LOCR: Location-Guided Transformer for Optical Character Recognition

Anonymous ACL submission

Abstract

Academic documents are packed with texts, equations, tables, and figures, requiring comprehensive understanding for accurate Optical Character Recognition (OCR). While end-to-end OCR methods offer improved accuracy over layout-based approaches, they often grapple with significant repetition issues, especially with complex layouts in Out-Of-Domain (OOD) documents. To tackle this, we propose LOCR[1], a model integrating location guiding into the transformer architecture during autoregression, training on a dataset comprising over 77M text-location pairs from 125K academic document pages, including bounding boxes for words, tables and mathematical symbols. LOCR adeptly handles various formatting elements and generates content in Markdown language. It outperforms existing models in our testset with an edit distance of 0.125, BLEU of 0.827 and F1 of 0.897. LOCR also reduces repetition frequency from 51% to 2% in the arXiv dataset and from 56% to 7% in OOD documents. Additionally, LOCR features an interactive OCR mode, facilitating the generation of complex documents through a few location prompts from human.

1 Introduction

Academic literature comprises a wealth of high-quality content, yet much of it is provided in formats like PDF that are not machine-readable. Particularly, most academic documents of the previous centuries are scanned version. Digitizing academic documents are important for scientific research, literature retrieval, and large-language model training. However, academic document layout tends to be highly intricate, including text, equations, images, tables, and annotations, posing challenges for obtaining accurate OCR results.

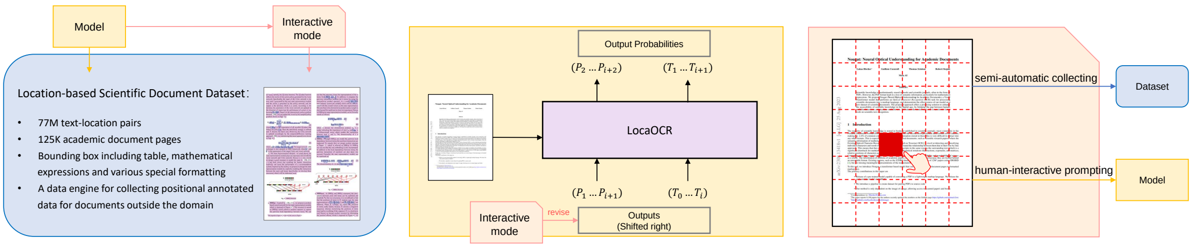
One approach to document OCR is to first analyze the layout of the document and then extract the text content (Zhu et al., 2022,mindee, 2023). While progresses has been made in any of the two stages or handling specific types of elements, such as table detection and recognition (Yang et al., 2022), handwritten formula recognition (Sakshi and Kukreja, 2023) and structured information extraction (Lu et al., 2022; Liao et al., 2023), it is very difficult for models to understand all the elements in an academic document and connect the different chunks into a coherent sequence.

Recently, an end-to-end transformer structure, Donut (Kim et al., 2022), was proposed for document understanding. It effectively address the complexity of combining multiple models and the issue of error propagation. Without to much changes in the model, Nougat (Blecher et al., 2023) processes academic PDFs into markup language. However, such methods are prone to hallucination and repetitions.

In fact, getting trapped in a repetitive loop is a common problem with Transformer-based models sampling with greedy search decoding (Holtzman et al., 2019). Thus, it is challenging for an autoregressive language model to accurately capture all the content of text-intensive documents. To make full use of the positional information in various layouts and address the issue of repetitive loop, we introduce LOCR, a location-guided document understanding model, together with an original large-scale dataset and an interactive OCR mode to align with human intension (see Figure 1 for an overview).

The most significant feature that distinguishes our model from previous works is the incorporation of positional autoregression alongside text autoregression. Different from two-stage OCR, LOCR simultaneously predicts the current token and the position of the next token, which is used to prompt

[1] Source codes and datasets will be available upon publication



(a) **Data**: dataset & data engine (b) **Model**: location-guided transformer (c) **Interactive**: align with human intent

Figure 1: An overview of three components of our work: a ***large-scale dataset*** with positional annotation and a data engine, a ***location-guided OCR model*** for various layouts, and an ***interactive mode*** for humans to prompt the model and modify data collection.

the decoding of the next token. Taking document images as input, our model output document content in Markdown format, including special formats such as superscripts and subscripts.

Besides, we introduce effective strategies for anti-repetition. With positional supervision, we perform importance decay on positions that have been visited during the autoregressive process or are blank in the image. The repetition behavior decreases from 51% of documents to 2% in the arXiv test set, and from 56% to 7% for out-of-domain documents. For documents with complex layouts, we also introduce an interactive OCR mode. In this mode, the model would continue to decode the text where the user has dragged a box. With these enhancement strategies, the generation ability of the model is significantly improved.

Additionally, we propose a data engine for constructing academic document OCR dataset with positional annotations. We collect a large-scale dataset of 125K academic document pages with 77M text-location pairs. To the best of our knowledge, it is the first dataset including bounding box of each mathematical symbol in academic documents.

In summary, the main contributions of this paper are:

- We introduce LOCR, a transformer-structured OCR model with positional supervision. Our model achieves the state-of-the-art score in academic document understanding task in arXiv test set (see Section 5.2) and alleviates the repetitive degradation problem to a great extent (see Section 5.3).
- We innovatively introduce an interactive OCR mode, enabling the model to handle any out-of-domain documents. Humans only need to provide the position box for the next word without any cumbersome operations (see Section 5.4).
- We will release a large-scale dataset composed of 125K pages of academic documents. Each piece of data contains a page image, the corresponding texts in Markdown format, and the bounding boxes of all words and mathematical symbols (see Section 3).

2 Related Work

2.1 General-purpose OCR

Optical Character Recognition (OCR) caters to a diverse array of applications, including document digitization (Smith, 2007; Moysset et al., 2017), handwriting recognition, and scene text recognition (Bautista and Atienza, 2022; Hernandez Diaz et al., 2021; Li et al., 2021). The classic OCR methods consist of two stages: text detection and text recognization. The text detection algorithm obtains the position of text boxes from the image, and then the recognition algorithm recognizes the content within the text boxes. Researches in these sub-fields has achieved satisfactory results, such as EAST (Zhou et al., 2017) for text detection, CRNN (Shi et al., 2015) for text recognization, and LayoutLM family (Xu et al., 2019; Xu et al., 2020; Huang et al., 2022) for document element identification. There also has been various intergrated toolbox to connect above functions, such as DocXChain (Yao, 2023) and EffOCR (Bryan et al., 2023).

2.2 Academic document OCR

For academic document understanding, additional tasks like table and mathematical equation parsing are also involved. Marker (Paruchuri and Lampa, 2023) is a pipeline of text extracting, layout detection and blocks combination, which converts PDF, EPUB, and MOBI to Markdown with a series of deep learning models. Such OCR-based approaches have shown promising performance but suffer from complexity and error propagation to the subsequent process. To address this issue, document understanding models based on transformer structure were proposed. Donut (Kim et al., 2022) is an encoder-decoder model that directly decode the expected sequences from visual inputs. Nougat (Blecher et al., 2023) is a specific model trained on academic documents to process academic PDFs into markup language. It combines an image encoder and a token decoder, with the ability to parse images of math equations and tables.

With the emergence of general large model, some Large Vision-Language Models (LVLMs) marks a significant milestone across a range of OCR tasks. MEGVII proposed Vary (Wei et al., 2023), a document parsing method by scaling up the vision vocabulary of LVLMs, equipping the large model with the fine-grained perception and understanding. As the state-of-the-art multimodality model, GPT-4v (Yang et al., 2023) performs well in recognizing and understanding Latin contents. But it shows limitations when dealing with complex tasks such as table structure recognition and end-to-end semantic entity recognition (Shi et al., 2023). When it comes to unstructured layouts or inconsistent text distribution, GPT-4v tends to omit lengthy tables and only reconstruct the short beginning of that.

Without the box detection of two-stage OCR, such methods are prone to hallucination and repetitions. This phenomenon indicates that it is crucial for the model to find the correct position in order to generate the correct sequences, especially for ambiguous layouts and out-of-domain documents.

2.3 Promptable model

Interactive models play a significant role in aligning behavior of artifical intelligence with human intentions, which have shown promising performance within a variety of domains. SAM(Kirillov et al., 2023) presents an interactive segmentation model capable of accommodating point, box, and text-based input. DINOv (Li et al., 2023) achieves visual in-context prompting in both referring and general segmentation. T-Rex (Jiang et al., 2023) explores object detection and counting, which can interactively refine the counting results by prompting on missing or falsely-detected objects. In contrast, the field of OCR revolves less interactive explorations, despite the dealing with complex layout has an urge for human prompts and interactions.

3 Dataset

3.1 Data collection

To the best of our knowledge, there is no paired dataset containing markup-formatted document contents along with corresponding bounding boxes (bbox) for each word and mathematical symbol. We proposed a data engine to collect such paired data. The process is shown in Figure 2.

We get the Tex source files of academic papers from arXiv. In the first step, we assign a unique RGB color identifier to each word or mathematical symbol automatically by using xcolor package in LaTeX (see Step1). In the second step, we follow the same pipeline as Nougat (Blecher et al., 2023) and compile LaTeX files into PDF and Markdown files respectively. Since PDF is a rich text format that supports color changes, we obtain colorful PDF files. Meanwhile Markdown is a plain text format and the RGB identifiers are compiled into text forms (see Step2). In the third step, we use the PyMuPDF package of python to parse the colorful PDF files and extract the pair of (color, bbox). At the same time, we parse the Markdown file with regular expressions to get the paired (color, text) data. Finally, we merge the two pairs of data by the key of RGB color to get paired (text, bbox) data (see Step3).

We collected academic papers released on arXiv from 2007 to 2023. During data processing, some articles failed the conversion due to user-defined macros or non-standardized formats. After all conversion and data cleaning, our dataset is composed of 125738 pages, which include, but are not limited to, the bounding box of plain text, Greek letters, arithmetic symbols, superscripts, subscripts, and tabular symbols. For invisible Markdown symbols like title symbols or line breaks, we assign the position of the next visible token to them. Examples of our dataset is available in Appedix A1.

