintaining state-of-the-art performance.

ProteinBERT Architecture

Unlike standard Transformers, ProteinBERT features dual processing paths: local representations (sequence-level) and global representations (protein-level). The architecture comprises six stacked blocks with convolutional layers processing sequence information and dense layers handling global annotations.



Key innovations include:

- **Global Attention**: Linear complexity mechanism replacing quadratic-complexity self-attention
- **Dual Paths**: Local (B×L×d_local) and global (B×d_global) representations with controlled information exchange
- **Efficiency**: Only 16M parameters versus 38M-3B parameters in competing models
- **Length Flexibility**: Linear scaling with sequence length enables processing proteins of any length

Conclusions and Impact

ProteinBERT provides an efficient framework for protein prediction with several advantages:

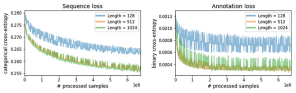
- **Efficiency**: Fine-tuning takes only 14 minutes on average per benchmark on a single GPU
- **Universality**: Performs well across diverse protein properties with the same architecture
- **Accessibility**: Much smaller model size enables broader adoption in protein research
- **Flexibility**: Effectively handles sequences of any length without performance degradation

By providing a compact yet powerful model for protein sequence analysis, ProteinBERT democratizes deep learning approaches in protein research. Code and pretrained weights are available at https://github.com/nadavbra/protein_bert.

Pretraining Methodology

ProteinBERT was pretrained on 106M non-redundant proteins from UniRef90 using two simultaneous tasks:

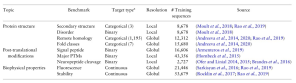
1. **Bidirectional Language Modeling**: Recovering corrupted amino acid tokens (5% randomly replaced)
2. **GO Annotation Prediction**: Predicting 8,943 Gene Ontology terms covering molecular functions, biological processes and subcellular locations



Training was performed for 6.4 epochs, periodically switching between sequence lengths of 128, 512, and 1024 tokens to improve generalization. The loss function combined categorical cross-entropy for sequence recovery and binary cross-entropy for GO annotation prediction.

Benchmarks and Datasets

We evaluated ProteinBERT on nine diverse benchmarks covering protein structure, post-translational modifications, and biophysical properties:



Protein Structure:

- Secondary structure prediction (categorical, local)
- Disorder prediction (binary, local)
- Remote homology detection (categorical, global)
- Fold class prediction (categorical, global)

Post-translational Modifications:

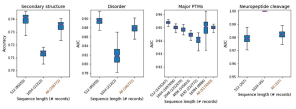
- Signal peptide prediction (binary, global)
- Major PTMs (binary, local)
- Neuropeptide cleavage (binary, local)

Biophysical Properties:

- Fluorescence prediction (continuous, global)
- Stability prediction (continuous, global)

Cross-Length Generalization

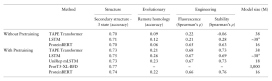
ProteinBERT effectively generalizes across protein lengths without significant performance degradation. Box plots demonstrate that prediction metrics remain robust across various sequence lengths:



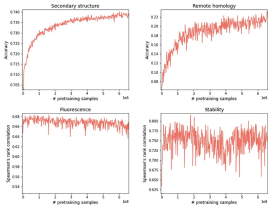
- Secondary structure prediction performs best with 512-length sequences
- PTM prediction maintains high AUC (0.93-0.96) across all sequence lengths
- Neuropeptide cleavage prediction reaches near-perfect performance (>0.99 AUC) with 1024-length sequences
- This generalization eliminates the need to split long sequences into smaller chunks, a limitation of traditional transformer models

Experimental Results

Despite its smaller size, ProteinBERT achieves performance comparable to or exceeding larger models across diverse benchmarks:



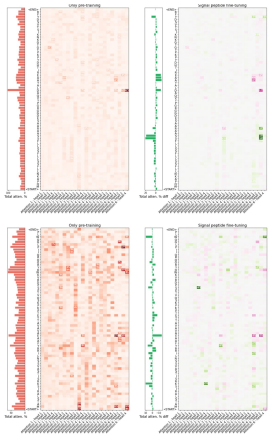
Pretraining significantly improved performance, with secondary structure accuracy increasing from 0.70 to 0.74 and remote homology detection from 0.06 to 0.22.



Longer pretraining (up to 600M samples) progressively improved performance on structure prediction tasks while engineering tasks showed more variable responses.

Global Attention Analysis

Global attention visualization reveals how ProteinBERT focuses on biologically relevant regions. Comparison of attention patterns before and after fine-tuning for signal peptide prediction shows:



- Pre-training attention (left panels) shows baseline patterns across protein sequences
- Fine-tuning (right panels) shifts attention toward the signal peptide region
- In positive examples, attention increases significantly at the signal peptide cleavage site
- Different attention heads specialize in different regions (e.g., head #3 in block 3 focuses on sequence beginnings)

This interpretable attention mechanism provides insights into model predictions and biological features.

Figure 11: Examples of academic poster (Brandes et al., 2022), powered by P2P.