3.2 Data augmentation

Image augmentation To simulate the imperfections and variability of scanned documents, we follow (Simard et al., 2003) to apply data augmen-

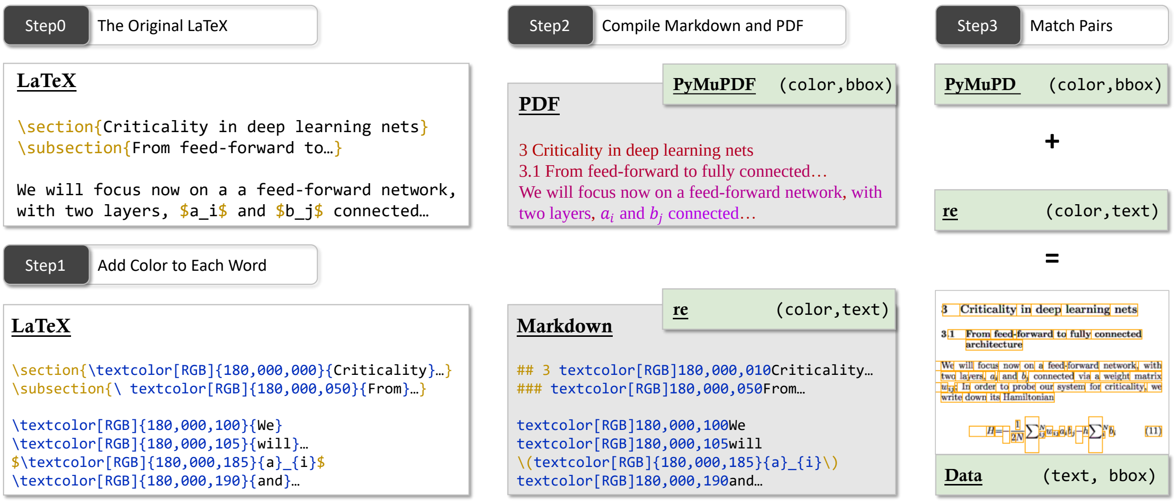


Figure 2: Data Processing. Step1: Add a unique RGB identifier to each word by parsing the Tex file. Step2: Convert source file into Markdown and PDF formats respectively. Step3: Extract color-bbox pairs from colored PDF, color-text pairs from Markdown, and merge the two to get the text-bbox pairs.

tation to document images, including of erosion, dilation, gaussian noise, gaussian blur, bitmap conversion, image compression, grid distortion and elastic transform. Each of the transformations is applied with a certain probability.

Text augmentation To address the issue of the model getting stuck in repetitive loops, we randomly skip 0 to 5 tokens and their corresponding positions in the ground truth labels. Different from the perturbation method in Nougat, which randomly replaces tokens rather than skip tokens, our method shows a more pronounced effect (see Section 5.3).

Position augmentation Since bounding boxes are involved in the autoregressive process, there may be some imprecise output. In some cases, a user may also draw a loose box in the interactive mode. Therefore, it is reasonable to add noise to the bounding boxes during the training phase. We add Gaussian noise with a mean of 0 and a standard deviation of 0.5 times the side length to each box.

4 Methodology

4.1 Model structure

The over view of our model is shown in Figure 3, with a transformer-based backbone and an additional prompt module to process positional information. Given an image as input, the image encoder transforms it as image embedding. Semantic information and visual information are integrated within the decoder, enabling simultaneous prediction of the current token and its next position.

Backbone Theoretically, our prompt module can be applied to any multimodal models with transformer structure, consisting of an image encoder and a text decoder. When no positional information is provided, the backbone model would autonomously generate sequences. In this paper, we choose Nougat (Blecher et al., 2023) as the backbone, which uses the implementation of Swin Transformer (Liu et al., 2021) as image encoder and mBART (Lewis et al., 2019) as decoder. Given an image of $x \in R^{3,H_0,W_0}$, the image encoder transfers it into dense embedding $h_{img} \in R^{H,W,d}$, which is then decoded into a sequence of token embeddings $h_t \in R^d$. Finally, the sequence of token embeddings is projected into a logit matrix with the size of the vocabulary v.

Prompt Module Without location guiding, the backbone model may get confused about where to find the next token. The prompt module is designed to perceive spatial information prompted by previous steps or human, so that the model can find the next token successfully. The prompt module consists of two-dimensional positional encoding and position detection heads.

We opt for positional encodings with Fourier Features (Tancik et al., 2020) to represent the positions of both token bounding boxes and the image. The token bounding box, defined by its top-left and bottom-right corners, is transformed into a dense position embedding $h_{box} \in R^d$. For the image em-

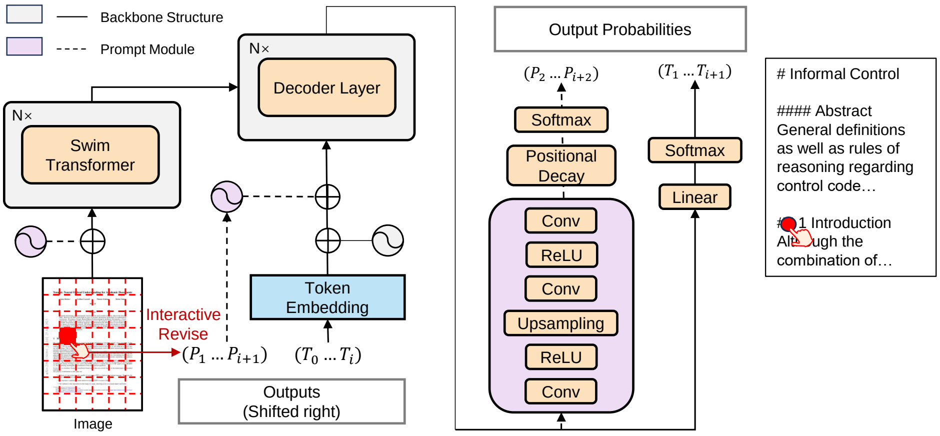


Figure 3: Model Architecture. Left: Image encoder and decoder of transformer structure. Right: Position detection head and token projection. Purple: Prompt module consisting of positional encodings and position detection head. Red: Interactive mode with human-reviewed input.

bedding $h_{img} \in R^{H,W,d}$, we divide it into grids of size (H, W) (shown in Figure 3), and apply positional encodings to each grid box.

The position detection heads are used to predict the position of the next token. Given that the weights of the cross-attention layers indicate the similarity between image grids and the current token, we utilize them as input for position detection. Inspired by CenterNet (Duan et al., 2019), an effective object detection algorithm, we use three convolutional heads with similar structure to predict the position of a token. The first convolution head conducts a classification task to find the grid containing the next token. The second and third convolution heads regress the size and center offset of the next bounding box respectively. Finally, the coordinates of the bounding box are calculated based on the center point and the width and height. To improve prediction accuracy, we upsample the image grid output by decoder, allowing finer-grained posititiion prediction.

Information fusion The token information and spatial information is fused in cross-attention layers of decoder. In backbone models without prompt module, the cross-attention layers take solely image embedding as encoder hidden states and solely token embedding as hidden states input. Instead, we use the sum of the image embedding $h_{img} \in R^{H,W,d}$ and its position embedding $H_{grid} \in R^{H,W,d}$ as the encoder hidden states, and the sum of token embedding $h_t \in R^d$ and position embedding $h_{box} \in R^d$ as the hidden states input. As a consequence, in cross-attention layers where token information interacts with the image contents, the positional information of tokens also interacts with that of the image.

4.2 Decay strategy for anti-repetition

During the inference stage, we introduce position decay strategies based on prior knowledge to guide the prediction of positions.

Accumulation Decay The core of the accumulation decay strategy is to record the count of tokens that have appeared in each grid. When the position detection head predicts subsequent positions, grids where many tokens have already been located will be penalized with a decay rate. The heatmap for predicting the next grid is adjusted as follows:

$$hm = hm + log(\sigma) \cdot cnt \tag{1}$$

Where the $\sigma \in (0, 1]$ denotes decay rate and cnt is the accumulative counts. When σ is set to 1, the decay function is deactivated. Smaller σ value means stronger decay effect. We recommend using a decay rate between 0.75 and 0.95, depending on the density of text in the target documents and the formatting style.

Blank Decay Another intuitive idea is to apply positional decay to blank grids. We calculate the standard deviation for pixels within each grid, where grids with smaller standard deviations (in extreme cases, containing no characters at all) are considered less likely to contain the next token. Together with blank decay strategy, the heatmap is adjusted as follows:

$$hm = hm + log(\sigma) \cdot cnt + log(\eta \cdot std) \quad (2)$$

4.3 Loss function

Our loss function consists of two parts: token loss and position loss.

Token loss We use the cross-entropy loss of tokens to train the language decoder.

Position loss For the three convolutional heads in the position detection module, we apply cross-entropy loss to the first classification head and supervise the subsequent two heads using the Intersection over Union (IOU) metric. Additionally, we integrate the normalized Euclidean distance between the center of the predicted box and that of the target box to mitigate the shortcomings of slow convergence and inaccurate regression inherent in IOU (Zheng et al., 2019). The position loss function is as follows:

$$L_p = \alpha L_p^{ce} + \beta(1 - iou + \gamma d^2) \quad (3)$$

As the prediction of the text at the beginning of a page is much more challenging and important, we assigned a higher weight θ for the initial text than the subsequent text.

The final loss function is as follows:

$$l = \theta(L_p^{init} + L_t^{init}) + L_p^{sub} + L_t^{sub} \quad (4)$$

4.4 Human interaction

As a complement to our location-guided OCR method, we provide an interactive mode, which serves both for improving the model's performance and as a part of our data construction engine.

Model Assistant In the interactive mode, We provide a browser-based tool to enable users to give real-time position prompts by simply dragging a box. LOCR takes the i-th token and the $(i + 1)$-th position as inputs, simultaneously predicting the $(i + 1)$-th token and the $(i + 2)$-th position. When the autoregressive process encounters a state of confusion, characterized by a predicted token or position confidence lower than a predetermine threshold, users can opt to provide a positional prompt. With the correct position provided, the autoregressive process would go on more smoothly (see Section 5.4 for results).

Data construction With the model automatically predicting positions, minimal human intervention is required to acquire additional out-of-domain data. Our positional encoding and detection modules can smoothly convert the bounding box between human-readable coordinate representations and machine-friendly dense embedding, making the idea easy to implement. This paves the way for broader applications of location-based OCR method.

5 Result and Evaluation

5.1 Implementation details

Baseline We use both the state-of-the-art intergrated toolbox Marker and end-to-end generation model Nougat as our baselines.

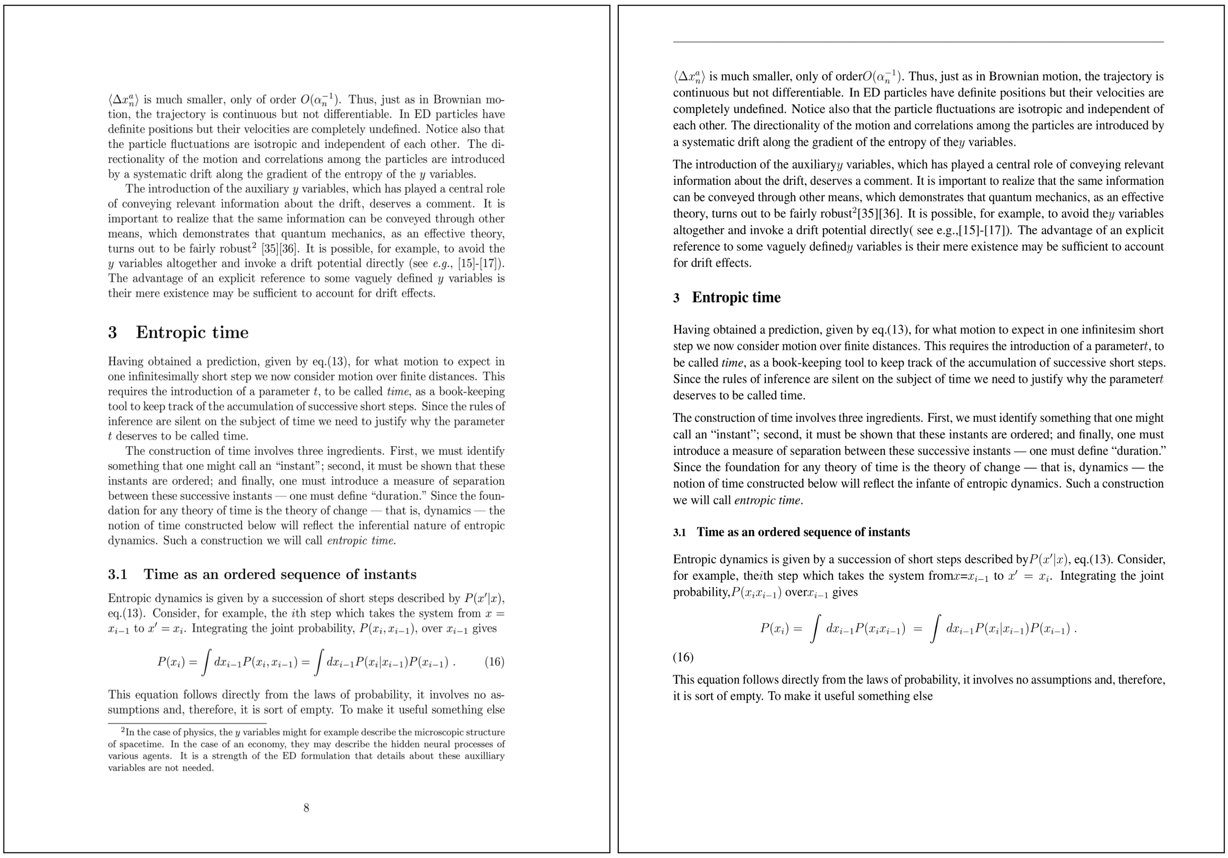
Dataset Since our main baseline model, Nougat, does not provide an open resource dataset, we evaluate our method with the dataset introduced in Section 3, which shares the same data source and processing pipeline as Nougat. The test set contains 1000 pages of academic documents and each piece of data consists of a triplet (image, text, bounding box). In the testing phase, only images are used as inputs, while the text and bounding boxes serve solely for evaluating model performance. Therefore, our evaluation method is fair and reasonable.

Setup We resize the input dimensions of the images to (H_0, W_0) = (896, 672), an aspect ratio that accommodates the majority of academic paper sizes. The maximal sequence length of transformer decoder is set to 4096 to allow the output of intensive text in academic research papers. During inference the text is generated using greedy decoding.

Training details We initialize the backbone parameters using the pretrained Nougat small model, while the prompt module is initialized randomly. Our model has 248 M parameters and was trained for three days using 128 A100 80GB GPUs, with a total batch size of 256. The maximum learning rate is set to 5×10^{-4}, with exponential decay until reaching 1×10^{-5}.

5.2 Metrics

Following Nougat (Blecher et al., 2023), we use Edit distance, BLEU, METEOR, Precision, Recall and F-measure to characterize the quality of output



$\langle \Delta x_n^a \rangle$ is much smaller, only of order $O(\alpha_n^{-1})$. Thus, just as in Brownian motion, the trajectory is continuous but not differentiable. In ED particles have definite positions but their velocities are completely undefined. Notice also that the particle fluctuations are isotropic and independent of each other. The directionality of the motion and correlations among the particles are introduced by a systematic drift along the gradient of the entropy of the y variables.

The introduction of the auxiliary y variables, which has played a central role of conveying relevant information about the drift, deserves a comment. It is important to realize that the same information can be conveyed through other means, which demonstrates that quantum mechanics, as an effective theory, turns out to be fairly robust[2] [35][36]. It is possible, for example, to avoid the y variables altogether and invoke a drift potential directly (see *e.g.*, [15]-[17]). The advantage of an explicit reference to some vaguely defined y variables is their mere existence may be sufficient to account for drift effects.

3 Entropic time

Having obtained a prediction, given by eq.(13), for what motion to expect in one infinitesimally short step we now consider motion over finite distances. This requires the introduction of a parameter t, to be called *time*, as a book-keeping tool to keep track of the accumulation of successive short steps. Since the rules of inference are silent on the subject of time we need to justify why the parameter t deserves to be called time.

The construction of time involves three ingredients. First, we must identify something that one might call an "instant"; second, it must be shown that these instants are ordered; and finally, one must introduce a measure of separation between these successive instants — one must define "duration." Since the foundation for any theory of time is the theory of change — that is, dynamics — the notion of time constructed below will reflect the inferential nature of entropic dynamics. Such a construction we will call *entropic time*.

3.1 Time as an ordered sequence of instants

Entropic dynamics is given by a succession of short steps described by $P(x'|x)$, eq.(13). Consider, for example, the ith step which takes the system from $x = x_{i-1}$ to $x' = x_i$. Integrating the joint probability, $P(x_i, x_{i-1})$, over x_{i-1} gives

$$P(x_i) = \int dx_{i-1} P(x_i, x_{i-1}) = \int dx_{i-1} P(x_i|x_{i-1}) P(x_{i-1}) \ . \qquad (16)$$

This equation follows directly from the laws of probability, it involves no assumptions and, therefore, it is sort of empty. To make it useful something else

[2]In the case of physics, the y variables might for example describe the microscopic structure of spacetime. In the case of an economy, they may describe the hidden neural processes of various agents. It is a strength of the ED formulation that details about these auxilliary variables are not needed.

8

$\langle \Delta x_n^a \rangle$ is much smaller, only of order$O(\alpha_n^{-1})$. Thus, just as in Brownian motion, the trajectory is continuous but not differentiable. In ED particles have definite positions but their velocities are completely undefined. Notice also that the particle fluctuations are isotropic and independent of each other. The directionality of the motion and correlations among the particles are introduced by a systematic drift along the gradient of the entropy of the$y$ variables.

The introduction of the auxiliary$y$ variables, which has played a central role of conveying relevant information about the drift, deserves a comment. It is important to realize that the same information can be conveyed through other means, which demonstrates that quantum mechanics, as an effective theory, turns out to be fairly robust[2][35][36]. It is possible, for example, to avoid the$y$ variables altogether and invoke a drift potential directly(see e.g.,[15]-[17]). The advantage of an explicit reference to some vaguely defined$y$ variables is their mere existence may be sufficient to account for drift effects.

3 Entropic time

Having obtained a prediction, given by eq.(13), for what motion to expect in one infinitesim short step we now consider motion over finite distances. This requires the introduction of a parameter$t$, to be called *time*, as a book-keeping tool to keep track of the accumulation of successive short steps. Since the rules of inference are silent on the subject of time we need to justify why the parameter$t$ deserves to be called time.

The construction of time involves three ingredients. First, we must identify something that one might call an "instant"; second, it must be shown that these instants are ordered; and finally, one must introduce a measure of separation between these successive instants — one must define "duration." Since the foundation for any theory of time is the theory of change — that is, dynamics — the notion of time constructed below will reflect the infante of entropic dynamics. Such a construction we will call *entropic time*.

3.1 Time as an ordered sequence of instants

Entropic dynamics is given by a succession of short steps described by$P(x'|x)$, eq.(13). Consider, for example, the$i$th step which takes the system from$x$=x_{i-1} to $x' = x_i$. Integrating the joint probability,$P(x_i x_{i-1})$ overx_{i-1} gives

$$P(x_i) = \int dx_{i-1} P(x_i x_{i-1}) \ = \int dx_{i-1} P(x_i|x_{i-1}) P(x_{i-1}) \ .$$

(16)

This equation follows directly from the laws of probability, it involves no assumptions and, therefore, it is sort of empty. To make it useful something else

Figure 4: Examples of our model output. Left: Origin image of document page. Right: Model output converted to Markdown and rendered back into a PDF. More detailed examples are available in Appendix B

text.