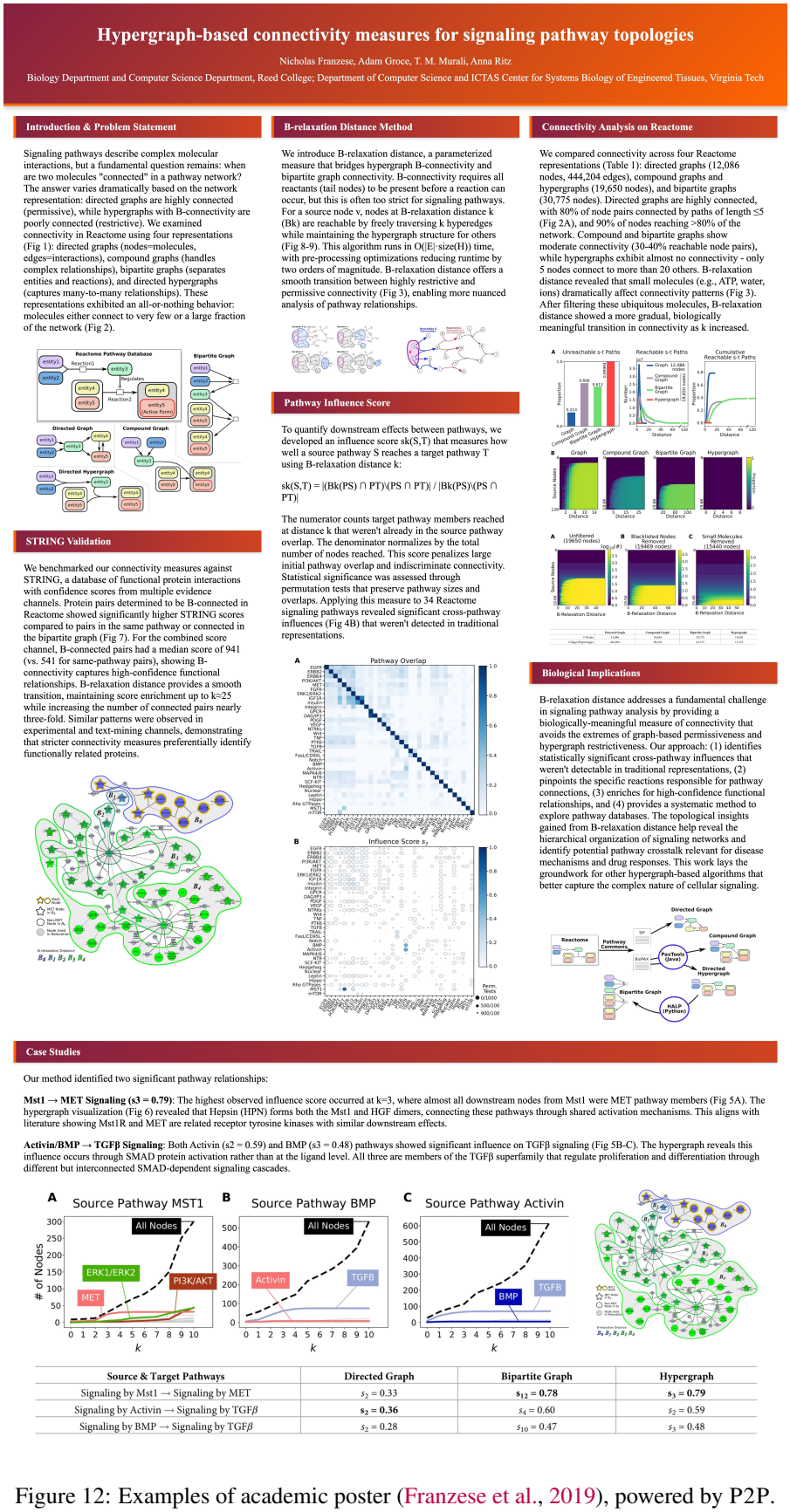
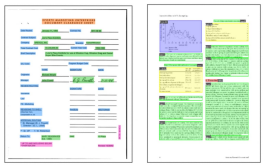


Figure 12: Examples of academic poster (Franzese et al., 2019), powered by P2P.

XFORMPARSER: A Simple and Effective Multimodal Multilingual Semi-structured Form Parser

Xianfu Cheng†, Hang Zhang†, Jian Yang, Xiang Li, Weixiao Zhou, Fei Liu, Kui Wu, Xiangyuan Guan, Tao Sun, Xianjie Wu, Tongliang Li*, Zhoujun Li*
CCSE, Beihang University, Beijing Language and Culture University, Beijing Information Science and Technology University, Shenzhen Intelligent Strong Technology Co.,Ltd.

Problem and Motivation



(a) Text-centric form understanding on FUNSD (b) Image-centric layout analysis on PubLayNet

In Document AI, parsing semi-structured forms remains challenging despite advances in pre-trained multimodal models. Current approaches face significant limitations in multilingual parsing capabilities and show diminished recall in industrial contexts with rich text and visual elements. Extracting structured information from diverse document formats (PDF, Word, images) requires models that can comprehend both textual content and spatial layout information.

Form parsing is essentially an entity relation mining task connecting Named Entity Recognition (NER) and Key Information Extraction (KIE). The diversity of layouts, poor quality of scanned documents, and complexity of template structures make it difficult to represent and understand unstructured information using generic rules.