As shown in Table 1, while the number of LOCR's parameters is only slightly more than the small version of Nougat, our model outperforms the base version of Nougat in all evaluation metrics when the decay rate is set to 0.85. In contrast, Marker, as a multi-stage pipeline, dose not convert all equations to LaTeX and not all lines are joined properly. For the autogressive method without position supervision, Nougat prones to hallucination and repetition. These results firmly demonstrate the effectiveness of LOCR model and the positional decay strategy.

Besides, we use IOU metrics to measure the performance of our prompt module. LOCR achieves a IOU score of 0.702. Our method successfully handles various layouts, including pages with multiple subfigures, tables, mathematical formulas, and references (Examples are available in Appendix B).

5.3 Repetition

We evaluate the generation ability of our model and present the frequency of repetitive degeneration in Table 2. To cover as much subject content and layout as possible, we selected 100 papers each from natural sciences (quantum physics) and social sciences (marketing), as out-of-domain test set.

As Table 2 shows, our method significantly reduces repetitions. For arXiv test set, the frequency of repetition reduces from 51.0% to 2.0%. For the out-of-domain documents with subject of quantum physics, where the document content is more challenging to comprehend, with longer and more complex formulas, the frequency of repetition reduces from 56.0% to 7.0%.

5.4 Interaction

Although the problem of repetitive degeneration has been largely alleviated, we aim to complete the remaining layouts in the interactive mode. When the model encounters a layout that is difficult to judge and the confidence of the predicted position is lower than the threshold, simply dragging

Method	Edit dist↓	BLEU↑	METEOR↑	Precision↑	Recall↑	F1↑
Marker	0.221	0.696	0.783	0.838	0.804	0.814
Nougat small (247M*)	0.209	0.789	0.851	0.887	0.874	0.867
Nougat base (348M*)	0.201	0.801	0.856	0.893	0.880	0.876
LOCR (248M*, $\sigma = 1$)	0.153	0.786	0.864	0.890	0.871	0.880
LOCR ($\sigma = 0.85$)	**0.125**	**0.827**	**0.893**	**0.898**	**0.897**	**0.897**
LOCR ($\sigma = 0.75$)	0.127	0.824	0.890	0.895	0.894	0.894

Table 1: Comparative performance results on the arXiv test set. Our LOCR method demonstrates superior performance across multiple metrics, significantly outperforming the baseline methods. Notably, LOCR with = 0.85 shows the best overall balance of high precision, recall, and F1 scores, along with the lowest edit distance and the highest BLEU and METEOR scores, confirming the effectiveness of our approach, especially when positional decay is finely tuned ($\sigma = 0.85$). The first entry for LOCR indicates performance without positional decay, illustrating the impact of this feature on the model's accuracy. * Number of parameters.

Method	arXiv	quantum	marketing
Nougat base	51.0%	56.0%	55.0%
LOCR	2.0%	7.0%	22.0%

Table 2: Robustness of LOCR across diverse domains, showcasing the significant reduction in generation failures with our LOCR model. The model exhibits an impressive decrease in repetition-induced failures, achieving a substantial improvement over the Nougat base across the arXiv, quantum, and marketing test sets. Specifically, LOCR reduces the failure rates to 2% for arXiv, 7% for quantum, and 22% for marketing documents, indicating a marked increase in reliability and accuracy in document generation tasks. These results underscore the efficacy of our model in handling complex document structures with a high degree of success.

a bounding box allows the model to automatically return to the expected position and continue outputting correct results. Interactive examples are available in

6 Discussion

In document OCR, each generated token corresponds to a specific location in the input image. In our work, we introduce LOCR, which incorporates location guiding into the language model to enhance the performance of OCR tasks. Moreover, our approach significantly mitigates the problem of repetitive loops often encountered by transformer-based models during greedy search. LOCR also allows for interactive correction in cases of errors or low confidence outputs, particularly when dealing with OOD complex layouts. Users can prompt the location interactively, guiding the model generates accurate OCR results.

We believe that LOCR and interactive tool can be applied to digitize documents from various fields with complex layouts, thereby assisting academic research, literature retrieval, and large language model training. Furthermore, the OCR datasets with location guiding can facilitate the community develops better OCR models. In turn, the interactive semi-automatic data engine can be utilized to construct datasets for fine-tuning OCR models to specific domain literature, and enhancing the generalization capability of our model. We hope this work can help the development of the area of OCR.

7 Limitations

Although the frequency of repetition has significantly mitigated, it has not been entirely eradicated. Secondly, our model hinges upon manual adjustments to the decay rate parameter. Additionally, our model encounters difficulties when the initial word on a page is incomplete, leading to imperfect handling.

References

Darwin Bautista and Rowel Atienza. 2022. Scene Text Recognition with Permuted Autoregressive Sequence Models. *arXiv e-prints*, page arXiv:2207.06966.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Tom Bryan, Jacob Carlson, Abhishek Arora, and Melissa Dell. 2023. EfficientOCR: An Extensible, Open-Source Package for Efficiently Digitizing World Knowledge. *arXiv e-prints*, page arXiv:2310.10050.

Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578.

Daniel Hernandez Diaz, Siyang Qin, Reeve Ingle, Yasuhisa Fujii, and Alessandro Bissacco. 2021. Rethinking Text Line Recognition Models. *arXiv e-prints*, page arXiv:2104.07787.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. *arXiv e-prints*, page arXiv:2204.08387.

Qing Jiang, Feng Li, Tianhe Ren, Shilong Liu, Zhaoyang Zeng, Kent Yu, and Lei Zhang. 2023. T-rex: Counting by visual prompting. *arXiv preprint arXiv:2311.13596*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Chunyuan Li, Jianwei Yang, et al. 2023. Visual in-context prompting. *arXiv preprint arXiv:2311.13601*.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. *arXiv e-prints*, page arXiv:2109.10282.

Haofu Liao, Aruni RoyChowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen Tu, Ravi Kumar Satzoda, R Manmatha, and Vijay Mahadevan. 2023. Doctr: Document transformer for structured information extraction in documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19584–19594.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. *arXiv preprint arXiv:2203.12277*.

mindee. 2023. doctr: Document text recognition. https://github.com/mindee/doctr.

Bastien Moysset, Christopher Kermorvant, and Christian Wolf. 2017. Full-Page Text Recognition: Learning Where to Start and When to Stop. *arXiv e-prints*, page arXiv:1704.08628.

Vik Paruchuri and Samuel Lampa. 2023. Marker: Convert pdf to markdown quickly with high accuracy. https://github.com/VikParuchuri/marker?tab=readme-ov-file.

Sakshi Sakshi and Vinay Kukreja. 2023. Recent trends in mathematical expressions recognition: An lda-based analysis. *Expert Systems with Applications*, 213:119028.

Baoguang Shi, Xiang Bai, and Cong Yao. 2015. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition. *arXiv e-prints*, page arXiv:1507.05717.

Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. 2023. Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation. *arXiv preprint arXiv:2310.16809*.

P.Y. Simard, D. Steinkraus, and J.C. Platt. 2003. Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 958–963.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. 2020. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *arXiv e-prints*, page arXiv:2006.10739.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2023. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2020. LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. *arXiv e-prints*, page arXiv:2012.14740.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2019. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. *arXiv e-prints*, page arXiv:1912.13318.

Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. 2022. Tableformer: Robust transformer modeling for table-text encoding. *arXiv preprint arXiv:2203.00274*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv e-prints*, page arXiv:2309.17421.

Cong Yao. 2023. DocXChain: A Powerful Open-Source Toolchain for Document Parsing and Beyond. *arXiv e-prints*, page arXiv:2310.12430.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. 2019. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *arXiv e-prints*, page arXiv:1911.08287.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. EAST: An Efficient and Accurate Scene Text Detector. *arXiv e-prints*, page arXiv:1704.03155.

Wenzhen Zhu, Negin Sokhandan, Guang Yang, Sujitha Martin, and Suchitra Sathyanarayana. 2022. DocBed: A Multi-Stage OCR Solution for Documents with Complex Layouts. *arXiv e-prints*, page arXiv:2202.01414.

A Dataset Examples

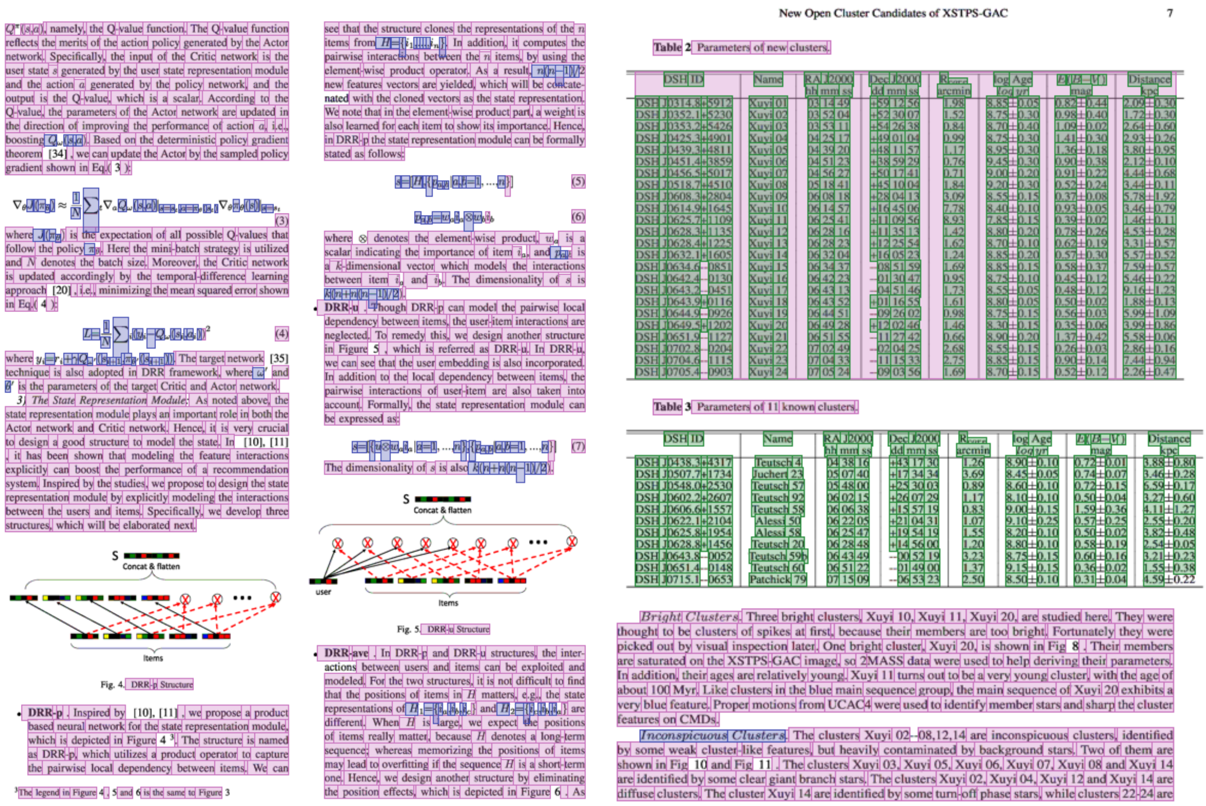
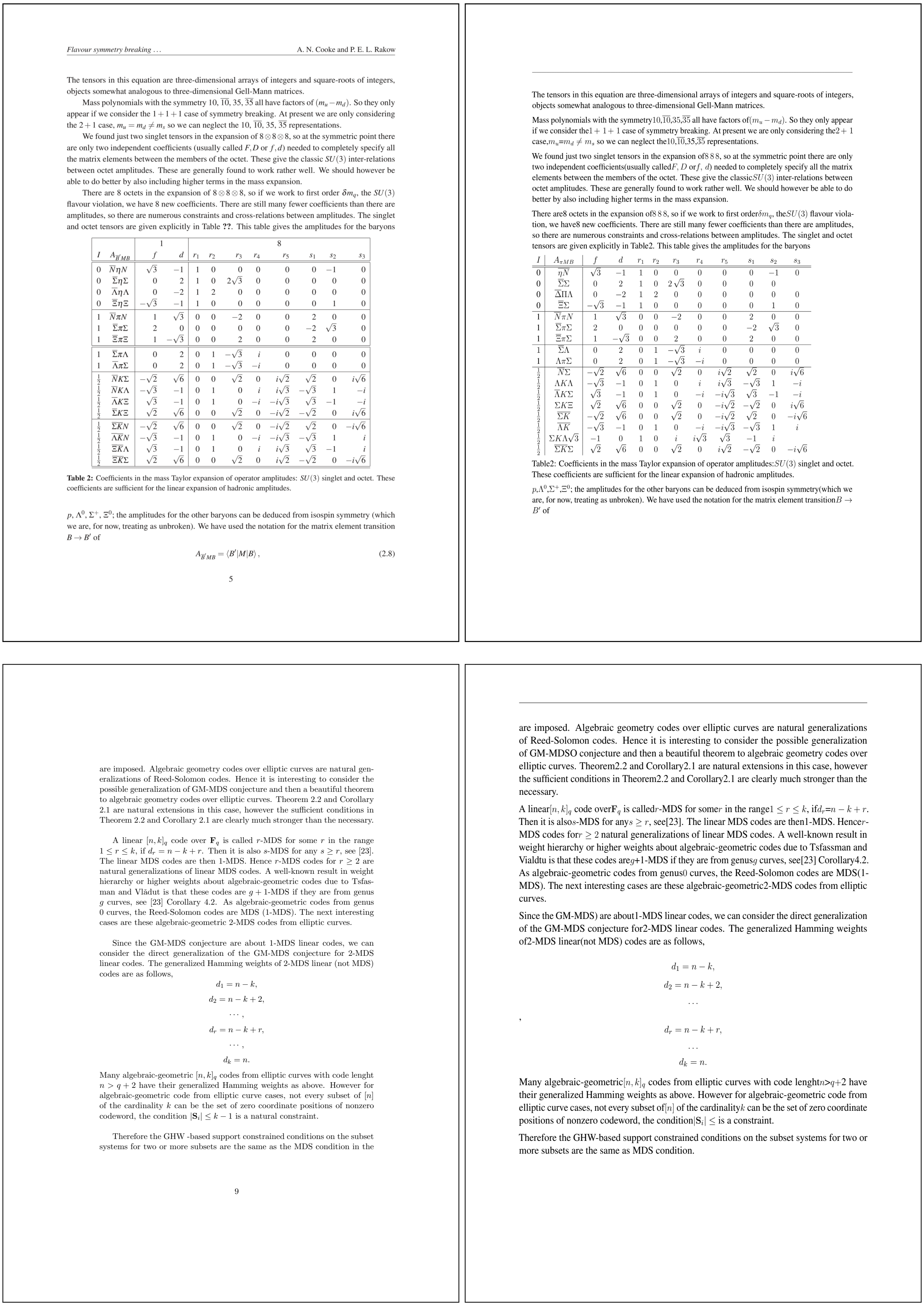


Figure A1: Dataset example. Bounding boxes of texts are highlighted in pink, mathematical expressions in blue, and tables in green.

B Output Examples

C Interactive Mode



Flavour symmetry breaking ... A. N. Cooke and P. E. L. Rakow

The tensors in this equation are three-dimensional arrays of integers and square-roots of integers, objects somewhat analogous to three-dimensional Gell-Mann matrices.

Mass polynomials with the symmetry 10, $\overline{10}$, 35, $\overline{35}$ all have factors of $(m_u - m_d)$. So they only appear if we consider the $1+1+1$ case of symmetry breaking. At present we are only considering the $2+1$ case, $m_u = m_d \neq m_s$ so we can neglect the 10, $\overline{10}$, 35, $\overline{35}$ representations.

We found just two singlet tensors in the expansion of $8 \otimes 8 \otimes 8$, so at the symmetric point there are only two independent coefficients (usually called F, D or f, d) needed to completely specify all the matrix elements between the members of the octet. These give the classic $SU(3)$ inter-relations between octet amplitudes. These are generally found to work rather well. We should however be able to do better by also including higher terms in the mass expansion.

There are 8 octets in the expansion of $8 \otimes 8 \otimes 8$, so if we work to first order δm_q, the $SU(3)$ flavour violation, we have 8 new coefficients. There are still many fewer coefficients than there are amplitudes, so there are numerous constraints and cross-relations between amplitudes. The singlet and octet tensors are given explicitly in Table **??**. This table gives the amplitudes for the baryons

		1		8							
I	$A_{\overline{B}'MB}$	f	d	r_1	r_2	r_3	r_4	r_5	s_1	s_2	s_3
0	$\overline{N}\eta N$	$\sqrt{3}$	-1	1	0	0	0	0	0	-1	0
0	$\overline{\Sigma}\eta\Sigma$	0	2	1	0	$2\sqrt{3}$	0	0	0	0	0
0	$\overline{\Lambda}\eta\Lambda$	0	-2	1	2	0	0	0	0	0	0
0	$\overline{\Xi}\eta\Xi$	$-\sqrt{3}$	-1	1	0	0	0	0	0	1	0
1	$\overline{N}\pi N$	1	$\sqrt{3}$	0	0	-2	0	0	2	0	0
1	$\overline{\Sigma}\pi\Sigma$	2	0	0	0	0	0	0	-2	$\sqrt{3}$	0
1	$\overline{\Xi}\pi\Xi$	1	$-\sqrt{3}$	0	0	2	0	0	2	0	0
1	$\overline{\Sigma}\pi\Lambda$	0	2	0	1	$-\sqrt{3}$	i	0	0	0	0
1	$\overline{\Lambda}\pi\Sigma$	0	2	0	1	$-\sqrt{3}$	$-i$	0	0	0	0
$\frac{1}{2}$	$\overline{N}K\Sigma$	$-\sqrt{2}$	$\sqrt{6}$	0	0	$\sqrt{2}$	0	$i\sqrt{2}$	$\sqrt{2}$	0	$i\sqrt{6}$
$\frac{1}{2}$	$\overline{N}K\Lambda$	$-\sqrt{3}$	-1	0	1	0	i	$i\sqrt{3}$	$-\sqrt{3}$	1	$-i$
$\frac{1}{2}$	$\overline{\Lambda}K\Xi$	$\sqrt{3}$	-1	0	1	0	$-i$	$-i\sqrt{3}$	$\sqrt{3}$	-1	$-i$
$\frac{1}{2}$	$\overline{\Sigma}K\Xi$	$\sqrt{2}$	$\sqrt{6}$	0	0	$\sqrt{2}$	0	$-i\sqrt{2}$	$-\sqrt{2}$	0	$i\sqrt{6}$
$\frac{1}{2}$	$\overline{\Sigma}\overline{K}N$	$-\sqrt{2}$	$\sqrt{6}$	0	0	$\sqrt{2}$	0	$-i\sqrt{2}$	$\sqrt{2}$	0	$-i\sqrt{6}$
$\frac{1}{2}$	$\overline{\Lambda}\overline{K}N$	$-\sqrt{3}$	-1	0	1	0	$-i$	$-i\sqrt{3}$	$-\sqrt{3}$	1	i
$\frac{1}{2}$	$\overline{\Xi}\overline{K}\Lambda$	$\sqrt{3}$	-1	0	1	0	i	$i\sqrt{3}$	$\sqrt{3}$	-1	i
$\frac{1}{2}$	$\overline{\Xi}\overline{K}\Sigma$	$\sqrt{2}$	$\sqrt{6}$	0	0	$\sqrt{2}$	0	$i\sqrt{2}$	$-\sqrt{2}$	0	$-i\sqrt{6}$

Table 2: Coefficients in the mass Taylor expansion of operator amplitudes: $SU(3)$ singlet and octet. These coefficients are sufficient for the linear expansion of hadronic amplitudes.

p, Λ^0, Σ^+, Ξ^0; the amplitudes for the other baryons can be deduced from isospin symmetry (which we are, for now, treating as unbroken). We have used the notation for the matrix element transition $B \to B'$ of

$$A_{\overline{B}'MB} = \langle B'|M|B\rangle \,, \tag{2.8}$$

5

The tensors in this equation are three-dimensional arrays of integers and square-roots of integers, objects somewhat analogous to three-dimensional Gell-Mann matrices.