XFormParser addresses these challenges by integrating semantic entity recognition and relation extraction into a unified framework, improving performance for industrial applications across languages.

Training Methodology

XFormParser integrates two tasks through joint training:

1. **Joint Loss Function**:

Loss = Loss_SER + Loss_RE

Where both components use cross-entropy.

2. **Warm-up Soft Label Mechanism**:

epoch start	epoch warm	RE F1
✗	✗	91.20
10	✓	92.67
20	✓	92.74
30	✗	92.44
30	✓	93.14
40	✗	91.06
40	✓	91.35
50	✗	90.93
50	✓	91.43

- Early training: Hard labels for quick convergence
- Mid-to-late training: Soft labels providing richer information
- Transition period: Gradually increasing soft label weight

α = min(1, (ep - ep_start)/ep_warm)

LE = LE_hard_label (ep ≤ ep_start)

LE = α·LE_soft_label + (1-α)·LE_hard_label (ep > ep_start)

3. **Entity Representation**:

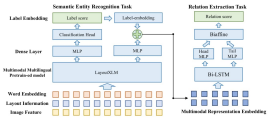
e_i = pooling(H_re) ⊕ p_i

where p_i is the predicted label embedding from SER, enabling information sharing between tasks.

Method	Task		Component		SER F1 Accuracy		RE F1 Accuracy↑	
	SER	RE	Decoder	soft label	EN	ZH	EN	ZH
XFormParser	✓	✓	✓	✓	**92.46**	**93.42**	**91.2**	**93.14**
1 w/o Task RE	✓	✗	✓	✗	91.89	92.86	-	-
2 w/o Task SER	✗	✓	✓	✓	-	-	90.90	92.64
3 w/o Decoder	✓	✓	✗	✓	91.40	92.75	79.79	81.73
4 w/o soft label	✓	✓	✓	✗	91.19	92.19	90.75	91.20

The ablation study demonstrates that removing the decoder significantly impacts RE performance (-11%), while removing soft labels decreases overall model effectiveness.

XFormParser Architecture



XFormParser employs a multimodal Transformer architecture that processes three input modalities:

- **Text inputs**: Word embeddings processed through tokenization
- **Position information**: 2D coordinates of text blocks obtained via OCR
- **Visual features**: Image embeddings from document regions

The architecture consists of:

1. **LayoutXLM backbone**: Pre-trained multimodal multilingual model that generates contextualized representations
2. **Semantic Entity Recognition**: MLP classification layers to identify entity types (HEADER, QUESTION, ANSWER, OTHER)
3. **Relation Extraction**: Bi-LSTM decoder with head/tail MLPs and a biaffine classifier to identify relationships between entities

Setup	Multi-language model
Optimizer	AdamW
Weight ratio	0.1
Lr scheduler	LINEAR
vocab size	250002
Max Steps	512
Batch size	8
Initial learning rate	5e-5
Training epochs	100
Evaluation metric	ref. B.4

This design enables effective information extraction without additional inference overhead, making it suitable for deployment in resource-constrained environments.

InDFormSFT Dataset

We introduce **InDFormSFT**, a new supervised fine-tuning dataset specifically addressing industrial form parsing needs:

- **Composition**: 562 form images across 8 major industrial application scenarios
- **Languages**: Chinese and English documents
- **Partition**: 422 training samples, 70 validation samples, 70 test samples

Dataset partition	Data volume	Question Entity	Answer Entity	Single Entity	Title Entity	Continuous char entity
Training set	422	8702	6825	422	370	186
Validation set	70	1535	1441	65	18	83
Testing set	70	1440	1844	78	58	18

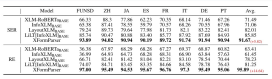
Annotation: Semi-automatically generated using tools like GPT4o and carefully verified by humans

Dataset partition	One-to-one	One-to-two	One-to-three	One-to-many
Training set	11806	265	96	171
Validation set	2366	55	38	39
Testing set	2626	76	53	72

- **Relation Types**: One-to-one (dominant), one-to-two, one-to-three, one-to-many relationships
- **Entity Types**: Question (6702), Answer (6825), Single (422), Title (370), Continuous character (194)

The dataset follows the format of XFUND with cell granularity, including absolute coordinates, text information, label information, and inter-cell linking.

Experimental Results

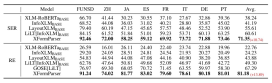


Language-specific Fine-tuning:
XFormParser achieves significant improvements over previous SOTA models:

- SER: 89.04% avg F1 (+6.53% improvement)
- RE: 91.65% avg F1 (+9.64% improvement)

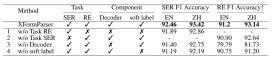
Multi-language Fine-tuning (trained on all 8 languages):

- SER: 91.67% avg F1 (+5.82% improvement)
- RE: 95.89% avg F1 (+14.64% improvement)



Zero-shot Transfer (trained on FUNSD, tested on 7 other languages):

- SER: 71.35% avg F1 (+10.74% improvement)
- RE: 81.18% avg F1 (+13.89% improvement)

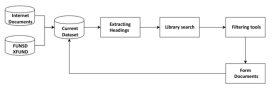


Ablation Study demonstrates contribution of each component:

- Removing Decoder: RE F1 drops by ~11 points
- Removing Soft Labels: Performance decreases for both tasks
- Optimal performance achieved with warm-up at epoch 30

The model demonstrates robust cross-lingual generalization and superior performance across both SER and RE tasks.

Key Innovations and Impacts



1. **Simple Yet Effective Architecture**

- Integration of SER and RE tasks with joint loss function and staged warm-up soft label mechanism
- Bi-LSTM decoder significantly enhances relation extraction capabilities
- Achieves state-of-the-art performance without additional inference overhead

2. **Industrial-strength Dataset**

- InDFormSFT dataset with 562 form images across 8 real-world scenarios
- Diverse relationship types and entity categories
- Semi-automatic annotation with human verification ensures quality

3. **Strong Multilingual Performance**

- Superior results across language-specific, multilingual, and zero-shot settings
- Significant improvements in both SER (+10.74%) and RE (+13.89%) for zero-shot transfer
- Robust cross-lingual generalization capabilities

XFormParser demonstrates that a thoughtfully designed architecture with appropriate training methodology can significantly outperform more complex models in multimodal multilingual document understanding tasks.

Figure 13: Examples of academic poster (Cheng et al., 2024a), powered by P2P.

Figure 14: Examples of academic poster (Caffagni et al., 2024), powered by P2P.

Evidence for Evolutionary and Nonevolutionary Forces Shaping the Distribution of Human Genetic Variants near Transcription Start Sites

Giovanni Scala, Ornella Affinito, Gennaro Miele, Antonella Monticelli, Sergio Cocozza

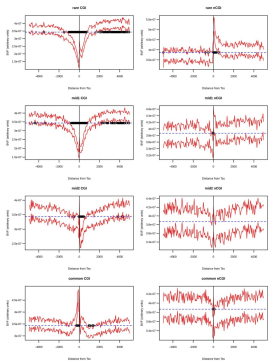
Gruppo Interdipartimentale di Bioinformatica e Biologia Computazionale, Università degli Studi di Napoli "Federico II", Naples, Italy

Introduction

Transcription start sites (TSSs) and their surrounding regions are critical for gene regulation but are also subject to transcription-related mutagenic processes. This study investigates the genome-wide distribution of single nucleotide polymorphisms (SNPs) in the 10kb regions flanking human TSSs to understand the forces that create and maintain genetic variability in these functionally important regions.

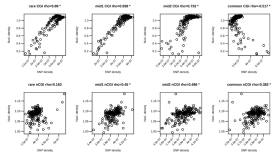
Methodology

We analyzed 27,487 human TSSs, categorizing them into TSSs located inside CpG islands (CGI-TSSs, 53%) and those outside (nCGI-TSSs, 47%). We classified ~2.6 million SNPs into four frequency groups based on minor allele frequency (MAF): rare (MAF≤4.59×10^{-4}), mid1 (4.59×10^{-4}-0.01). For each TSS, the surrounding 10kb region was divided into 200 bins of 50bp each to calculate normalized mean variant frequency (BVF). We examined correlations with nucleosome positioning scores, evolutionary conservation (GERP), GC-biased gene conversion (gBGC), and variant deleteriousness (CADD).



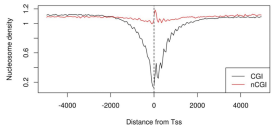
Distribution of Variants by Frequency

Our analysis revealed that variant distribution depends on their frequency and location relative to TSSs, with distinct patterns between CGI-TSSs and nCGI-TSSs. CGI-TSSs showed a significant positional effect for all frequency classes, with a marked depression of rare variants near the TSS but a relative 1.7-fold increase in the first 200bp downstream. Conversely, nCGI-TSSs showed positional effects only for rare variants, with a smaller 1.15-fold downstream increase. Common variants showed a sharp peak near CGI-TSSs, completely absent in nCGI-TSSs, suggesting influences beyond random mutation.



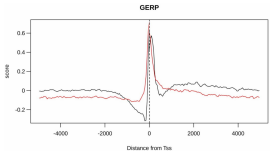
Nucleosome Occupancy and Rare Variants

We found a strong positive correlation between nucleosome positioning and rare variant density in CGI-TSSs (ρ=0.89), and a weaker but significant correlation in nCGI-TSSs. CGI-TSSs exhibited a pronounced nucleosome depletion directly at the TSS, while nCGI-TSSs maintained relatively consistent nucleosome density. These findings suggest that transcription-related mutational phenomena could be linked to reduced DNA repair efficiency in nucleosome-occupied regions, as nucleosomes can limit the accessibility of repair proteins, with damage within nucleosome cores repaired at approximately 10% the rate of naked DNA.



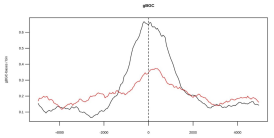
Evolutionary Forces

Using GERP conservation scores, we identified signatures of purifying selection around TSSs. For nCGI-TSSs, we found a strong positive correlation (ρ=0.725) between GERP scores and BVF-delta (difference between rare and common variant frequencies). For CGI-TSSs, we observed a complex pattern with a strong positive correlation (ρ=0.774) between GERP and BVF-delta in regions >700bp from TSSs, suggesting purifying selection preserves functional regions by preventing deleterious mutations from reaching common frequencies. These evolutionary constraints are stronger in CGI-TSSs, consistent with their association with housekeeping genes that require stringent conservation.