Mass polynomials with the symmetry10,$\overline{10}$,35,$\overline{35}$ all have factors of$(m_u - m_d)$. So they only appear if we consider the$1+1+1$ case of symmetry breaking. At present we are only considering the$2+1$ case,m_u=$m_d \neq m_s$ so we can neglect the10,$\overline{10}$,35,$\overline{35}$ representations.

We found just two singlet tensors in the expansion of$8\,8\,8$, so at the symmetric point there are only two independent coefficients(usually called$F$, D or$f$, d) needed to completely specify all the matrix elements between the members of the octet. These give the classic$SU(3)$ inter-relations between octet amplitudes. These are generally found to work rather well. We should however be able to do better by also including higher terms in the mass expansion.

There are8 octets in the expansion of$8\,8\,8$, so if we work to first orderδm_q, the$SU(3)$ flavour violation, we have8 new coefficients. There are still many fewer coefficients than there are amplitudes, so there are numerous constraints and cross-relations between amplitudes. The singlet and octet tensors are given explicitly in Table2. This table gives the amplitudes for the baryons

I	$A_{\pi MB}$	f	d	r_1	r_2	r_3	r_4	r_5	s_1	s_2	s_3
0	$\overline{\eta N}$	$\sqrt{3}$	-1	1	0	0	0	0	0	-1	0
0	$\overline{\Sigma}\Sigma$	0	2	1	0	$2\sqrt{3}$	0	0	0	0	
0	$\overline{\Delta}\Pi\Lambda$	0	-2	1	2	0	0	0	0	0	0
0	$\overline{\Xi}\Sigma$	$-\sqrt{3}$	-1	1	0	0	0	0	0	1	0
1	$\overline{N}\pi N$	1	$\sqrt{3}$	0	0	-2	0	0	2	0	0
1	$\overline{\Sigma}\pi\Sigma$	2	0	0	0	0	0	0	-2	$\sqrt{3}$	0
1	$\overline{\Xi}\pi\Sigma$	1	$-\sqrt{3}$	0	0	2	0	0	2	0	0
1	$\overline{\Sigma}\Lambda$	0	2	0	1	$-\sqrt{3}$	i	0	0	0	0
1	$\Lambda\pi\Sigma$	0	2	0	1	$-\sqrt{3}$	$-i$	0	0	0	0
$\frac{1}{2}$	$\overline{N}\Sigma$	$-\sqrt{2}$	$\sqrt{6}$	0	0	$\sqrt{2}$	0	$i\sqrt{2}$	$\sqrt{2}$	0	$i\sqrt{6}$
$\frac{1}{2}$	$\Lambda K\Lambda$	$-\sqrt{3}$	-1	0	1	0	i	$i\sqrt{3}$	$-\sqrt{3}$	1	$-i$
$\frac{1}{2}$	$\overline{\Lambda}K\Sigma$	$\sqrt{3}$	-1	0	1	0	$-i$	$-i\sqrt{3}$	$\sqrt{3}$	-1	$-i$
$\frac{1}{2}$	$\Sigma K\Xi$	$\sqrt{2}$	$\sqrt{6}$	0	0	$\sqrt{2}$	0	$-i\sqrt{2}$	$-\sqrt{2}$	0	$i\sqrt{6}$
$\frac{1}{2}$	$\overline{\Sigma K}$	$-\sqrt{2}$	$\sqrt{6}$	0	0	$\sqrt{2}$	0	$-i\sqrt{2}$	$\sqrt{2}$	0	$-i\sqrt{6}$
$\frac{1}{2}$	$\overline{\Lambda K}$	$-\sqrt{3}$	-1	0	1	0	$-i$	$-i\sqrt{3}$	$-\sqrt{3}$	1	i
$\frac{1}{2}$	$\Sigma K\Lambda\sqrt{3}$	-1	0	1	0	i	$i\sqrt{3}$	$\sqrt{3}$	-1	i	
$\frac{1}{2}$	$\overline{\Sigma K}\Sigma$	$\sqrt{2}$	$\sqrt{6}$	0	0	$\sqrt{2}$	0	$i\sqrt{2}$	$-\sqrt{2}$	0	$-i\sqrt{6}$

Table2: Coefficients in the mass Taylor expansion of operator amplitudes:$SU(3)$ singlet and octet. These coefficients are sufficient for the linear expansion of hadronic amplitudes.

p,Λ^0,Σ^+,Ξ^0; the amplitudes for the other baryons can be deduced from isospin symmetry(which we are, for now, treating as unbroken). We have used the notation for the matrix element transition$B \to B'$ of

are imposed. Algebraic geometry codes over elliptic curves are natural generalizations of Reed-Solomon codes. Hence it is interesting to consider the possible generalization of GM-MDS conjecture and then a beautiful theorem to algebraic geometry codes over elliptic curves. Theorem 2.2 and Corollary 2.1 are natural extensions in this case, however the sufficient conditions in Theorem 2.2 and Corollary 2.1 are clearly much stronger than the necessary.

A linear $[n,k]_q$ code over $\mathbf{F}_q$ is called r-MDS for some r in the range $1 \leq r \leq k$, if $d_r = n-k+r$. Then it is also s-MDS for any $s \geq r$, see [23]. The linear MDS codes are then 1-MDS. Hence r-MDS codes for $r \geq 2$ are natural generalizations of linear MDS codes. A well-known result in weight hierarchy or higher weights about algebraic-geometric codes due to Tsfasman and Vlădut is that these codes are $g+1$-MDS if they are from genus g curves, see [23] Corollary 4.2. As algebraic-geometric codes from genus 0 curves, the Reed-Solomon codes are MDS (1-MDS). The next interesting cases are these algebraic-geometric 2-MDS codes from elliptic curves.

Since the GM-MDS conjecture are about 1-MDS linear codes, we can consider the direct generalization of the GM-MDS conjecture for 2-MDS linear codes. The generalized Hamming weights of 2-MDS linear (not MDS) codes are as follows,

$$d_1 = n-k,$$
$$d_2 = n-k+2,$$
$$\cdots,$$
$$d_r = n-k+r,$$
$$\cdots,$$
$$d_k = n.$$

Many algebraic-geometric $[n,k]_q$ codes from elliptic curves with code lenght $n > q+2$ have their generalized Hamming weights as above. However for algebraic-geometric code from elliptic curve cases, not every subset of $[n]$ of the cardinality k can be the set of zero coordinate positions of nonzero codeword, the condition $|\mathbf{S}_i| \leq k-1$ is a natural constraint.

Therefore the GHW -based support constrained conditions on the subset systems for two or more subsets are the same as the MDS condition in the

9

are imposed. Algebraic geometry codes over elliptic curves are natural generalizations of Reed-Solomon codes. Hence it is interesting to consider the possible generalization of GM-MDSO conjecture and then a beautiful theorem to algebraic geometry codes over elliptic curves. Theorem2.2 and Corollary2.1 are natural extensions in this case, however the sufficient conditions in Theorem2.2 and Corollary2.1 are clearly much stronger than the necessary.

A linear$[n,k]_q$ code over$\mathbf{F}_q$ is called$r$-MDS for some$r$ in the range$1 \leq r \leq k$, ifd_r=$n-k+r$. Then it is also$s$-MDS for any$s \geq r$, see[23]. The linear MDS codes are then1-MDS. Hence$r$-MDS codes for$r \geq 2$ are natural generalizations of linear MDS codes. A well-known result in weight hierarchy or higher weights about algebraic-geometric codes due to Tsfassman and Vialdtu is that these codes are$g$+1-MDS if they are from genus$g$ curves, see[23] Corollary4.2. As algebraic-geometric codes from genus0 curves, the Reed-Solomon codes are MDS(1-MDS). The next interesting cases are these algebraic-geometric2-MDS codes from elliptic curves.

Since the GM-MDS) are about1-MDS linear codes, we can consider the direct generalization of the GM-MDS conjecture for2-MDS linear codes. The generalized Hamming weights of2-MDS linear(not MDS) codes are as follows,

$$d_1 = n-k,$$
$$d_2 = n-k+2,$$
$$\cdots$$
,
$$d_r = n-k+r,$$
$$\cdots$$
$$d_k = n.$$

Many algebraic-geometric$[n,k]_q$ codes from elliptic curves with code lenght$n$>q+2 have their generalized Hamming weights as above. However for algebraic-geometric code from elliptic curve cases, not every subset of$[n]$ of the cardinality$k$ can be the set of zero coordinate positions of nonzero codeword, the condition$|\mathbf{S}_i| \leq$ is a constraint.

Therefore the GHW-based support constrained conditions on the subset systems for two or more subsets are the same as MDS condition.

Figure B1: Examples of our model output. Left: Origin image of document page with tables and equations. Right: Model output converted to Markdown and rendered back into a PDF.