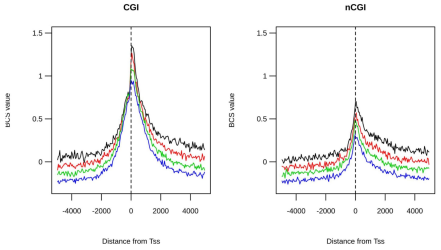
Non-evolutionary Forces

We identified GC-biased gene conversion (gBGC) as a significant non-evolutionary force affecting allele frequencies near TSSs. In CGI-TSSs, we found a strong negative correlation (ρ=-0.734) between gBGC scores and BVF-delta within the inner 700bp region around TSSs. This indicates that gBGC competes with purifying selection, particularly in CpG-rich regions, by preferentially resolving GC/AT heterozygotes to GC/GC homozygotes during gene conversion, thereby increasing the frequency of variants that might otherwise be selected against.



Variant Deleteriousness

Analysis of CADD scores revealed that variants closer to TSSs are potentially more deleterious than those more distant, with deleteriousness increasing toward the TSS from both sides. Rare variants consistently showed higher deleteriousness scores than common variants across all positions, supporting the notion that purifying selection prevents deleterious mutations from reaching high frequencies. Furthermore, variants in CGI-TSSs exhibited significantly higher deleteriousness scores than those in nCGI-TSSs within approximately 1300bp of the TSS, highlighting the functional importance of CGI regions.



Conclusions

This study provides a detailed view of how human genetic variants are distributed around TSSs, revealing that both evolutionary (purifying selection) and non-evolutionary (gBGC) forces shape genetic variability in these critical regulatory regions. Rare variants show strong correlations with nucleosome positioning, suggesting transcription-related mutagenic processes influence their distribution. The competing effects of purifying selection and gBGC create distinctive frequency patterns, particularly in CGI-TSSs, while the higher deleteriousness of variants near TSSs underscores the functional importance of these regions. These findings enhance our understanding of the complex interplay between mutational processes and selective forces in shaping human genomic diversity.

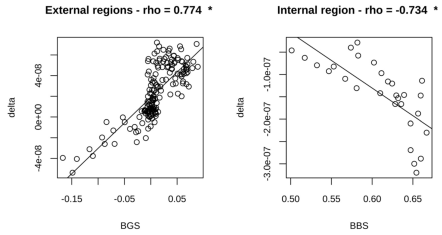


Figure 15: Examples of academic poster (Scala et al., 2014), powered by P2P.

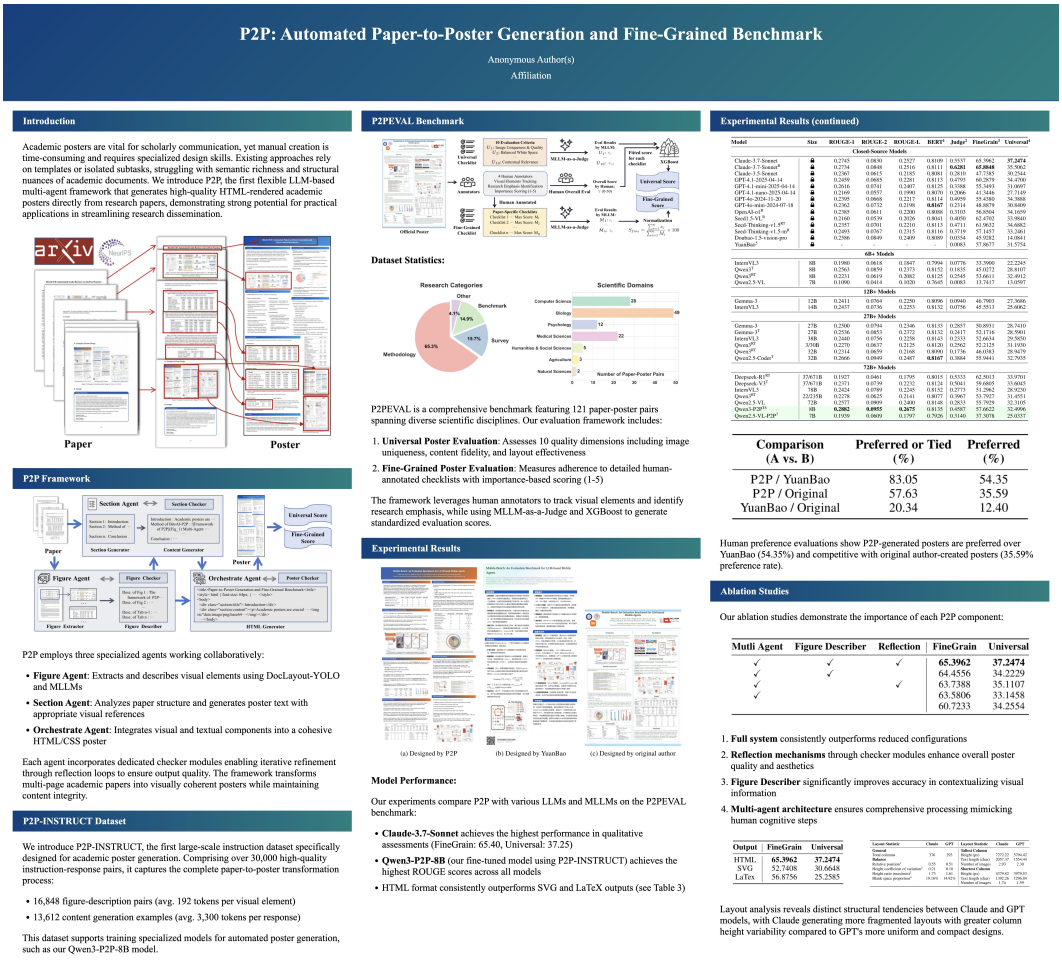


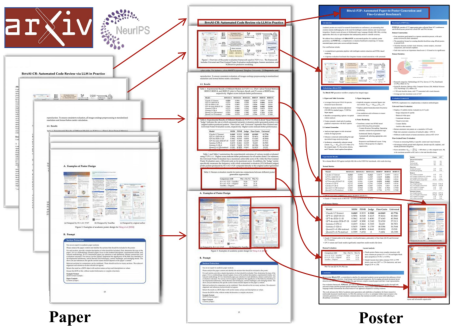
Figure 16: The horizontal poster for this paper, powered by P2P.

P2P: Automated Paper-to-Poster Generation and Fine-Grained Benchmark

Anonymous Author(s)
Affiliation

Introduction

Academic posters are vital for scholarly communication but their manual creation is time-consuming and skill-intensive. Existing approaches primarily rely on template-based methods that struggle to capture semantic richness and structural nuances of academic papers. While recent Large Language Models (LLMs) show promise in understanding document structures, their application to poster generation remains limited due to insufficient quality control and lack of standardized benchmarks.

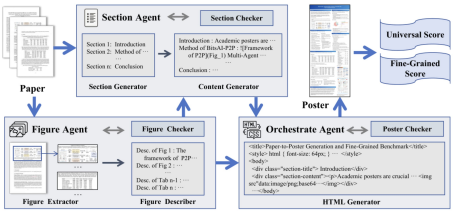


P2P Framework

P2P employs three specialized agents working collaboratively:

- **Figure Agent**: Extracts visual elements using DocLayout-YOLO, generates descriptions via MLLM, and validates visual-caption pairings
- **Section Agent**: Analyzes paper structure, generates cohesive textual content with strategic integration of visual references
- **Orchestrate Agent**: Combines visual and textual components into HTML/CSS posters following principles of content-structure decoupling, institutional identity alignment, and responsive layout

Each agent operates with dedicated checker modules enabling iterative refinement to ensure output quality.



P2P-INSTRUCT Dataset

P2P-INSTRUCT comprises 30,460 high-quality instruction-response pairs spanning the complete poster generation workflow:

- 16,848 figure-description pairs averaging 192 tokens per visual element
- 13,612 instruction-response pairs from Section Generator, Content Generator, and HTML Generator components averaging 3,300 tokens per response

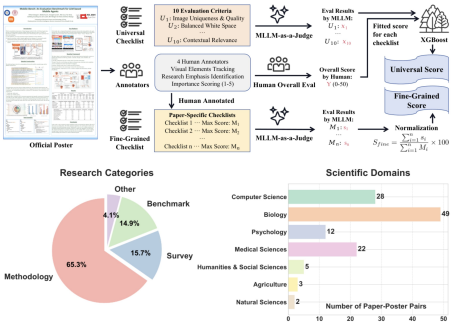
This dataset supports fine-tuning models like Qwen3-P2P-8B for enhanced poster generation capabilities.

P2PEVAL Benchmark

P2PEVAL includes 121 paper-poster pairs from diverse scientific domains with 1,738 fine-grained checklist items. The evaluation framework features:

- **Universal Poster Evaluation**: Assesses 10 criteria (authorship accuracy, image quality, white space balance, etc.) using LLM-as-a-Judge validated against human annotations
- **Fine-Grained Poster Evaluation**: Measures fidelity to specific content using human-annotated checklists with importance-based scoring (1-5)

XGBoost combines these evaluations to produce comprehensive quality metrics ($R^2 = 0.92$).



Experimental Setup

Experiments evaluated P2P against various MLLMs on P2PEVAL, including Claude, GPT, Qwen, InternVL, Gemma, and Deepseek models. All models used temperature=1 and 8000 maximum tokens. P2P was compared to Tencent's YuanBao application and original human-created posters. Metrics included ROUGE scores, BERTScore, LLM-as-a-Judge preferences, and P2PEVAL's Universal and Fine-Grained evaluations. Qwen3-P2P-8B and Qwen2.5-VL-P2P-7B were fine-tuned on P2P-INSTRUCT for 3 epochs.