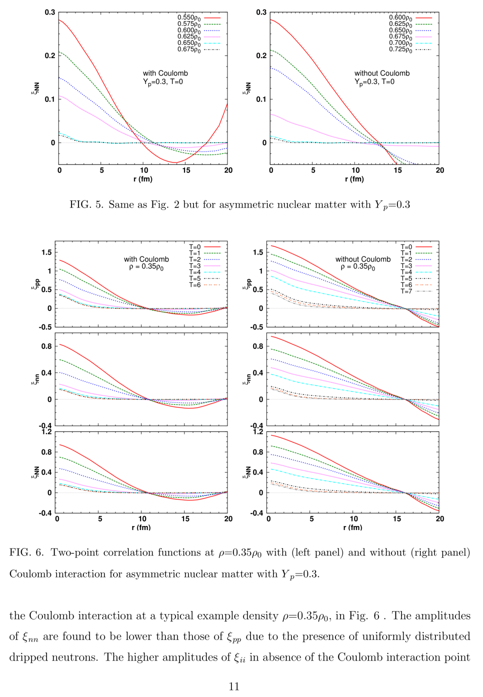


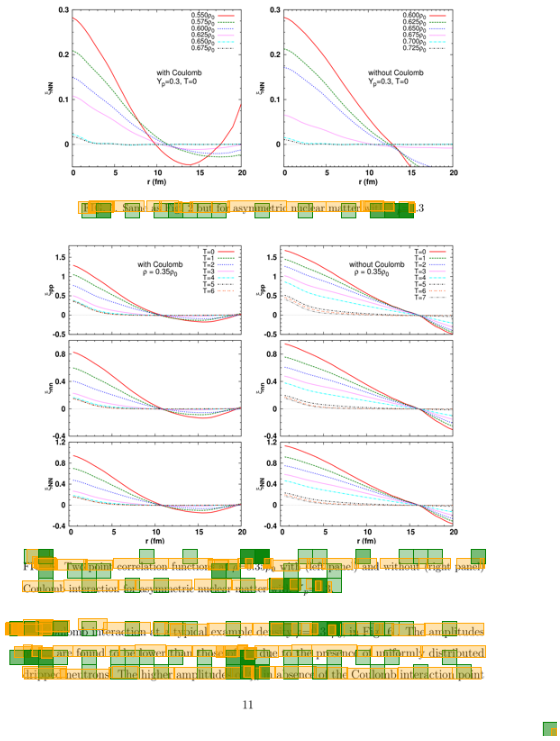
FIG. 5. Same as Fig. 2 but for asymmetric nuclear matter with Y_p=0.3

FIG. 6. Two-point correlation functions at ρ=0.35ρ_0 with (left panel) and without (right panel) Coulomb interaction for asymmetric nuclear matter with Y_p=0.3.

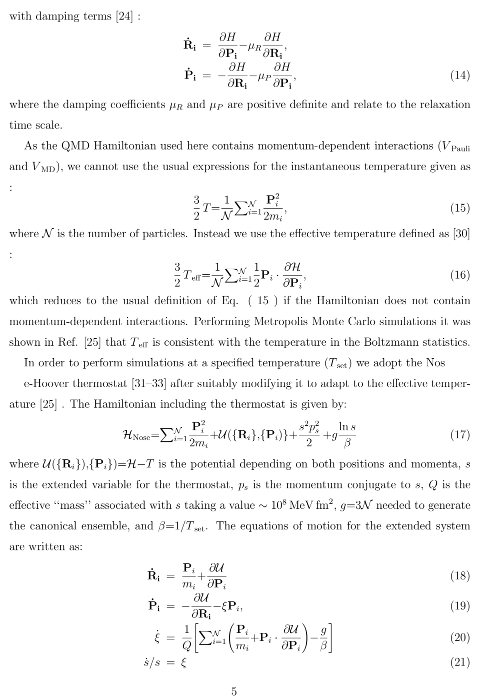
the Coulomb interaction at a typical example density ρ=0.35ρ_0, in Fig. 6 . The amplitudes of ξ_{nn} are found to be lower than those of ξ_{pp} due to the presence of uniformly distributed dripped neutrons. The higher amplitudes of ξ_{ii} in absence of the Coulomb interaction point

11

(a) Origin page with figures



(b) Result



with damping terms [24] :

$$\dot{\mathbf{R}}_{\mathbf{i}} = \frac{\partial H}{\partial \mathbf{P}_{\mathbf{i}}} - \mu_R \frac{\partial H}{\partial \mathbf{R}_{\mathbf{i}}},$$
$$\dot{\mathbf{P}}_{\mathbf{i}} = -\frac{\partial H}{\partial \mathbf{R}_{\mathbf{i}}} - \mu_P \frac{\partial H}{\partial \mathbf{P}_{\mathbf{i}}}, \tag{14}$$

where the damping coefficients μ_R and μ_P are positive definite and relate to the relaxation time scale.

As the QMD Hamiltonian used here contains momentum-dependent interactions (V_{Pauli} and V_{MD}), we cannot use the usual expressions for the instantaneous temperature given as :

$$\frac{3}{2}T = \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}} \frac{\mathbf{P}_i^2}{2m_i}, \tag{15}$$

where $\mathcal{N}$ is the number of particles. Instead we use the effective temperature defined as [30] :

$$\frac{3}{2}T_{\mathrm{eff}} = \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}} \frac{1}{2}\mathbf{P}_i \cdot \frac{\partial \mathcal{H}}{\partial \mathbf{P}_i}, \tag{16}$$

which reduces to the usual definition of Eq. (15) if the Hamiltonian does not contain momentum-dependent interactions. Performing Metropolis Monte Carlo simulations it was shown in Ref. [25] that T_{eff} is consistent with the temperature in the Boltzmann statistics.

In order to perform simulations at a specified temperature (T_{set}) we adopt the Nos

e-Hoover thermostat [31–33] after suitably modifying it to adapt to the effective temperature [25] . The Hamiltonian including the thermostat is given by:

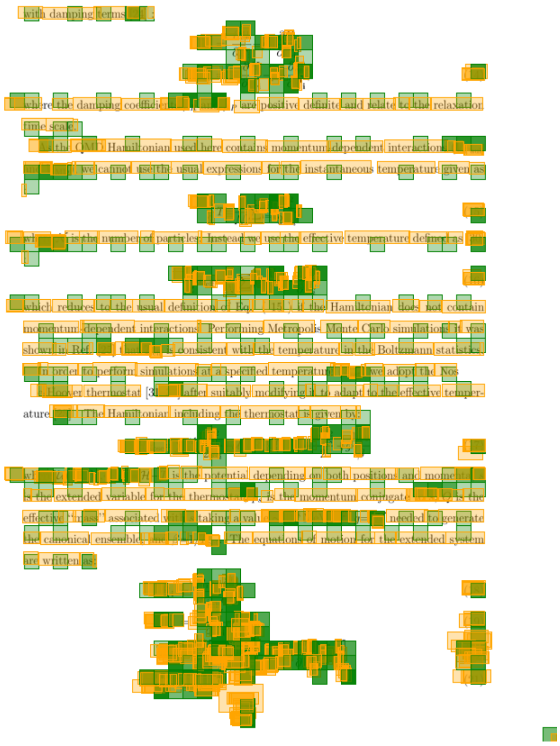
$$\mathcal{H}_{\mathrm{Nose}} = \sum_{i=1}^{\mathcal{N}} \frac{\mathbf{P}_i^2}{2m_i} + \mathcal{U}(\{\mathbf{R}_i\},\{\mathbf{P}_i)\} + \frac{s^2 p_s^2}{2} + g\frac{\ln s}{\beta} \tag{17}$$

where $\mathcal{U}(\{\mathbf{R}_i\}),\{\mathbf{P}_i\}) = \mathcal{H} - T$ is the potential depending on both positions and momenta, s is the extended variable for the thermostat, p_s is the momentum conjugate to s, Q is the effective "mass" associated with s taking a value $\sim 10^8$ MeV fm^2, g=3$\mathcal{N}$ needed to generate the canonical ensemble, and β=1/T_{set}. The equations of motion for the extended system are written as:

$$\dot{\mathbf{R}}_{\mathbf{i}} = \frac{\mathbf{P}_i}{m_i} + \frac{\partial \mathcal{U}}{\partial \mathbf{P}_i} \tag{18}$$
$$\dot{\mathbf{P}}_{\mathbf{i}} = -\frac{\partial \mathcal{U}}{\partial \mathbf{R}_{\mathbf{i}}} - \xi \mathbf{P}_i, \tag{19}$$
$$\dot{\xi} = \frac{1}{Q}\left[\sum_{i=1}^{\mathcal{N}}\left(\frac{\mathbf{P}_i}{m_i} + \mathbf{P}_i \cdot \frac{\partial \mathcal{U}}{\partial \mathbf{P}_i}\right) - \frac{g}{\beta}\right] \tag{20}$$
$$\dot{s}/s = \xi \tag{21}$$

5

(c) Origin page with mathematical formulas



(d) Result

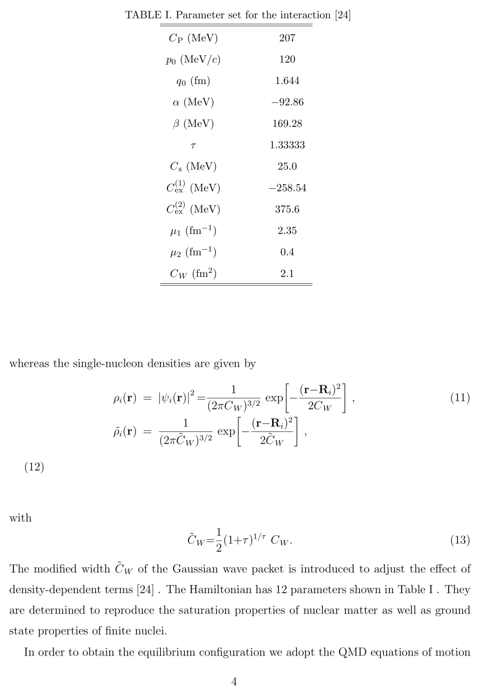


TABLE I. Parameter set for the interaction [24]

C_{P} (MeV)	207
p_0 (MeV/c)	120
q_0 (fm)	1.644
α (MeV)	−92.86
β (MeV)	169.28
τ	1.33333
C_{s} (MeV)	25.0
$C_{\mathrm{ex}}^{(1)}$ (MeV)	−258.54
$C_{\mathrm{ex}}^{(2)}$ (MeV)	375.6
μ_1 (fm^{-1})	2.35
μ_2 (fm^{-1})	0.4
C_W (fm^2)	2.1

whereas the single-nucleon densities are given by

$$\rho_i(\mathbf{r}) = |\psi_i(\mathbf{r})|^2 = \frac{1}{(2\pi C_W)^{3/2}} \exp\left[-\frac{(\mathbf{r}-\mathbf{R}_i)^2}{2C_W}\right], \quad (11)$$

$$\tilde{\rho}_i(\mathbf{r}) = \frac{1}{(2\pi \tilde{C}_W)^{3/2}} \exp\left[-\frac{(\mathbf{r}-\mathbf{R}_i)^2}{2\tilde{C}_W}\right],$$