(a) Designed by P2P (b) Designed by YuanBao (c) Designed by original author

Results and Analysis

Key findings from experiments:

- **Model Performance**: Claude-3.7-Sonnet achieved highest qualitative scores (FineGrain: 65.40, Universal: 37.25); Qwen3-P2P-8B led in ROUGE scores through fine-tuning on P2P-INSTRUCT
- **Human Preference**: P2P was preferred over YuanBao (54.35%) and original posters (35.59%) in pairwise evaluations
- **Output Format**: HTML consistently outperformed SVG and LaTeX formats across all metrics
- **Ablation Study**: Removing individual components (Multi-Agent, Figure Describer, Reflection) progressively degraded performance, confirming the value of each element

Model	Size	ROUGE-1	ROUGE-2	ROUGE-L	BERT[1]	Judge[2]	FineGrain[3]	Universal[4]
Closed-Source Models								
Claude-3.7-Sonnet	🔒	0.2745	0.0830	0.2527	0.8109	0.5537	65.3962	**37.2474**
Claude-3.7-Sonnet^R	🔒	0.2734	0.0848	0.2516	0.8111	**0.6281**	**65.8848**	35.5062
Claude-3.5-Sonnet	🔒	0.2367	0.0615	0.2185	0.8081	0.2810	47.7385	30.2544
GPT-4.1-2025-04-14	🔒	0.2459	0.0685	0.2281	0.8113	0.4793	60.2879	34.4700
GPT-4.1-mini-2025-04-14	🔒	0.2616	0.0741	0.2407	0.8125	0.3388	55.3493	31.0697
GPT-4.1-nano-2025-04-14	🔒	0.2169	0.0557	0.1990	0.8070	0.2066	41.3446	27.7149
GPT-4o-2024-11-20	🔒	0.2395	0.0668	0.2217	0.8114	0.4959	55.4380	34.3888
GPT-4o-mini-2024-07-18	🔒	0.2362	0.0732	0.2198	**0.8167**	0.2314	48.8879	30.8409
OpenAI-o1^R	🔒	0.2385	0.0611	0.2200	0.8088	0.3103	56.8504	34.1659
Seed1.5-VL^R	🔒	0.2160	0.0539	0.2026	0.8041	0.4050	62.4702	33.9840
Seed-Thinking-v1.5^{RT}	🔒	0.2357	0.0701	0.2210	0.8113	0.4711	61.9632	34.6882
Seed-Thinking-v1.5-m^R	🔒	0.2493	0.0767	0.2315	0.8116	0.3719	57.1457	33.2461
Doubao-1.5-vision-pro	🔒	0.2586	0.0849	0.2409	0.8089	0.0354	45.9282	14.0841
YuanBao[5]	🔒	-	-	-	-	0.0083	57.8677	31.5754
6B+ Models								
InternVL3	8B	0.1980	0.0618	0.1847	0.7994	0.0776	33.3900	22.2245
Qwen3^T	8B	0.2563	0.0859	0.2373	0.8152	0.1835	45.0272	28.8107
Qwen3^{RT}	8B	0.2231	0.0619	0.2082	0.8125	0.2545	53.6611	32.4912
Qwen2.5-VL	7B	0.1090	0.0414	0.1020	0.7645	0.0083	13.7417	13.0597
12B+ Models								
Gemma-3	12B	0.2411	0.0764	0.2250	0.8096	0.0940	46.7903	27.3686
InternVL3	14B	0.2437	0.0736	0.2253	0.8132	0.0756	45.5513	25.6062
27B+ Models								
Gemma-3	27B	0.2500	0.0794	0.2346	0.8133	0.2857	50.8931	28.7410
Gemma-3^T	27B	0.2536	0.0853	0.2372	0.8132	0.2417	52.1716	28.5901
InternVL3	38B	0.2440	0.0756	0.2258	0.8143	0.2333	52.6634	29.5850
Qwen3^{RT}	3/30B	0.2270	0.0637	0.2125	0.8120	0.2562	52.2125	31.1930
Qwen3^{RT}	32B	0.2314	0.0659	0.2168	0.8090	0.1736	46.0383	28.9479
Qwen2.5-Coder^T	32B	0.2666	0.0949	0.2487	**0.8167**	0.3884	55.9441	32.7935
72B+ Models								
Deepseek-R1^{RT}	37/671B	0.1927	0.0461	0.1795	0.8015	0.5333	62.5013	33.9701
Deepseek-V3^T	37/671B	0.2371	0.0739	0.2232	0.8124	0.5041	59.6805	33.6045
InternVL3	78B	0.2424	0.0789	0.2245	0.8152	0.2773	51.2962	28.9230
Qwen3^{RT}	22/235B	0.2278	0.0625	0.2141	0.8077	0.3967	53.7927	31.4551
Qwen2.5-VL	72B	0.2577	0.0909	0.2400	0.8148	0.2833	55.7929	32.3105
Qwen3-P2P^{FT}	8B	**0.2882**	**0.0955**	**0.2675**	0.8135	0.4587	57.6622	32.4996
Qwen2.5-VL-P2P^T	7B	0.1939	0.0609	0.1797	0.7926	0.3140	37.3078	25.0337

Comparison (A vs. B)	Preferred or Tied (%)	Preferred (%)
P2P / YuanBao	83.05	54.35
P2P / Original	57.63	35.59
YuanBao / Original	20.34	12.40

Mutli Agent	Figure Describer	Reflection	FineGrain	Universal
✓	✓	✓	**65.3962**	**37.2474**
✓	✓		64.4556	34.2229
✓		✓	63.7388	35.1107
✓			63.5806	33.1458
			60.7233	34.2554

Layout Analysis

Comparative analysis of layouts generated by Claude and GPT models revealed distinct structural tendencies:

- Claude produced more fragmented structures (376 vs 293 columns) with greater height variation (CV: 0.21 vs 0.18)
- Claude's layouts showed taller maximum columns (7272px vs 5794px) and higher blank space proportion (19.16% vs 14.92%)
- GPT generated more uniform and compact layouts with better balance

These differences highlight challenges in achieving consistent content allocation across poster layouts.

Layout Statistic	Claude	GPT	Layout Statistic	Claude	GPT
General			**Tallest Column**		
Total columns	376	293	Height (px)	7272.22	5794.42
Balance			Text length (char)	2057.37	1554.44
Relative position[1]	0.55	0.51	Number of images	2.93	2.30
Height coefficient of variation[2]	0.21	0.18	**Shortest Column**		
Height ratio (max/min)[3]	1.73	1.61	Height (px)	4379.82	3979.53
Blank space proportion[4]	19.16%	14.92%	Text length (char)	1392.26	1296.84
			Number of images	1.74	1.59

Figure 17: The vertical poster for this paper, powered by P2P.

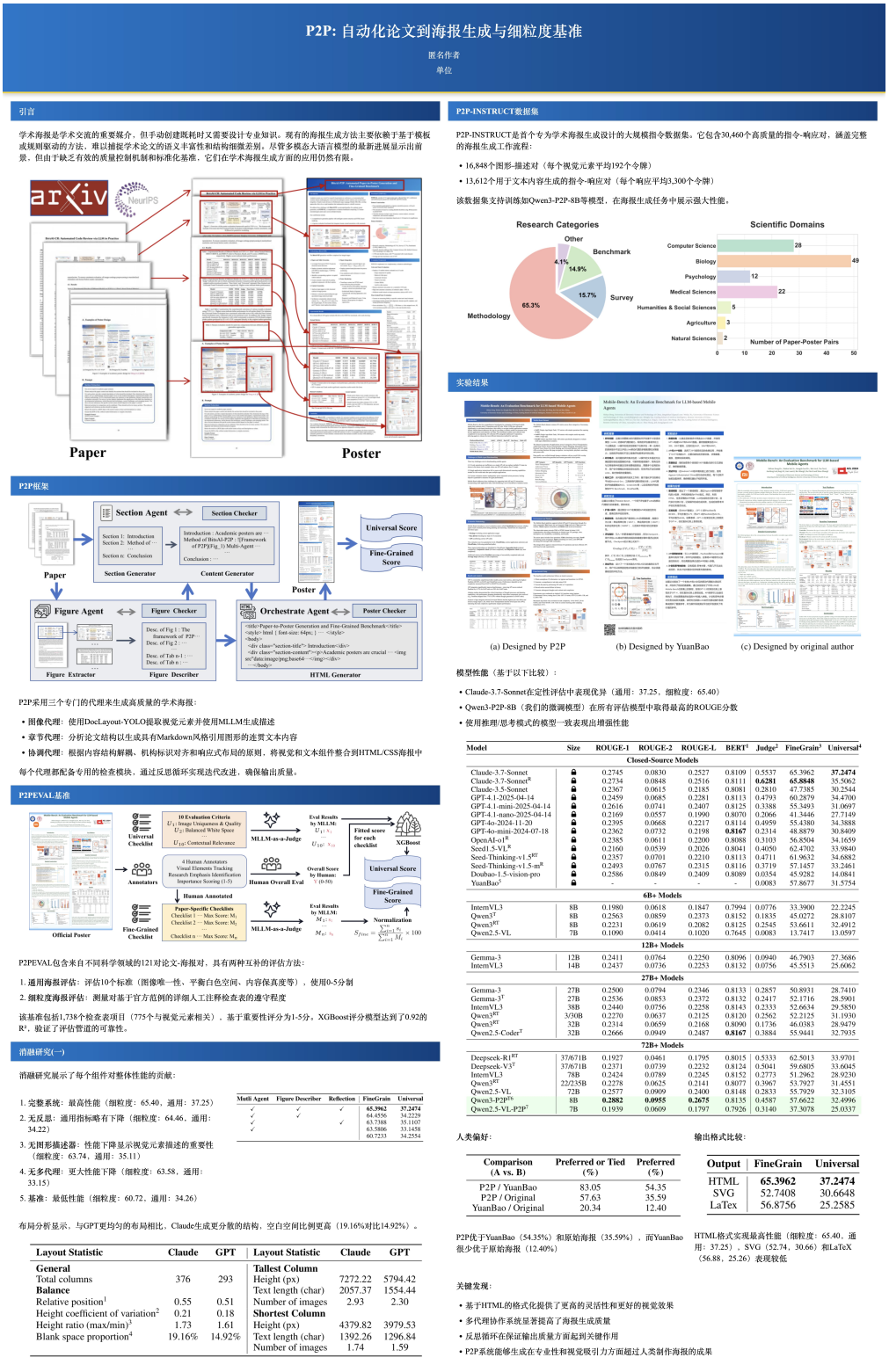


Figure 18: The poster for this paper in another language (Chinese), powered by P2P.