(12)

with

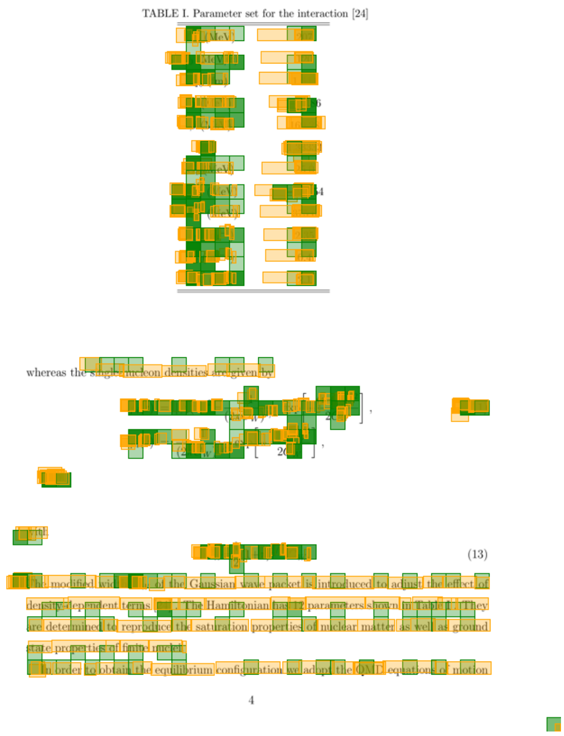
$$\tilde{C}_W = \frac{1}{2}(1+\tau)^{1/\tau}\, C_W. \quad (13)$$

The modified width $\tilde{C}_W$ of the Gaussian wave packet is introduced to adjust the effect of density-dependent terms [24] . The Hamiltonian has 12 parameters shown in Table I . They are determined to reproduce the saturation properties of nuclear matter as well as ground state properties of finite nuclei.

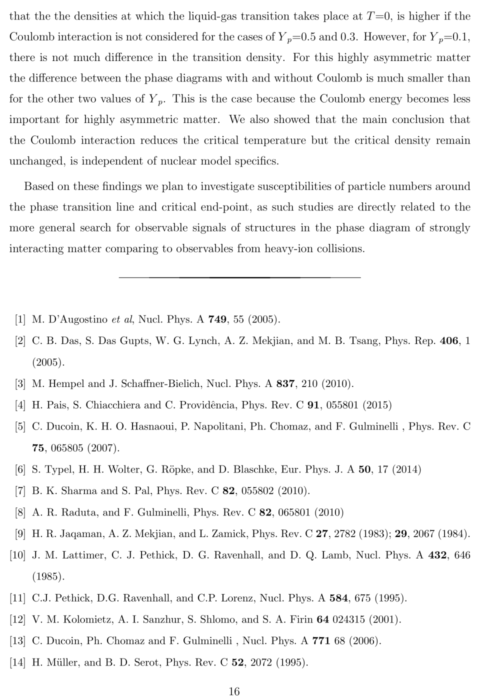
In order to obtain the equilibrium configuration we adopt the QMD equations of motion

4

(e) Origin page with tables



(f) Result



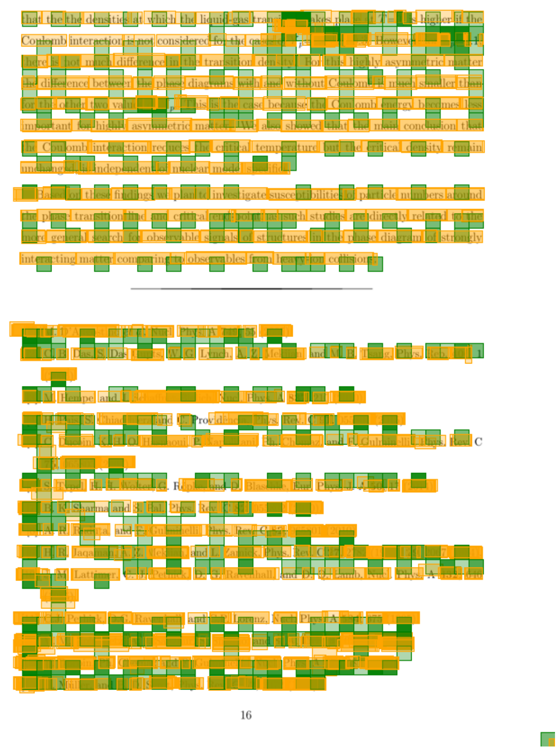
that the the densities at which the liquid-gas transition takes place at T=0, is higher if the Coulomb interaction is not considered for the cases of Y_p=0.5 and 0.3. However, for Y_p=0.1, there is not much difference in the transition density. For this highly asymmetric matter the difference between the phase diagrams with and without Coulomb is much smaller than for the other two values of Y_p. This is the case because the Coulomb energy becomes less important for highly asymmetric matter. We also showed that the main conclusion that the Coulomb interaction reduces the critical temperature but the critical density remain unchanged, is independent of nuclear model specifics.

Based on these findings we plan to investigate susceptibilities of particle numbers around the phase transition line and critical end-point, as such studies are directly related to the more general search for observable signals of structures in the phase diagram of strongly interacting matter comparing to observables from heavy-ion collisions.

[1] M. D'Augostino *et al*, Nucl. Phys. A **749**, 55 (2005).
[2] C. B. Das, S. Das Gupts, W. G. Lynch, A. Z. Mekjian, and M. B. Tsang, Phys. Rep. **406**, 1 (2005).
[3] M. Hempel and J. Schaffner-Bielich, Nucl. Phys. A **837**, 210 (2010).
[4] H. Pais, S. Chiacchiera and C. Providência, Phys. Rev. C **91**, 055801 (2015)
[5] C. Ducoin, K. H. O. Hasnaoui, P. Napolitani, Ph. Chomaz, and F. Gulminelli , Phys. Rev. C **75**, 065805 (2007).
[6] S. Typel, H. H. Wolter, G. Röpke, and D. Blaschke, Eur. Phys. J. A **50**, 17 (2014)
[7] B. K. Sharma and S. Pal, Phys. Rev. C **82**, 055802 (2010).
[8] A. R. Raduta, and F. Gulminelli, Phys. Rev. C **82**, 065801 (2010)
[9] H. R. Jaqaman, A. Z. Mekjian, and L. Zamick, Phys. Rev. C **27**, 2782 (1983); **29**, 2067 (1984).
[10] J. M. Lattimer, C. J. Pethick, D. G. Ravenhall, and D. Q. Lamb, Nucl. Phys. A **432**, 646 (1985).
[11] C.J. Pethick, D.G. Ravenhall, and C.P. Lorenz, Nucl. Phys. A **584**, 675 (1995).
[12] V. M. Kolomietz, A. I. Sanzhur, S. Shlomo, and S. A. Firin **64** 024315 (2001).
[13] C. Ducoin, Ph. Chomaz and F. Gulminelli , Nucl. Phys. A **771** 68 (2006).
[14] H. Müller, and B. D. Serot, Phys. Rev. C **52**, 2072 (1995).

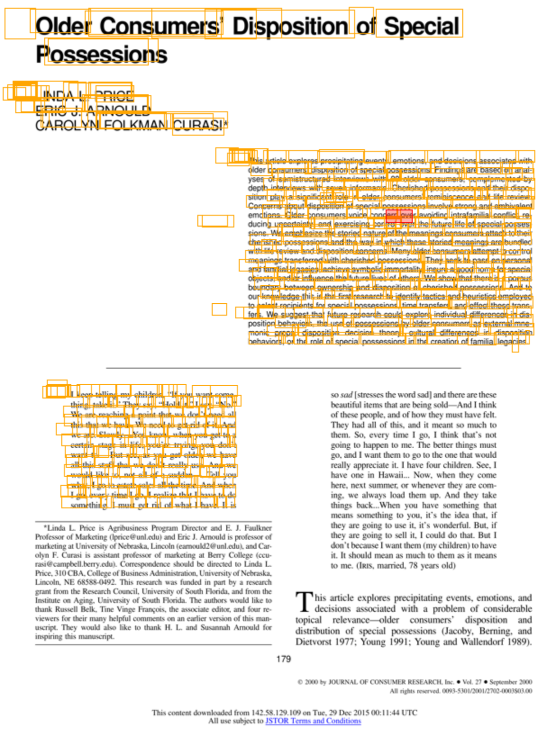
16

(g) Origin page with references

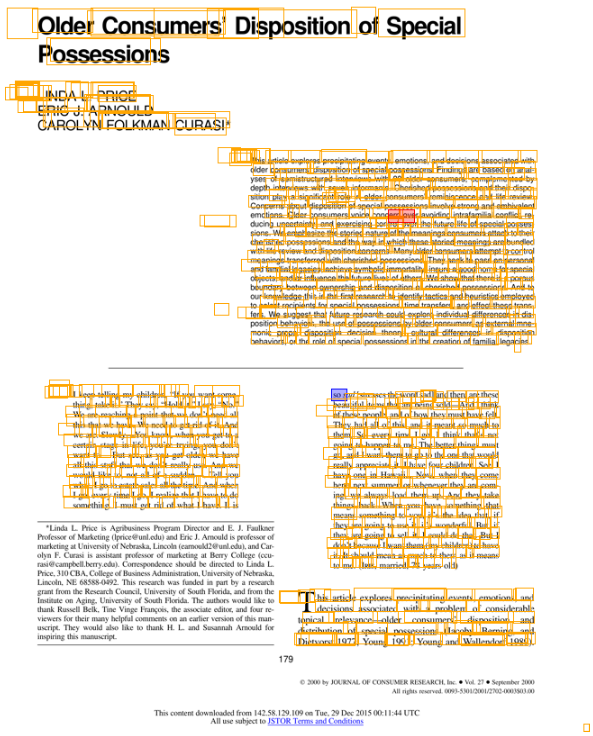


(h) Result

Figure B2: Example of position prediction. Green box: Rough result of grid classification. Yellow: Final result of box regression.



(i) A case model predicting wrong position



(j) Result

Figure C1: Visualization of interaction on out-of-domain documents. Red box: Wrong position. Blue box: Human prompt input.